Lifetime Data Analysis manuscript No.
(will be inserted by the editor)

Bayes Factors For Choosing Among Six Common Survival Models

**Jiajia Zhang · Timothy Hanson ·
Haiming Zhou**

the date of receipt and acceptance should be inserted later

J. Zhang
Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC 29208, USA
E-mail: jzhang@mailbox.sc.edu

T. Hanson
Senior Principal Statistician, Medtronic Inc., Minneapolis, Minnesota, U.S.A.

H. Zhou
Division of Statistics, Northern Illinois University, DeKalb, IL 60115, USA

Abstract A super model that includes proportional hazards, proportional odds, accelerated failure time, accelerated hazards, and extended hazards models, as well as the model proposed in Diao et al. (2013) accounting for crossed survival as special cases is proposed for the purpose of testing and choosing among these popular semiparametric models. Efficient methods for fitting and computing fast, approximate Bayes factors are developed using a nonparametric baseline survival function based on a transformed Bernstein polynomial. All manner of censoring is accommodated including right, left, and interval censoring, as well as data that are observed exactly and mixtures of all of these; current status data are included as a special case. The method is tested on simulated data and two real data examples. The approach is easily carried out via a new function in the `spBayesSurv` R package.

Keywords Interval censoring · Model choice · Bernstein polynomial · Bayes factor

1 Introduction

One of the central interests in health sciences research is to identify and quantify the association between the mortality/incidence of a certain disease and its potential risk factors, so that risk factors can be used in disease prevention and control. Survival models serve as the major statistical tools in analyzing mortality/incidence data. Among them, the Cox proportional hazards model (PH) (Cox, 1992) is unquestionably the most popular one in practice, where the risk factor has a multiplicative association with hazard risk. Let $H_{\mathbf{x}}(t)$ be the cumulative hazard function for a subject with covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ and $H_0(t)$ be the baseline cumulative hazard function for those with $\mathbf{x} = \mathbf{0}$. The proportional hazards model can be written as

$$H_{\mathbf{x}}(t) = e^{\boldsymbol{\beta}'\mathbf{x}} H_0(t)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a vector of unknown coefficients, and e^{β_j} represents the hazard ratio corresponding to a one unit increase of the jth covariate. When the proportional hazard assumption is invalid, the accelerated failure time (AFT) model (Kalbfleisch and Prentice, 2011), and proportional odds (PO) model can be considered as alternative models. Let $S_{\mathbf{x}}(t)$ and $S_0(t)$ denote the survival function and baseline survival function, and $h_{\mathbf{x}}(t), h_0(t)$ denote the hazard and baseline hazard function corresponding to $H_{\mathbf{x}}(t), H_0(t)$. The accelerated failure time model can be written as

$$S_{\mathbf{x}}(t) = S_0\left\{e^{\boldsymbol{\beta}'\mathbf{x}}\, t\right\},$$

where e^{β_j} represents the time scale change due to the jth covariate in survival probability. The proportional odds model can be written as

$$\frac{1 - S_{\mathbf{x}}(t)}{S_{\mathbf{x}}(t)} = e^{\boldsymbol{\beta}'\mathbf{x}} \frac{1 - S_0(t)}{S_0(t)},$$

where e^{β_j} represents the change in failure odds by time t due to the jth covariate.

All of the models mentioned above do not account for crossing survival curves for different covariate combinations. Failure to capture crossing survival may incorrectly characterize the association between risk factors and mortality/incidence. Chen and Wang (2000) and Chen et al. (2014) consider the accelerated hazards model (AH), which is

$$h_{\mathbf{x}}(t) = h_0 \left\{ e^{\boldsymbol{\beta}'\mathbf{x}}\ t \right\},$$

where e^{β_j} represents the time scale change in hazard risk due to the jth covariate. Zhang and Peng (2009) discuss properties of the hazard function under PH, AH and AFT models. Etezadi-Amoli and Ciampi (1987), Chen and Jewell (2001), and Li et al. (2015) consider the extended hazards model (EH)

$$h_{\mathbf{x}}(t) = \exp(\boldsymbol{\beta}'\mathbf{x}) h_0\{\exp(\boldsymbol{\gamma}'\mathbf{x})t\}. \tag{1}$$

Here, $\boldsymbol{\beta} = \boldsymbol{\gamma}$ gives AFT, $\boldsymbol{\gamma} = \mathbf{0}$ gives PH, and $\boldsymbol{\beta} = \mathbf{0}$ gives the AH model.

Quantin et al. (1996) consider a generalization of PH that allows for crossing survival curves $H_{\mathbf{x}}(t) = e^{\boldsymbol{\beta}'\mathbf{x}} H_0(t)^{\exp(\mathbf{x}'\boldsymbol{\gamma})}$; the PH model is formally nested within when $\boldsymbol{\gamma} = \mathbf{0}$. Also see Devarajan and Ebrahimi (2011) and references therein. Diao et al. (2013) add covariates to the model of Yang and Prentice (2005) (YP) yielding the hazard model

$$h_{\mathbf{x}}(t) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x} + \boldsymbol{\gamma}'\mathbf{x}) h_0(t)}{\exp(\boldsymbol{\beta}'\mathbf{x}) F_0(t) + \exp(\boldsymbol{\gamma}'\mathbf{x}) S_0(t)}. \tag{2}$$

They point out that

$$\lim_{t\to 0^+} \frac{h_{\mathbf{x}_1}(t)}{h_{\mathbf{x}_2}(t)} = e^{\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)}, \quad \lim_{t\to\infty} \frac{h_{\mathbf{x}_1}(t)}{h_{\mathbf{x}_2}(t)} = e^{\boldsymbol{\gamma}'(\mathbf{x}_1 - \mathbf{x}_2)},$$

thus $\boldsymbol{\beta}$ gives a short-term relative risk interpretation whereas $\boldsymbol{\gamma}$ gives a long-term relative risk interpretation. Note that $\boldsymbol{\beta} = \boldsymbol{\gamma}$ and $\boldsymbol{\gamma} = \mathbf{0}$ give PH and PO as a formally nested special cases, respectively.

Each model listed above can capture different characteristics of survival data. However, choosing which model is the most appropriate and accurate in reflecting the association between the potential risk factors and mortality/incidence is a challenge and an important question that needs to be addressed. The YP model (2) and EH model (1) augment $\boldsymbol{\beta}$ with an entirely new set of p regression effects, say $\boldsymbol{\gamma}$, to formally nest simpler models within a larger super model. Such augmentations allow for simpler models to be special cases arising from standard linear constraints on the parameters, thus the likelihood ratio tests for frequentist models, or efficient computation of Bayes factors for Bayesian models can be used.

Another aspect of survival analysis is that survival times can be censored in myriad ways, including right, left, and interval censoring (Sun, 2006), as well as data that are observed exactly and mixtures of all of these; current status data are included as a special case. It is challenge to handle all types of censoring simultaneously from frequentist approaches.

This paper develops a super model that includes the PH, PO, AFT, AH, YP and EH models as formally nested special cases. As such, model choice among these models can be carried out by computing approximate Bayes factors based on the Savage-Dickey ratio (Verdinelli and Wasserman, 1995). A transformed Bernstein polynomial prior proposed by Chen et al. (2014) is used to model baseline survival S_0 and a multivariate normal g-prior for regression coefficients is developed. All manner of censoring is accommodated and the approach is implemented via a new function in the `spBayesSurv` R package. Once a model is chosen, any of PH, PO, or AFT can be fitted through many existing R packages including the `spBayesSurv` R package. The remaining paper is organized as follows: Section 2 describes the proposed super model; Section 3 lists details about the Bayesian estimation procedure, including prior development, posterior sampling, and Bayes factor computation; Section 4 presents a simulation and two real data analyses with software implementation. Conclusions are made in Section 5.

2 Model

The super model proposed has the following closed form

$$S_{\mathbf{x}}(t) = \left[1 + e^{(\boldsymbol{\beta}_o - \boldsymbol{\beta}_h + \boldsymbol{\beta}_q)'\mathbf{x}} \frac{F_0\left\{e^{\boldsymbol{\beta}_q'\mathbf{x}}\, t\right\}}{S_0\left\{e^{\boldsymbol{\beta}_q'\mathbf{x}}\, t\right\}}\right]^{-\exp\{(\boldsymbol{\beta}_h - \boldsymbol{\beta}_q)'\mathbf{x}\}}, \qquad (3)$$

where the baseline cumulative distribution and survival functions are $F_0(\cdot)$ and $S_0(\cdot)$. The hazard is computed to be

$$h_{\mathbf{x}}(t) = \frac{e^{(\boldsymbol{\beta}_o + \boldsymbol{\beta}_h + \boldsymbol{\beta}_q)'\mathbf{x}} h_0\left\{e^{\boldsymbol{\beta}_q'\mathbf{x}}\, t\right\}}{e^{(\boldsymbol{\beta}_o + \boldsymbol{\beta}_q)'\mathbf{x}} F_0\left\{e^{\boldsymbol{\beta}_q'\mathbf{x}}\, t\right\} + e^{\boldsymbol{\beta}_h'\mathbf{x}} S_0\left\{e^{\boldsymbol{\beta}_q'\mathbf{x}}\, t\right\}}. \qquad (4)$$

Then $f_{\mathbf{x}}(t) = h_{\mathbf{x}}(t)S_{\mathbf{x}}(t)$ through (3) and (4).

The super model includes PH, AFT, PO, AH, EH and YP models as special cases. One can show if $H_h : \boldsymbol{\beta}_q = \mathbf{0},\ \boldsymbol{\beta}_o = \boldsymbol{\beta}_h$ is true, then

$$S_{\mathbf{x}}(t) = S_0(t)^{\exp(\boldsymbol{\beta}_h'\mathbf{x})},$$

the PH model obtains. Similarly, assuming $H_o : \boldsymbol{\beta}_q = \boldsymbol{\beta}_h = \mathbf{0}$ implies

$$\frac{F_{\mathbf{x}}(t)}{S_{\mathbf{x}}(t)} = e^{\boldsymbol{\beta}_o'\mathbf{x}} \frac{F_0(t)}{S_0(t)},$$

the PO model. Assuming $H_q : \boldsymbol{\beta}_o = \mathbf{0},\ \boldsymbol{\beta}_h = \boldsymbol{\beta}_q$ implies

$$S_{\mathbf{x}}(t) = S_0\left\{e^{\boldsymbol{\beta}_q'\mathbf{x}}\, t\right\},$$

the AFT model (proportional quantiles). Assuming $H_a : \boldsymbol{\beta}_h = \mathbf{0},\ \boldsymbol{\beta}_q + \boldsymbol{\beta}_o = \mathbf{0}$ implies

$$h_{\mathbf{x}}(t) = h_0 \left\{ e^{\boldsymbol{\beta}_q' \mathbf{x}}\ t \right\},$$

the AH model obtains. YP model (2) occurs as a special case when $H_y : \boldsymbol{\beta}_q = \mathbf{0}$; EH model (1) is a special case when $H_e : \boldsymbol{\beta}_h = \boldsymbol{\beta}_q + \boldsymbol{\beta}_o$.

We seek to fit model (3) assuming a transformed Bernstein polynomial prior on $S_0(\cdot)$, and test the adequacy of the formally nested hypotheses H_h, H_o, H_q, H_a, H_y and H_e via Bayes factors.

3 Priors and Bayes factors

3.1 Transformed Bernstein polynomial prior on baseline survival S_0

For a given positive integer J, the Bernstein polynomial of degree $J-1$ is defined by

$$b(x|J, \boldsymbol{\xi}_J) = \sum_{j=1}^{J} \xi_{Jj} \beta(x|j, J-j+1), \qquad (5)$$

where $\boldsymbol{\xi}_J = (\xi_{J1}, \ldots, \xi_{JJ})'$ is a vector of positive weights satisfying $\sum_{j=1}^{J} \boldsymbol{\xi}_{Jj} = 1$ and $\beta(\cdot|a, b)$ denotes a beta density with parameters (a, b). Clearly $b(x|J, \boldsymbol{\xi}_J)$ is a density function and is very flexible, so that, in fact, any smooth density with support $(0, 1)$ can be well approximated by a Bernstein polynomial (Ghosal, 2001). More precisely, if $f(x)$ is any continuously differentiable density with support $(0, 1)$ and bounded second derivative, it can be shown that, with suitable choice of $\boldsymbol{\xi}_J$,

$$\sup_{0<x<1} |f(x) - b(x|J, \boldsymbol{\xi}_J)| = O(J^{-1}).$$

Integrating (5) gives the corresponding cumulative distribution function (cdf)

$$B(x|J, \boldsymbol{\xi}_J) = \sum_{j=1}^{J} \xi_{Jj} I_x(j, J-j+1), \qquad (6)$$

where $I_x(a, b)$ is the cdf associated with $\beta(x|a, b)$. Note that one can calculate (6) without too much computational cost through the recursive relation

$$I_x(j+1, J-j) = I_x(j, J-j+1) - \frac{\Gamma(J+1)}{\Gamma(j+1)\Gamma(J-j+1)} x^j (1-x)^{J-j}.$$

A referee has brought up the issue of consistency and the choice of J. Note at the outset that *none* of the semiparametric models under consideration support the "truth," as they are all first-order approximations to reality formulated to provide readily interpretable regression coefficients. However, some assurance that the Bernstein polynomial supports a wide range of density shapes and is consistent over this range is comforting. Petrone and Wasserman

(2002) show that under mild conditions on the true underlying density and suitable priors on $J \in \mathbb{N}^+$ and $\boldsymbol{\xi}_J$, the posterior predictive density (i.e. Bayes estimate of the density with respect to quadratic loss) is Hellinger consistent. Fitting such a model is complicated and typically done via reversible jump MCMC (**?**). As such, the vast majority of authors simply fix J at some "reasonable" value, truncating the estimate; Chen et al. (2014) suggest $J = 15$ based on simulations involving the random L_1 distance between the prior and the truth. Accordingly, Petrone and Wasserman (2002) further argue that a truncated Bernstein polynomial will converge to a Bayes estimate that minimizes the Kullback-Liebler distance between the truncated estimate and the truth. Certainly larger values $J > 15$ can be chosen to achieve more flexible estimates of the baseline survival density; the supplemental material inChen et al. (2014) can provide a guide in terms of L_1. However, there is a "law of diminishing returns", also observed by Hanson (2006), in that the LPML tends to level off and not increase after some K for $J \geq K$. Restated, the cross-validated predictive ability of the model does *not increase* after some K. In this spirit, and similar to the use of AIC in choosing the number of mixands in finite mixture models, one could choose $J = K$ based on when LPML levels off. However, each computation of the LPML requires a separate MCMC run.

A remarkably useful result is that any Bernstein polynomial can be written in terms of Bernstein polynomials of higher degree through the relation

$$\beta(x|j, J-j) = \frac{J-j}{J}\beta(x|j, J-j+1) + \frac{j}{J}\beta(x|j+1, J-j).$$

It follows that $b(x|J-1, \boldsymbol{\xi}_{J-1})$ can be written as $b(x|J, \boldsymbol{\xi}_J^*)$ with suitable choice of $\boldsymbol{\xi}_J^*$. Since every lower order Bernstein polynomial $J < K$ is included as a special case of $J = K$, one only need pick one reasonable $J = K$; a prior on $1 \leq J \leq K$ is superfluous.

Regarding the prior for $\boldsymbol{\xi}_J$, we consider a Dirichlet distribution,

$$\boldsymbol{\xi}_J|J \sim \text{Dirichlet}(\alpha, \ldots, \alpha), \tag{7}$$

where $\alpha > 0$ is a parameter. An attractive property of the BP prior specified above is that $E[b(x|J, \boldsymbol{\xi}_J)] = \sum_{j=1}^{J} \beta(x|j, J-j+1)/J = 1$ and $E[B(x|J, \boldsymbol{\xi}_J)] = x$ for $x \in (0, 1)$.

We next describe how we define a random survival function S_0 based on (5). Let $\{S_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ denote a parametric family of survival functions with support on positive reals $\mathbb{R}^+$. For example, a log-logistic family is defined as $S_{\boldsymbol{\theta}}(t) = \{1+(e^{\theta_1}t)^{\exp(\theta_2)}\}^{-1}$ in our R function, where $\boldsymbol{\theta} = (\theta_1, \theta_2)'$. Weibull and log-normal families are also implemented in the function. In our experience all three parametric distribution families yield similar results across many data sets. Note that $S_{\boldsymbol{\theta}}(t)$ always lies in the interval $(0, 1)$ for $0 < t < \infty$, so a natural prior on S_0, termed the transformed Bernstein polynomial (TBP) prior, is

$$S_0(t) = B(S_{\boldsymbol{\theta}}(t)|J, \boldsymbol{\xi}_J), \tag{8}$$

with density

$$f_0(t) = b(S_{\boldsymbol{\theta}}(t)|J, \boldsymbol{\xi}_J) f_{\boldsymbol{\theta}}(t), \tag{9}$$

where $f_{\boldsymbol{\theta}}$ is the density associated with $S_{\boldsymbol{\theta}}$. Clearly, the random distribution S_0 is centered at S_θ, that is, $E[S_0(t)] = S_{\boldsymbol{\theta}}(t)$ and $E[f_0(t)] = f_{\boldsymbol{\theta}}(t)$. The weight parameters $\boldsymbol{\xi}_J$ "adjust" the shape of the baseline survival S_0 relative to the prior guess $S_{\boldsymbol{\theta}}$. If all ξ_{Jj}s are equal to $1/J$ then $S_0 \equiv S_{\boldsymbol{\theta}}$. This adaptability makes the TBP prior attractive in its flexibility, but also anchors the random S_0 firmly about $S_{\boldsymbol{\theta}}$. Moreover, unlike a mixture of Polya trees prior, the TBP prior always selects smooth densities, leading to more efficient posterior sampling.

The TBP parameter α acts much like the precision in a Dirichlet process (Ferguson, 1973), controlling how stochastically "pliable" S_0 is relative to $S_{\boldsymbol{\theta}}$. Large values of α indicate a strong belief that S_0 can be modeled using $S_{\boldsymbol{\theta}}$, since as α tends to infinity, the random S_0 is $S_{\boldsymbol{\theta}}$ with probability one. On the other hand, a smaller values of α allow more pronounced deviations of S_0 from $S_{\boldsymbol{\theta}}$. The choice of $\alpha = 1$ has been advocated by many authors, e.g. recently (Chen et al., 2014). Similar to Dirichlet processes we consider a gamma prior on α, say, $\alpha \sim \Gamma(a_0, b_0)$, where a_0 is the shape parameter and b_0 is the rate parameter. Through L_1 considerations, Chen et al. (2014) provide some guidance on choosing an informative prior for α, but this is not pursued here; in our experience different priors for α leads to very similar posterior inference in reasonably large sample sizes.

3.2 Prior on regression coefficients

The g-prior (Zellner, 1983) has been widely considered for model selection in Bayesian regression models. Hanson et al. (2014) develop an informative g-prior for logistic regression; we consider their approach adapted for use in the semiparametric survival models considered here. The prior is

$$\boldsymbol{\beta} \sim N_p(\mathbf{0}, gn(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}), \tag{10}$$

where n is the sample size, $\mathbf{X}^*$ is the usual $n \times p$ design matrix only with mean-centered predictors, i.e. $\mathbf{1}_n'\mathbf{X}^* = \mathbf{0}_p'$. Derivations in Hanson et al. (2014) imply that for covariates $\mathbf{x}$ generated from some distribution H with support on $\mathcal{X} \subset \mathbb{R}^p$ and $\boldsymbol{\beta}$ assigned in (10),

$$e^{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\beta}} \overset{\bullet}{\sim} \log N(\mathbf{0}, gp),$$

where $\boldsymbol{\mu} = \int_{\mathcal{X}} \mathbf{x} H(d\mathbf{x})$. Thus, *a priori*, the relative risks (PH), acceleration factors (AFT), and odds factors (PO) of random individuals $\mathbf{x}$ relative to their sample mean $\bar{\mathbf{x}}$ approximately follow a log-normal distribution in reasonably large samples. A simple method for choosing g is to pick a number M such that any random quantity $e^{(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\beta}}$ is less than M with probability r. A simple calculation reveals that

$$g = \left[\frac{\log M}{\Phi^{-1}(r)}\right]^2 \frac{1}{r}.$$

For example, choosing $M = 10$ and $r = 0.9$ yields $g = \frac{3.228}{p}$; these are the values considered here. Concisely,

$$\boldsymbol{\beta}_h, \boldsymbol{\beta}_o, \boldsymbol{\beta}_q \overset{iid}{\sim} N_p(\mathbf{0}, \mathbf{S}_0^*), \text{ where } \mathbf{S}_0^* = \tfrac{3.228n}{p}(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}.$$

3.3 Likelihood construction and MCMC

Let t_i be a random survival time for the ith individual and $\mathbf{x}_i$ be a related p-dimensional vector of covariates, $i = 1, \ldots, n$. Assume the survival time t_i lies in the interval (a_i, b_i), $0 \le a_i \le b_i \le \infty$. Here left-censored data are of the form $(0, b_i)$, right-censored (a_i, ∞), interval-censored (a_i, b_i) and uncensored values simply have $a_i = b_i$, i.e., we define $(x, x) = \{x\}$.

Denote by $\mathcal{D} = \{(\mathbf{x}_i, a_i, b_i); i = 1, \ldots, n\}$ the set of observed data. Assume $t_i \sim S_{\mathbf{x}_i}(\cdot)$, where $S_{\mathbf{x}_i}(t)$ is given by (3) with the TBP prior on $S_0(t)$ and $f_0(t)$ defined in (8) and (9). Set $\boldsymbol{\beta} = (\boldsymbol{\beta}_h', \boldsymbol{\beta}_o', \boldsymbol{\beta}_q')'$. The likelihood for $(\boldsymbol{\xi}_J, \boldsymbol{\theta}, \boldsymbol{\beta})$ is given by

$$L(\boldsymbol{\xi}_J, \boldsymbol{\theta}, \boldsymbol{\beta}) = \prod_{i=1}^{n} [S_{\mathbf{x}_i}(a_i) - S_{\mathbf{x}_i}(b_i)]^{I\{a_i < b_i\}} f_{\mathbf{x}_i}(a_i)^{I\{a_i = b_i\}}. \tag{11}$$

Markov chain Monte Carlo (MCMC) is carried out through an empirical Bayes approach coupled with adaptive Metropolis-Hastings updating (Haario et al., 2001). The posterior density given the data $\mathcal{D}$ is

$$p(\boldsymbol{\xi}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \alpha | \mathcal{D}) \propto L(\boldsymbol{\xi}_J, \boldsymbol{\theta}, \boldsymbol{\beta}) p(\boldsymbol{\xi}_J | \alpha) p(\alpha) p(\boldsymbol{\theta}) p(\boldsymbol{\beta}_q) p(\boldsymbol{\beta}_o) p(\boldsymbol{\beta}_h),$$

where $p(\boldsymbol{\xi}_J|\alpha)$ is the density of the Dirichlet distribution in (7) and the remaining terms are prior densities for α, $\boldsymbol{\theta}$, $\boldsymbol{\beta}_h$, $\boldsymbol{\beta}_o$, and $\boldsymbol{\beta}_q$. Here we assume $\alpha \sim \Gamma(a_0, b_0)$, $\boldsymbol{\theta} \sim N_2(\boldsymbol{\theta}_0, \mathbf{V}_0)$ and $\boldsymbol{\beta}_h, \boldsymbol{\beta}_o, \boldsymbol{\beta}_q \overset{iid}{\sim} N_p(\mathbf{0}, \mathbf{S}_0^*)$.

Note that when $\boldsymbol{\xi}_{Jj} = 1/J$ the underlying parametric model with $S_0(t) = S_{\boldsymbol{\theta}}(t)$ is obtained and $\mathcal{L}(\boldsymbol{\xi}_J, \boldsymbol{\theta}, \boldsymbol{\beta})$ is equal to the corresponding parametric likelihood function. A fit from a standard parametric survival model can provide starting values for the TBP survival model. Consider a standard fit $\log t_i = \tau_0 + \boldsymbol{\tau}'\mathbf{x}_i + \sigma\epsilon_i$ using the `survreg` function in the `survival` package for R, where $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} F(\epsilon)$. For log-logistic data $F(\epsilon) = \frac{e^\epsilon}{1+e^\epsilon}$ (standard logistic), for Weibull $F(\epsilon) = 1 - \exp(-e^\epsilon)$ (extreme value), and for log-normal $F(\epsilon)$ is the standard normal cdf. This model has a scale σ, intercept τ_0, and regression coefficients $\boldsymbol{\tau}' = (\tau_1, \ldots, \tau_p)$. We parametrize $S_{\boldsymbol{\theta}}(t)$ so that $\theta_1 = -\tau_0$ and $\theta_2 = -\log\sigma$. Let $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{V}}$ be the point and asymptotic variance estimates for $\boldsymbol{\theta}$ via the `survreg` fit. To choose starting values for $\boldsymbol{\beta}$, we fit both the Weibull and log-logistic `survreg` models. Noting that the Weibull model has both PH and AFT representations and the log-logistic model has both PO and AFT representations, the `survreg` fits with Weibull and log-logistic will provide us coefficient estimates under each of the PH, PO and AFT, denoted by $\boldsymbol{\beta}_{h0}$, $\boldsymbol{\beta}_{o0}$ and $\boldsymbol{\beta}_{q0}$, respectively, and let $\mathbf{S}_{h0}$,

$\mathbf{S}_{o0}$ and $\mathbf{S}_{q0}$ be their covariance estimates. If the Weibull model has smaller AIC, we set $\hat{\boldsymbol{\beta}} = (\boldsymbol{\beta}_{q0}', \mathbf{0}', \boldsymbol{\beta}_{q0}')'$ and $\hat{\mathbf{S}} = \text{diag}(\mathbf{S}_{q0}, \mathbf{S}_{o0}, \mathbf{S}_{q0})$; otherwise, we set $\hat{\boldsymbol{\beta}} = (\mathbf{0}', \boldsymbol{\beta}_{o0}', \mathbf{0}')'$ and $\hat{\mathbf{S}} = \text{diag}(\mathbf{S}_{h0}, \mathbf{S}_{o0}, \mathbf{S}_{q0})$.

For ease of posterior sampling, we work with $\mathbf{z} = (z_1, \ldots, z_{J-1})'$ through the relation $\xi_{Jj} = e^{z_j}/(\sum_{k=1}^J e^{z_j})$ for $j = 1, \ldots, J$, where $z_J = 0$. Under the Dirichlet prior (7), the induced prior on $\mathbf{z}$ is:

$$p(\mathbf{z}|\alpha) = \frac{\Gamma(\alpha J)}{\Gamma(\alpha)^J} \prod_{j=1}^J \left[\frac{e^{z_j}}{\sum_{k=1}^J e^{z_j}} \right]^\alpha .$$

The vector $\mathbf{z}$ can be updated using adaptive Metropolis-Hastings. Suppose we are currently in iteration l and have sampled the states $\mathbf{z}^{(0)}, \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(l-1)}$. We select an index l_0 (e.g., $l_0 = 5000$) for the length of an initial period and define

$$\boldsymbol{\Sigma}_l = \begin{cases} \boldsymbol{\Sigma}_0, & l \leq l_0 \\ \frac{(2.4)^2}{J-1}(\mathcal{C}_l + 10^{-10}\mathbf{I}_{J-1}) & l > l_0. \end{cases}$$

Here $\mathcal{C}_l$ is the sample variance of $\mathbf{z}^{(0)}, \mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(l-1)}$, and $\boldsymbol{\Sigma}_0$ is an initial diagonal covariance matrix of $\mathbf{z}$, defined so that the variance of z_j is 0.16. The choice of 0.16 is based on extensive simulation studies; other choices (as long as it is not too small or large) will have little impact on posterior inferences. We generate $\mathbf{z}^* = (z_1^*, \ldots, z_{J-1}^*)'$ from $N_{J-1}(\mathbf{z}^{(l-1)}, \boldsymbol{\Sigma}_l)$ and accept it with probability

$$\min\left\{1, \frac{L(\boldsymbol{\xi}_J^*, \boldsymbol{\theta}, \boldsymbol{\beta}) \prod_{j=1}^J (\xi_{Jj}^*)^\alpha}{L(\boldsymbol{\xi}_J^{(l-1)}, \boldsymbol{\theta}, \boldsymbol{\beta}) \prod_{j=1}^J (\xi_{Jj}^{(l-1)})^\alpha}\right\},$$

where $\boldsymbol{\xi}_J^*$ and $\boldsymbol{\xi}_J^{(l-1)}$ are defined corresponding to $\mathbf{z}^*$ and $\mathbf{z}^{(l-1)}$, respectively.

The centering distribution parameters $\boldsymbol{\theta}$ are updated via adaptive Metropolis-Hastings. At iteration l, each candidate is sampled as $\boldsymbol{\theta}^* \sim N_2(\boldsymbol{\theta}^{(l-1)}, \boldsymbol{\Sigma}_l)$ and accepted with probability

$$\min\left\{1, \frac{L(\boldsymbol{\xi}_J, \boldsymbol{\theta}^*, \boldsymbol{\beta})\phi_2(\boldsymbol{\theta}^*|\boldsymbol{\theta}_0, \mathbf{V}_0)}{L(\boldsymbol{\xi}_J, \boldsymbol{\theta}^{(l-1)}, \boldsymbol{\beta})\phi_2(\boldsymbol{\theta}^{(l-1)}|\boldsymbol{\theta}_0, \mathbf{V}_0)}\right\}.$$

where $\phi_2(\cdot|\boldsymbol{\theta}_0, \mathbf{V}_0)$ denotes the density of $N_2(\boldsymbol{\theta}_0, \mathbf{V}_0)$, and $\boldsymbol{\Sigma}_l$ is defined similarly as above, but with $\boldsymbol{\Sigma}_0$ set to be $\hat{\mathbf{V}}$.

The survival model coefficients $\boldsymbol{\beta} \in \{\boldsymbol{\beta}_o, \boldsymbol{\beta}_h, \boldsymbol{\beta}_q\}$ are updated via adaptive Metropolis-Hastings as well with proposal $\boldsymbol{\beta}^* \sim N_p(\boldsymbol{\beta}^{(l-1)}, \boldsymbol{\Sigma}_l)$ and acceptance probability

$$\min\left\{1, \frac{L(\boldsymbol{\xi}_J, \boldsymbol{\theta}, \boldsymbol{\beta}^*)\phi_p(\boldsymbol{\beta}^*|\boldsymbol{\beta}_0, \mathbf{S}_0)}{L(\boldsymbol{\xi}_J, \boldsymbol{\theta}, \boldsymbol{\beta}^{(l-1)})\phi_p(\boldsymbol{\beta}^{(l-1)}|\boldsymbol{\beta}_0, \mathbf{S}_0)}\right\},$$

where $\boldsymbol{\Sigma}_l$ is defined similarly as above with $\boldsymbol{\Sigma}_0 = \hat{\mathbf{S}}$.

Finally, the precision parameter α is updated via adaptive Metropolis-Hastings with normal proposal $\alpha^* \sim N_1(\alpha^{(l-1)}, \boldsymbol{\Sigma}_l)$ with $\boldsymbol{\Sigma}_l$ defined as above but taking $\boldsymbol{\Sigma}_0 = 0.16$, and the acceptance probability is

$$\min\left\{1, \frac{(\alpha^*)^{a-1}e^{-b\alpha^*}\Gamma(\alpha^* J)\Gamma(\alpha^{(l-1)})^J \prod_{j=1}^J (\xi_{Jj})^{\alpha^*-1}}{(\alpha^{(l-1)})^{a-1}e^{-b\alpha^{(l-1)}}\Gamma(\alpha^*)^J\Gamma(\alpha^{(l-1)}J)\prod_{j=1}^J (\xi_{Jj})^{\alpha^{(l-1)}-1}}\right\}.$$

Regarding default choices for hyperparameters, we set $a_0 = b_0 = 1$, $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$, and $\mathbf{V}_0 = 10\hat{\mathbf{V}}$. Note here we assume a relatively informative prior on $\boldsymbol{\theta}$ to avoid potential instability of MCMC and obviate confounding between $S_{\boldsymbol{\theta}}$ and the Bernstein polynomial.

3.4 Approximate Bayes factors for model selection

Once the model is fitted via MCMC, the triples $\{(\boldsymbol{\beta}_q^m, \boldsymbol{\beta}_o^m, \boldsymbol{\beta}_h^m)\}_{m=1}^M$ are obtained after burnin and thinning. Let BF_q, BF_o, BF_h, BF_a, BF_y, BF_e be the Bayes factors for testing the AFT, PO, PH, AH, YP and EH assumptions relative to the full model, respectively. A large-sample approximation to the Savage-Dickey ratio based on approximate normality is proposed to compute these Bayes factors (Li et al., 2015, Zhou et al., 2017); see Appendix A of the online material for details.

The BF for PH relative to the super model is

$$BF_h \approx \frac{N_{2p}(\mathbf{0}; \mathbf{m}_h, \mathbf{V}_h)}{N_p(\mathbf{0}; \mathbf{0}, \mathbf{S}_0^*)N_p(\mathbf{0}; \mathbf{0}, 2\mathbf{S}_0^*)},$$

where $\mathbf{m}_h$ and $\mathbf{V}_h$ are the posterior sample mean and variance of $(\boldsymbol{\beta}_q', \boldsymbol{\beta}_o' - \boldsymbol{\beta}_h')'$, respectively. The BF for AFT relative to the super model is

$$BF_q \approx \frac{N_{2p}(\mathbf{0}; \mathbf{m}_q, \mathbf{V}_q)}{N_p(\mathbf{0}; \mathbf{0}, \mathbf{S}_0^*)N_p(\mathbf{0}; \mathbf{0}, 2\mathbf{S}_0^*)},$$

where $\mathbf{m}_q$ and $\mathbf{V}_q$ are the posterior sample mean and variance of $(\boldsymbol{\beta}_o', \boldsymbol{\beta}_h' - \boldsymbol{\beta}_q')'$, respectively. The BF for PO relative to the super model is

$$BF_o \approx \frac{N_{2p}(\mathbf{0}; \mathbf{m}_o, \mathbf{V}_o)}{N_p(\mathbf{0}; \mathbf{0}, \mathbf{S}_0^*)N_p(\mathbf{0}; \mathbf{0}, \mathbf{S}_0^*)},$$

where $\mathbf{m}_o$ and $\mathbf{V}_o$ are the posterior sample mean and variance of $(\boldsymbol{\beta}_q', \boldsymbol{\beta}_h')'$, respectively. The BF for AH relative to the super model is

$$BF_a \approx \frac{N_{2p}(\mathbf{0}; \mathbf{m}_a, \mathbf{V}_a)}{N_p(\mathbf{0}; \mathbf{0}, \mathbf{S}_0^*)N_p(\mathbf{0}; \mathbf{0}, 2\mathbf{S}_0^*)},$$

where $\mathbf{m}_a$ and $\mathbf{V}_a$ are the posterior sample mean and variance of $(\boldsymbol{\beta}_h', \boldsymbol{\beta}_q' + \boldsymbol{\beta}_o')'$, respectively. The BF for YP relative to the super model is

$$BF_y \approx \frac{N_p(\mathbf{0}; \mathbf{m}_y, \mathbf{V}_y)}{N_p(\mathbf{0}; \mathbf{0}, \mathbf{S}_0^*)},$$

where $\mathbf{m}_y$ and $\mathbf{V}_y$ are the posterior sample mean and variance of $\boldsymbol{\beta}_q$, respectively. The BF for EH relative to the super model is

$$BF_e \approx \frac{N_p(\mathbf{0}; \mathbf{m}_e, \mathbf{V}_e)}{N_p(\mathbf{0}; \mathbf{0}, 3\mathbf{S}_0^*)},$$

where $\mathbf{m}_e$ and $\mathbf{V}_e$ are the posterior sample mean and variance of $\boldsymbol{\beta}_h - \boldsymbol{\beta}_q - \boldsymbol{\beta}_o$, respectively.

4 Illustrations

4.1 Simulated data

To show that the method correctly picks the right model most of the time, we generate 500 data sets of size $n = 200$, 500, and 1000 from the super model under six scenarios: (1) $\boldsymbol{\beta}_q = \mathbf{0}$, $\boldsymbol{\beta}_o = \boldsymbol{\beta}_h = \mathbf{1}$, i.e the PH, (2) $\boldsymbol{\beta}_q = \boldsymbol{\beta}_h = \mathbf{0}$, $\boldsymbol{\beta}_o = \mathbf{1}$, i.e. the PO, (3) $\boldsymbol{\beta}_o = \mathbf{0}$, $\boldsymbol{\beta}_h = \boldsymbol{\beta}_q = \mathbf{1}$, i.e. the AFT, (4) $\boldsymbol{\beta}_h = \mathbf{0}$, $\boldsymbol{\beta}_o = -\boldsymbol{\beta}_q = \mathbf{1}$, i.e. the AH, (5) $\boldsymbol{\beta}_q = \mathbf{0}$, $\boldsymbol{\beta}_o = -\boldsymbol{\beta}_h = \mathbf{1}$, i.e the YP, and (6) $\boldsymbol{\beta}_h = \mathbf{1}$, $\boldsymbol{\beta}_o = \boldsymbol{\beta}_q = (0.5, 0.5)'$, i.e. the EH. In each case, we consider three baseline survival functions: lognormal $S_0(t) = 1 - \Phi(\log t)$, mixture of two lognormals $S_0(t) = 1 - [0.5\Phi((\log t + 1)/0.5) + 0.5\Phi((\log t - 1)/0.5)]$, and Weibull $S_0(t) = 1 - \exp\{-(0.5t)^{0.8}\}$. The covariate vector is chosen as $\mathbf{x}_i = (x_{i1}, x_{i2})$ with $x_{i1} \overset{iid}{\sim}$ Bernoulli(0.5) and $x_{i2} \overset{iid}{\sim} N(0, 1)$. Finally, a noninformative censoring scheme is used, where we apply right censoring to half of the sample data and interval censoring to the other half. Here the right censoring times are independently simulated from a Uniform$(2, 6)$ distribution. For interval censoring, each subject is assumed to have N observation times, say, $O_1, O_2, \ldots, O_N$, where $(N-1) \sim$ Poisson(2) and $(O_k - O_{k-1})|N \overset{iid}{\sim}$ Exp(1) with $O_0 = 0$, $k = 1, \ldots, N$. A censoring interval has endpoints which are the two adjacent observation times (possibly 0 or ∞) that include the true survival time. The final data yield around 20% right censored, 40% uncensored, 25% left censored and 15% interval censored under all settings. Models were fit with $J = 15$, a loglogistic TBP and the default priors introduced in Section 3. For each MCMC run, 5,000 scans were thinned from 50,000 after a burn-in period of 10,000 iterations. Table 1 reports the proportion (out of 500 replicated data sets) of times each model is picked. The model picked is the one with the largest value among BF_h, BF_o, BF_q, BF_a, BF_y and BF_e relative to the super model.

Table 1: Proportion of times Bayes factor selects each model when truth is known out of 500 replicated data sets.

		Model picked					
Baseline	n	AFT	PH	PO	AH	EH	YP
		True AFT model					

Table 1: Proportion of times Bayes factor selects each model when truth is known out of 500 replicated data sets.

		Model picked					
Baseline	n	AFT	PH	PO	AH	EH	YP
Lognormal	200	0.918	0.034	0.024	0.000	0.024	0.000
	500	0.956	0.000	0.030	0.000	0.014	0.000
	1000	0.964	0.000	0.030	0.000	0.004	0.000
Mixture	200	0.970	0.004	0.000	0.000	0.026	0.000
	500	0.966	0.000	0.000	0.000	0.034	0.000
	1000	0.972	0.000	0.000	0.000	0.028	0.000
Weibull	200	0.432	0.552	0.000	0.000	0.016	0.000
	500	0.356	0.618	0.000	0.000	0.024	0.000
	1000	0.310	0.640	0.000	0.000	0.005	0.000
		True PH model					
Lognormal	200	0.030	0.950	0.002	0.000	0.018	0.000
	500	0.000	0.982	0.000	0.000	0.018	0.000
	1000	0.000	0.980	0.000	0.000	0.020	0.000
Mixture	200	0.000	0.948	0.040	0.000	0.012	0.000
	500	0.000	0.986	0.012	0.000	0.002	0.000
	1000	0.000	0.992	0.002	0.000	0.002	0.004
Weibull	200	0.414	0.558	0.014	0.000	0.014	0.000
	500	0.396	0.524	0.000	0.000	0.080	0.000
	1000	0.324	0.526	0.000	0.000	0.150	0.000
		True PO model					
Lognormal	200	0.878	0.068	0.044	0.000	0.010	0.000
	500	0.748	0.006	0.240	0.000	0.006	0.000
	1000	0.418	0.000	0.578	0.000	0.004	0.000
Mixture	200	0.002	0.150	0.842	0.000	0.000	0.006
	500	0.000	0.012	0.980	0.000	0.000	0.008
	1000	0.000	0.000	0.998	0.000	0.000	0.002
Weibull	200	0.816	0.024	0.146	0.000	0.014	0.000
	500	0.000	0.012	0.980	0.000	0.000	0.006
	1000	0.062	0.002	0.930	0.000	0.000	0.006
		True AH model					
Lognormal	200	0.008	0.000	0.000	0.982	0.008	0.002
	500	0.000	0.000	0.000	0.982	0.014	0.004
	1000	0.000	0.000	0.000	0.974	0.020	0.006
Mixture	200	0.000	0.000	0.000	0.968	0.032	0.000
	500	0.000	0.000	0.000	0.946	0.054	0.000
	1000	0.000	0.000	0.000	0.860	0.140	0.000
Weibull	200	0.388	0.062	0.000	0.544	0.006	0.000
	500	0.546	0.176	0.004	0.272	0.002	0.000
	1000	0.500	0.376	0.008	0.114	0.002	0.000
		True EH model					

Table 1: Proportion of times Bayes factor selects each model when truth is known out of 500 replicated data sets.

		Model picked					
Baseline	n	AFT	PH	PO	AH	EH	YP
Lognormal	200	0.522	0.358	0.026	0.000	0.094	0.000
	500	0.288	0.134	0.016	0.000	0.562	0.000
	1000	0.040	0.004	0.008	0.000	0.940	0.006
Mixture	200	0.092	0.026	0.030	0.000	0.852	0.000
	500	0.000	0.000	0.000	0.000	1.000	0.000
	1000	0.000	0.000	0.000	0.000	1.000	0.000
Weibull	200	0.390	0.582	0.008	0.002	0.018	0.000
	500	0.338	0.624	0.002	0.000	0.036	0.000
	1000	0.356	0.550	0.000	0.000	0.092	0.002
		True YP model					
Lognormal	200	0.000	0.000	0.000	0.972	0.024	0.004
	500	0.000	0.000	0.000	0.848	0.076	0.076
	1000	0.000	0.000	0.000	0.534	0.182	0.284
Mixture	200	0.000	0.000	0.000	0.024	0.004	0.972
	500	0.000	0.000	0.000	0.000	0.002	0.998
	1000	0.000	0.000	0.000	0.000	0.000	1.000
Weibull	200	0.000	0.000	0.000	0.046	0.700	0.254
	500	0.000	0.000	0.000	0.000	0.280	0.720
	1000	0.000	0.000	0.000	0.000	0.052	0.948

When the baseline is the mixture of lognormal distributions, our method works very well even for the smallest sample size $n = 200$; for larger sample sizes $n = 500$ and $n = 1000$ the correct classification rates are all approaching one except for the AH model. When AH is the truth, the proportion picking AH decreases (from 97% to 86%) as n increases while the proportion choosing EH increases. To confirm this observation, we also tried the size of $n = 2000$, and the proportions of choosing AH and EH are 57% and 43%, respectively. In other words, as the sample size increases, our method tends to favor the more complex EH model against the special case of AH. Since EH includes AH as a special case the choice is not incorrect, but is more complex than necessary.

When the baseline is lognormal, our method also works well for most cases except when the true model is PO or YP. For instance, when PO is the truth with $n = 1000$ the method has a 58% chance of picking PO and a 42% chance of choosing AFT. However, picking AFT does not mean that a wrong model is picked if one notices that lognormal can be well approximated by loglogistic and loglogistic AFT is also a PO model. When lognormal YP is the truth with $n = 1000$, our method only has a 28% chance to pick YP with the remaining % allocated to AH or EH. In this case, we also tried the size of $n = 2000$, resulting in AH, EH or YP being picked with proportions being 12%, 35% and 53%, respectively. One reason to explain such a low correct classification rate is that lognormal YP considered here could be very close to a AH and/or a EH model; the baseline distribution plays a large role in how "close" competing semiparametric models actually are and several models may predict equally well. In addition, when lognormal EH is the truth, we need a sample size of $n = 1000$ or larger to identify the correct model. Otherwise, our method tends to select the simpler models AFT or PH, both special cases of EH.

To see how our method performs when the baseline is Weibull, first note that the Weibull AFT, Weibull PH and Weibull AH are all equivalent models, and they are also special cases of EH. Keeping that in mind, we can see that our method has overall low misclassification proportions across most scenarios with the following several exceptions. First, when EH is the truth, both AFT and PH have high chance to be picked; this can be explained by the fact that simulation scenario (6) is not only an EH but also an AFT and a PH. Second, when PO or YP is the truth, we need sample size $n = 1000$ or larger to identify the correct model. When sample size is small like $n = 200$ under true PO (or YP), our method picks AFT (or EH) with 82% (or 70%) chance. This may be because the estimated baseline function hardly deviates from the TBP's centering loglogistic distribution with the small sample size leading to the fitted model close to an AFT (or EH).

To study the impact of the informative g-prior, we also compared two cases $M = 10$ and $M = 50$ for part of the simulation scenarios in Appendix C.1 of the online material, and the two different values yielded almost identical results.

The proposed super model can also be used for survival function estimates when all six Bayes factors are less than 1, i.e., none of the six models fit the data better than the super model. We next demonstrate its finite sample performance. We generate 500 data sets of size $n = 500$ from the super model with $\boldsymbol{\beta}_h = \boldsymbol{\beta}_o = \boldsymbol{\beta}_q = \mathbf{1}$ which is none of the six models. All other simulation settings are the same as before. Table 2 shows the posterior inference results for the regression coefficients. We can see that all coefficient estimates are nearly unbiased with the coverage probabilities around the nominal level 95% when the true baseline is mixture of two lognormals. However, these encouraging results do not hold when the baseline survival function is lognormal or Weibull. This is not surprising, since the super model with Weibull baseline becomes non-identifiable if one notices that the AFT, PH and AH models with Weibull baselines are all equivalent with appropriate reparametrizations. The same argument also applies to the lognormal baseline, since lognormal can be well approximated with a scaled loglogistic and loglogistic AFT is equivalent to loglogistic PO. Figure 1 presents the average, across the 500 MC replicates, of fitted (posterior means over a grid of time points) survival functions when $\mathbf{x} = (0, 0)'$ and $\mathbf{x} = (0, 1)'$; the super model capably estimates complex (here bimodal) survival curves very accurately even for the lognormal and Weibull baselines. Therefore, the super model can still be used for survival/density estimates, even though interpretation of the three sets of regression coefficients is challenging.

Table 2 Simulated data when the super model is the truth and sample size is $n = 500$. Averaged bias (BIAS) and posterior standard deviation (PSD) of each point estimate, standard deviation (across 500 MC replicates) of the point estimate (SD-Est), and coverage probability (CP) for the 95% credible interval.

Parameter	BIAS	PSD	SD-Est	CP
		Lognormal baseline		
$\beta_{h,1} = 1$	-0.021	0.882	0.397	0.996
$\beta_{h,2} = 1$	0.014	0.447	0.238	0.990
$\beta_{o,1} = 1$	0.079	1.504	0.624	0.990
$\beta_{o,2} = 1$	0.059	0.753	0.391	0.990
$\beta_{q,1} = 1$	-0.056	0.843	0.357	0.988
$\beta_{q,2} = 1$	-0.042	0.417	0.222	0.988
		Mixture baseline		
$\beta_{h,1} = 1$	0.062	0.305	0.259	0.972
$\beta_{h,2} = 1$	0.079	0.189	0.171	0.954
$\beta_{o,1} = 1$	-0.027	0.315	0.302	0.962
$\beta_{o,2} = 1$	-0.020	0.194	0.189	0.952
$\beta_{q,1} = 1$	0.003	0.109	0.099	0.974
$\beta_{q,2} = 1$	-0.003	0.066	0.065	0.948
		Weibull baseline		
$\beta_{h,1} = 1$	-0.093	0.780	0.465	0.990
$\beta_{h,2} = 1$	-0.098	0.444	0.275	0.988
$\beta_{o,1} = 1$	0.156	0.921	0.531	0.990
$\beta_{o,2} = 1$	0.175	0.491	0.304	0.980
$\beta_{q,1} = 1$	-0.160	0.943	0.541	0.990
$\beta_{q,2} = 1$	-0.193	0.504	0.322	0.976

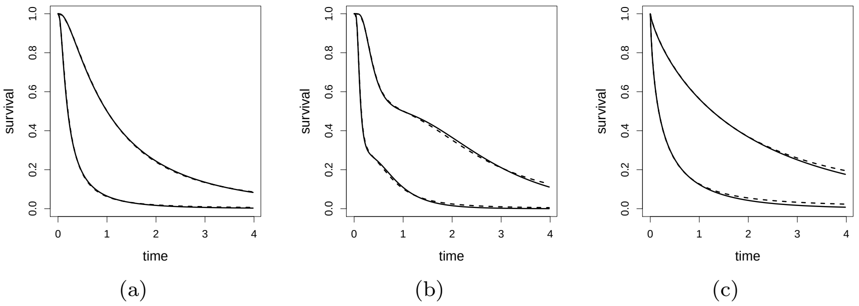


Fig. 1 Simulated data when true mode is none of the six models and sample size is $n = 500$. Mean, across the 500 MC replicates, of the posterior mean of the survival functions when $\mathbf{x} = (0, 0)'$ (upper lines) and $\mathbf{x} = (0, 1)'$ (lower lines). The true curves are represented by continuous lines and the fitted curves are represented by dashed lines.

4.2 Veterans Administration Lung Cancer Trial

The data considered is the well-known Veterans Administration lung cancer trial (Prentice, 1973), which has been incorporated into `MASS` package in R. As in Cheng et al. (1995), Murphy et al. (1997), Yang and Prentice (1999), and Hanson (2006) we consider a subgroup of $n = 97$ patients with no prior therapy. Two covariates considered are performance status, a measure that is a multiple of 10 and ranges from 0 to 100, and the tumor type, a factor with four levels (large=1, adeno=2, small=3, squamous=4). Six of the 97 survival times are censored. Cheng et al. (1995) used the transformation model; Murphy et al. (1997) and Yang and Prentice (1999) considered the PO model; Hanson (2006) considered the AFT, PH and PO model.

The proposed super model is fit with $J = 15$, a Weibull TBP, and the hyperparameter settings in Section 3.3; see Appendix B.1 of the online material for R commands. The Bayes factors for testing AFT, PH, PO, AH, EH and YP vs. the super model are 115, 27, 97, 0.25, 123, and 11, respectively.

The AFT, PH, PO, EH, and YP fit better than the super model; the EH, AFT and PO models fit about the same and are about four times better than PH. The LPML for the super model compares favorably to those observed in Hanson and Yang (2007), most notably the log-logistic regression model had the best LPML of about -509. Since the parametric log-logistic model has both PO and AFT properties, seeing that these semiparametric models are favored about the same makes sense. Other centering distributions gave roughly the same results, log-logistic gave -509.6 and lognormal gave -511.5.

Since the EH model has the highest Bayes factor, the super model can be used as a model in its own for prediction. Figure 2 presents the predictive survival densities for squamous with score equal to 40, 60 and 80; the code is available in Appendix B.1 of the online material. These plots can be compared to Figure 1 in Hanson and Yang (2007), which have much rougher densities. The Polya tree encourages spikiness in densities, whereas the transformed Bernstein allows multimodality but tends to smooth over spurious spikes.

Notice that the BF comparing EH to AFT is $123.0/115.1 \approx 1.07$. Thus the AFT model may be considered adequate and can be fitted parametrically via `survreg` or semiparametrically by the `lss` package in R. Other R packages for fitting semiparametric AFT models are reviewed in Zhou and Hanson (2015) including `spBayesSurv`.

4.3 Breast Cancer Study

Beadle et.al (Beadle et al., 1984) reported a retrospective study to compare the cosmetic effects of radiotherapy alone versus radiotherapy and adjuvant chemotherapy on 94 women with early breast cancer. There are 46 patients in radiation only group and 48 patients in radiation plus chemotherapy group. Patients were observed initially every 46 months, but, as their recovery progressed, the interval between visits lengthened. The event of interest was the

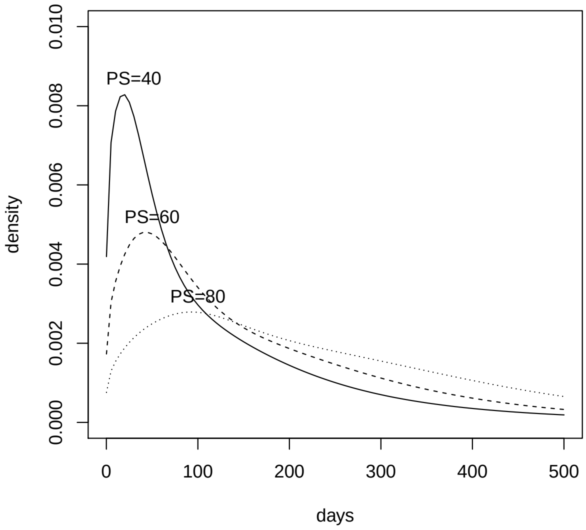


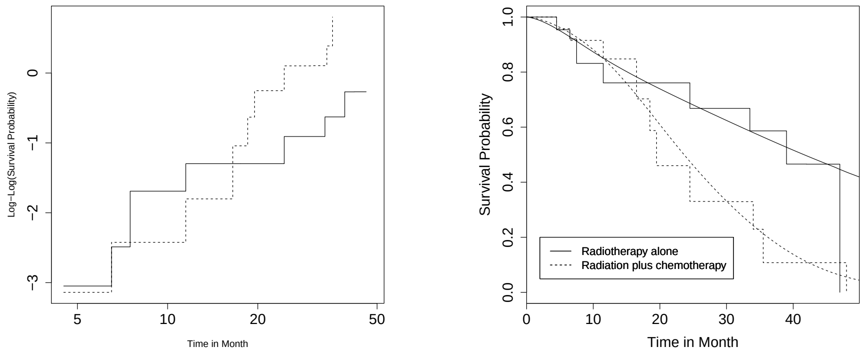
Fig. 2 Preditive densities for squamous, score=40, 60, 80.

time to first appearance of moderate or severe breast retraction. There are 5.3% of the women who were left censored, 55.8% were interval censored and 38.9% were right censored. The dataset is available in the R package `KMsurv`.

The proposed super model is fit with $J = 15$, a Weibull TBP and the hyperparameter settings in Section 3.3; see Appendix B.2 of the online material for R commands. The Bayes factors for testing AFT, PH, PO, AH, EH and YP vs. the super model are 18, 32, 4, 24, 8 and 8, respectively. All models fit better than the super model; the PH and AH models fit about the same and are about seven times better than PO. In choosing between PH and AH, log-log survival plots can help. Figure 3(a) shows crossing lines based on Turnbull's estimator (Turnbull, 1976), suggesting that the AH may be more appropriate for these data. In fact, the estimated survival curves in Figure 3(b) from the super model show crossing survival, which is disallowed under PH.

5 Discussion

We proposed a new super model which includes PH, PO, AFT, AH, EH and YP models as specials cases. Bayes factors have been developed under the transformed Bernstein polynomial prior. Simulation studies demonstrate the appropriate model can be selected based on this approach; the proposed model appears to work especially well for choosing among the mostly widely-used PH, PO, and AFT models. The R package `spBayesSurv` implements the proposed method directly as demonstrated via two real data analyses.



(a) Proportional Hazards Assumption

(b) Predictive survival functions

Fig. 3 Breast Cancer Data

Note that the AFT, PH and AH models are equivalent under the Weibull distribution. The AFT and PO models are equivalent under the loglogistic distribution. The EH model includes PH and AH as special cases, and the YP model includes PH and PO as special cases. In practice, the small sample size may cause a lot of uncertainty. If we look at the Table 1 closely for the sample size of $n = 200$, we can see that the proportions that the "correct" model (including all equivalent models) is picked are all nearly 95% or above when the true model is AFT, PH, PO or AH. When EH (or YP) is the truth with small sample size, our method tends to select a simpler model (one of AFT, PH, PO or AH) that is closest to EH (or YP). Therefore, for small n, we recommend choosing a model only among AFT, PH, PO and AH; the EH or YP models may be poorly identified in such cases depending on the true baseline survival function. Additionally, in smaller sample sizes several models may fit similarly; in such cases a final model can be chosen based on the most suitable assumption for answering clinical questions of interest (e.g. proportional hazards), interpretability (e.g. hazard ratios) and simplicity.

When none of the six simpler models is picked, the proposed super model can be used for accurate survival estimates although the regression coefficients do not have useful interpretation. Other alternatives are to consider a general linear transformation model (Zeng and Lin, 2007), or Bayesian nonparametric model, e.g. De Iorio et al. (2009); however, just as in the proposed super model, there is no easy interpretation of model coefficients. The latter model can be fit using the function `anovaDDP` in `spBayesSurv`.

The approach we have taken is to formally nest commonly used semiparametric models into a large, encompassing 'super model.' An alternative approach is parametric transformations. In terms of cumulative hazards $H_{\mathbf{x}}(\cdot)$ and baseline cumulative hazard $H_0(\cdot)$, semiparametric linear transformation models can be written as

$$H_{\mathbf{x}}(t) = G\{e^{\boldsymbol{\beta}'\mathbf{x}} H_0(t)\}.$$

Zeng and Lin (2007) note that $G(x) = \frac{1}{\rho}[(1+x)^\rho - 1]$ gives PH when $\rho = 1$ or PO as $\rho \to 0+$; also $G(x) = \frac{1}{\rho}\log(1+\rho x)$ gives PO when $\rho = 1$ or PH as $\rho \to 0+$. The latter model is equivalent to the generalized odds rate model of Scharfstein et al. (1998). Yin and Ibrahim (2005) instead consider

$$\tfrac{1}{\rho}[h_{\mathbf{x}}(t)^\rho - 1] = \tfrac{1}{\rho}[h_0(t)^\rho - 1] + \mathbf{x}'\boldsymbol{\beta}.$$

Here, $\rho = 1$ gives the additive hazards model whereas $\rho \to 0^+$ gives PH. It has been generally noted by these authors that estimation of ρ is problematic and inference proceeds typically by fitting several values of ρ and choosing the value closest to one or zero that maximizes a likelihood or posterior density. In all of these models one of two common models is obtained on the boundary of the parameter space, i.e. $\rho \to 0^+$, which presents unique challenges to model selection and estimation.

References

Beadle, G. F., Come, S., Henderson, I. C., Silver, B., Hellman, S., and Harris, J. R. (1984). The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer. *International Journal of Radiation Oncology Biology Physics*, 10(11):2131–2137.

Chen, Y., Hanson, T., and Zhang, J. (2014). Accelerated hazards model based on parametric families generalized with bernstein polynomials. *Biometrics*, 70(1):192–201.

Chen, Y. Q. and Jewell, N. P. (2001). On a general class of semiparametric hazards regression models. *Biometrika*, 88(3):687–702.

Chen, Y. Q. and Wang, M.-C. (2000). Analysis of accelerated hazards models. *Journal of the American Statistical Association*, 95(450):608–618.

Cheng, S., Wei, L., and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845.

Cox, D. R. (1992). Regression models and life-tables. In *Breakthroughs in Statistics*, pages 527–541. Springer.

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). Bayesian nonparametric nonproportional hazards survival modeling. *Biometrics*, 65(3):762–771.

Devarajan, K. and Ebrahimi, N. (2011). A semi-parametric generalization of the Cox proportional hazards regression model: Inference and applications. *Computational Statistics & Data Analysis*, 55(1):667–676.

Diao, G., Zeng, D., and Yang, S. (2013). Efficient semiparametric estimation of short-term and long-term hazard ratios with right-censored data. *Biometrics*, 69(4):840–849.

Etezadi-Amoli, J. and Ciampi, A. (1987). Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function. *Biometrics*, 43(2):181–192.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Ghosal, S. (2001). Convergence rates for density estimation with Bernstein polynomials. *Annals of Statistics*, 29(5):1264–1280.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242.

Hanson, T. and Yang, M. (2007). Bayesian semiparametric proportional odds models. *Biometrics*, 63(1):88–95.

Hanson, T. E. (2006). Inference for mixtures of finite Polya tree models. *Journal of the American Statistical Association*, 101(476):1548–1565.

Hanson, T. E., Branscum, A. J., Johnson, W. O., et al. (2014). Informative g-priors for logistic regression. *Bayesian Analysis*, 9(3):597–612.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Li, L., Hanson, T., and Zhang, J. (2015). Spatial extended hazard model with application to prostate cancer survival. *Biometrics*, 71(2):313–322.

Murphy, S., Rossini, A., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976.

Petrone, S. and Wasserman, L. (2002). Consistency of bernstein polynomial posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(1):79–100.

Prentice, R. L. (1973). Exponential survivals with censoring and explanatory variables. *Biometrika*, 60(2):279–288.

Quantin, C., Moreau, T., Asselain, B., Maccario, J., and Lellouch, J. (1996). A regression survival model for testing the proportional hazards hypothesis. *Biometrics*, 52(3):874–885.

Scharfstein, D. O., Tsiatis, A. A., and Gilbert, P. B. (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime Data Analysis*, 4(4):355–391.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*. Springer.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, 90(430):614–618.

Yang, S. and Prentice, R. (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. *Biometrika*, 92(1):1–17.

Yang, S. and Prentice, R. L. (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Associa-*

tion, 94(445):125–136.

Yin, G. and Ibrahim, J. G. (2005). Bayesian frailty models based on Box-Cox transformed hazards. *Statistica Sinica*, 15(3):781–794.

Zellner, A. (1983). Applications of Bayesian analysis in econometrics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32:23–34.

Zeng, D. and Lin, D. (2007). Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association*, 102(477):167–180.

Zhang, J. and Peng, Y. (2009). Crossing hazard functions in common survival models. *Statistics & Probability Letters*, 79(20):2124–2130.

Zhou, H. and Hanson, T. (2015). Bayesian spatial survival models. In *Nonparametric Bayesian Inference in Biostatistics*, pages 215–246. Springer.

Zhou, H., Hanson, T., and Zhang, J. (2017). Generalized accelerated failure time spatial frailty model for arbitrarily censored data. *Lifetime Data Analysis*, 23(3):495–515.