Penalized Likelihood Hazard Estimation: Efficient Approximation and Bayesian Confidence Intervals

Pang Du and Chong Gu
Purdue University

Abstract

Penalized likelihood method can be used for hazard estimation with lifetime data that are right-censored, left-truncated, and possibly with covariates. In this article, we are concerned with more scalable computation of the method and with the derivation and assessment of certain interval estimates. The asymptotic convergence rates are preserved when the estimation is restricted to certain q-dimensional spaces with q increasing at a much slower rate than the sample size n, and simulation studies are performed to determine default values of q for practical use; the computation cost is of the order $O(nq^2)$. Through a quadratic approximation of the log likelihood, approximate Bayesian confidence intervals can be derived for log hazard, and empirical studies are conducted to assess their properties. The techniques are implemented in open-source R code and real-data example is presented to illustrate the applications of the techniques through the use of the software.

KEY WORDS: Bayesian confidence interval; Computation; Hazard; Penalized likelihood.

1 Introduction

Censored lifetime data are common in life testing, medical follow-up and other studies. Let T_i be the lifetime of an item, Z_i be the left-truncation time at which the item enters the study, and C_i be the right-censoring time beyond which the item is dropped from the study, independent of each other. One observes $(Z_i, X_i, \delta_i, U_i)$, $i = 1, \ldots, n$, where $X_i = \min(T_i, C_i)$, $\delta_i = I_{[T_i \leq C_i]}$, $Z_i < X_i$, and U_i is a covariate. Assume that $T_i|U_i$ follow a survival function $S(t,u) = \text{Prob}(T > t|U = u)$. Of interest is the estimation of the hazard function $e^{\eta(t,u)} = -\partial \log S(t,u)/\partial t$.

Penalized likelihood method estimates $\eta(t,u)$ via the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\{\delta_i \eta(X_i, U_i) - \int_{Z_i}^{X_i} e^{\eta(t,U_i)} dt\} + \frac{\lambda}{2} J(\eta), \tag{1}$$

where the first term is the minus log likelihood, $J(\eta)$ is a roughness functional, and the smoothing parameter λ controls the tradeoff between the goodness-of-fit and the smoothness of η. The formulation, found in Gu (1996) along with an asymptotic theory, evolved from the work of Anderson and Senthilselvan (1980), O'Sullivan (1988a, 1988b), and Zucker and Karr (1990), among others. The minimizer of (1) is actually the maximum likelihood estimate (MLE) under a soft constraint of the form $J(\eta) \leq \rho$ for some $\rho > 0$, with λ being the Lagrange multiplier. With $\lambda = \infty$ (i.e. $\rho = 0$), one enforces a parametric model in the null space of $J(\eta)$, $\{\eta : J(\eta) = 0\}$, and as $\lambda \rightarrow 0$ (i.e. $\rho \rightarrow \infty$), one approaches the nonparametric MLE; the latter is the Kaplan-Meier in the absence of the covariate U. The practical performance of the estimate hinges on the proper selection of λ, for which an effective cross-validation procedure can be found in Gu (2002, Section 7.2).

The minimization of (1) is in a function space in which $J(\eta)$ is defined and finite, and under conditions, one can establish the asymptotic convergence rates for the minimizer of (1); see Gu (1996). It can be shown that the minimizers of (1) in certain data-adaptive q-dimensional spaces share the same convergence rates, with q increasing with n but at a much slower rate, say $q \asymp n^{2/9}$; the computation is of the order $O(nq^2)$. For computational efficiency, one prefers a small q, but to maintain the inferential efficiency of the method, one needs to keep a large enough q so as to leave model selection primarily with λ selection. One purpose of this article is to devise through simulation studies some default q values, that are "minimally sufficient," for use in practice.

For penalized likelihood regression, Bayesian confidence intervals can be derived through the associated Bayes models; see, e.g., Wahba (1983), Silverman (1985), and Gu (1992). For approximations in q-dimensional spaces, the corresponding Bayes models were discussed in Kim and Gu (2004). In this article, we derive similar interval estimates for the hazard estimation of (1), by using a quadratic approximation of the log likelihood at the minimizer of (1) and the Bayes models of Kim and Gu (2004); empirical studies will be performed to assess the properties of such intervals. Bayesian confidence intervals for penalized likelihood hazard estimation were previously discussed by Joly, Commenges, and Letenneur (1998), for models parametric in the covariate U.

The rest of the article is organized as follows. In Section 2, some technical details concerning (1) are briefly reviewed to set the stage for subsequent developments. Simulation studies for the empirical choice of q are presented in Section 3. Bayesian confidence intervals are derived in Section 4 and their properties are assessed in Section 5. Real-data examples are given in Section 6 using open-source R code that implements the techniques developed. A few remarks in Section 7 conclude the article.

2 Penalized Likelihood Estimation

We shall fill in some details concerning the method under study. The minimization of (1) is done in a Hilbert space $\mathcal{H}$ on the product domain $\mathcal{T} \times \mathcal{U}$ of time and covariate and $J(\eta)$ is taken as a square seminorm in $\mathcal{H}$ with a finite dimensional null space $\mathcal{N}_J \subset \mathcal{H}$, where a finite dimensional

$\mathcal{N}_J$ prevents interpolation, the conceptual equivalent of a delta sum. The evaluation functional $[t,u]f = f(t,u)$ is assumed to be continuous in $f \in \mathcal{H}$, which is necessary for (1) to be continuous in its argument η. When $\mathcal{U}$ is a singleton (i.e., with no covariate), the formulation reduces to that of O'Sullivan (1988a).

A space $\mathcal{H}$ in which the evaluation functional is continuous is called a reproducing kernel Hilbert space (RKHS) possessing a reproducing kernel (RK) $R(\cdot,\cdot)$, a non-negative definite function satisfying $R_x(\cdot) = R(x,\cdot) \in \mathcal{H}$, $\forall x = (t,u) \in \mathcal{T} \times \mathcal{U}$, and $\langle R(x,\cdot), f(\cdot)\rangle = f(x)$, $\forall f \in \mathcal{H}$, where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathcal{H}$; the RK $R(\cdot,\cdot)$ and the space $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ determine each other uniquely. Typically, $\langle\cdot,\cdot\rangle = J(\cdot,\cdot) + \tilde{J}(\cdot,\cdot)$, where $J(\cdot,\cdot)$ is the semi inner product associated with $J(\cdot)$ and $\tilde{J}(\cdot,\cdot)$ is an inner product in the null space $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ of $J(\eta)$ when restricted therein. There exists a tensor sum decomposition $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$, where the space $\mathcal{H}_J$ has $J(\eta)$ as its square norm and an RK R_J satisfying $J(R_J(x,\cdot), f(\cdot)) = f(x)$, $\forall f \in \mathcal{H}_J$. See, e.g., Gu (2002, Section 2.1).

Example 1 (Cubic Spline) *For $\mathcal{U}$ a singleton and $\mathcal{T} = [0,1]$, a choice of $J(\eta)$ is $\int_0^1 \ddot{\eta}^2 dt$, which yields the popular cubic splines. A choice of $\tilde{J}(f,g)$ is $(\int_0^1 f\,dt)(\int_0^1 g\,dt) + (\int_0^1 \dot{f}\,dt)(\int_0^1 \dot{g}\,dt)$, yielding $\mathcal{H}_J = \{\eta : \int_0^1 \eta dt = \int_0^1 \dot{\eta} dt = 0, J(\eta) < \infty\}$ and the RK $R_J(t_1,t_2) = k_2(t_1)k_2(t_2) - k_4(t_1 - t_2)$, where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials. The null space $\mathcal{N}_J$ has a basis $\{1, k_1(t)\}$, where $k_1(t) = t - 0.5$. See, e.g., Gu (2002, Section 2.3.3).* □

Example 2 (Tensor Product Cubic Spline) *For $\mathcal{U} = \mathcal{T} = [0,1]$, the construction of Example 1 can be used to build tensor product cubic splines. On $x \in [0,1]$, one has*

$$\begin{aligned}\left\{\eta : \textstyle\int_0^1 \ddot{\eta}^2 dx < \infty\right\} &= \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1 \\ &= span\{1\} \oplus span\{k_1(x)\} \oplus \left\{\eta : \textstyle\int_0^1 \eta dx = \int_0^1 \dot{\eta} dx = 0,\ \int_0^1 \ddot{\eta}^2 dx < \infty\right\},\end{aligned}$$

with RKs $R_{00}(x_1,x_2) = 1$, $R_{01}(x_1,x_2) = k_1(x_1)k_1(x_2)$, and $R_1 = k_2(x_1)k_2(x_2) - k_4(x_1 - x_2)$. Taking tensor product, one obtains nine tensor sum terms $\mathcal{H}_{\nu,\mu} = \mathcal{H}_\nu^{(t)} \otimes \mathcal{H}_\mu^{(u)}$ on $\mathcal{T} \times \mathcal{U}$, $\nu, \mu = 00, 01, 1$, with RKs $R_{\nu,\mu}(x_1,x_2) = R_\nu(t_1,t_2)R_\mu(u_1,u_2)$, where $x = (t,u)$. The four subspaces with $\nu,\mu = 00, 01$ are of one-dimension each, and can be lumped together as $\mathcal{N}_J$. The other five subspaces can be put together as $\mathcal{H}_J$ with the RK

$$R_J = \theta_{00,1}R_{00,1} + \theta_{1,00}R_{1,00} + \theta_{01,1}R_{01,1} + \theta_{1,01}R_{1,01} + \theta_{1,1}R_{1,1},$$

where $\theta_{\nu,\mu}$ are a set of extra smoothing parameters adjusting the relative weights of the roughness of different components.

For interpretation, the nine subspaces readily define an ANOVA decomposition

$$\eta(t,u) = \eta_\emptyset + \eta_t(t) + \eta_u(u) + \eta_{t,u}(t,u)$$

for functions on $\mathcal{T} \times \mathcal{U}$, with $\eta_{\emptyset} \in \mathcal{H}_{00}^{(t)} \otimes \mathcal{H}_{00}^{(u)}$ being the constant term, $\eta_t \in \{\mathcal{H}_{01}^{(t)} \oplus \mathcal{H}_{1}^{(t)}\} \otimes \mathcal{H}_{00}^{(u)}$ the t main effect, $\eta_u \in \mathcal{H}_{00}^{(t)} \otimes \{\mathcal{H}_{01}^{(u)} \oplus \mathcal{H}_{1}^{(u)}\}$ the u main effect, and $\eta_{t,u} \in \{\mathcal{H}_{01}^{(t)} \oplus \mathcal{H}_{1}^{(t)}\} \otimes \{\mathcal{H}_{01}^{(u)} \oplus \mathcal{H}_{1}^{(u)}\}$ the interaction. One may obtain an additive model for log hazard $\eta(t,u)$, i.e., a proportional hazard model, by setting $\eta_{y,u} = 0$. See, e.g., (Gu 2002, Example 2.8). □

Let $N = \sum_{i=1}^{n} \delta_i$ be the number of events and (X_i^*, U_i^*), $i = 1, \ldots, N$, be the observed lifetimes along with the associated covariates. The space $\mathcal{H}$ is usually infinite dimensional, and the minimizer of (1) in $\mathcal{H}$ is in general not computable. To circumvent the problem, Gu (1996) proposed to use the minimizer of (1) in an adaptive finite dimensional space $\mathcal{H}_N = \mathcal{N}_J \oplus \mathrm{span}\{R_J((X_i^*, U_i^*), \cdot), i = 1, \ldots, N\}$. Under mild conditions, the minimizer of (1) in $\mathcal{H}_N$ was shown to share the same asymptotic convergence rates as the minimizer in $\mathcal{H}$. A careful look at the theory reveals that one could actually achieve the same convergence rates in a space $\mathcal{H}_q = \mathcal{N}_J \oplus \{R_J(v_j, \cdot), j = 1, \ldots, q\}$ with $q \asymp n^{2/(pr+1)+\epsilon}$ for some $p \in [1,2]$, $r > 1$, $\forall \epsilon > 0$, where $\{v_j\}$ is a random subset of $\{(X_i^*, U_i^*)\}$; the proof, which we shall omit to minimize repetition, builds on the calculus of Gu (1996) and parallels the treatment for density estimation in Gu and Wang (2003) and that for regression in Gu and Kim (2002). The constant r quantifies the smoothness imposed by $J(\eta)$: $r = 4$ for the cubic spline of Example 1 and $r = 4 - \delta$, $\forall \delta > 0$, for the tensor product cubic spline of Example 2. The constant p depends on how smooth the "true" η is: for the cubic spline of Example 1, $p = 1$ if $\ddot{\eta}^2$ is "barely" integrable, and $p = 2$ if $\eta^{(4)}$ is square integrable.

Write $\xi_j = R_J(v_j, \cdot)$ and let $\{\phi_\nu\}_{\nu=1}^{m}$ be a basis of $\mathcal{N}_J$. By definition, a function in $\mathcal{H}_q$ has an expression

$$\eta = \sum_{\nu=1}^{m} d_\nu \phi_\nu + \sum_{j=1}^{q} c_j \xi_j = \boldsymbol{\phi}^T \mathbf{d} + \boldsymbol{\xi}^T \mathbf{c}, \tag{2}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\xi}$ are vectors of functions and $\mathbf{d}$ and $\mathbf{c}$ are vectors of coefficients. Substituting (2) into (1), noting that $J(\eta) = \left\langle \sum_{j=1}^{q} c_j \xi_j \, , \sum_{k=1}^{q} c_k \xi_k \right\rangle = \sum_{j=1}^{q} \sum_{k=1}^{q} c_j c_k R_J(v_j, v_k)$, one calculates the minimizer η_λ of (1) in $\mathcal{H}_q$ by minimizing

$$-\frac{1}{n} \mathbf{1}^T (S\mathbf{d} + R\mathbf{c}) + \frac{1}{n} \sum_{i=1}^{n} \int_{Z_i}^{X_i} \exp(\boldsymbol{\phi}_i^T \mathbf{d} + \boldsymbol{\xi}_i^T \mathbf{c}) dt + \frac{\lambda}{2} \mathbf{c}^T Q \mathbf{c} \tag{3}$$

with respect to $\mathbf{d}$ and $\mathbf{c}$, where S is $N \times m$ with the (i,ν)th entry $\phi_\nu(X_i^*, U_i^*)$, R is $N \times q$ with the (i,j)th entry $\xi_j(X_i^*, U_i^*) = R_J((X_i^*, U_i^*), v_j)$, Q is $q \times q$ with the (j,k)th entry $\xi_j(v_k) = R_J(v_j, v_k)$, $\boldsymbol{\phi}_i$ is $m \times 1$ with the νth entry $\phi_\nu(t, U_i)$, and $\boldsymbol{\xi}_i$ is $q \times 1$ with the jth entry $\xi_j(t, U_i)$.

Write $\mu_f(g) = n^{-1} \sum_{i=1}^{n} \int_{Z_i}^{X_i} g(t, U_i) e^{f(t, U_i)} dt$ and $V_f(g,h) = \mu_f(gh)$. The minimization of (3) for fixed smoothing parameters can be done through Newton iteration, which updates the coefficients from the current iterate $\tilde{\eta} = \boldsymbol{\phi}^T\tilde{\mathbf{d}} + \boldsymbol{\xi}^T\tilde{\mathbf{c}}$ through

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} \\ V_{\xi,\phi} & V_{\xi,\xi} + \lambda Q \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S^T\mathbf{1}/n - \mu_\phi + V_{\phi,\eta} \\ R^T\mathbf{1}/n - \mu_\xi + V_{\xi,\eta} \end{pmatrix}, \qquad (4)$$

where $V_{\phi,\phi} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\phi}^T)$, $V_{\phi,\xi} = V_{\xi,\phi}^T = V_{\tilde{\eta}}(\boldsymbol{\phi}, \boldsymbol{\xi}^T)$, $V_{\xi,\xi} = V_{\tilde{\eta}}(\boldsymbol{\xi}, \boldsymbol{\xi}^T)$, $\mu_\phi = \mu_{\tilde{\eta}}(\boldsymbol{\phi})$, $\mu_\xi = \mu_{\tilde{\eta}}(\boldsymbol{\xi})$, $V_{\phi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\phi}, \tilde{\eta})$, and $V_{\xi,\eta} = V_{\tilde{\eta}}(\boldsymbol{\xi}, \tilde{\eta})$; see, e.g., Gu (2002, Section 7.1). The selection of the smoothing parameters can be done through an outer-loop optimization of a cross-validation score derived in Gu (2002, Section 7.2). The computational cost for each step of the Newton iteration is of the order $O(nq^2) + O(ldq^2)$, where l is the number of distinctive U_i and d is the quadrature size for numerical integration.

3 Empirical Choice of q

As noted in Section 2, a dimension of the order $q \asymp n^{2/(pr+1)+\epsilon}$, $\forall\epsilon > 0$, is sufficient for asymptotic efficiency, where $r = 4 - \delta$, $\forall\delta > 0$, for (tensor product) cubic splines. Since $\epsilon, \delta > 0$ can be arbitrarily small, one may use $q = kn^{2/(4p+1)}$ in practice. We now present some simulation results to suggest adequate values of k for practical use.

With $u \in (0,1)$, consider gamma mixtures for the distribution of $T^3|U = u$ such that

$$S(t,u) = u\int_{t^{1/3}}^{\infty} \frac{s^{0.8}e^{-s/5}}{\Gamma(1.8)5^{1.8}}\,ds + (1-u)\int_{t^{1/3}}^{\infty} \frac{s\,e^{-s/40}}{\Gamma(2)40^2}\,ds.$$

Two sets of simulations will be presented, one with U_i fixed at $u = 0.95$ (thus no covariate), another with $U_i \sim U(0.01, 0.99)$. Censoring times were generated with distribution functions $P(C < t) = \int_0^{t^{1/3}} s\,e^{-s/12.5}ds/12.5^2$ for the first set without covariate and $P(C < t) = \int_0^{t^{1/3}} s^{0.9}e^{-s/50}ds/\Gamma(1.9)50^{1.9}$ for the second set with covariate; for both sets, the censoring rate was around 20%. Z_i were set to 0 so no left-truncation was present. The cubic spline of Example 1 was used in the first set and the tensor product cubic spline of Example 2 was used in the second set.

The test hazards are sufficiently smooth so $p = 2$. Samples of sizes $n = 150, 300, 600$ were generated for both cases. For each of the six samples and every k on the grid 5(1)20, 30 different random subsets $\{v_j\} \subset \{(X_i^*, U_i^*)\}$ of size $q = kn^{2/9}$ were selected to form 30 different $\mathcal{H}_q$, and 30 different estimates were calculated based on the same data. The Kullback-Leibler loss,

$$L(\lambda) = \mathrm{KL}(\eta, \eta_\lambda) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{X_i} \big\{e^{\eta(t,U_i)}(\eta(t,U_i) - \eta_\lambda(t,U_i)) - (e^{\eta(t,U_i)} - e^{\eta_\lambda(t,U_i)})\big\}dt,$$

was calculated for the 30 estimates; see, e.g., Gu (2002, Sections 7.2) for the derivation of $\mathrm{KL}(\eta, \eta_\lambda)$ in this context. The results are summarized in Figure 1 in box plots. The fact that the box width

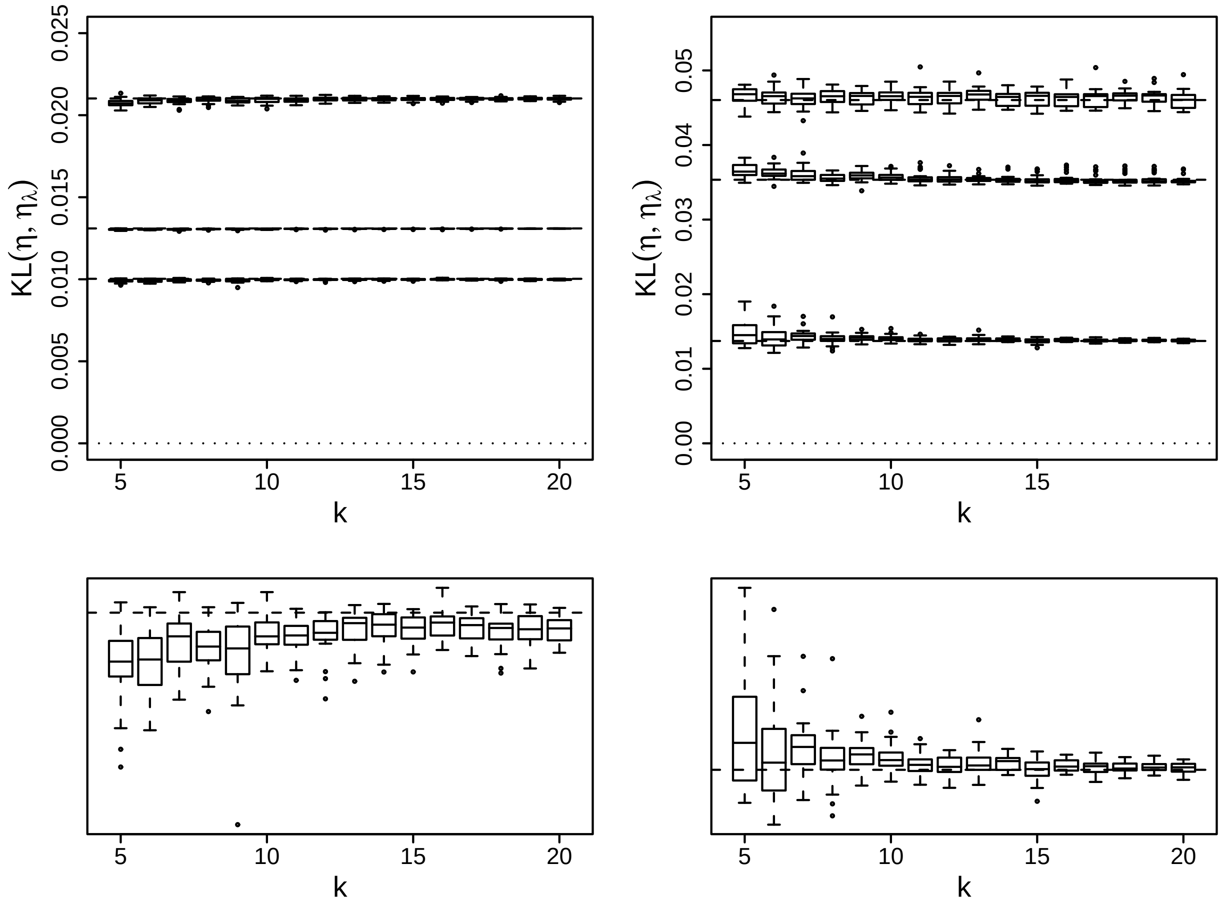

Figure 1: Effect of q on Estimation Consistency. Boxplots of $L(\lambda)$ with 30 different random subsets $\{v_j\}$ of size q, for each of $q = kn^{2/9}$. Left: without covariate. Right: with covariate. Top: from high to low, $n = 150, 300, 600$. Bottom: $n = 600$ with better resolution. The dashed lines correspond to $q = n$.

gradually decreases as k increases indicates that $q \asymp n^{2/9}$ is the "correct" scale. The plots suggest that a k around $10 \sim 15$ could be stable enough for practical use.

In practice, we suggest the use of $q = kn^{2/9}$ with k around 10 to 15 for tensor product splines; examples with "barely" square integrable second derivatives may be artificially constructed but we doubt there are many such "true" functions in the real world.

4 Bayesian Confidence Intervals

We now derive Bayesian confidence intervals for the hazard estimates of (1). Empirical coverage properties of these intervals will be assessed in Section 5.

Following Kim and Gu (2004), consider $\eta = \eta_0 + \eta_1$, where η_0 has a diffuse prior in $\mathcal{N}_J$ and η_1

has a mean 0 Gaussian process prior with the covariance function

$$E[\eta_1(x_1)\eta_1(x_2)] = bR_J(x_1, \mathbf{v}^T)Q^+R_J(\mathbf{v}, x_2),$$

where $x_1, x_2 \in \mathcal{T} \times \mathcal{U}$ and Q^+ is the Moore-Penrose inverse of $Q = R_J(\mathbf{v}, \mathbf{v}^T)$; one may parameterize $\eta_0 = \sum_{\nu=0}^{m} d_\nu \phi_\nu$ with d_ν diffuse and $\eta_1 = \sum_{j=1}^{q} c_j \xi_j$ with $\mathbf{c} \sim N(\mathbf{0}, bQ^+)$. Setting $b = 1/n\lambda$, the minimizer η_λ of (1) in $\mathcal{H}_q$ is seen to be the posterior mode under this prior. The Bayesian confidence intervals are based on a quadratic approximation of the log likelihood at this posterior mode.

We use the parameterization of (2) and refer the function $\eta = \boldsymbol{\phi}^T\mathbf{d} + \boldsymbol{\xi}^T\mathbf{c} = \boldsymbol{\psi}^T\mathbf{b}$ and the coefficients $(\mathbf{d}^T, \mathbf{c}^T)^T = \mathbf{b}$ interchangeably. Through a second order Taylor expansion of the second term of (3), its quadratic approximation at $\tilde{\eta} = \eta_\lambda$ is seen to be

$$\frac{1}{2n}(\mathbf{b} - \tilde{\mathbf{b}})^T(nH)(\mathbf{b} - \tilde{\mathbf{b}}) + C, \tag{5}$$

where H is the left-hand side matrix in (4), $\tilde{\mathbf{b}}^T = (\tilde{\mathbf{d}}^T, \tilde{\mathbf{c}}^T)$ is the solution of (4), and C is a constant; detailed calculations are tedious but straightforward, which we omit here. Note that (1) and (3) are the minus log posterior divided by n, thus the approximate posterior of $\mathbf{b} = (\mathbf{d}^T, \mathbf{c}^T)^T$ through (5) is Gaussian with mean $\tilde{\mathbf{b}}$ and covariance H^+/n. It follows that the approximate posterior mean of $\eta(x) = \eta(t, u)$ is $\tilde{\eta}(x) = \boldsymbol{\phi}^T(x)\tilde{\mathbf{d}} + \boldsymbol{\xi}^T(x)\tilde{\mathbf{c}} = \boldsymbol{\psi}^T(x)\tilde{\mathbf{b}}$ and the approximate posterior variance is $s^2(x) = \boldsymbol{\psi}^T(x)H^+\boldsymbol{\psi}(x)/n$, which are to be used to construct the Bayesian confidence intervals $\tilde{\eta}(x) \pm z_{\alpha/2}s(x)$.

5 Coverage Properties

For penalized likelihood regression, the Bayesian confidence intervals derived from the Bayes models demonstrate a certain frequentist across-the-function coverage property; see, e.g., Wahba (1983), Nychka (1988) and Gu (1992). In this section, we present simple simulation results to assess the coverage properties of the intervals derived in Section 4.

Two sets of experiments were carried out with the simulation settings of Section 3, one with and one without covariate. For each set, one hundred replicates were generated with sample size $n = 300$, and cross-validated estimates were calculated using $q = 15n^{2/9} = 54$. The nominal 95% Bayesian confidence intervals $\eta_\lambda \pm 1.96s$ for log hazard were calculated on the sampling points and the coverage of the intervals were recorded. For the set without covariate, the average coverage was 95.55% over all the sampling points and was 96.55% over uncensored points; the corresponding numbers for the set with covariate were 90.63% and 91.62%, respectively. The empirical pointwise mean coverage along the time line for cases without covariate is shown in Figure 2, with the magnitude of the curvature $|\ddot{\eta}|$ superimposed. Lower coverage appears to roughly track high

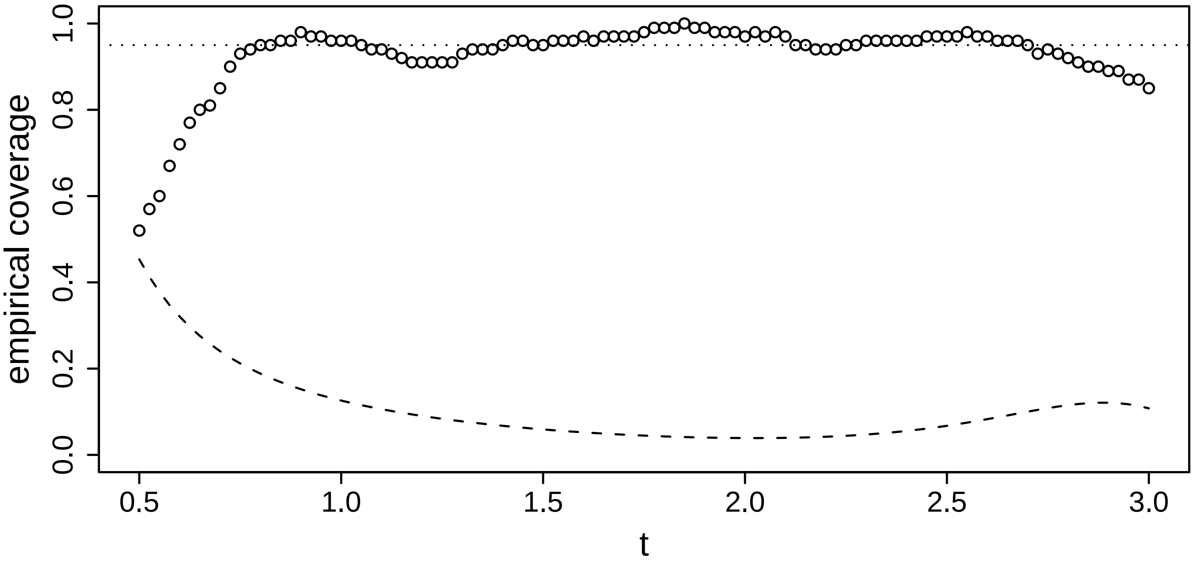

Figure 2: Point-Wise Coverage In Simulations Without Covariate. The circles are point-wise coverage of 95% Bayesian confidence intervals. The dotted line is the nominal coverage 95%. The dashed line depicts the magnitude of $|\ddot{\eta}|$, off scale.

curvature, similar to the observations of Wahba (1983), Nychka (1988), and Gu (1992) in regression settings.

Figure 3 illustrates the Bayesian confidence intervals in one of the cases without covariate, superimposed with the test hazard function, the raw data in the form of discrete empirical hazards, and the size of the at-risk set. Note that the band is connected point-wise intervals, with no simultaneous coverage property intended. The widths of the intervals appear to be of the proper magnitude, and it is reassuring to see the widening of the intervals towards the upper end of the time axis where information from the data is vanishing.

6 Example: Australian AIDS Data

We now apply the techniques developed to analyze a real data set and illustrate along the way the associated R code from the `gss` package by the second author; timing results are also provided.

The data were extracted from the Australian AIDS Data listed in Venables and Ripley (1999), available in R through Brian Ripley's `MASS` package as a data frame `Aids2`; the data were originally collected by the Australian National Center in HIV Epidemiology and Clinical Research. The data consists of 7 variables on 2843 individuals diagnosed with AIDS before July 1, 1991: `state` (state of origin), `sex`, `diag` (date of diagnosis), `death` (date of event or censoring), `status` (censoring indicator), `T.categ` (transmission category), and `age` (age at diagnosis). The subset we consider

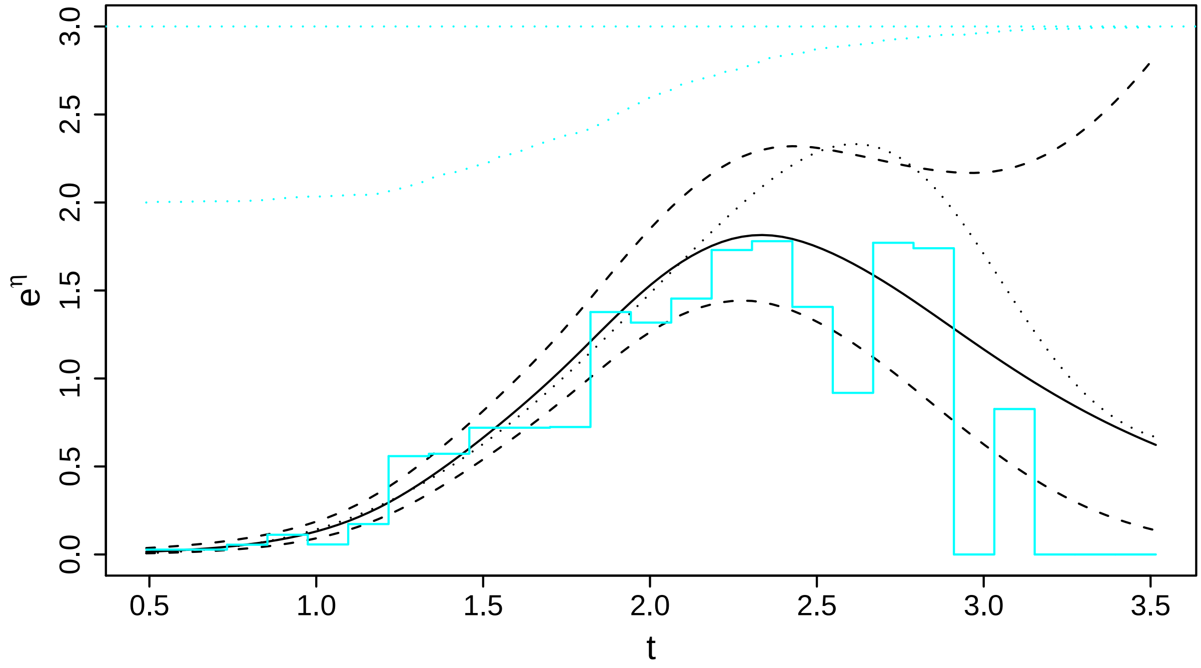

Figure 3: Bayesian Confidence Intervals In Simulation Without Covariate. The cross-validated estimate e^{η_λ} is in solid line, the 95% intervals $e^{\eta_\lambda \pm 1.96 s}$ are in dashed lines, and the test hazard e^{η} is in dotted line. The faded steps plot the empirical hazards calculated from discretized data. The faded dotted line from above depicts the size of the at-risk set, off scale.

here contains the 2443 patients whose transmission was through male homosexual or bisexual contact and who were not dead at diagnosis; all but one patient in the subset is male, so `sex` is also removed from the covariate list along with `T.categ`. The time origin is set at the date of diagnosis and we create a variable `futime` by subtracting `diag` from `death`. A total of 932 of the 2443 subjects, or about 38%, were censored.

With the data of 4 variables `futime`, `status`, `state` and `age` on 2443 subjects stored in a data frame `aids`, the following R command fits a "full factorial" log hazard model of 3 variables with a constant, 3 main effects, 3 two-way interactions, and a three-way interaction:

```
    fit0 <- sshzd(Surv(futime,status)~futime*state*age,data=aids,nbasis=85)
```

where `state` is a factor of 4 levels and $q = 15n^{2/9} = 85$ is specified via `nbasis`; see Example 2 for ANOVA decomposition of functions with 2 variables. The model diagnostics of Gu (2004) based on the Kullback-Leibler projection could be used to trim the model a bit; for example, executing the command

```
    project(fit0,include=c("futime","state","age"))
```

reveals that about 29% of the structure in `fit0` is lost when the fit is projected to an additive model for log hazard, $\eta(t,s,a) = \eta_{\emptyset} + \eta_t + \eta_s + \eta_a$; see Gu (2004) for the quantification of the amount of structure in a fit. After explorations with a variety of term inclusion/exclusion combinations, it was concluded that a model containing the terms `"futime"`, `"state"`, `"age"`, `"futime:state"`, and `"futime:age"` plus the constant would be appropriate, losing about 3% of the structure; such a model can be fitted via

```
    fit <- sshzd(Surv(futime,status)~futime*(state+age),data=aids,nbasis=85)
```

Hazard proportionality does not hold in this model, but the two covariates do not interact at fixed time point. To evaluate the fit on a grid `time` at `state="NSW"` and `age=35`, use

```
    est <- hzdcurve.sshzd(fit,time,cov=data.frame(state=factor("NSW"),
                                 age=35),se=TRUE)
```

where `se=TRUE` asks for the standard errors $s(x)$ derived in Section 4. Bayesian confidence limits with a nominal 95% coverage for the hazard function are to be constructed via `est$fit*exp(1.96*est$se)` and `est$fit/exp(1.96*est$se)`.

Shown in Figure 4 are 8 slices of the hazard fit along with the corresponding 95% Bayesian confidence intervals, at the quartiles of the `age` variable, `age=31` and `age=43`, and at the four levels of the `state` variable, `state="NSW"` (New South Wales), `state="QLD"` (Queensland), `state="VIC"` (Victoria), and `state="Other"`.

For the record, `fit0` and `fit` both took about 135 CPU minutes to compute, on an Athlon MP2900+ workstation with 4 GB RAMS running FreeBSD 4.4 and R 1.8.1.

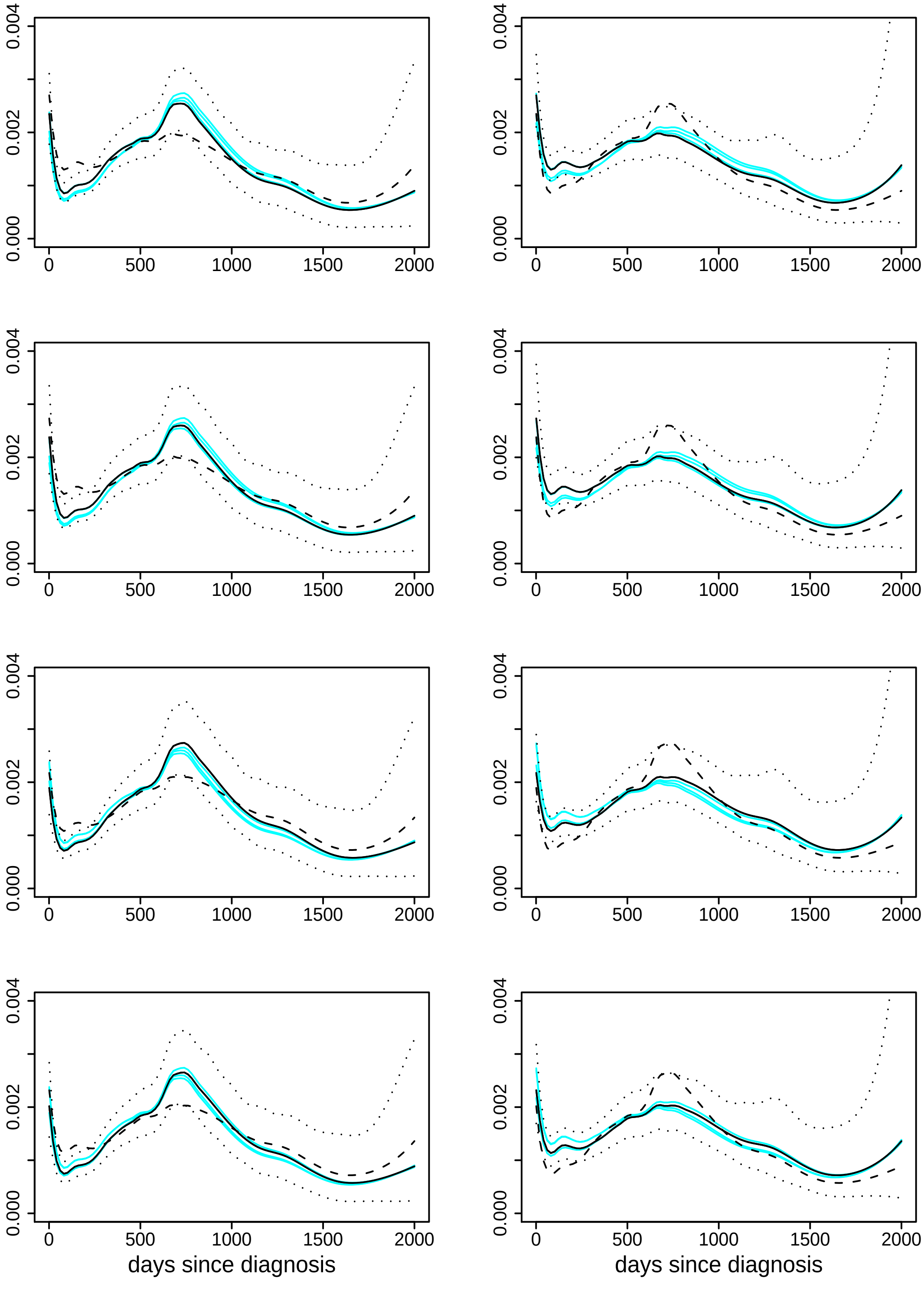

Figure 4: Hazard Estimates for Australian AIDS Data. Left: `age=31`. Right: `age=43`. From top to bottom: New South Wales, Queensland, Victoria, and others. The hazard fit is in solid line and the 95% Bayesian confidence intervals are in dotted lines; superimposed are the other hazard fit in the same row as dashed line and those in the same column as faded lines.

7 Remarks

In this article, we have studied the practical computation of penalized hazard estimation via efficient approximation and derived and illustrated Bayesian confidence intervals for use with the estimates. Open-source R code has been developed to implement the techniques and its usage is demonstrated through real-data example.

With no covariate, the linear algebra calculations of order $O(nq^2)$ comprise the main computational load, but with covariate, the multiple integrations of order $O(ldq^2)$ typically dominate, so the number of distinctive covariate values is a major factor determining the overall computational cost.

While the method takes only static covariates, time-varying covariates that are piece-wise constants can be accommodated by breaking a subject into multiple ones with non-overlapping at-risk periods through left truncation and right censoring.

Acknowledgements

The authors thank John Fieberg for communications concerning the Bayesian confidence intervals. This research was supported by U.S. National Institutes of Health under Grant R33HL68515.

References

Anderson, J. A. and A. Senthilselvan (1980). Smooth estimates for the hazard function. *J. Roy. Statist. Soc. Ser. B 42*, 322–327.

Gu, C. (1992). Penalized likelihood regression: A Bayesian analysis. *Statist. Sin. 2*, 255–264.

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

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

Gu, C. (2004). Model diagnostics for smoothing spline ANOVA models. *The Canadian Journal of Statistics 00*, 000–000.

Gu, C. and Y.-J. Kim (2002). Penalized likelihood regression: General formulation and efficient approximation. *Can. J. Statist. 30*, 619–628.

Gu, C. and J. Wang (2003). Penalized likelihood density estimation: Direct cross validation and scalable approximation. *Statist. Sin. 13*, 811–826.

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

Kim, Y.-J. and C. Gu (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *J. Roy. Statist. Soc. Ser. B 66*, 337–356.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc. 83*, 1134–1143.

O'Sullivan, F. (1988a). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput. 9*, 363–379.

O'Sullivan, F. (1988b). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM J. Sci. Statist. Comput. 9*, 531–542.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J. Roy. Statist. Soc. Ser. B 47*, 1–52 (with discussions).

Venables, W. N. and B. D. Ripley (1999). *Modern Applied Statistics with S-PLUS* (3rd ed.). New York: Springer-Verlag.

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

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