Statistica Sinica **20** (2010), 561-580

FRAILTY MODEL WITH SPLINE ESTIMATED NONPARAMETRIC HAZARD FUNCTION

Pang Du and Shuangge Ma

Virginia Tech and Yale University

Abstract: Frailty has been introduced as a group-wise random effect to describe the within-group dependence for correlated survival data. In this article, we propose a penalized joint likelihood method for nonparametric estimation of hazard function. With the proposed method, the frailty variance component and the smoothing parameters become the tuning parameters that are selected to minimize a loss function derived from the Kullback-Leibler distance through delete-one cross-validation. Confidence intervals for the hazard function are constructed using the Bayes model of the penalized likelihood. Combining the functional ANOVA decomposition and the Kullback-Leibler geometry, we also derive a model selection tool to assess the covariate effects. We establish that our estimate is consistent and its nonparametric part achieves the optimal convergence rate. We investigate finite sample performance of the proposed method with simulations and data analysis.

Key words and phrases: Bayesian confidence intervals, cross-validation, frailty, hazard, model selection, penalized likelihood.

1. Introduction

With grouped survival data, the correlation between subjects within the same group cannot be ignored. To model this dependence, a common approach is to introduce a group-wise random effect called frailty. The frailty models have been widely used in analyzing different types of grouped survival data. Examples include Sastry (1997), Fine, Glidden, and Lee (2003), and Therneau, Grambsch, and Pankratz (2003). In this article, we consider the shared frailty model, where all subjects within the same group share a common frailty and the frailties of different groups are independent.

The proportional hazards model has been commonly adopted in the study of frailty. As a semiparametric approach, the proportional structure essentially assumes a linear covariate effect on the log hazard function. When this assumption is violated, the corresponding inferences become questionable. Moreover, estimation with such frailty models relies on EM algorithms, whose convergence is often slow, and further computation is needed for variance estimates; see Therneau et al. (2003). These difficulties motivate development of the penalized likelihood method in this article.

In the absence of frailty, hazard estimation using penalized likelihood has been studied. Some recent examples are Gu (1996), Joly, Commenges, and Letenneur (1998), Joly and Commenges (1999), and Du and Gu (2006). In those studies, the hazard estimate was obtained as the optimizer of a penalized likelihood function, which consists of a likelihood part representing the goodness-of-fit, a roughness penalty specifying the smoothness, and the smoothing parameters balancing the tradeoff. Compared with their parametric and semiparametric peers, nonparametric hazard estimates possess certain advantages. First, they are more robust against violations of parametric assumptions commonly made in parametric and semiparametric models. Second, they are usually smooth functions, which are appealing for practitioners. Last, they can be more reliable exploratory tools for selecting appropriate models.

Random effects have been introduced to penalized likelihood methods in regression problems; see, e.g., Wang (1998), Ke and Wang (2001), and Gu and Ma (2005). However, much less attention has been paid to penalized likelihood hazard estimation with frailties. One exception was Rondeau, Commenges, and Joly (2003), where a penalized marginal likelihood with frailty integration was optimized to obtain the hazard estimate. This is in the line of Wang (1998) and Ke and Wang (2001). A link between the penalized partial likelihood and the proportional hazards frailty model was noted in Therneau et al. (2003).

In this paper, we propose a penalized joint likelihood hazard estimation method. The joint likelihood consists of the conditional likelihood given frailties and the likelihood of frailties, with the latter acting as a penalty on frailties. The penalized joint likelihood penalizes both the frailties and the hazard function, earning the name "doubly penalized" estimation (Lin and Zhang (1999)). The hazard function and the frailties are then estimated through the minimization of the penalized joint likelihood. The variance component of the frailties is treated as an additional tuning parameter, and jointly selected with the smoothing parameters by minimizing a loss function derived through delete-one cross-validation. To assess variability of the hazard estimate, point-wise confidence intervals are constructed using the Bayes model for penalized likelihood. To assess the covariate effects, a model selection tool is developed based on the functional ANOVA decomposition and Kullback-Leibler geometry. We also establish the asymptotic properties of the estimates.

The rest of the article is organized as follows. Section 2 gives details of the proposed penalized likelihood method, including the model (Sec. 2.1), the estimation procedure (Sec. 2.2), smoothing parameter selection (Sec. 2.3), confidence intervals (Sec. 2.4), model selection (Sec. 2.5), and asymptotic properties (Sec. 2.6). Section 3 is devoted to simulation studies, where the cross-validation score, Bayesian confidence intervals, and model selection tool are examined. In

Section 4, we apply the proposed methods to studies on lung and prostate cancer. Remarks in the last section conclude the article.

2. Penalized Likelihood Frailty Model

2.1. Model

In a survival study, one usually observes $X = \min(T, C)$ and $\delta = I_{[T \leq C]}$ for a subject, where T is the failure time and C is the right-censoring time. Sometimes a left truncation time Z is also present, representing the time when the subject entered the study. Let U be the covariate. In studies where subjects are divided into p groups, a vector $\mathbf{b} = (b_1, \ldots, b_p)^T$ of unknown group frailties is used to represent the heterogeneity across groups and capture the correlation between subjects within the same group. Assume that T, C, and Z are mutually independent conditional on U and $\mathbf{b}$. Suppose $T|(U, \mathbf{b})$ follows a survival function $S(t, u; \mathbf{b})$. One is interested in estimation of the hazard function $h(t, u; \mathbf{b}) = -\partial \log S(t, u; \mathbf{b})/\partial t$. Here the frailties $\mathbf{b} \sim N(0, B)$ are unknown additive random effects on the log hazard function, i.e.,

$$\log h(t, u; \mathbf{b}) = \eta(t, u) + \mathbf{z}^T \mathbf{b},$$

where $\mathbf{z}$ is the group indicator vector.

The observations are $(Z_i, X_i, \delta_i, U_i, \mathbf{z}_i)$, $i = 1, \ldots, n$. Let $\Sigma = B^{-1}$. A common choice of Σ is $\sigma^{-2} I$, which corresponds to independent normal frailties with unknown variance σ^2. We propose to estimate η and $\mathbf{b}$ jointly through minimization of

$$-\frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i (\eta(X_i, U_i) + \mathbf{z}_i^T \mathbf{b}) - \int_{Z_i}^{X_i} e^{\eta(t, U_i) + \mathbf{z}_i^T \mathbf{b}} dt \right\} + \frac{1}{2n} \mathbf{b}^T \Sigma \mathbf{b} + \frac{\lambda}{2} J(\eta). \quad (2.1)$$

Here the first two terms form the negative log joint likelihood of η and $\mathbf{b}$. $J(\eta)$ is a roughness penalty, and the smoothing parameter λ controls the trade-off between the goodness-of-fit and the smoothness of η.

In survival analysis, η is often a function of time and covariate. With a generic covariate u, a functional ANOVA decomposition of η is

$$\eta(t, u) = \eta_0 + \eta_t(t) + \eta_u(u) + \eta_{t,u}(t, u), \quad (2.2)$$

where η_0 is a constant, η_t is the main effect of time t, η_u is the main effect of covariate u, and $\eta_{t,u}(t, u)$ is the interaction between time and covariate. When $\eta_{t,u} = 0$, (2.2) reduces to an additive model for η, or the well-known proportional hazards model. Various side conditions through averaging operators are needed to ensure the identifiability of the terms in (2.2); see Wahba (1990) and Gu (2002).

2.2. Estimation

Consider the Hilbert space $\mathcal{H} = \{\eta : J(\eta) < \infty\}$ on the product domain $\mathcal{T} \times \mathcal{U}$ of time and covariate, in which J is a square semi-norm. Assume that the evaluation functional $[(t,u)]f = f(t,u)$ is continuous in $\mathcal{H}$. Then $\mathcal{H}$ becomes a reproducing kernel Hilbert space. It possesses a reproducing kernel $R(\cdot,\cdot)$, a nonnegative definite function with the reproducing property that $\langle R((t,u),\cdot), f(\cdot)\rangle = f(t,u)$ for any $f \in \mathcal{H}$, where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{H}$. Let $\mathcal{N}_J$ be the null space of J in $\mathcal{H}$. Then $\mathcal{H}$ can be decomposed into a tensor sum $\mathcal{N}_J \oplus \mathcal{H}_J$, with $\mathcal{H}_J$ possessing a reproducing kernel $R_J(\cdot,\cdot)$.

Although $\mathcal{H}$ is infinite dimensional, in practice, minimization of (2.1) is performed in a data-adaptive finite dimensional space. This can be done because asymptotically the minimizer of (2.1) belongs to the same Sobolev space as the true unknown parameter. Gu (1996) considered $\mathcal{H}_n = \mathcal{N}_J \oplus \text{span}\{R_J((X_j,U_j),\cdot) : \delta_j = 1, j = 1,\ldots,n\}$. Du and Gu (2006) considered a much smaller space $\mathcal{H}_q = \mathcal{N}_J \oplus \text{span}\{R_J(v_j,\cdot) : j = 1,\ldots,q\}$, where $\{v_j\}_{j=1}^q$ is a random subset of $\{(X_i,U_i) : \delta_i = 1, i = 1,\ldots,n\}$ and $q \to \infty$ at a much slower rate of n (e.g., $q \asymp n^{2/9}$ suffices for cubic splines). Without loss of generality, we use the expression

$$\eta(t,u) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(t,u) + \sum_{j=1}^{q} c_j R_J(v_j,(t,u)) = \boldsymbol{\phi}^T\mathbf{d} + \boldsymbol{\xi}^T\mathbf{c}, \tag{2.3}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are vectors of basis functions in, respectively, $\mathcal{N}_J$ and $\mathcal{H}_q \ominus \mathcal{N}_J$, and $\boldsymbol{d}$ and $\boldsymbol{c}$ are vectors of coefficients. Substituting (2.3) into (2.1), one calculates the minimizer of (2.1) in $\mathcal{H}_q$ by minimizing

$$\begin{aligned} &A_\lambda(\boldsymbol{b},\boldsymbol{c},\boldsymbol{d}) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i\Big(\boldsymbol{\phi}_i^T\boldsymbol{d} + \boldsymbol{\xi}_i^T\boldsymbol{c} + \boldsymbol{z}_i^T\boldsymbol{b}\Big) - \int_{Z_i}^{X_i} \exp\Big(\boldsymbol{\phi}(t,U_i)^T\boldsymbol{d} + \boldsymbol{\xi}(t,U_i)^T\boldsymbol{c} + \boldsymbol{z}_i^T\boldsymbol{b}\Big)dt\right\} \\ &\quad + \frac{1}{2n}\boldsymbol{b}^T\Sigma\boldsymbol{b} + \frac{\lambda}{2}\boldsymbol{c}^T Q\boldsymbol{c} \end{aligned} \tag{2.4}$$

with respect to $\boldsymbol{d}$, $\boldsymbol{c}$ and $\boldsymbol{b}$, where $\boldsymbol{\phi}_i$ is $m \times 1$ with the νth entry $\phi_\nu(X_i,U_i)$, $\boldsymbol{\xi}_i$ is $q \times 1$ with the kth entry $\xi_k(X_i,U_i)$, and Q is $q \times q$ with the (j,k)th entry $R_J(v_j,v_k)$.

Let $\Theta = (\eta,\boldsymbol{b})$, and write $\mu_\Theta(g) = (1/n)\sum_{i=1}^n \int_{Z_i}^{X_i} g(t,U_i)e^{\eta(t,U_i)+\boldsymbol{z}_i^T\boldsymbol{b}}dt$, $V_\Theta(g,h) = \mu_\Theta(gh)$, $V_\Theta(g) = V_\Theta(g,g)$, and $e(g) = (1/n)\sum_{i=1}^n \delta_i g(X_i,U_i)$. For a function z with $z(t, U_i) = z_i$, we define $\mu_\Theta(z)$, $V_\Theta(z, z)$, $V_\Theta(z, h)$, $V_\Theta(g, z)$, and $e(z)$ similarly except that the terms $g(t, U_i)$, $h(t, U_i)$, and $g(X_i, U_i)$ are replaced by the constant z_i. With the smoothing parameters determined using the procedure in Sec. 2.3, (2.4) is then minimized through Newton iterations with the current estimate $\tilde{\Theta} = (\tilde{\eta}, \tilde{\boldsymbol{b}})$ updated by

$$\begin{pmatrix} V_{z,z}+\frac{\Sigma}{n} & V_{z,\xi} & V_{z,\phi} \\ V_{\xi,z} & V_{\xi,\xi}+\lambda Q & V_{\xi,\phi} \\ V_{\phi,z} & V_{\phi,\xi} & V_{\phi,\phi} \end{pmatrix} \begin{pmatrix} \boldsymbol{b} \\ \boldsymbol{c} \\ \boldsymbol{d} \end{pmatrix} = \begin{pmatrix} e_z - \mu_z + V_{z,\Theta} \\ e_\xi - \mu_\xi + V_{\xi,\Theta} \\ e_\phi - \mu_\phi + V_{\phi,\Theta} \end{pmatrix}, \tag{2.5}$$

where, with g and h running through (z, ξ, ϕ), $V_{g,h} = V_{\tilde{\Theta}}(\boldsymbol{g}, \boldsymbol{h}^T)$, $\mu_g = \mu_{\tilde{\Theta}}(\boldsymbol{g})$, $e_g = e(\boldsymbol{g})$, and $V_{g,\Theta} = V_{\tilde{\Theta}}(\boldsymbol{g}, \tilde{\eta} + \boldsymbol{z}^T\tilde{\boldsymbol{b}})$.

2.3. Smoothing parameter selection

With varying smoothing parameters $\Lambda = (\lambda, \Sigma)$, the minimizers of (2.1) define an array of possible estimates. Define the Kullback-Leibler distance between the true $\Theta = (\eta, \mathbf{b})$ and the estimate $\Theta_\Lambda = (\eta_\Lambda, \mathbf{b}_\Lambda)$ as

$$\begin{aligned} \mathrm{KL}((\eta, \mathbf{b}), (\eta_\Lambda, \mathbf{b}_\Lambda)) = E\bigg[\int_{\mathcal{T}} Y(t) \Big\{ &(\eta(t, U) + \mathbf{z}^T\mathbf{b} - \eta_\Lambda(t, U) - \mathbf{z}^T\mathbf{b}_\Lambda \Big) e^{\eta(t,U)+\mathbf{z}^T\mathbf{b}} \\ &-(e^{\eta(t,U)+\mathbf{z}^T\mathbf{b}} - e^{\eta_\Lambda(t,U)+\mathbf{z}^T\mathbf{b}_\Lambda}) \Big\} dt \bigg], \end{aligned} \tag{2.6}$$

where the expectation is with respect to Z, X, U, and $\mathbf{b}$. Estimating (2.6) through delete-one cross-validation and the counting process theory for frailty model, one ends up with the score

$$\begin{aligned} V_\alpha(\Lambda) = &-\frac{1}{n} \sum_{i=1}^{n} \Big\{ \delta_i \Big(\eta_\Lambda(X_i, U_i) + \boldsymbol{z}_i^T \boldsymbol{b}_\Lambda \Big) - \int_{Z_i}^{X_i} e^{\eta_\Lambda(t,U_i)+\boldsymbol{z}_i^T\boldsymbol{b}_\Lambda} dt \Big\} \\ &+\alpha \frac{1}{n(n-1)} \sum_{i=1}^{n} \Big\{ \delta_i \boldsymbol{\psi}(X_i, U_i)^T H^{-1} \Big(\boldsymbol{\psi}(X_i, U_i) - \frac{K\mathbf{1}}{n} \Big) \Big\}, \end{aligned} \tag{2.7}$$

which is minimized to select $\Lambda = (\lambda, \Sigma)$. Here H is the Hessian matrix on the left side of (2.5), K is a $(p + q + m) \times n$ matrix with columns $\boldsymbol{\psi}(X_i, U_i) = (\boldsymbol{z}_i^T, \boldsymbol{\xi}(X_i, U_i)^T, \boldsymbol{\phi}(X_i, U_i)^T)^T$, and $\mathbf{1}$ is a $(p + q + m)$-vector of all 1's. See Appendix A (online) for detailed derivation. In (2.7), the constant $\alpha > 1$ is added to prevent occasional under-smoothing. Too big an α may introduce too much bias into V_α, thus an α between 1 and 2 is often used in practice. When frailty is absent, an α around 1.4 was suggested by Gu (2002) to ensure little loss of effectiveness. A similar range is indicated by our empirical studies in the presence of frailty.

2.4. Bayesian confidence intervals

In this section, we give confidence intervals for the hazard estimate of (2.1) using the Bayes model of smoothing splines (Wahba (1983)). To connect (2.1) with Bayes models, consider $\eta = \eta_0 + \eta_1$, where η_0 has a diffuse prior in $\mathcal{N}_J$ and η_1 has a Gaussian process prior with mean zero and covariance function $E[\eta_1(x_1)\eta_1(x_2)] = (1/(n\lambda))R_J(x_1, \boldsymbol{v}^T)Q^+R_J(\boldsymbol{v}, x_2)$, $x_1 = (t_1, u_1), x_2 = (t_2, u_2) \in \mathcal{T} \times \mathcal{U}$, $Q = R_J(\boldsymbol{v}, \boldsymbol{v}^T)$ is as defined in (2.4), and Q^+ is the Moore-Penrose inverse of Q. Under these priors, the posterior mean and variance of $\eta(x) + \boldsymbol{z}^T\boldsymbol{b}$ can be approximated, respectively, by $\eta_\Lambda(x) + \boldsymbol{z}^T\boldsymbol{b}_\Lambda$ and $\boldsymbol{\psi}^T(x)H^{-1}\boldsymbol{\psi}(x)/n$, where $\boldsymbol{\psi}$ and H are defined in Section 2.3; see Appendix B (online) for a detailed proof. Hence, for any given $(x, \boldsymbol{z})$, the $100(1-\alpha)\%$ confidence interval for $\eta(x) + \boldsymbol{z}^T\boldsymbol{b}$ is

$$\Big(\eta_\Lambda(x) + \boldsymbol{z}^T\boldsymbol{b}_\Lambda\Big) \pm z_{\alpha/2}\sqrt{\boldsymbol{\psi}^T(x)H^{-1}\boldsymbol{\psi}(x)n^{-1}},$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. Setting $\boldsymbol{z} = \boldsymbol{0}$ yields the confidence interval for $\eta(x)$.

2.5. Model selection

In this section, we develop a model selection tool based on the Kullback-Leibler geometry and use it to assess the covariate effect. Suppose the estimation of η has been done in space $\mathcal{H}_1$, but in fact $\eta \in \mathcal{H}_2 \subset \mathcal{H}_1$. Let $\hat{\eta}$ be the estimate of η in $\mathcal{H}_1$ and $\hat{\boldsymbol{b}}$ be the corresponding estimate of frailties. Treating $\hat{\boldsymbol{b}}$ as an offset in the model, we define the Kullback-Leibler distance between two log hazard estimates η_1 and η_2 as

$$\begin{aligned}\mathrm{KL}(\eta_1, \eta_2) = \frac{1}{n}\sum_{i=1}^{n}\int_{Z_i}^{X_i} \Big\{ & e^{\eta_1(t,U_i)+\boldsymbol{z}_i^T\hat{\boldsymbol{b}}}\Big(\eta_1(t, U_i) + \boldsymbol{z}_i^T\hat{\boldsymbol{b}} - \eta_2(t, U_i) - \boldsymbol{z}_i^T\hat{\boldsymbol{b}}\Big) \\ & -\Big(e^{\eta_1(t,U_i)+\boldsymbol{z}_i^T\hat{\boldsymbol{b}}} - e^{\eta_2(t,U_i)+\boldsymbol{z}_i^T\hat{\boldsymbol{b}}}\Big)\Big\}dt. \end{aligned} \tag{2.8}$$

Let $\tilde{\eta}$ be the Kullback-Leibler projection of $\hat{\eta}$ in $\mathcal{H}_2$ (i.e., the minimizer of $\mathrm{KL}(\hat{\eta}, \eta)$ for $\eta \in \mathcal{H}_2$), and η_c be the estimate from the constant model. Set $\eta = \tilde{\eta} + \alpha(\tilde{\eta} - \eta_c)$ for α real. Differentiating $\mathrm{KL}(\hat{\eta}, \eta)$ with respect to α and evaluating at $\alpha = 0$, one has

$$\frac{1}{n}\sum_{i=1}^{n}\int_{Z_i}^{X_i}\Big(e^{\tilde{\eta}(t,U_i)+\boldsymbol{z}_i^T\hat{\boldsymbol{b}}} - e^{\hat{\eta}(t,U_i)+\boldsymbol{z}_i^T\hat{\boldsymbol{b}}}\Big)\Big(\tilde{\eta}(t, U_i) - \eta_c(t, U_i)\Big)dt = 0,$$

which, through straightforward calculation, yields $\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c)$. Hence the ratio $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ can be used to diagnose the feasibility of a reduced model $\eta \in \mathcal{H}_2$: the smaller the ratio is, the more feasible the reduced model is. In the absence of frailty, Gu (2004) suggests a threshold of 0.05 for the ratio. Our empirical study indicates that this threshold also works for the case with frailty.

2.6. Asymptotic properties

In the proposed doubly penalized estimation, the variance component $\mathbf{b}$ is folded into, and estimated as part of, the mean component. Thus, $\mathbf{b}$ can be treated as "fixed" group effects satisfying $\Sigma\mathbf{b} = 0$. The constraint $\Sigma\mathbf{b} = 0$ is necessary here for identifiability. For computation, this extra constraint is not needed since it comes naturally from minimizing (2.1).

The model parameters are thus $\mathbf{b}$ and η. Let $l = \delta(\eta + \mathbf{z}^T\mathbf{b}) - \int_Z^X \exp(\eta + \mathbf{z}^T\mathbf{b})dt$ be the log-likelihood function. Let P and P_n be the expectation and the empirical measure based on n observations, respectively. Take

$$d((\mathbf{b}, \eta), (\mathbf{b}^*, \eta^*)) = \left\{ \|\mathbf{b} - \mathbf{b}^*\|^2 + \int (\eta - \eta^*)^2 d\mu(t, U) \right\}^{1/2}$$

as the distance between $(\mathbf{b}, \eta)$ and $(\mathbf{b}^*, \eta^*)$. We make the following assumptions.

Assumptions A1. Covariate U is bounded.

Assumptions A2. Let $(\mathbf{b}_T, \eta_T)$ be the unknown true value of $(\mathbf{b}, \eta)$.

(A2.1) For any fixed p, $\mathbf{b}_T = (b_{T,1}, \ldots, b_{T,p})^T$ is an internal point of a bounded set and $\sup_{1\le j\le p} |b_{T,j}| < B_1 < \infty$ for a fixed B_1 and any p.

(A2.2) η_T belongs to the Sobolev space $\mathbb{F}_{s_0}$ indexed by the order of derivative s_0, commonly $s_0 = 2$. There exists $M_1 > 0$ such that $|\eta_T| \le M_1$ and $J(\eta_T) < \infty$.

Assumptions A3. There exist constants $M_3 > 0$ and $M_4 > 0$ such that, when $d((\mathbf{b}_T, \eta_T), (\mathbf{b}, \eta))$ is small enough,

$$M_4 d^2((\mathbf{b}_T, \eta_T), (\mathbf{b}, \eta)) \ge \mathrm{P}[l(\mathbf{b}_T, \eta_T) - l(\mathbf{b}, \eta)] \ge M_3 d^2((\mathbf{b}_T, \eta_T), (\mathbf{b}, \eta)). \quad (2.9)$$

Assumptions A4. $\lambda = O_p(n^{-[(2s_0)/(2s_0+1)]})$ and all the eigenvalues of Σ are of the order $O_p(n^{[2/(3(2s_0+1))]})$.

Assumptions A5. $p = O(n^{[1/(3(2s_0+1))]})$.

Assumption A1 has been commonly made. Assumption A2.1 has the true parameter not on the boundary of a set, and $\mathbf{b}_T$ component-wise bounded.

In Assumption A2.2, η_T resides in the Sobolev space $\mathbb{F}_{s_0}$ defined by $J(\eta) = \int [\eta^{(s_0)}(x)]^2 dx$. The property of penalized likelihood also ensures $\hat{\eta} \in \mathbb{F}_{s_0}$; see Wahba (1990). Assumption A3 can be checked by Taylor expansion of Pl. The first inequality can be derived from the boundedness Assumptions A1 and A2. For the second inequality, we note that Pl is maximized at $(\mathbf{b}_T, \eta_T)$. Thus $\partial \mathrm{P}l/\partial \mathbf{b}\big|_{\mathbf{b}=\mathbf{b}_T} = 0$ component-wise. We also have

$$\frac{\partial^2 \mathrm{P}l}{\partial \mathbf{b} \partial \mathbf{b}^T}\Big|_{\mathbf{b}=\mathbf{b}_T} = -\mathrm{P}\left(\mathbf{z}\mathbf{z}^T \int_Z^X \exp(\eta_T + \mathbf{z}^T \mathbf{b}_T) dt\right). \tag{2.10}$$

Under the boundedness assumptions, if $Pr(X > Z)$ is bounded away from 0, then the right hand side of (2.10) is bounded away from 0. With (2.10) and a similar equation for η, A3 can be satisfied. Assumption A4 on the tuning parameter λ matches that in standard spline studies. The assumption on Σ is made so that the two penalties have equal orders. With A5, the proposed method can accommodate two scenarios. The first has finite p; an example is the lung cancer study in Section 4.1, where patients are grouped according to the histological types of their tumors and there are a finite number of tumor histological types. The second scenario has p increasing at a rate slower than n; here both the number of groups and the number of subjects per group increase with n. A representative example is the SEER data presented in Section 4.2, where the grouping variable is the year of cancer diagnosis. With data being collected regularly at the registry, the number of groups (years of diagnosis) increases, while the number of people diagnosed per year (and hence the total sample size) can grow at a much faster rate.

We note that some studies assume the number of groups $p = O(n)$. An example is recurrent event data, where each subject is considered as a group. A major difference of such data from those satisfying Assumption A5 is that they have bounded group sizes. As evident from our proof in Appendix C (online), establishing the asymptotic properties under the assumption $p = O(n)$ would require significantly different techniques. Thus we focus on data satisfying Assumption A5.

Theorem 1. *Under* A1 *and* A2*, the proposed model is identifiable. If* A3–A5 *also hold,* $d((\hat{\mathbf{b}}, \hat{\eta}), (\mathbf{b}_T, \eta_T)) = O_p(n^{-[(s_0)/(2s_0+1)]})$ *and* $J(\hat{\eta}) = O_p(1)$ *for the minimizer* $(\hat{\mathbf{b}}, \hat{\eta})$ *of* (2.1)*.*

Besides identifiability, Theorem 1 shows that η is estimable at the optimal rate for a spline function. In addition, $J(\hat{\eta}) = O_p(1)$, i.e., $\hat{\eta}$ has the “proper” amount of smoothness. We refer to Appendix C (online) for the proof.

3. Simulation Studies

We conducted simulations to evaluate the proposed methods. Let $\mathcal{W}(a, b)$ denote the Weibull distribution with density function $f(t) = (a/b^a)t^{a-1}e^{-(t/b)^a}$. The corresponding hazard function and the log hazard function are respectively $h(t) = at^{a-1}/b^a$ and $\eta(t) = \log a + (a-1)\log t - a \log b$.

3.1. Performance of cross validation

To gauge performance of the cross validation score developed in Sec. 2.3, we carried out two sets of simulations. In the first simulation, covariate U_i was the scale parameter, $U_i = 1$ for $i = 1, \ldots, n/2$ and 1.5 for $i = n/2+1, \ldots, n$, and $T_i|U_i = u \sim \mathcal{W}(3, u)$. In the second simulation, covariate U_i was the shape parameter, $U_i = 1.5$ for $i = 1, \ldots, n/2$ and 4.5 for $i = n/2+1, \ldots, n$, $T_i|U_i = u \sim \mathcal{W}(u, 1)$. In both simulations, the censoring time $C_i \sim \mathcal{W}(3, 2)$ and truncation time $Z_i \sim \mathcal{W}(5, 0.3)$. In each simulation, one hundred replicates were generated. Each data set was of size $n = 100$ and divided into $p = 25$ groups, with group frailty $b_k \sim N(0, 1)$. So $\Sigma = \sigma^2 I$ with $\sigma^2 = 1$.

The full model in (2.2) was used for data with the shape covariate and the additive model with $\eta_{t,u}(t, u) = 0$ in (2.2) was used for data with the scale covariate. Let $\mathrm{KL}_e((\eta, \boldsymbol{b}), (\eta_\Lambda, \boldsymbol{b}_\Lambda))$ be the empirical Kullback-Leibler loss in (2.6). For each replicate, the minimum Kullback-Leibler loss $\mathrm{KL}_e((\eta, \boldsymbol{b}), (\eta_\Lambda, \boldsymbol{b}_\Lambda))$ achievable by (2.1) was computed, along with the losses of the estimates when Λ was selected by $V_\alpha(\Lambda)$ with $\alpha = 1, 1.2, 1.4, 1.6, 1.8, 2.0$. The estimate $\hat{\sigma}^2$ for each α was also recorded.

The performance of $V_\alpha(\lambda)$ is summarized in Figure 1. The box plots in the top panels are $\hat{\sigma}^2$ from six CV scores in both simulations, with the true value superimposed as the faded line. The box plots in the bottom panels are the relative efficacy of all six CV scores in both simulations, defined as the ratios of the minimum loss KL_e to the loss KL_e corresponding to the CV scores. These plots suggest the best performance of $V_\alpha(\lambda)$ around $\alpha = 1.4$, which agrees with Gu (2002).

3.2. Bayesian confidence intervals

To assess coverage property of the Bayesian confidence intervals developed in Sec. 2.4, we carried out four simulations with sample sizes $n = 100, 400$ and two levels of censoring. In all the simulations, the failure time $T_i \sim \mathcal{W}(3, 1)$ and truncation time $Z_i \sim \mathcal{W}(5, 0.3)$, which led to a 5% truncation rate. In two simulations, censoring time $C_i \sim \mathcal{W}(3, 2)$, which gave an average censoring rate of 15%; in the other two simulations, $C_i \sim \mathcal{W}(2, 1.5)$ and the average censoring rate was 35%. Each simulation had 500 replicates with $p = 10$ and group frailties

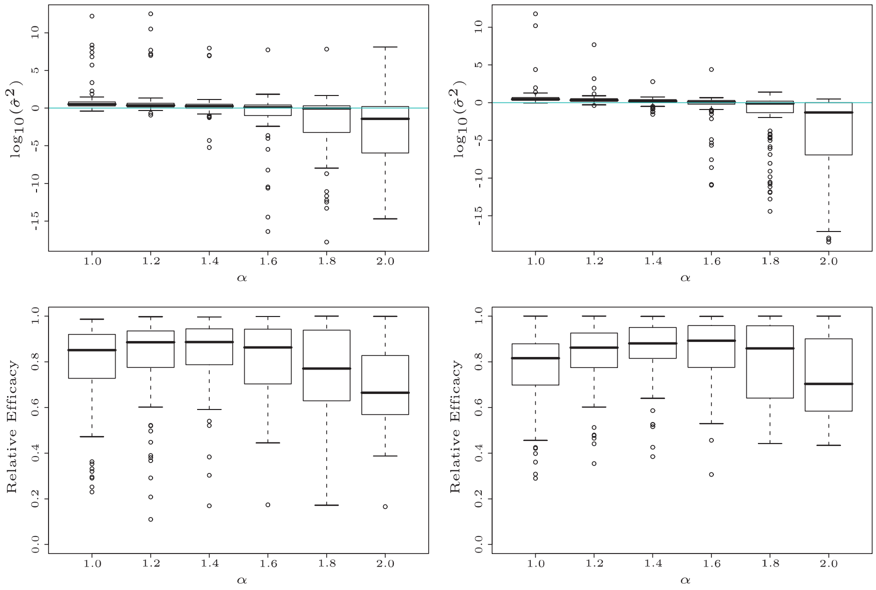

Figure 1. Performance of Cross-Validation Scores $V_\alpha(\lambda)$. Top: Estimates of σ^2 from $V_\alpha(\lambda)$ with $\alpha = 1, 1.2, 1.4, 1.6, 1.8, 2.0$. Faded lines denote the true value $\sigma^2 = 1$. Bottom: Relative efficacy of $V_\alpha(\lambda)$ with $\alpha = 1, 1.2, 1.4, 1.6, 1.8, 2.0$. Left: Data with scale covariate. Right: Data with shape covariate.

$b_k \sim N(0, 1)$. The smoothing parameters were selected using cross-validation scores with $\alpha = 1.4$. For each replicate and each point on the grid $t = 0.5$ to 1.5 by 0.01, a 95% confidence interval of η was calculated and compared with the true value. The left panels in Figure 2 plot the point-wise coverage against the time for the four simulations. The right panels in Figure 2 plot the estimates against the true log hazard function (faded solid line). The plotted estimates include the connected point-wise averages of the log hazard estimates (solid line), the connected medians of the point-wise 95% confidence interval limits (dashed lines), and the point-wise 0.975 and 0.025 quantiles of the log hazard estimates (faded dashed lines). We can see that the quantiles of the log hazard estimates matched well with the medians of confidence interval limits, indicating a proper magnitude of standard errors. Also increasing the sample size appears to stabilize the point-wise coverage, while increasing the level of censoring seems to pull down the coverage a little bit. And it is reassuring to see the decrease of coverage towards the upper end of the time axis where information from the data is vanishing.

3.3. Model selection

To assess the model selection tool developed in Sec. 2.5, we also carried out

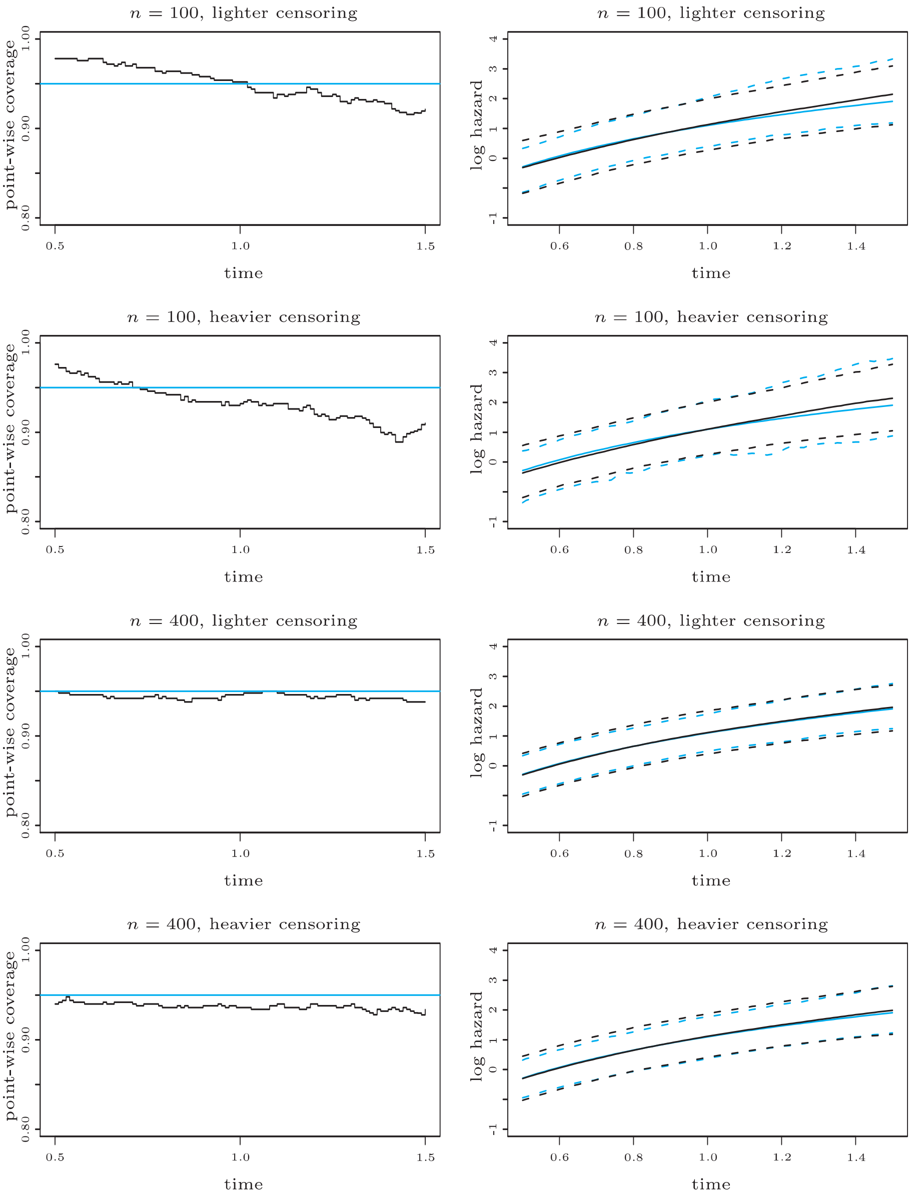

Figure 2. Bayesian Confidence Intervals. Left: Point-wise coverage against time, faded line is the nominal confidence level 0.95. Right: Connected point-wise averages of the log hazard estimates (solid line), connected medians of the point-wise 95% confidence interval limits (dashed lines), connected point-wise 0.975 and 0.025 quantiles of the log hazard estimates (faded dashed lines), and the true log hazard (faded solid line).

four simulations with sample sizes $n = 100, 400$ and two levels of censoring. In each simulation, two sets of 500 data replicates were generated with $p = 10$ and group frailties $b_k \sim N(0, 1)$. One set was a scale-covariate model with $T_i \sim \mathcal{W}(3, U_i)$ and $U_i \in \{1, 1.5\}$, and the other set was a shape-covariate model with $T_i \sim \mathcal{W}(U_i, 1)$ and $U_i \in \{1.5, 4.5\}$. In two simulations, the censoring time $C_i \sim \mathcal{W}(3, 2)$ and truncation time $Z_i \sim \mathcal{W}(5, 0.3)$, leading to an average censoring rate around 25% and an average truncation rate around 8%; in the other two simulations, no truncation was imposed and the censoring time $C_i \sim \mathcal{W}(3, 3.5)$, which led to an average censoring rate around 8%. For all the data replicates, we first fitted the full model (2.2), and then used the model selection tool to check whether it could be reduced to the additive model. The ratio $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ was recorded for each data replicate. For replicates with a scale covariate the real model is additive, so we should expect the reduction detected by the model selection tool with small $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$. For replicates with a shape covariate, the real model is non-additive, so we should expect the reduction rejected with large $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$. A box plot of the ratios is presented in Figure 3 with the line $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c) = 0.05$ superimposed. For replicates with a scale covariate, the number of ratios above 0.05 (out of the total 500 ratios in each simulation) were respectively 82, 40, 2, 0 for simulations ($n = 100$, heavier censoring), ($n = 100$, lighter censoring), ($n = 400$, heavier censoring), ($n = 400$, lighter censoring). The success rate of detecting the hazard proportionality increased with the sample size and decreased with the level of censoring. For replicates with a shape covariate, all the ratios of the four simulations were above 0.05, which led to successful detection of the non-proportionality. Hence the model selection tool with a threshold of 0.05 was satisfactory for identifying the correct model.

3.4. Comparison with stratified proportional hazards model

When the number of groups is fixed at p, a popular alternative method is the stratified proportional hazards (SPH) model

$$h_k(t, u) = h_{0k}(t) e^{\beta^T u}, \quad k = 1, \ldots, p, \tag{3.1}$$

where $h_{0k}(t)$ is the baseline hazard function for the kth stratum and β is the common coefficient vector. The coefficient β in (3.1) is estimated by maximizing the partial likelihood. Smooth estimates of $h_{0i}(t)$ can be obtained with a kernel estimator $\hat{h}_{0k}(t) = \int K_\tau(t - u) d\hat{H}_{0k}(t)$, where $K_\tau(t - u) = (1/\tau)K((t - u)/\tau)$ is a kernel function with bandwidth τ, and $\hat{H}_{0k}(t)$ is the Breslow estimator of the cumulative baseline hazard function. The relationship between our model and the SPH model is discussed in Section 5. In this section, we present simulations

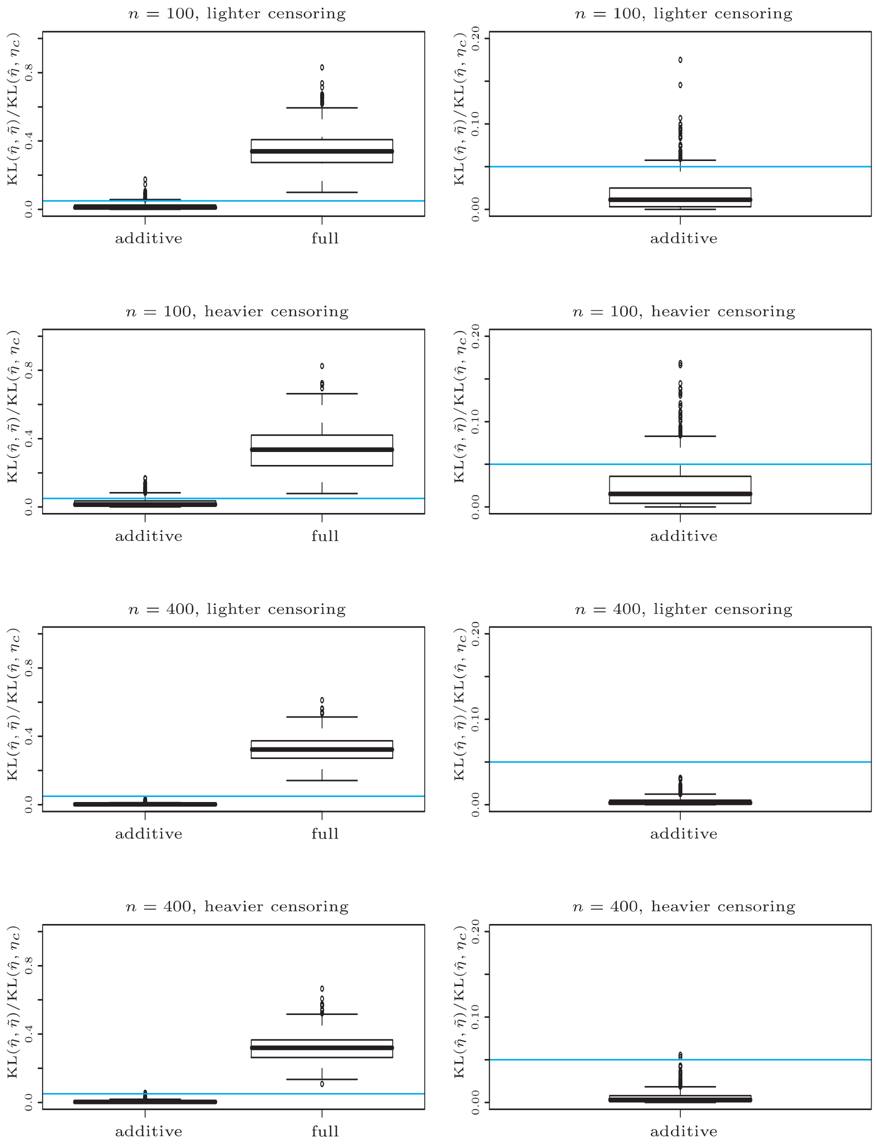

Figure 3. Model Selection. Left: Box plots of the ratios $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ for replicates with scale covariate (additive model) and replicates with shape covariate (full model). Right: Zoom-in of the ratios $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ for replicates with scale covariate. The faded lines represent $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c) = 0.05$.

comparing these two models under their common playground: the case of stratified proportional hazards with $h_{0k}(t) = e^{b_k} h_0(t)$ for some constants b_k and an unknown function $h_0(t)$.

Table 1. Summary statistics of empirical Kullback-Leibler losses KL_e

	Mean	Std. Dev.	1st Quartile	Median	3rd Quartile
Proposed Model	0.045	0.014	0.035	0.043	0.053
Stratified PH	1.543	0.409	1.258	1.464	1.749

We simulated 500 data replicates, each with $n = 200$ and $p = 8$. The failure times were $T_i|U_i = u \sim \mathcal{W}(3, u)$ with U_i randomly generated from $\{0.5, 1.0, 1.5\}$, and the frailties $b_k \sim N(0, 1)$. Thus the real hazard function has the form $h_k(t, u) = (3e^{b_k}t^2)\exp(-3\log u)$. The censoring times $C_i \sim \mathcal{W}(3, 2)$, which led to an average censoring rate around 25%. There was no truncation. Both models were applied to each data replicate with the empirical Kullback-Leibler loss KL_e between the truth and the estimate recorded. The proposed model was applied under the assumption of an additive log hazard $\eta_k(t, u) = \eta_0 + \eta_t(t) + \eta_u(u) + b_k$. The SPH model was implemented using $\log u$ as the only covariate and the Epanechnikov kernel $K(x) = 0.75(1 - x^2)$, $-1 \leq x \leq 1$ with a fixed bandwidth $\tau = 0.5$ (Wells (1994)), the bandwidth $\tau = 0.5$ being selected empirically such that the resulting estimate was close to optimal in terms of minimizing KL_e. The summary statistics for the 500 KL_e losses of the proposed model and those of the SPH model are listed in Table 1. Clearly, the proposed model yielded better estimates than the SPH model with kernel smoothed baseline hazards. A possible improvement over the kernel estimate is to generalize the local bandwidth selection criterion in Wells (1994). We do not pursued that here.

4. Applications

4.1. VA lung cancer study

In a clinical trial reported in Prentice (1973), 137 veteran males with advanced inoperable lung cancer were randomized to either a standard or test chemotherapy. Eight of the 137 survival times were censored. Besides the survival time (in days) and the treatment, also recorded are the performance status (completely hospitalized, partial confinement, or able to care for self), the duration time of disease up to randomization, age, indicator of having prior therapy or not, and histological type of tumor (squamous, small cell, adeno, or large cell). Age and disease duration have been found insignificant in studies such as Kalbfleisch and Prentice (2002), so they were not included in the analysis here.

Patients with different histological types of tumor are expected to respond differently to the treatments. However, the actual estimation of such effects is of less interest. We thus grouped the patients according to the histological type of tumor. For an intensively studied cancer like lung cancer, the number of histological types of tumor is generally assumed fixed. This then corresponds to the fixed p scenario described in Section 2.6.

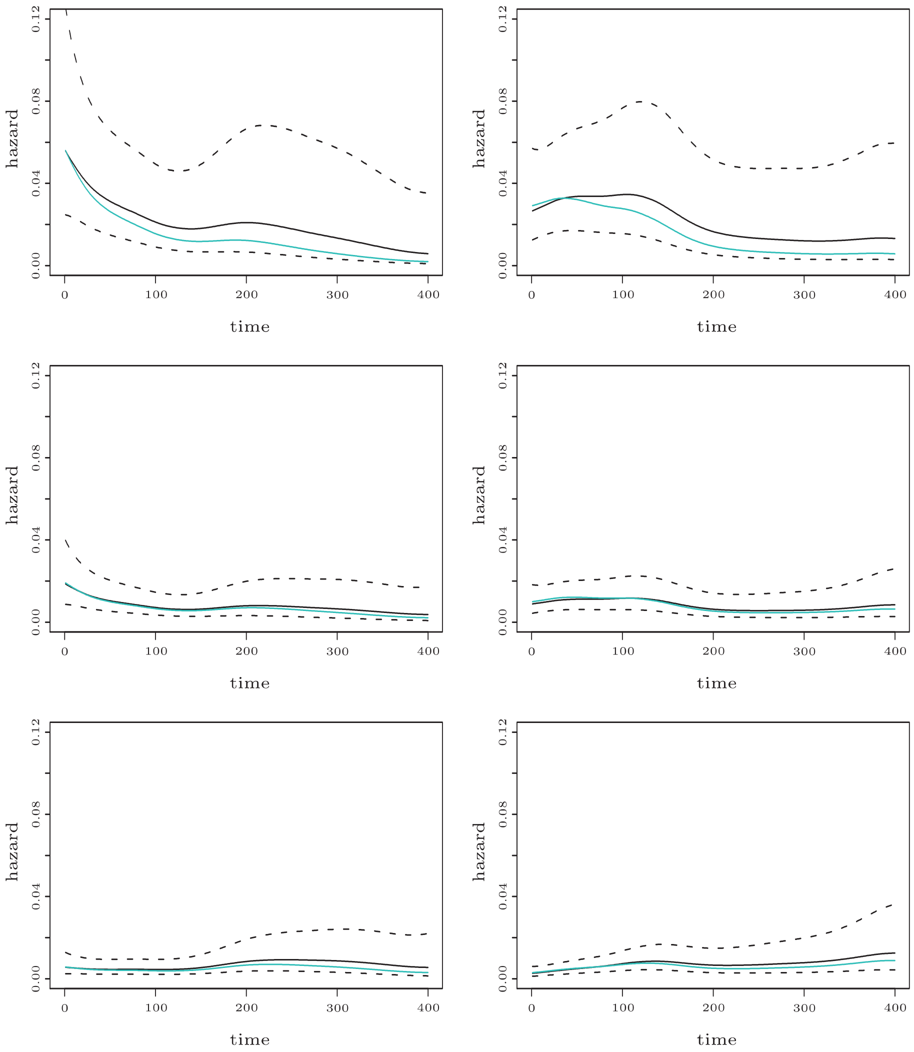

Figure 4. Hazard Estimates and Their Confidence Intervals for VA Lung Cancer Data. Left: no prior therapy history ($u_{\langle h\rangle} = 0$). Right: with prior therapy history ($u_{\langle h\rangle} = 1$). Top: completely hospitalized ($u_{\langle p\rangle} = 1$). Middle: partial confinement ($u_{\langle p\rangle} = 2$). Bottom: able to care for self ($u_{\langle p\rangle} = 3$). Solid lines are the estimates at the specified covariate levels from model (4.1) with frailties, dashed lines are the connected point-wise confidence intervals, and faded lines are the estimates from model (4.1) without frailties.

Let $\eta(t, u)$ be the log hazard at time t given the covariate u, where $u = (u_{\langle t\rangle}, u_{\langle p\rangle}, u_{\langle h\rangle})$ consists of the treatment $u_{\langle t\rangle}$, performance status $u_{\langle p\rangle}$, and indicator of prior therapy history $u_{\langle h\rangle}$. Our initial model consists of all the main effects and all the two-way interactions (including the time-covariate interaction and covariate-covariate interaction). After applying the model selection tool repeatedly, we ended up with the model

$$\begin{aligned}\eta(t_i, u_i, \boldsymbol{z}_i) = \eta_\emptyset + \eta_1(t_i) + \eta_2(u_{i\langle p\rangle}) + \eta_3(u_{i\langle h\rangle}) \\ +\eta_{1,2}(t_i, u_{i\langle p\rangle}) + \eta_{1,3}(t_i, u_{i\langle h\rangle}) + \boldsymbol{z}_i^T\boldsymbol{b}.\end{aligned} \tag{4.1}$$

There are a couple of messages in this final model. First, all the terms involving treatment were dropped out of the model, indicating a negligible treatment effect. This matches findings in Kalbfleisch and Prentice (2002). The second message is that the proportional hazards model may not be valid here due to the presence of the two time-covariate interactions. The non-proportionality is also demonstrated by Figure 4, the plots of the hazard estimates and their confidence intervals from the final model (4.1). The hazard estimates at the first level (completely hospitalized, the top panels) of the performance status $u_{\langle p\rangle}$ show drastically different trends from the estimates at the other two levels (the middle and bottom panels). There are also visible differences between the hazard estimates at the two levels of the prior therapy history $u_{\langle h\rangle}$ (the left panels versus the right ones).

We also fitted the final model (4.1) without the frailties. The fitted hazard rates were imposed in Figure 4 as faded lines. These two sets of hazard estimates are slightly different for completely hospitalized patients (top panels), but indistinguishable for patients who were partially confined (middle panels) and patients who were able to care for themselves (bottom panels). Also the estimated frailty variance $\hat{\sigma}^2 = 0.21$ in model (4.1), confirming a fair degree of frailty effect.

4.2. SEER prostate cancer study

Hamilton and Ries (2007) and references therein show that survival of prostate cancer patients may be affected by (a) clinical variables, such as stage of disease and tumor grade; (b) demographic variables, such as age, race, and geographic location; and (c) "environmental" variables, such as time of diagnosis. Specifically, time of diagnosis may carry information such as the advancement of medical treatments. In our study, we are more interested in the effects of the clinical and the demographic variables, and less interested in estimating the effect of time of diagnosis. Thus we treat the latter as a random effect and group the patients by their time of diagnosis. Since more patient records are regularly added to the registry database, the number of groups here (i.e., number of diagnosis time periods) increases over time, as well as the number of patients per group. This corresponds to the second scenario described in Section 2.6.

The working dataset contains 800 patients diagnosed with prostate cancer between 1975 and 2004 and reported to the Metropolitan Atlanta Registry of the

Surveillance Epidemiology and End Results (SEER) Program. The survival time was recorded as the number of months between the date of diagnosis and one of the following: date of death, date last known to be alive, or follow-up cutoff date. Covariates collected included year of diagnosis, age at diagnosis $u_{\langle a\rangle}$, marital status $u_{\langle m\rangle}$ (married or not), race $u_{\langle r\rangle}$ (white or non-white), tumor grade $u_{\langle g\rangle}$ (3 levels), and county $u_{\langle c\rangle}$ (5 levels). Following common practice for analysis of SEER data, we divided the year of diagnosis into six groups (1975−1979, …, 2000−2004) and assumed that patients in the same year group shared a common frailty.

Our first model for the log hazard function η was an additive model with the main effects of the five covariates. Application of the model selection tool in Section 2.5 suggested that the effects of $u_{\langle r\rangle}$ and $u_{\langle c\rangle}$ were negligible. Thus, our final model was

$$\eta(t_i, u_i, \boldsymbol{z}_i) = \eta_\emptyset + \eta_1(t_i) + \eta_2(u_{i\langle a\rangle}) + \eta_3(u_{i\langle m\rangle}) + \eta_4(u_{i\langle g\rangle}) + \boldsymbol{z}_i^T\boldsymbol{b}. \tag{4.2}$$

We plotted in Figure 5 the hazard components e^{η_k}, $k = 1, 2, 3, 4$, together with their 95% confidence intervals. Clearly, higher hazard was associated with longer survival time, older age at diagnosis, unmarried status and higher grade of tumor. Such findings are consistent with the published results.

To see whether the inclusion of frailty was worthwhile, we fitted model (4.2) without the frailties. The corresponding hazard components were superimposed in Figure 5 as faded lines/stars. There is clear difference between the two sets of estimates. The estimated frailty variance $\hat{\sigma}^2 = 0.32$ in model (4.2), confirming a non-negligible level of frailty effect.

5. Discussion

We have proposed a frailty model to analyze correlated survival data with penalized joint likelihood. By including the frailties $\boldsymbol{b}$ into the estimation, the proposed penalized joint likelihood method turns the “variance component” into “mean component”. This simplifies the model structure and reduces computational burden otherwise required, e.g., for the extra integrations over the frailties when a marginal likelihood is penalized. Our treatment of $\boldsymbol{b}$ as “fixed” group effects in Sec. 2.6 also fits this ”mean component” motivation of doubly penalized methods.

Although we have only considered $\Sigma = \sigma^{-2}I$ in the simulations and applications, our main results are applicable to the other settings with more complex Σ. One example has group frailties independent but with different variances, resulting in a diagonal Σ with p distinct diagonal elements. Another example has frailties time dependent as in Yau and McGilchrist (1998), say, following a

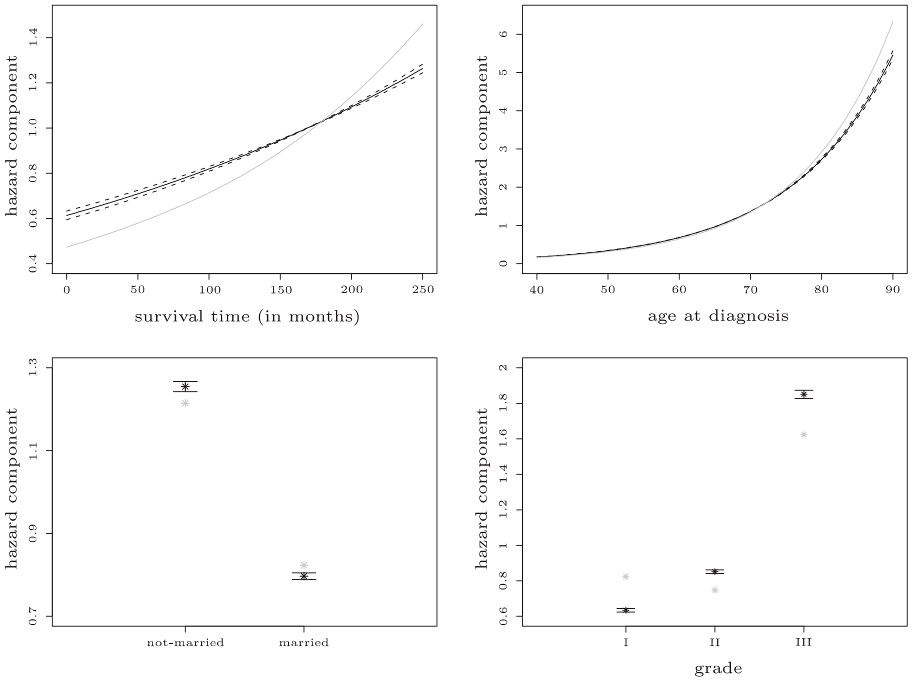

Figure 5. Hazard Components and Confidence Intervals for SEER Data. Solid lines/stars are the estimated hazard components from model (4.2) with frailties, dashed lines in the top frames and error bars in the bottom frames are the corresponding 95% confidence intervals, and faded lines/stars are the estimates from model (4.2) without frailties.

first-order autoregressive process; then a Σ incorporating both the variance and the autocorrelation parameter would be used in (2.1).

Besides normal random effects, there are other common choices of frailty distributions in the literature. In principle, for any other type of random effects, one can replace the term $(1/2)\mathbf{b}^T\Sigma\mathbf{b}$ in (2.1) by the corresponding negative log density to obtain the new penalized likelihood. However, normal random effects are preferred here due to an unrestricted covariance matrix and more tractable computation in the penalized likelihood setting.

When the number of groups p is considered fixed, as in the lung cancer example, a popular alternative method is the stratified proportional hazards (SPH) model (3.1) with kernel smoothed baseline hazards. By having stratum-specific baseline hazards $h_{0i}(t)$, the SPH model allows more flexibility in modeling the time effect difference between the group (or stratum) hazards, but the proportionality assumption limits its flexibility in modeling the covariate effect that can be nonlinear and interact with time in practice. On the other hand, the group hazard in the proposed model is the product of a group-wise frailty and a nonparametric hazard function. The nonparametric part allows more general covariate effects including non-proportional and nonlinear effects. Our simulation in Section 3.4, comparing the two approaches under their common applicable setting, demonstrates a favorable performance of the proposed model.

Acknowledgement

We are grateful to an associate editor and the referees for their helpful remarks that have significantly improved the paper. The second author has been partly supported by NSF DMS 0805984.

References

Du, P. and Gu, C. (2006). Penalized likelihood hazard estimation: efficient approximation and Bayesian confidence intervals. *Statist. Probab. Lett.* **76**, 244-254.

Fine, J. P., Glidden, D. V. and Lee, K. E. (2003). A simple estimator for a shared frailty regression model. *J. Roy. Statist. Soc. Ser. B* **65**, 317-329.

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

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

Gu, C. (2004). Model diagnostics for smoothing spline ANOVA models. *Canad. J. Statist.* **32**, 347-358.

Gu, C. and Ma, P. (2005). Optimal smoothing in nonparametric mixed-effect models. *Ann. Statist.* **33**, 1357-1379.

Hamilton, A. and Ries, L. A. G. (2007). Cancer of the prostate. In *SEER Survival Monograph: Cancer Survival Among Adults: U.S. SEER Program,* 1988−2001*, Patient and Tumor Characteristics* (Edited by L. A. G. Ries, J. L. Young, G. E. Keel, M. P. Eisner, Y. D. Lin and M.-J. Horner), 171-180. National Cancer Institute, SEER Program, NIH Pub. No. 07-6215, Bethesda, MD.

Joly, P. and Commenges, D. (1999). A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to AIDS. *Biometrics* **55**, 887-890.

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.* Wiley, New York.

Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications (with discussion). *J. Amer. Statist. Assoc.* **96**, 1272-1298.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *J. Roy. Statist. Soc. Ser. B* **61**, 381-400.

Nielsen, G. G., Gill, R. D., Andersen, P. K. and Sørensen, T. I. (1992). A counting process approach to maximum likelihood estimation of frailty models. *Scand. J. Statist.* **19**, 25-43.

Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika* **60**, 279-288.

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

Sastry, N. (1997). A nested frailty model for survival data, with an application to the study of child survival in northeast Brazil. *J. Amer. Statist. Assoc.* **92**, 426-435.

Therneau, T. M., Grambsch, P. M. and Pankratz, V. S. (2003). Penalized survival models and frailty. *J. Comput. Graph. Statist.* **12**, 156-175.

van de Geer, S. A. (2000). *Empirical Processes in M-Estimation.* Cambridge University Press, Cambridge.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45**, 133-150.

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

Wang, Y. (1998). Mixed-effects smoothing spline ANOVA. *J. Roy. Statist. Soc. Ser. B* **60**, 159-174.

Wells, M. T. (1994). Nonparametric kernel estimation in counting processes with explanatory variables. *Biometrika* **81**, 795-801.

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

Department of Statistics, Virginia Tech, Blacksburg, VA 24061, U.S.A.

E-mail: pangdu@vt.edu

Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, U.S.A.

E-mail: shuangge.ma@yale.edu

(Received June 2008; accepted February 2009)