Biometrics 67, 1330–1339
December 2011
DOI: 10.1111/j.1541-0420.2011.01584.x

Smoothing Spline ANOVA Frailty Model for Recurrent Event Data

Pang Du,[1,*] Yihua Jiang,[2] and Yuedong Wang[3]

[1]Department of Statistics, Virginia Tech, Blacksburg, Virginia 24061, U.S.A.
[2]StubHub Inc., San Francisco, California 94105, U.S.A.
[3]Department of Statistics and Applied Probability, University of California at Santa Barbara, Santa Barbara, California 93106, U.S.A.
[*]*email:* pangdu@vt.edu

Summary. Gap time hazard estimation is of particular interest in recurrent event data. This article proposes a fully nonparametric approach for estimating the gap time hazard. Smoothing spline analysis of variance (ANOVA) decompositions are used to model the log gap time hazard as a joint function of gap time and covariates, and general frailty is introduced to account for between-subject heterogeneity and within-subject correlation. We estimate the nonparametric gap time hazard function and parameters in the frailty distribution using a combination of the Newton–Raphson procedure, the stochastic approximation algorithm (SAA), and the Markov chain Monte Carlo (MCMC) method. The convergence of the algorithm is guaranteed by decreasing the step size of parameter update and/or increasing the MCMC sample size along iterations. Model selection procedure is also developed to identify negligible components in a functional ANOVA decomposition of the log gap time hazard. We evaluate the proposed methods with simulation studies and illustrate its use through the analysis of bladder tumor data.

Key words: Frailty; Gap time hazard; Model selection; Nonparametric hazard model; Recurrent event data; Smoothing spline ANOVA; Stochastic approximation.

1. Introduction

Recurrent event data are common in engineering, economics, public health, and medical studies. Examples include tumor recurrences, recurrent infections, repeated hospitalizations, warranty claims, and repeated failures of a mechanical system. Often these recurrent events are treated as realizations of a counting process with multiple jumps. Methods for analyzing such data can be roughly divided into two categories based on their choices of the time scale. One category of methods model the intensity function of the counting process as a function of calendar time, that is, time since the start of follow-up. These models investigate the trend of event recurrences over the whole calendar time domain and its dependence on the covariates. An extensive review of such methods can be found in Cook and Lawless (2009) and Kalbfleisch and Prentice (2002). The other category of methods focus on gap time or interoccurrence time. Models based on gap time provide different perspectives on recurrent events. Some examples in the absence of covariates are Lin, Sun, and Ying (1999); Wang and Chang (1999); and Peña, Strawderman, and Hollander (2001). This article concerns gap time hazard in the presence of covariates, for which parametric and semiparametric models were proposed in Chang and Wang (1999); Huang (2002); Rondeau, Commenges, and Joly (2003); Strawderman (2005); Peña, Slate, and González (2007); and Clement and Strawderman (2009), among others. These existing models generally assume a parametric form of covariate effects, together with a nonparametric nuisance component such as baseline hazard in a proportional hazards (PH) model or error distribution in an accelerated failure time model. Powerful diagnostic tools for general departures from these semiparametric models and flexible model building procedures without these assumptions are lacking. In this article, we propose a fully nonparametric procedure for gap time hazard that offers more flexibility by allowing a general form of covariate effects as well as a general form of the interaction between gap time and covariates.

Our nonparametric gap time hazard model is constructed under the framework of smoothing spline ANOVA (SS ANOVA) decompositions. An SS ANOVA decomposition decomposes a multivariate function into a sum of orthogonal components as main effects and interactions. When applied to log hazard function, such a functional ANOVA decomposition yields main effect of gap time, main effects of covariates, and interaction between gap time and covariates. A PH model is a special case when the interaction equals to zero. Since all components are modeled nonparametrically, the proposed general models and methods may be used as exploratory and diagnostic tools. See Section 2 for more details.

Earlier smoothing spline hazard estimation methods focused mostly on single event data. Some examples are Anderson and Senthilselvan (1980); Zucker and Karr (1990); O'Sullivan (1993); Gu (1996); and Joly, Commenges, and Letenneur (1998). For recurrent event data, Rondeau et al. (2003) proposed a gamma-frailty model where the baseline hazard was estimated by smoothing splines while the covariate effect was modeled parametrically. This approach was then extended to nested frailty models in Rondeau, Filleul, and Joly (2006) and to joint modeling of recurring events and death in

Rondeau et al. (2007). Du (2009) developed a fully nonparametric gap time hazard model for recurrent event data. But the drawback of his approach was the lack of subject-specific frailty. Thus potential correlation between recurrences from the same subject was ignored. We note that the inclusion of frailty is highly nontrivial since the additional computational difficulty requires both new methods and novel algorithms. Our method provides a framework to include general frailty. Du and Ma (2010) also introduced frailty to their nonparametric hazard model for clustered survival data. They adopted the so-called "double penalization" approach with an additional penalty on individual cluster frailties. Their approach treated cluster frailties as fixed parameters and estimated them together with the nonparametric hazard function. Since each cluster frailty is essentially determined by the observations in the cluster, their approach requires cluster size to increase with sample size in certain order to guarantee "consistent estimation" of the cluster frailty. This assumption, however, is generally not satisfied by recurrent event data where each subject is a "cluster" and the number of recurrences per subject is usually of order $O(1)$.

In this article, we propose a fully nonparametric procedure for gap time hazard with a general form of frailty. To the best of our knowledge, this is the first fully nonparametric approach with frailty for gap time hazard function in the presence of covariates. It offers practitioners the ability to investigate a general shape of the hazard function and extract useful information that might be missed by parametric or semiparametric models. The frailties in our model are general in terms of both distribution and structure. For example, the nested frailties in Rondeau et al. (2006) can be easily incorporated into our model. We propose a simple model selection procedure based on the Kullback–Leibler geometry developed in Gu (2004). It will be used to identify negligible effects in a model. In particular, it can be applied to check the validity of the PH assumption by examining the gap-time-covariate interaction. Our numerical experiments show satisfactory performance of the proposed methods.

We use an iterative procedure to obtain the estimates for the gap time hazard function and the parameters in the frailty distribution. For each iteration, given the frailty distribution, the gap time hazard is estimated through the minimization of a penalized likelihood (PL), extending the standard PL estimation for smoothing splines. This optimization is carried out by the Newton–Raphson procedure. One difficulty is that the derivatives of the PL involve expectations with respect to the posterior distribution of frailty given data that do not have closed forms. A Markov chain Monte Carlo (MCMC) approach will be applied to estimate these expectations. The next step within the iteration is the estimation of frailty parameters by a stochastic search algorithm modified from the SAA in Gu and Kong (1998). We will reduce the step size of parameter update and/or increase the MCMC sample size along iterations in a certain way such that the algorithm is guaranteed to converge to the expected estimator.

The rest of the article is laid out as follows. Section 2 introduces the SS ANOVA frailty model. Section 3 describes our estimation procedure and algorithm. Section 4 derives a model selection procedure. Section 5 presents simulations to evaluate the performance of the proposed estimation and model selection methods. Section 6 illustrates the methods through the bladder tumor data. Section 7 concludes the paper with some discussions.

2. SS ANOVA Frailty Model for Recurrent Event Data

Consider m independently selected subjects. For the ith subject, let $0 = S_{i0} < S_{i1} < S_{i2} < \cdots$ be the successive calendar times of event occurrences and $\widetilde{t}_{ij} = S_{ij} - S_{i,j-1}$ be the gap times between successive event occurrences. Let $\mathbf{x}_i$ be the covariates associated with subject i and C_i be the right censoring time for subject i. We assume that C_i is independent of $\widetilde{t}_{ij}$'s given the covariate $\mathbf{x}_i$. Let $n_i = \max\{j \in \{0, 1, 2, \ldots\} : S_{ij} \leqslant C_i\}$ be the number of uncensored gap times observed over the study period $[0, C_i]$. To simplify the notation, we introduce $t_{ij} = \widetilde{t}_{ij}$ for $j = 1, \ldots, n_i$, and $t_{i,n_i+1} = C_i - S_{in_i}$. Hence the observable entities over the study period $[0, C_i]$ are $(\mathbf{y}_i, \mathbf{x}_i)$, where $\mathbf{y}_i = (n_i, C_i, t_{i1}, \ldots, t_{in_i}, t_{i,n_i+1})$. For subject i with covariates $\mathbf{x}_i$, we assume that $\widetilde{t}_{ij}$'s are realizations of random variables with survival function $\overline{F}_i(t, \mathbf{x}_i)$ where t is the time since the last recurrence. Of interest is the estimation of the gap time hazard function $h_i(t, \mathbf{x}_i) = -\partial \log \overline{F}_i(t, \mathbf{x}_i)/\partial t$ or its logarithm $f_i(t, \mathbf{x}_i) = \log h_i(t, \mathbf{x}_i)$. To account for heterogeneity between subjects and potential correlation within a subject, we introduce a vector $\mathbf{b}$ that collects random effects from all the subjects. The random effects $\mathbf{b}$ are often referred to as frailties in survival analysis. Let $\mathbf{z}_i$ be the covariate vector for random effects associated with subject i. We consider the following frailty model:

$$f_i(t, \mathbf{x}_i) = f(t, \mathbf{x}_i) + \mathbf{z}_i^T \mathbf{b}, \quad i = 1, \ldots, m, \tag{1}$$

where f is an unknown function of gap time t and covariates $\mathbf{x}$. A typical model for frailty $\mathbf{b}$ is a multivariate normal model. For generality, in this article we will not assume any specific structure or distribution for frailty $\mathbf{b}$. The well-known shared frailty model corresponds to $\mathbf{z}_i^T \mathbf{b} = b_i$ for a scalar b_i. Time-varying frailty may be incorporated into $\mathbf{b}$ where an autoregressive process or a Gaussian stochastic process based on smoothing spline may be used to model the dependence structure (Wang, 1998; Yau and McGilchrist, 1998). We assume that $\mathbf{b}$ follows a distribution with density $p_b(\mathbf{b}; \boldsymbol{\tau})$ where $\boldsymbol{\tau}$ is a vector of parameters.

We now describe how to construct SS ANOVA models for f. Let $\mathcal{T}$ and $\mathcal{X}$ be the domains of t and $\mathbf{x}$, respectively, and consider f as a joint function on the product domain $\mathcal{T} \times \mathcal{X}$. Let $\mathcal{H}_{\langle t \rangle}$ and $\mathcal{H}_{\langle x \rangle}$ be RK Hilbert spaces (RKHS) on $\mathcal{T}$ and $\mathcal{X}$, respectively. Consider the tensor product RKHS space $\mathcal{H} = \mathcal{H}_{\langle t \rangle} \otimes \mathcal{H}_{\langle x \rangle}$. Suppose that the component spaces can be further decomposed as tensor sums $\mathcal{H}_{\langle t \rangle} = \{1\} \oplus \mathcal{H}_{\langle 1,t \rangle}$ and $\mathcal{H}_{\langle x \rangle} = \{1\} \oplus \mathcal{H}_{\langle 1,x \rangle}$, where $\{1\}$ is the one-dimension space of constant functions. Then the tensor product space can be decomposed into

$$\mathcal{H} = \big(\{1\} \oplus \mathcal{H}_{\langle 1,t \rangle}\big) \otimes \big(\{1\} \oplus \mathcal{H}_{\langle 1,x \rangle}\big) = \mathcal{H}_\emptyset \oplus \mathcal{H}_t \oplus \mathcal{H}_x \oplus \mathcal{H}_{tx}, \tag{2}$$

where $\mathcal{H}_\emptyset = \{1\} \otimes \{1\}, \mathcal{H}_t = \mathcal{H}_{\langle 1,t \rangle} \otimes \{1\}, \mathcal{H}_x = \{1\} \otimes \mathcal{H}_{\langle 1,x \rangle}$, and $\mathcal{H}_{tx} = \mathcal{H}_{\langle 1,t \rangle} \otimes \mathcal{H}_{\langle 1,x \rangle}$. In terms of the function, the decomposition in (2) is equivalent to decomposing f as

$$f(t, \mathbf{x}) = f_\emptyset + f_1(t) + f_2(\mathbf{x}) + f_{12}(t, \mathbf{x}), \tag{3}$$

where $f_\emptyset$ is a constant, $f_1(t)$ is the main effect of gap time t, $f_2(\mathbf{x})$ is the main effect of covariates $\mathbf{x}$, and $f_{12}(t, \mathbf{x})$ is the interaction between gap time and covariates. The equation (3) is called an SS ANOVA decomposition since it is a natural extension of the classical analysis of variance (ANOVA) model from discrete domains to general domains. Notice that the PH assumption is equivalent to $f_{12}(t, \mathbf{x}) = 0$. Therefore, the general departure from the PH assumption may be checked by examining f_{12} in the model. Most existing semi-parametric models assume that $f_{12}(t, \mathbf{x}) = 0$ and a parametric model for f_2. Therefore the SS ANOVA decompositions provides a general framework for model building and diagnostics.

Example 1 (Tensor Product Cubic Spline). We now provide an example of the SS ANOVA decomposition that will be used in the simulation studies in Section 5. Consider the situation where $\mathbf{x}$ is a univariate continuous covariate x. For simplicity, assume the domains $\mathcal{X} = \mathcal{T} = [0, 1]$. Consider $\mathcal{H}_{\langle t\rangle} = \mathcal{H}_{\langle x\rangle} = W_2^2[0, 1]$ where

$$W_2^2[0,1] = \left\{ f : f \text{ and } f' \text{ absolutely continuous}, \int_0^1 (f'')^2 \, dt < \infty \right\} \tag{4}$$

is the cubic smoothing spline model space. The space $W_2^2[0, 1]$ can be decomposed into

$$W_2^2[0,1] = \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1, \tag{5}$$

where $\mathcal{H}_{00} = \text{span}\{1\}, \mathcal{H}_{01} = \text{span}\{k_1(t)\}, \mathcal{H}_1 = \{f : \int_0^1 f \, dt = \int_0^1 f' \, dt = 0, \int_0^1 (f'')^2 \, dt < \infty\}$, and $k_\nu(t) = B_\nu(t)/\nu!$ are scaled Bernoulli polynomials. The RKs for subspaces $\mathcal{H}_{00}, \mathcal{H}_{01}$, and $\mathcal{H}_1$ are $R_{00}(t_1, t_2) = 1$, $R_{01}(t_1, t_2) = k_1(t_1)k_1(t_2)$, and $R_1(t_1, t_2) = k_2(t_1)k_2(t_2) - k_4(|t_1 - t_2|)$. Denote the decompositions corresponding to (5) for marginal spaces $\mathcal{H}_{\langle t\rangle}$ and $\mathcal{H}_{\langle x\rangle}$ as $\mathcal{H}_{\langle t\rangle} = \mathcal{H}_{00\langle t\rangle} \oplus \mathcal{H}_{01\langle t\rangle} \oplus \mathcal{H}_{1\langle t\rangle}$ and $\mathcal{H}_{\langle x\rangle} = \mathcal{H}_{00\langle x\rangle} \oplus \mathcal{H}_{01\langle x\rangle} \oplus \mathcal{H}_{1\langle x\rangle}$. Taking tensor product, one obtains nine tensor sum terms $\mathcal{H}_{\nu,\mu} = \mathcal{H}_\nu^{(t)} \otimes \mathcal{H}_\mu^{(x)}$ on $\mathcal{T} \times \mathcal{X}$ for $\nu, \mu = 00, 01, 1$, with RKs $R_{\nu,\mu}(z_1, z_2) = R_\nu(t_1, t_2)R_\mu(x_1, x_2)$ where $z = (t, x)$. The decomposition in (2) can be derived by combining subspaces such that $\mathcal{H}_{\langle 1,t\rangle} = \mathcal{H}_{01\langle t\rangle} \oplus \mathcal{H}_{1\langle t\rangle}$ and $\mathcal{H}_{\langle 1,x\rangle} = \mathcal{H}_{01\langle x\rangle} \oplus \mathcal{H}_{1\langle x\rangle}$. See Gu (2002) and Wang (2011) for more details about the SS ANOVA decomposition.

We note that the decomposition of the tensor product space in (2) is just one form of the SS ANOVA decomposition. For a multivariate covariate vector $\mathbf{x}$, the component $f_2(\mathbf{x})$ (or the space $\mathcal{H}_x$) can be further decomposed into main effects and interactions between covariates, and the component $f_{12}(t, \mathbf{x})$ (or the space $\mathcal{H}_{tx}$) can be further decomposed into two-way and higher order interactions between gap time and one or more covariates. Any subset of the full SS ANOVA decomposition may be considered as a model space for f and will be referred to as an SS ANOVA model. After a model space $\mathcal{H}$ is chosen, it can be represented as (Wahba, 1990; Gu, 2002; Wang, 2011)

$$\mathcal{H} = \mathcal{H}_0 \oplus \sum_{j=1}^{q} \mathcal{H}_j, \tag{6}$$

where $\mathcal{H}_0$ collects all unpenalized subspaces.

3. Estimation

The development of the estimation procedure will involve the following sequence of steps. For the estimation of the nonparametric function f, we first define an estimation criterion based on the PL and approximate its solution in a data-adaptive finite-dimensional space. Then the problem reduces to estimating coefficients in the solution by minimizing the PL. The minimization is performed through two nested loops: for fixed smoothing parameters, the inner loop minimizes the PL through a combination of the Newton–Raphson and MCMC procedures; and the outer loop selects the optimal smoothing parameters by minimizing an approximation to the Kullback–Leibler distance based on a delete-one-observation cross-validation (CV). For the estimation of the frailty parameters, we adopt the MCMC SAA in Gu and Kong (1998) to find the MLE of $\boldsymbol{\tau}$. We now provide details about the estimation procedure.

3.1 *The Penalized Likelihood*

Given $\mathbf{b}$, the conditional likelihood of $\mathbf{y}_i \,|\, \mathbf{b}$ is

$$p_{t|b}(\mathbf{y}_i \,|\, \mathbf{b}) = \exp\left\{ -\int_0^{t_{i,n_i+1}} h_i(s, \mathbf{x}_i)\, ds \right\} \times \prod_{j=1}^{n_i} \left\{ h_i(t_{ij}, \mathbf{x}_i) \exp\left(-\int_0^{t_{ij}} h_i(s, \mathbf{x}_i)\, ds \right) \right\}.$$

Hence the log marginal likelihood of $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_m^T)^T$ is

$$l(\boldsymbol{\tau}, f) = \sum_{i=1}^{m} \log \int p_{t|b}(\mathbf{y}_i \,|\, \mathbf{b}) p_b(\mathbf{b}; \boldsymbol{\tau}) \, d\mathbf{b}. \tag{7}$$

We first consider the estimation of f given $\boldsymbol{\tau}$. Let $N = \sum_{i=1}^m (n_i + 1)$ be the total number of observations. For $f \in \mathcal{H}$ where $\mathcal{H}$ is an SS ANOVA model space given in (6), we estimate f as the minimizer of the PL

$$\text{PL}(f) = -l(\boldsymbol{\tau}, f) + \frac{N\lambda}{2} \sum_{j=1}^{q} \theta_j^{-1} ||P_j f||^2, \tag{8}$$

where the negative log likelihood $-l(\boldsymbol{\tau}, f)$ measures the goodness-of-fit, P_j is the orthogonal projector in $\mathcal{H}$ onto $\mathcal{H}_j$, and $\lambda\theta_j^{-1}$ are smoothing parameters controlling the trade-off between the goodness of fit and the departure from the null space $\mathcal{H}_0$.

Choices for the model space and penalty depend on several factors including the domain $\mathcal{X}$, assumption about the function f and purpose of the study. For the SS ANOVA decomposition given in Example 1, the four subspaces with $\nu, \mu = 00, 01$ are unpenalized and lumped together as $\mathcal{H}_0$. The other five subspaces form the penalized subspaces and have penalties $||P_{00,1}f||^2 = \int_{\mathcal{X}} \{\int_{\mathcal{T}} f_{xx}(t, x) \, dt\}^2 dx$, $||P_{1,00}f||^2 = \int_{\mathcal{T}} \{\int_{\mathcal{X}} f_{tt}(t, x) \, dx\}^2 \, dt$, $||P_{01,1}f||^2 = \int_{\mathcal{X}} \{\int_{\mathcal{T}} f_{txx}(t, x) \, dt\}^2 dx$, $||P_{1,01}f||^2 = \int_{\mathcal{T}} \{\int_{\mathcal{X}} f_{ttx}(t, x) \, dx\}^2 \, dt$, $||P_{1,1}f||^2 = \int_{\mathcal{X}} \int_{\mathcal{T}} \{f_{ttxx}(t, x)\}^2 dt dx$, where $f_{tt} = \partial^2 f/\partial t^2$, $f_{xx} = \partial^2 f/\partial x^2$, $f_{txx} = \partial^3 f/\partial t \partial x^2$, $f_{ttx} = \partial^3 f/\partial t^2 \partial x$, and $f_{ttxx} = \partial^4 f/\partial t^2 \partial x^2$.

Let R_j be the reproducing kernel (RK) of $\mathcal{H}_j$ for $j = 1, \ldots, q$, and $R_{\boldsymbol{\theta}} = \sum_{j=1}^q \theta_j R_j$. The solution to the PL (8) does not fall in a finite-dimensional space. Therefore, certain approximation to the solution is necessary. Let $M = \sum_{i=1}^m n_i$

be the total number of events. Reindex the recurrence observations $\{(t_{ij}, \mathbf{x}_i), j = 1,\ldots,n_i, i = 1,\ldots,m\}$ as $\mathcal{V} = \{\mathbf{v}_k, k = 1,\ldots,M\}$. Instead of $\mathcal{H}$, we solve the minimization problem (8) in the following data-adaptive finite-dimensional subspace $\mathcal{H}_L = \mathcal{H}_0 \oplus \text{span}\{R_{\boldsymbol{\theta}}(\mathbf{v}_{k_l}, \cdot) : l = 1,\ldots,L\}$, where the set $\mathcal{V}_L = \{\mathbf{v}_{k_1},\ldots,\mathbf{v}_{k_L}\}$ of "knots" is a random subset of $\mathcal{V}$. When $L = M$, all elements in $\mathcal{V}$ are selected as the knots (Gu, 1996). This approximation is motivated by the Kimeldorf-Wahba representer theorem that states that the minimizer of a wide range of RKHS-based regularizatrion problems resides in the finite-dimensional space $\mathcal{H}_M$ (Wang, 2011). Kim and Gu (2004) showed that an L of the order $N^{2/9}$ is sufficient for estimating a reasonably smooth multivariate function in the sense that the estimates in $\mathcal{H}_L$ and $\mathcal{H}$ have the same convergence rate. Without loss of generality, we express the solution to (8) in $\mathcal{H}_L$ as a linear combination of the corresponding basis functions

$$f(t,\mathbf{x}) = \sum_{\nu=1}^{p} d_\nu \phi_\nu(t,\mathbf{x}) + \sum_{l=1}^{L} c_l R_{\boldsymbol{\theta}}(\mathbf{v}_{k_l},(t,\mathbf{x})) = \boldsymbol{\phi}^T\mathbf{d} + \boldsymbol{\xi}^T\mathbf{c}, \tag{9}$$

where $\phi_1,\ldots,\phi_p$ is a basis of $\mathcal{H}_0, \xi_l(t,\mathbf{x}) = R_{\boldsymbol{\theta}}(\mathbf{v}_{k_l},(t,\mathbf{x}))$ for $l = 1,\ldots,L, \boldsymbol{\phi} = (\phi_1,\ldots,\phi_p)^T$ and $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_L)^T$ are vectors of functions, and $\mathbf{d} = (d_1,\ldots,d_p)^T$ and $\mathbf{c} = (c_1,\ldots,c_L)^T$ are vectors of coefficients. We need to estimate $\mathbf{c}$ and $\mathbf{d}$ from data. Substituting (9) into (8), the PL reduces to

$$\text{PL}(\boldsymbol{\tau},\mathbf{c},\mathbf{d}) = -l(\boldsymbol{\tau},\mathbf{c},\mathbf{d}) + \frac{N\lambda}{2}\mathbf{c}^T Q_{\boldsymbol{\theta}}\mathbf{c}, \tag{10}$$

where $Q_{\boldsymbol{\theta}} = \{R_{\boldsymbol{\theta}}(\mathbf{v}_{k_i},\mathbf{v}_{k_j})\}_{i,j=1}^{L}$.

3.2 *Estimation of* $\mathbf{c}$ *and* $\mathbf{d}$

To find the $\mathbf{c}$ and $\mathbf{d}$ that minimizes (10), we apply the Newton–Raphson algorithm, using an MCMC procedure to evaluate certain expectation terms that appear in the derivatives of PL. The derivatives, whose expressions are given in Web Appendix A, contain the terms $E_{b|t}(e^{\mathbf{z}_i^T\mathbf{b}})$ for $i = 1,\ldots,m$. These terms are not readily available and thus estimated through the MCMC procedure. Specifically, at the kth iteration, denote by $\mathbf{b}^{(k;1)},\ldots,\mathbf{b}^{(k;m_k)}$ an MCMC sample of size m_k generated by the single component Metropolis–Hastings procedure described in Web Appendix D. Then we estimate $E_{b|t}(e^{\mathbf{z}_i^T\mathbf{b}})$ at iteration k by $\widetilde{E}_{b|t}(e^{\mathbf{z}_i^T\mathbf{b}}) \equiv \sum_{j=1}^{m_k}\exp(\mathbf{z}_i^T\mathbf{b}^{(k;j)})/m_k$.

To simplify our notation, for functions u, w, w_1, and w_2 on the domain $\mathcal{T}\times\mathcal{X}$, we write $e(w) = \sum_{i=1}^{m}\sum_{j=1}^{n_i+1} w(t_{ij},\mathbf{x}_i)/N$, $\mu_u(w) = \sum_{i=1}^{m} E_{b|t}(e^{\mathbf{z}_i^T\mathbf{b}})\sum_{j=1}^{n_i+1}\int_0^{t_{ij}} e^{u(s,\mathbf{x}_i)}w(s,\mathbf{x}_i)ds/N$ and $V_u(w_1,w_2) = \mu_u(w_1w_2)$. Given the current estimate $\widetilde{f} = \boldsymbol{\phi}^T\widetilde{\mathbf{d}} + \boldsymbol{\xi}^T\widetilde{\mathbf{c}}$, the Newton–Raphson procedure updates coefficients $\mathbf{c}$ and $\mathbf{d}$ with the following equation:

$$\begin{pmatrix} V_{\xi,\xi} + \lambda Q_{\boldsymbol{\theta}} & V_{\xi,\phi} \\ V_{\phi,\xi} & V_{\phi,\phi} \end{pmatrix}\begin{pmatrix}\mathbf{c}\\ \mathbf{d}\end{pmatrix} = \begin{pmatrix} e_\xi - \mu_\xi + V_{\xi,f} \\ e_\phi - \mu_\phi + V_{\phi,f}\end{pmatrix}, \tag{11}$$

where, with u and w each running through $\{\xi,\phi\}, V_{u,w} = V_{\widetilde{f}}(\mathbf{u},\mathbf{w}^T), \mu_w = \mu_{\widetilde{f}}(\mathbf{w}), e_w = e(\mathbf{w})$, and $V_{u,f} = V_{\widetilde{f}}(\mathbf{u},\widetilde{f})$.

3.3 *Smoothing Parameter Selection*

The smoothing parameters $\lambda\theta_j^{-1}$ are fixed in the above Newton–Raphson procedure. In practice, it is crucial to estimate these smoothing parameters based on data. Let $\boldsymbol{\lambda} = (\lambda\theta_1^{-1},\ldots,\lambda\theta_q^{-1})$ and $f_{\boldsymbol{\lambda}}$ be the log hazard estimate obtained from minimizing (8). Define the Kullback–Leibler distance between the true log hazard f_0 and its estimate $f_{\boldsymbol{\lambda}}$ as

$$\begin{aligned}\text{KL}_0(f_0,f_{\boldsymbol{\lambda}}) &= E\Bigg[\int_{\mathcal{T}} Y(t)\big\{(f_0(t,\mathbf{x}) - f_{\boldsymbol{\lambda}}(t,\mathbf{x}))e^{f_0(t,\mathbf{x})+\mathbf{z}^T\mathbf{b}} \\ &\qquad - (e^{f_0(t,\mathbf{x})+\mathbf{z}^T\mathbf{b}} - e^{f_{\boldsymbol{\lambda}}(t,\mathbf{x})+\mathbf{z}^T\mathbf{b}})\big\}\,dt\Bigg] \\ &= E\Bigg[e^{\mathbf{z}^T\mathbf{b}}\int_{\mathcal{T}} Y(t)\big\{(f_0(t,\mathbf{x}) - f_{\boldsymbol{\lambda}}(t,\mathbf{x}))e^{f_0(t,\mathbf{x})} \\ &\qquad - (e^{f_0(t,\mathbf{x})} - e^{f_{\boldsymbol{\lambda}}(t,\mathbf{x})})\big\}\,dt\Bigg],\end{aligned} \tag{12}$$

where $Y(t)$ is the at-risk process defined in Web Appendix B. Estimating (12) through a combination of gap time counting process and a deleting one observation CV, one ends up with the CV score

$$\begin{aligned}V_\alpha(\boldsymbol{\lambda}) = &\Bigg\{-\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} f_{\boldsymbol{\lambda}}(t_{ij},\mathbf{x}_i) \\ &+ \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i+1}\widetilde{E}_{b|t}(e^{\mathbf{z}_i^T\mathbf{b}})\int_0^{t_{ij}} e^{f_{\boldsymbol{\lambda}}(t,\mathbf{x}_i)}\,dt\Bigg\} \\ &+ \alpha\left\{\frac{\text{tr}(\breve{Q}^T H^{-1}\breve{Q})}{N(N-1)} - \frac{\mathbf{1}^T\breve{Q}^T H^{-1}\breve{Q}\mathbf{1}}{N^2(N-1)}\right\},\end{aligned} \tag{13}$$

where H is the Hessian matrix appearing on the left side of (11), $\breve{Q}$ is a $(p+L)\times N$ matrix with columns $\boldsymbol{\psi}(t_{ij},\mathbf{x}_i) = (\boldsymbol{\xi}^T(t_{ij},\mathbf{x}_i),\boldsymbol{\phi}^T(t_{ij},\mathbf{x}_i))^T$, and $\mathbf{1}$ is a $(p+L)$-vector of all 1's. See Web Appendix B for detailed derivation. The smoothing parameters are estimated as the minimizer of the CV score (13). The constant $\alpha > 1$ is added in (13) to prevent occasional under-smoothing. An α around 1.4 was suggested for various hazard estimation problems; see, e.g., Gu (2002, Chapter 7), Du (2009), and Du and Ma (2010). This value appears to work well for our problem too as indicated by our simulation studies in Section 5.

3.4 *Estimation of the Frailty Parameters*

We now consider the estimation of the frailty parameters $\boldsymbol{\tau}$ for fixed f. We estimate $\boldsymbol{\tau}$ as the minimizer of the negative log likelihood $-l(\boldsymbol{\tau},f)$ by solving the following equation:

$$-\frac{\partial l(\boldsymbol{\tau},f)}{\partial\boldsymbol{\tau}} = E_{b|t}\big\{\mathbf{H}(\boldsymbol{\tau},\mathbf{b})\big\} = 0, \tag{14}$$

where $\mathbf{H}(\boldsymbol{\tau},\mathbf{b}) = \partial\log p_{\mathbf{b}}(\mathbf{b};\boldsymbol{\tau})/\partial\boldsymbol{\tau}$. Let $\mathbf{I}(\boldsymbol{\tau},\mathbf{b}) = -\partial\mathbf{H}(\boldsymbol{\tau},\mathbf{b})/\partial\boldsymbol{\tau}$. Applying the SAA procedure in Gu and Kong (1998), we solve equation (14) by iteratively updating $\boldsymbol{\tau}$ and a matrix $\boldsymbol{\Gamma}$ as follows, at the kth iteration:

$$\boldsymbol{\Gamma}_k = (1-\gamma_k)\boldsymbol{\Gamma}_{k-1} + \gamma_k\bar{\mathbf{I}}_k, \quad \widehat{\boldsymbol{\tau}}^{(k)} = \widehat{\boldsymbol{\tau}}^{(k-1)} + \gamma_k\boldsymbol{\Gamma}_k^{-1}\overline{\mathbf{H}}_k, \tag{15}$$

where $\boldsymbol{\Gamma}$ acts like a proxy of the Hessian matrix, $\{\gamma_k,\ k \geqslant 1\}$ is a sequence of positive step sizes, $\{m_k,\ k \geqslant 1\}$ is a sequence of MCMC sample sizes, and $\overline{\mathbf{H}}_k$ and $\overline{\mathbf{I}}_k$ are averages of $\mathbf{H}$ and $\mathbf{I}$ over the MCMC sample $\mathbf{b}^{(k;1)}, \ldots, \mathbf{b}^{(k;m_k)}$: $\overline{\mathbf{H}}_k = \sum_{j=1}^{m_k} \mathbf{H}(\widehat{\boldsymbol{\tau}}^{(k)}, \mathbf{b}^{(k;j)})/m_k$ and $\overline{\mathbf{I}}_k = \sum_{j=1}^{m_k} \mathbf{I}(\widehat{\boldsymbol{\tau}}^{(k)}, \mathbf{b}^{(k;j)})/m_k$.

We choose γ_k's and m_k's as follows. Fix three sufficiently large integers K_0, K, and m_0 where K_0 is the number of burn-in iterations. For the first $K_0 + K$ iterations, we set $\gamma_k = 1$ and $m_k = m_0 + k^2$. This corresponds to an MCMC procedure with the MCMC sample size increasing quadratically. The procedure with such choices of γ_k and m_k usually reaches the neighborhood of the estimates quickly. However, it converges slowly thereafter. To speed up the convergence, we switch to the following adaptive choices of γ_k and m_k after the first $K_0 + K$ iterations. Let $\widehat{r}_k$ be the correlation between the previous K estimates of $\boldsymbol{\tau}$ and their corresponding iteration numbers and $t_k = 1 - \widehat{r}_k^2$ for $k > K_0 + K$. Then, for $k > K_0 + K$, we set $\gamma_k = k^{-t_k}$ and $m_k = m_0 + k^{2(1-t_k)}$ (Jiang, Karcher, and Wang, 2011). Note that our choices of γ_k's and m_k's guarantee convergence (Gu and Kong, 1998). We use $K_0 = 300$, $K = 20$, and $m_0 = 300$ in all the analysis presented in the article. Our simulations with larger values of K_0, K, and m_0 (not presented) have not shown any significant improvement.

3.5 *The Complete Algorithm*

Combining pieces together, we have the following complete algorithm:

1. Provide initial values $\widehat{\boldsymbol{\tau}}^{(0)}$ and $\boldsymbol{\Gamma}^{(0)}$.
2. At iteration k,
   (a) generate MCMC samples $\mathbf{b}^{(k;1)}, \ldots, \mathbf{b}^{(k;m_k)}$ using the Metropolis–Hastings procedure in Web Appendix D.
   (b) Update $\mathbf{c}$ and $\mathbf{d}$ by solving (11) with $E_{b|t}(e^{\mathbf{z}_i^T \mathbf{b}})$ being approximated by $\widetilde{E}_{b|t}(e^{\mathbf{z}_i^T \mathbf{b}})$.
   (c) Update $\widehat{\boldsymbol{\tau}}$ and $\boldsymbol{\Gamma}$ by equations (15).
3. Repeat step 2 until convergence.

A simple choice for $\boldsymbol{\Gamma}^{(0)}$ is the identity matrix, which we have used throughout the article. $\widehat{\boldsymbol{\tau}}^{(0)}$ can be any reasonable user-specified value such as an estimate from a simpler model (e.g., the MLE ignoring covariates). Our algorithm has performed well with such choices of initial values.

4. Model Selection

In this section, we derive a model selection procedure based on the Kullback–Leibler geometry introduced in Gu (2004). This procedure can be used to identify negligible terms in an SS ANOVA model.

For two estimates f_1 and f_2 of the true log hazard function f_0, the empirical Kullback–Leibler distance between f_1 and f_2 is defined as

$$\begin{aligned} \mathrm{KL}(f_1, f_2) = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n_i+1} \widetilde{E}_{b|t}\{e^{\mathbf{z}_i^T \mathbf{b}}\} \\ \times \int_0^{t_{ij}} \left[e^{f_1(t,\mathbf{x}_i)}\{f_1(t,\mathbf{x}_i) - f_2(t,\mathbf{x}_i)\} \right. \\ \left. -(e^{f_1(t,\mathbf{x}_i)} - e^{f_2(t,\mathbf{x}_i)}) \right] dt, \end{aligned} \tag{16}$$

where $\widetilde{E}$ is the average over the MCMC sample generated at the last iteration of our algorithm in Section 3. Note that (16) essentially defines an empirical version of (12).

Suppose the estimation of f_0 has been done in a space $\mathcal{H}_1$ and we want to assess the possibility of reducing the model space to a subspace $\mathcal{H}_2 \subset \mathcal{H}_1$. Let $\widehat{f}$ be the estimate of f_0 in $\mathcal{H}_1$. Let $\widetilde{f}$ be the Kullback–Leibler projection of $\widehat{f}$ in $\mathcal{H}_2$, that is, the minimizer of $\mathrm{KL}(\widehat{f}, f)$ for $f \in \mathcal{H}_2$. Let f_c be the estimate from the model assuming that f is a constant independent of both t and $\mathbf{x}$. Then calculation in Web Appendix C yields

$$\mathrm{KL}(\widehat{f}, f_c) = \mathrm{KL}(\widehat{f}, \widetilde{f}) + \mathrm{KL}(\widetilde{f}, f_c).$$

Hence the ratio $\mathrm{KL}(\widehat{f}, \widetilde{f})/\mathrm{KL}(\widehat{f}, f_c)$ can be used to diagnose the suitability of the reduced model space $\mathcal{H}_2$. Specifically, the subspace $\mathcal{H}_2$ will be favored when $\mathrm{KL}(\widehat{f}, \widetilde{f})/\mathrm{KL}(\widehat{f}, f_c) < \kappa$. A small value $\kappa = 0.05$ has been used in Gu (2004); Du and Ma (2010); and Du, Ma, and Liang (2010). Our empirical study indicates that $\kappa = 0.05$ also works well for gap time hazard estimation.

5. Empirical Studies

In this section we present some simulations to evaluate the proposed estimation and model selection methods. We considered m subjects with one continuous covariate $x \in [0, 1]$. Denote $\mathcal{W}(\alpha(x), \beta(x))$ as the Weibull distribution with shape parameter $\alpha(x)$ and scale parameter $\beta(x)$. The corresponding log hazard function

$$f_0(x, t) = \log \alpha(x) + \{\alpha(x) - 1\} \log t - \alpha(x) \log \beta(x). \tag{17}$$

Note that f_0 is an additive model if $\alpha(x)$ is a constant.

We set $x_i = i/m$ for $i = 1, \ldots, m$ in all simulations. We generated the censoring times $D_1, \ldots, D_m$ as independent sample from Weibull distribution such that $D_i \sim \mathcal{W}(\alpha(x_i), c\beta(x_i))$ where the constant $c > 0$ was adjusted to achieve desired average number of observed recurrences per subject (denoted by $\#rec/sub$ from now on). For each fixed subject i with covariate x_i, gap times during the period $[0, D_i]$ were generated based on the shared frailty model $f_i(t, x_i) = f_0(t, x_i) + b_i$ where f_0 is given in (17) and random effects $b_i \overset{iid}{\sim} \mathrm{N}(0, 0.5^2)$.

We model the log hazard function $f(t, x)$ using the tensor product cubic splines in Example 1. The resulting SS ANOVA decomposition can be represented as

$$f(t, x) = f_\emptyset + f_1(t) + f_2(x) + f_{1,2}(t, x). \tag{18}$$

We consider two nested model spaces: the full model $\mathcal{H}_1$ that includes all components in (18), and the additive model $\mathcal{H}_2$ under the PH assumption tha excludes $f_{1,2}$ in (18). The space $\mathcal{H}_2$ is used in Section 5.1 and both spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are used in Section 5.2. We set $L = 10N^{2/9}$ in all simulations.

5.1 *Estimation Performance*

We evaluated the estimation performance using three sets of simulations. In these simulations, we considered an additive model with $\alpha(x) \equiv 3$ and $\beta(x) = \exp(6x(1 - x))$. Denote the SS ANOVA decomposition of f_0 in (17) for this setting as $f_0(t, x) = f_{\emptyset 0} + f_{10}(t) + f_{20}(x)$ where $f_{\emptyset 0} + f_{10}(t)$ corresponds to the baseline log hazard. It is not difficult to check that $f_{\emptyset 0} + f_{10}(t) = \log 3 - 3 + 2 \log t$ and $f_{20}(x) = -18x(1 -$

$x) + 3$. Note that, for identifiability, $\int_0^1 f_{20}(x)\, dx = 0$ which is satisfied by the SS ANOVA decomposition. The following three settings were used to assess the effects of sample size and censoring: (1) $m = 25$, $\#rec/sub \approx 3.0$, (2) $m = 50$, $\#rec/sub \approx 3.0$, and (3) $m = 50$, $\#rec/sub \approx 1.7$. The constant c in the censoring time distribution was fixed at 3.0 and 2.2, respectively, to achieve $\#rec/sub \approx 3.0$ and $\#rec/sub \approx 1.7$ here. We generated 500 data replicates for each setting.

For each data replicate, we fitted the nonparametric frailty model (1) with an additive model space $\mathcal{H}_2$ and computed the Kullback–Leibler distance $\mathrm{KL}(\widehat{f}, f_0)$ between the estimate $\widehat{f}$ and the true function f_0 as defined in (16). Under each setting, we selected three estimates whose $\mathrm{KL}(\widehat{f}, f_0)$'s are, respectively, the empirical 10th, 50th, and 90th percentiles of all the $\mathrm{KL}(\widehat{f}, f_0)$'s. For each of these estimates, we extracted the estimate of the baseline log hazard $\widehat{f}_\emptyset + \widehat{f}_1(t)$ and covariate effect $\widehat{f}_2(x)$, and plotted them in Figure 1 together with the true baseline log hazard $f_{\emptyset 0} + f_{10}(t)$ and covariate effect $f_{20}(t)$. Note that $\mathrm{KL}(\widehat{f}, f_0)$ measures the overall proximity of two bivariate functions $\widehat{f}(t, x)$ and $f_0(t,\, x)$. The ranking of $\mathrm{KL}(\widehat{f}, f_0)$ does not necessarily reflect the accuracy of the estimated components. Nevertheless, in all three sets of simulations, the fits of the log baseline hazard and covariate effects for the top 10% and 50% estimates are close to the true functions. Although the top 90% estimates are a little off, they are still within a reasonable range of the true function. Also, when sample size decreased or censoring rate increased, the estimation became more variable and less accurate as one would expect. We conclude that the proposed method provides good estimates to the overall hazard function as well as to its function components.

For the estimation of frailty parameter, the empirical quartiles of the $500\widehat{\sigma}$'s under the three settings were, respectively, (0.39, 0.55, 0.69), (0.42, 0.49, 0.68), and (0.30, 0.38, 0.59). As expected, the setting of $m = 50$ and $\#rec/sub \approx 3.0$ gave the most accurate estimate. Lower $\#rec/sub$ seemed to have a bigger impact on the estimation accuracy of σ than smaller number of subjects, possibly because the estimation of individual random effect was more difficult when $\#rec/sub \approx 1.7$.

5.2 *Model Selection Performance*

In this section, we evaluate the performance of the model selection procedure proposed in Section 4. We used a cutoff of $\kappa = 0.05$ for the ratio $\mathrm{KL}(\widehat{f}, \widetilde{f})/\mathrm{KL}(\widehat{f}, f_c)$ to determine the suitability of a reduced model. This corresponds to a 5% loss of structure when a larger model is reduced.

We carried out two groups of simulations for this purpose, with each group having the same three settings of m and $\#rec/sub$ as in Section 5.2. Gap times and censoring times in the first group were generated from the same distributions as in Section 5.1. In the second group of simulations, we used $\alpha(x) = 3 + 2\sin\{\pi(x + 1)\}$ and $\beta(x) \equiv 1$. Therefore, the true log hazard f_0 is additive in the first group of simulations and nonadditive in the second group of simulations. In the second group of simulations, the constant c in the censoring time distribution was fixed at 3.2 and 2.2, respectively to achieve $\#rec/sub \approx 3.0$ and $\#rec/sub \approx 1.7$. Five hundred data replicates were generated for each of the six combinations of three settings and two groups. For each data replicate, we fitted with a complete model $\mathcal{H}_1$, an additive model $\mathcal{H}_2$, and a model without any covariate effects. Thus for each replicate, two ratios of Kullback–Leibler distances were recorded and compared with $\kappa = 0.05$ to determine the suitability of the corresponding reduced models. The results are reported in Table 1 in the form of proportions of selecting main covariate effect x and interaction $t : x$, and proportions of under-fits defined as final models missing a covariate effect (main effect or interaction), correct fits, and over-fits defined in the true additive model simulations as selecting the interaction.

Table 1 shows good performance of the proposed model selection procedure. In the additive simulations, the main covariate effect was always selected and the interaction was incorrectly selected only in less than 10% of the cases. In the nonadditive simulations, the proportions of identifying the correct model was above 85% in all the settings and the proportion of under-fit was less than 15% in all settings. Also, we can notice a boost of performance in both groups of simulations when the number of subjects or the $\#rec/sub$ is increased.

6. Application to Bladder Tumor Data

We consider a recurrent event data set from a bladder tumor cancer study reported by Byar (1980). Between November 1971 and August 1976, 121 bladder tumor patients from 10 Veteran's Administration hospitals were admitted to the study according to the criteria described in Byar (1980). All patients had stage I bladder tumors at the time they entered the study but these tumors were completely removed by transurethral resection. Patients were then randomly assigned to one of three treatments: placebo, pyridoxine, or thiotepa. Patients were examined cystoscopically every three months for recurrence of tumor and any new tumors were removed. In all three treatment groups, the average follow-up was about 31 months, but some cases may have been followed as long as five years. A recurrence in this study was defined as a visit in which one or more tumors had reappeared in the bladder after having been previously removed completely by transurethral resection. We use the data consisting of the placebo and thiotepa groups that are publicly available. There were 48 patients in the placebo group with a total of 87 observed tumor recurrences and 38 patients in the thiotepa group with a total of 45 observed tumor recurrences. Besides the treatment and number of months to the event since last tumor occurrence, the number of initial tumors and diameter of the largest initial tumor were also recorded for each patient. The goal is to investigate covariate effect on gap time.

Denote by $x_{\langle t\rangle}$, $x_{\langle n\rangle}$, and $x_{\langle s\rangle}$ the three covariates: treatment, number of initial tumors, and size of the largest initial tumor in diameter. Then $\mathbf{x} = (x_{\langle t\rangle},\, x_{\langle n\rangle},\, x_{\langle s\rangle})$. We used the tensor product smoothing splines with $L = 10(207)^{2/9} = 33$ knots noting that $N = 207$. We first considered the following shared frailty model

$$f(t_{ij}, \mathbf{x}_i) = f + f_1(t_{ij}) + f_2(x_{i\langle t\rangle}) + f_3(x_{i\langle n\rangle}) + f_4(x_{i\langle s\rangle}) + f_{1,2}(t_{ij}, x_{i\langle t\rangle}) + f_{1,3}(t_{ij}, x_{i\langle n\rangle}) + f_{1,4}(t_{ij}, x_{i\langle s\rangle}) + b_i,$$

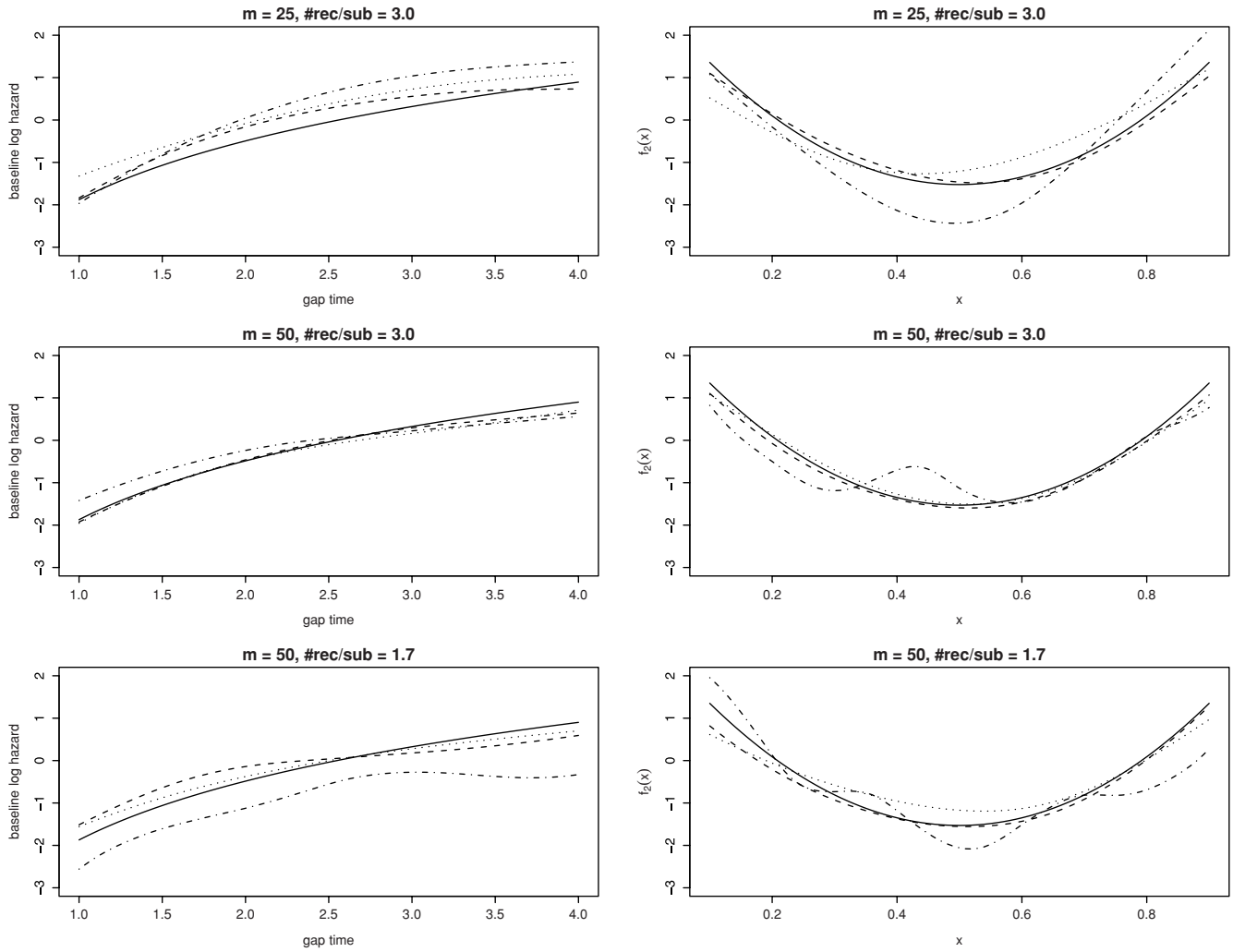

Figure 1. Hazard estimation performance: plots of estimates of baseline log hazard $f_\emptyset + f_1(t)$ and covariate effect function $f_2(x)$ from data with different numbers of subjects m and levels of censoring. Solid lines are true function values; dashed lines, dotted lines, and dash-dotted lines are estimated values from, respectively, the top 10%, 50%, and 90% function estimates in terms of their empirical KL distances to the true function.

Table 1

Performance of model selection (Section 5.2). Under-fit refers to the case when final model misses at least one effect in the true log hazard model, correct-fit refers to the case when final model matches the true model, and under-fit refers to the case when final model contains all the effects in the true model and at least one additional effect not in the true model.

	Proportion of selecting		Proportion of		
Setting	x	t:x	Under-fit	Correct-fit	Over-fit
True model: $\alpha(x) \equiv 3$, $\beta(x) = \exp(6x(1-x))$					
$m = 25$, $\#rec/sub \approx 3.0$	1.000	0.081	0.000	0.919	0.081
$m = 50$, $\#rec/sub \approx 3.0$	1.000	0.032	0.000	0.968	0.032
$m = 50$, $\#rec/sub \approx 1.7$	1.000	0.094	0.000	0.906	0.094
True model: $\alpha(x) = 3 + 2\sin(\pi\,(x+1))$, $\beta(x) \equiv 1$					
$m = 25$, $\#rec/sub \approx 3.0$	0.947	0.865	0.135	0.865	-
$m = 50$, $\#rec/sub \approx 3.0$	0.990	0.984	0.016	0.984	-
$m = 50$, $\#rec/sub \approx 1.7$	0.980	0.907	0.093	0.907	-

where random effects $b_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$. After applying the model selection procedure in Section 4 to all potential submodels using $\kappa = 0.05$, we ended up with the final model

$$f(t_{ij}, \mathbf{x}_i) = f_\emptyset + f_1(t_{ij}) + f_2(x_{i\langle t\rangle}) + f_3(x_{i\langle n\rangle}) + f_{1,2}(t_{ij}, x_{i\langle t\rangle}) + b_i. \quad (19)$$

Note that the final model (19) contains a gap-time-treatment interaction $f_{1,2}(t_{ij}, x_{i\langle t\rangle})$, indicating a possible violation of the PH assumption. Many earlier studies on these data indicated a negligible effect of $x_{\langle s\rangle}$ and a significant effect of $x_{\langle n\rangle}$, and some studies reported significance or marginal significance of the treatment $x_{\langle t\rangle}$; see, e.g., Kalbfleisch and Prentice (2002), Peña et al. (2007), and references therein. These are all reflected in our final model (19).

To assess the variability of the estimation, we introduce a bootstrap procedure to compute confidence interval estimates. Given the gap time log hazard estimate $\widehat{f}$ and the frailty variance estimate $\widehat{\sigma}^2$, the bootstrap procedure repeated the following steps $B = 1000$ times.

(1) Generate $m = 86$ frailties b_i^*, $i = 1, \ldots, 86$ from the distribution $\mathrm{N}(0, \widehat{\sigma}^2)$.
(2) For each subject, compute the gap time log hazard function as $b_i^* + \widehat{f}(t, \mathbf{x}_i)$ and its corresponding quantile function $q_i^*(\cdot)$. Generate gap times as $q_i^*(U_{ij})$, where U_{ij}'s are i.i.d. Uniform$(0, 1)$ random variables. Censor the gap times using the observed censoring time for the subject. These censored gap times together with the covariates form the observation for this subject in the resampling.
(3) Apply the proposed method to obtain the estimates $(\widehat{f}^*, \widehat{\sigma}^*)$ for the resample.

The corresponding empirical quantiles of all the resample estimates, $(\widehat{f}_k^*, \widehat{\sigma}_k^*), k = 1, \ldots, B$, were then used to construct confidence intervals for σ and f. González, Peña, and Delicado (2010) proposed a similar semiparametric bootstrap procedure for recurrent data without covariate where nonparametric estimate of survivor function and parametric estimate of frailty parameter were used. Our procedure generalizes their approach to the setting with covariate where nonparametric hazard estimate is a function of both gap time and covariate.

We computed the hazard estimates at a grid of gap time for both patient groups with fixed $x_{\langle n\rangle} = 2$, and at a grid of $x_{\langle n\rangle}$ for both patient groups with gap time fixed at 7 months. These estimates are plotted in Figure 2 as solid lines, with their corresponding pointwise 95% confidence intervals imposed as dashed lines. The top plots show clearly different and nonlinear gap time hazard trends for the two patient groups, indicating the interaction between gap time and treatment is not negligible. For the placebo group, hazard increased with gap time when gap time was within 10 months, dropped quickly when gap time was between 10 and 20 months, and then leveled off afterwards. For the thiotepa group, hazard displayed an overall decreasing trend, dropping quickly before gap time reached 10, stayed flat or showed a slight increase between 10 and 20 months, and then slowly decreased afterwards.

To see whether the inclusion of frailty was worthwhile, we also fitted the model (19) without the frailties. The fitted hazard rates were imposed in Figure 2 as dotted lines. These two sets of hazard estimates are slightly different. Also the estimated frailty standard deviation $\widehat{\sigma} = 0.581$ with a 95% confidence interval (0.313, 0.839) in model (19), confirming a fair degree of frailty effect.

For the record, our final model (19) took 47.3 CPU minutes, on a Dell PowerEdge 2950 workstation with dual Xeon dual core 3.0GHz CPUs and 16 GB RAMs running openSUSE 11.0 and R 2.9.2.

7. Discussion

For simplicity of notation we have limited our consideration to the situation where covariates are fixed for each subject. Consequently, conditional on the subject-wise frailty b_i, the gap times are limited to be identically distributed. However, this restriction can be relaxed to allow the covariates to change at each recurrence. For example, in some clinical trials, upon each recurrent event doctors will review the current treatment for a patient and may decide to change the treatment. In this case, the treatment will be a recurrence-specific covariate. Extensions to allow recurrence-specific covariates, and thus nonidentically distributed gap times conditional on frailty, is straightforward. It only requires the change of notation from $\mathbf{x}_i$ to $\mathbf{x}_{ij}$ in the description and minor changes in the program.

One interesting future research direction is to extend the modeling and estimation methods in this article to calendar time models. There are two possible directions, namely marginal model and intensity model methods, as described in Strawderman (2005). Marginal methods focus on the cumulative mean function of recurrences with Lin et al. (2000) as an example. In this case the function of interest is naturally defined on the calendar time domain. Intensity methods focus on the dependence of the probability of subsequent recurrences on the past event history. An example of intensity method whose target function is defined on the calendar time domain is Chang and Wang (1999). Extension of our methods in either direction would lead to a likelihood function that is much more complicated for optimization. Further research is merited.

The smoothing parameter in our method is selected to optimize a score (13) derived from delete-one-observation CV. In Du (2009), a delete-one-subject CV approximation was used. Our empirical experiments show little difference in the smoothing parameters selected by these two approaches. Although delete-one-subject CV seems more appropriate, our experience with both approximations suggests that it actually has two practical drawbacks: First, the score derived from delete-one-subject CV is hard to compute since it requires us to store more big matrices (subject-wise matrices). This consumes a large trunk of memory and slows down the computation quite a bit. Second, it may become unreliable in applications when a few subjects have many more recurrences than the others. Deleting one of these subjects can have a big impact on the estimation.

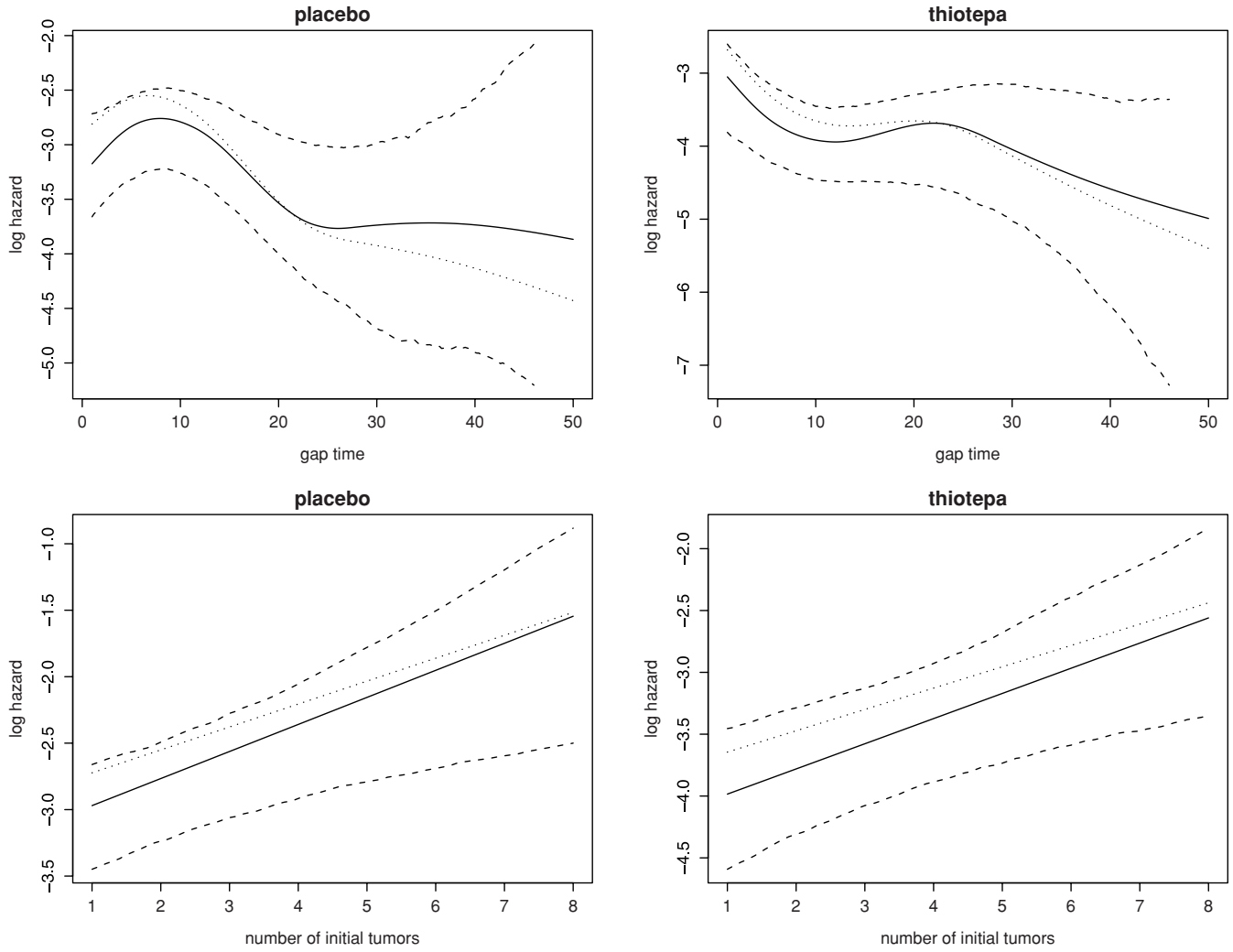

Figure 2. Hazard estimates for bladder tumor cancer data. Left: Placebo group. Right: Thiotepa group. Top: log gap time hazard when number of initial tumors is fixed at 2. Bottom: log gap time hazard versus number of initial tumors at $t = 7$ months. Solid lines are the estimates from model (19) with frailties, dotted lines are the estimates from model (19) without frailties, and dashed lines are the 95% bootstrap point-wise confidence interval estimates from model (19) with frailties.

8. Supplementary Materials

Web Appendices A-D, referenced in Sections 3 and 4, are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors thank the editor, the associate editor, and two referees for their insightful comments that have helped improve the article. Du's research is supported by NSF DMS-1007126. Wang's research was supported by NSF DMS-0706886.

References

Anderson, J. A. and Senthilselvan, A. (1980). Smooth estimates for the hazard function. *Journal of the Royal Statistical Society, Series B* **42,** 322–327.

Byar, D. P. (1980). The Veterans Administration study of chemoprophylaxis for recurrent stage I bladder tumors: Comparisons of placebo, pyridoxine, and topical thiotepa. In M. Pavone-Macaluso, P. H. Smith, and F. Edsmyn (eds.), *Bladder Tumors and Other Topics in Urological Oncology*, 363–370. New York: Plenum.

Chang, S.-H. and Wang, M.-C. (1999). Conditional regression analysis for recurrence time data. *Journal of the American Statistical Association* **94,** 1221–1230.

Clement, D. Y. and Strawderman, R. L. (2009). Conditional GEE for recurrent event gap times. *Biostatistics* **10,** 451–467.

Cook, R. J. and Lawless, J. F. (2009). *The Statistical Analysis of Recurrent Events*. New York: Springer.

Du, P. (2009). Nonparametric modeling of the gap time in recurrent event data. *Lifetime Data Analysis* **15,** 256–277.

Du, P. and Ma, S. (2010). Frailty model with spline estimated nonparametric hazard function. *Statistica Sinica* **20,** 561–580.

Du, P., Ma, S., and Liang, H. (2010). Penalized variable selection procedure for Cox models with semiparametric relative risk. *Annals of Statistics* **38,** 2092–2117.

González, J. R., Peña, E. A., and Delicado, P. (2010). Confidence intervals for median survival time with recurrent event data. *Computational Statistics and Data Analysis* **54,** 78–89.

Gu, C. (1996). Penalized likelihood hazard estimation: A general procedure. *Statistica Sinica* **6,** 861–876.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer-Verlag.

Gu, C. (2004). Model diagnostics for smoothing spline ANOVA models. *Canadian Journal of Statistics* **32,** 347–358.

Gu, M. G. and Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte Carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences* **95,** 7270–7274.

Huang, Y. (2002). Censored regression with the multistate accelerated sojourn times model. *Journal of the Royal Statistical Society, Series B* **64,** 17–29.

Jiang, Y., Karcher, P., and Wang, Y. (2011). On implementation of the Markov chain Monte Carlo stochastic approximation algorithm. In M. T. Wells and A. Sengupta (eds.), *Advances in Directional and Linear Statistics: A Festschrift for Sreenivasa Rao Jammalamadaka*, 97–111. Berlin: Springer-Verlag.

Joly, P., Commenges, D., and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: Application to age-specific incidence of dementia. *Biometrics* **54,** 185–194.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. New York: Wiley.

Kim, Y.-J. and Gu, C. (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society, Series B* **66,** 337–356.

Lin, D. Y., Sun, W., and Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika* **86,** 59–70.

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B* **62,** 711–730.

O'Sullivan, F. (1993). Nonparametric estimation in the Cox model. *Annals of Statistics* **21,** 124–145.

Peña, E., Strawderman, R., and Hollander, M. (2001). Nonparametric estimation with recurrent event data. *Journal of the American Statistical Association* **96,** 1299–1315.

Peña, E., Slate, E. H., and González, J. R. (2007). Semiparametric inference for a general class of models for recurrent events. *Journal of Statistical Planning and Inference* **137,** 1727–1747.

Rondeau, V., Commenges, D., and Joly, P. (2003). Maximum penalized likelihood estimation in a gamma-frailty model. *Lifetime Data Analysis* **9,** 139–153.

Rondeau, V., Filleul, L., and Joly, P. (2006). Nested frailty models using maximum penalized likelihood estimation. *Statistics in Medicine* **25,** 4036–4052.

Rondeau, V., Mathoulin-Pelissier, S., Jacqmin-Gadda, H., Brouste, V., and Soubeyran, P. (2007). Joint frailty models for recurring events and death using maximum penalized likelihood estimation: Application on cancer events. *Biostatistics* **8,** 708–721.

Strawderman, R. L. (2005). The accelerated gap times model. *Biometrika* **92,** 647–666.

Wahba, G. (1990). *Spline Models for Observational Data*, volume 59 of CBMS-NSF *Regional Conference Series in Applied Mathematics*. Philadelphia: SIAM.

Wang, M.-C. and Chang, S.-H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association* **94,** 146–153.

Wang, Y. (1998). Mixed-effects smoothing spline ANOVA. *Journal of the Royal Statistical Society, Series B* **60,** 159–174.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*. New York: Chapman and Hall.

Yau, K. K. W. and McGilchrist, C. A. (1998). ML and REML estimation in survival analysis with time dependent correlated frailty. *Statistics in Medicine* **17,** 1201–1213.

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,** 329–353.

Received June 2010. *Revised January* 2011.
Accepted January 2011.