Supplementary Materials for *"A unified framework for fitting Bayesian semiparametric models to arbitrarily censored survival data, including spatially-referenced data"* by Haiming Zhou and Timothy Hanson

Appendix 0 Notation and Prior Tables

Table S1 presents the notation symbols used in the main paper and their definitions. Table S2 lists the priors for all parameters and the reasons of choosing them, where $\text{TBP}_J(\alpha, S_{\boldsymbol{\theta}})$ is the TBP prior, $\text{ICAR}(\tau^2)$ is the ICAR prior, $\text{GRF}(\tau^2, \phi)$ is the GRF prior, and $\text{IID}(\tau^2)$ is the IID prior.

Table S1: List of Notations.

Notation	Definition
α	precision parameter of the TBP prior
$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$	p-vector of regression coefficients for survival models
$\boldsymbol{\beta}_0$	mean of the normal $N_p(\boldsymbol{\beta}_0, \mathbf{W}_0)$ prior on $\boldsymbol{\beta}$
$\hat{\boldsymbol{\beta}}$	estimate of $\boldsymbol{\beta}$ under the parametric survival model with $S_0 = S_{\boldsymbol{\theta}}$
$\delta_{j,J}(\cdot)$	beta density function with parameters $(j, J-1+1)$
$\Delta_{j,J}(\cdot)$	beta cumulative distribution function with parameters $(j, J-1+1)$
$\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$	latent binary variable with $\gamma_\ell = 1$ indicating the presence of the ℓth covariate, $\ell = 1, \ldots, p$
$\Gamma(a, b)$	gamma distribution with shape parameter a and rate parameter b
$\boldsymbol{\theta} = (\theta_1, \theta_2)'$	parameters of the centering distribution families $S_{\boldsymbol{\theta}}$
$\boldsymbol{\theta}_0$	mean of the normal $N_2(\boldsymbol{\theta}_0, \mathbf{V}_0)$ prior on $\boldsymbol{\theta}$
$\hat{\boldsymbol{\theta}}$	estimate of $\boldsymbol{\theta}$ under the parametric survival model with $S_0 = S_{\boldsymbol{\theta}}$
ν	powered exponential correlation function shape parameter, $\nu \in (0, 2]$
$\boldsymbol{\xi}_\ell = (\xi_{\ell 1}, \ldots, \xi_{\ell K})'$	coefficients of the cubic B-spline basis functions for the ℓth covariate, $\ell = 1, \ldots, p$
$\rho(\cdot, \cdot)$	correlation function; arguments are two spatial locations
$\rho(\cdot, \cdot; \phi)$	correlation function indexed by the range parameter ϕ
κ	shrinkage parameter used under the proper CAR
τ^2	scale parameter under the ICAR or GRF or IID frailty prior
ϕ	range parameter in the powered exponential correlation function
a_{ij}, b_{ij}	endpoints of the interval (a_{ij}, b_{ij}) that contains the survival time t_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$
a_α, b_α	shape and rate parameters of the $\Gamma(a_\alpha, b_\alpha)$ prior on α
a_τ, b_τ	shape and rate parameters of the $\Gamma(a_\tau, b_\tau)$ prior on τ^{-2}

Table S1: List of Notations.

Notation	Definition
a_ϕ, b_ϕ	shape and rate parameters of the $\Gamma(a_\phi, b_\phi)$ prior on ϕ
A	number of knots used in the FSA
B	number of blocks used in the FSA
$\mathbf{C}$	precision matrix of the vector of frailties $\mathbf{v} = (v_1, \ldots, v_m)'$
$d(\cdot\|J, \mathbf{w}_J)$	density function of Bernstein polynomial
$D(\cdot\|J, \mathbf{w}_J)$	cdf associated with density $d(\cdot\|J, \mathbf{w}_J)$
$\mathbf{F}_e$	$m \times m$ diagonal matrix with the i diagonal element being e_{i+}
e_{ij}	equals 1 if regions i and j share a common boundary and 0 otherwise, $i, j = 1, \ldots, m$; set $e_{ii} = 0$
e_{i+}	number of adjacent regions for region i, i.e. $e_{i+} = \sum_{j=1}^{m} e_{ij}$
$\mathbf{E}$	$m \times m$ adjacency matrix with the ijth element equal to e_{ij}
$f_{\mathbf{x}_{ij}}(\cdot)$	density function of the survival time t_{ij} given the covariate $\mathbf{x}_{ij}$
$f_0(\cdot)$	baseline density function in the survival models
g	parameter in the g-prior for variable selection
G	distribution of covariate vectors $\mathbf{x}$ with support on $\mathcal{X} \subseteq \mathbb{R}^p$
$h_{\mathbf{x}_{ij}}(\cdot)$	hazard function of the survival time t_{ij} given the covariate $\mathbf{x}_{ij}$
$h_0(\cdot)$	baseline hazard function in the survival models
$\mathbf{I}_p$	$p \times p$ identity matrix
$I(\cdot)$	the usual indicator function
J	number of Bernstein polynomials used for defining $d(\cdot\|J, \mathbf{w}_J)$
K	number of basis functions used for modeling the nonlinear function $u_\ell(\cdot)$
$L(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v})$	likelihood function for $(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v})$
m	number of distinct spatial locations
M	used to determine the g in the g-prior for variable selection
n_i	number of subjects within the ith spatial location, $i = 1 \ldots, m$
n	total number of subjects in the data, i.e. $n = \sum_{i=1}^{m} n_i$
$N(a, b^2)$	normal distribution with mean a and variable b^2
$N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	k-variate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
o_{ij}	number of observations for the time-dependent covariate vector $\mathbf{x}_{ij}(t)$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$
p	dimension of the covariate vector $\mathbf{x}_{ij}$
$p(\cdot)$	generic symbol for prior and posterior density functions
q	used to determine the g in the g-prior for variable selection
$r(t_{ij})$	Cox-Snell residual equal to $-\log\{S_{\mathbf{x}_{ij}}(t_{ij})\}$
$\mathbf{R}$	correlation matrix of $\mathbf{v}$ under the GRF prior
$S_{\mathbf{x}_{ij}}(\cdot)$	survival function of the survival time t_{ij} given the covariate $\mathbf{x}_{ij}$
$S_0(\cdot)$	baseline survival function in the survival models
$u_\ell(\cdot)$	nonlinear function for the ℓth covariate, $\ell = 1, \ldots, p$
$\mathbf{v} = (v_1, \ldots, v_m)'$	vector of frailties
$\mathbf{V}_0$	covariance matrix of the normal $N_2(\boldsymbol{\theta}_0, \mathbf{V}_0)$ prior on $\boldsymbol{\theta}$
$\hat{\mathbf{V}}$	estimate of the covariance of $\hat{\boldsymbol{\theta}}$ under the parametric survival model
$\mathbf{w}_J = (w_1, \ldots, w_J)'$	J-vector of positive weights used in the TBP prior
$\mathbf{W}_0$	covariance matrix of the normal $N_p(\boldsymbol{\beta}_0, \mathbf{W}_0)$ prior on $\boldsymbol{\beta}$
$\hat{\mathbf{W}}$	estimate of the covariance of $\hat{\boldsymbol{\beta}}$ under the parametric survival model

Table S1: List of Notations.

Notation	Definition
$\mathbf{x}_{ij}$	p-vector of covariates for subject ij, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$
$x_{ij\ell}$	ℓth element of the $\mathbf{x}_{ij}$, $\ell = 1, \ldots, p$
$\mathbf{x}$	generic symbol for p-vector of covariates
x_ℓ	ℓth element of the $\mathbf{x}$, $\ell = 1, \ldots, p$
$\mathbf{X}$	design matrix associated with $\{\mathbf{x}_{ij}\}$ with mean-centered columns
$\mathbf{X}_\ell$	design matrix associated with $u_\ell(x_\ell)$ with mean-centered columns $\ell = 1, \ldots, p$
z_j	equals $\log(w_j) - \log(w_J)$
$\mathbf{z}_{J-1}$	equals $(z_1, \ldots, z_{J-1})'$

Table S2: List of priors.

Parameter	Prior	Justification
$S_0(\cdot)$	$\text{TBP}(\alpha, S_{\boldsymbol{\theta}})$	Selects smooth densities and can be centered at a standard parametric family: one of log-logistic, log-normal, and Weibull.
α	$\Gamma(a_\alpha, b_\alpha)$	$\alpha > 0$ acts like the precision in a Dirichlet process controlling how stochastically pliable S_0 is close to $S_{\boldsymbol{\theta}}$. A gamma prior has been widely used for Dirichlet processes. Defaults: $a_\alpha = b_\alpha = 1$.
$\boldsymbol{\beta}$	$N_p(\boldsymbol{\beta}_0, \mathbf{W}_0)$	Gaussian is common for regression effects. Defaults: $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{W}_0 = 10^{10}\mathbf{I}_p$ or $\mathbf{W}_0 = gn(\mathbf{X}'\mathbf{X})^{-1}$ when the SSVS is performed.
$\boldsymbol{\theta}$	$N_2(\boldsymbol{\theta}_0, \mathbf{V}_0)$	Centering distribution $S_{\boldsymbol{\theta}}$ is parameterized so that $\boldsymbol{\theta}$ is defined on $\mathbb{R}^2$, so a Gaussian prior is appropriate. Defaults: $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$, $\mathbf{V}_0 = 10\hat{\mathbf{V}}$. Note here we assume a somewhat informative prior on $\boldsymbol{\theta}$ to obviate confounding between $\boldsymbol{\theta}$ and $\mathbf{w}_J$.
$\mathbf{v}$	$\text{ICAR}(\tau^2)$	When clusters are formed by spatial regions and spatial smoothing is of interest, the ICAR prior is commonly used for modeling the frailties in survival models.
$\mathbf{v}$	$\text{GRF}(\tau^2, \phi)$	Very common prior for georeferenced data.
$\mathbf{v}$	$\text{IID}(\tau^2)$	When spatial dependence among clusters is not of interest, the IID Gaussian frailties are commonly assumed.
τ^{-2}	$\Gamma(a_\tau, b_\tau)$	The gamma distribution is a conjugate prior on τ^{-2}. Defaults: $a_\tau = b_\tau = 0.001$.
ϕ	$\Gamma(a_\phi, b_\phi)$	The range parameter ϕ is positive and the gamma prior is a natural choice. Defaults: $a_\phi = 2$ and $b_\phi = (a_\phi - 1)/\phi_0$ so that the prior of ϕ has mode at ϕ_0, where ϕ_0 satisfies $\rho(\mathbf{s}', \mathbf{s}''; \phi_0) = 0.001$ with $\|\mathbf{s}', \mathbf{s}''\| = \max_{i,j}\|\mathbf{s}_i - \mathbf{s}_j\|$.
$\boldsymbol{\gamma}$	$\prod_{\ell=1}^{p} \text{Bern}(q_\ell)$	Commonly used for Bayesian variable selection (e.g. Kuo and Mallick, 1998). Defaults: $q_\ell = 0.5$, $\ell = 1, \ldots, p$.
$\boldsymbol{\xi}_\ell$	$N_K(\mathbf{0}, \mathbf{S}_{\boldsymbol{\xi}})$	Here $\mathbf{S}_{\boldsymbol{\xi}} = gn(\mathbf{X}_\ell'\mathbf{X}_\ell)^{-1}$ was chosen following the idea of informative g-prior introduced Appendix E.

Appendix A MCMC Sampling

The joint posterior distribution for all parameters is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\theta},\mathbf{w}_J,\alpha,\mathbf{v},\tau^2,\phi) \propto\ & L(\mathbf{w}_J,\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{v}) \\ & \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{W}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\} \\ & \times \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{V}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right\} \\ & \times \frac{\Gamma(\alpha J)}{\Gamma(\alpha)^J}\prod_{j=1}^{J}(w_j)^{\alpha-1}\times\alpha^{a_\alpha-1}\exp\{-b_\alpha\alpha\} \\ & \times \left(\tau^{-2}\right)^{\frac{\text{rank}(\mathbf{C})}{2}}\exp\left\{-\frac{1}{2\tau^2}\mathbf{v}'\mathbf{C}\mathbf{v}\right\}\times\left(\tau^{-2}\right)^{a_\tau-1}\exp\left\{-b_\tau\tau^{-2}\right\} \\ & \times p(\phi)|\mathbf{C}|^{1/2}\end{aligned} \tag{A.1}$$

For the GRF prior, $\mathbf{C}=\mathbf{R}^{-1}$ and $p(\phi)=\phi^{a_\phi-1}\exp\{-b_\phi\phi\}$. The ICAR prior does not need $p(\phi)|\mathbf{C}|^{1/2}$, and $\mathbf{C}=\mathbf{F}_e-\mathbf{E}$, where $\mathbf{F}_e$ is an $m\times m$ diagonal matrix with $\mathbf{F}_e[i,i]=e_{i+}$. For the IID prior, $p(\phi)|\mathbf{C}|^{1/2}$ is also not needed and $\mathbf{C}=\mathbf{I}_m$ is an identity matrix.

Note that when $w_j=1/J$ the underlying parametric model with $S_0(t)=S_{\boldsymbol{\theta}}(t)$ is obtained, so a fit from a standard parametric survival model can provide starting values for the TBP survival model. Let $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ denote the parametric point estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, and let $\hat{\mathbf{V}}$ and $\hat{\mathbf{W}}$ denote their asymptotic covariance matrices, respectively. These estimates can be easily obtained by running the proposed MCMC below with $w_j\equiv 1/J$ and relatively vague priors on $(\boldsymbol{\theta},\boldsymbol{\beta})$.

Step 1: Update $\mathbf{w}_J$.
Set $\mathbf{z}_{J-1}=(z_1,\ldots,z_{J-1})'$ with $z_j=\log(w_j)-\log(w_J)$. The full conditional distribution for $\mathbf{z}_{J-1}$ is

$$p(\mathbf{z}_{J-1}|\text{else})\propto L(\mathbf{w}_J,\boldsymbol{\theta},\boldsymbol{\beta},\mathbf{v})\times\prod_{j=1}^{J}\left[\frac{e^{z_j}}{\sum_{k=1}^{J}e^{z_j}}\right]^{\alpha},$$

where $z_J=0$. The vector $\mathbf{z}_{J-1}$ can be updated using adaptive Metropolis samplers (Haario et al., 2001). Suppose we are currently in iteration l and have sampled the states $\mathbf{z}_{J-1}^{(1)},\ldots,\mathbf{z}_{J-1}^{(l-1)}$. We select an index l_0 (e.g., $l_0=5000$) for the length of an initial period and define

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

Here $\mathcal{C}_l$ is the sample variance of $\mathbf{z}_{J-1}^{(1)},\ldots,\mathbf{z}_{J-1}^{(l-1)}$, $d=J-1$ is the dimension of $\mathbf{z}_{J-1}$, 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 $\mathbf{z}_{J-1}^{*}=(z_1^*,\ldots,z_{J-1}^*)'$ from $N_{J-1}(\mathbf{z}_{J-1}^{(l-1)},\boldsymbol{\Sigma}_l)$ and accept it with probability

$$\min\left\{1,\frac{p(\mathbf{z}_{J-1}^{*}|\text{else})}{p(\mathbf{z}_{J-1}^{(l-1)}|\text{else})}\right\}.$$

Step 2: Update $\boldsymbol{\theta}$.
The full conditional distribution for $\boldsymbol{\theta}$ is

$$p(\boldsymbol{\theta}|\text{else}) \propto L(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}) \times \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\mathbf{V}_0^{-1}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right\}.$$

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

$$\min\left\{1, \frac{p(\boldsymbol{\theta}^*|\text{else})}{p(\boldsymbol{\theta}^{(l-1)}|\text{else})}\right\}.$$

where $\boldsymbol{\Sigma}_l$ is defined similarly as above, but with $\boldsymbol{\Sigma}_0$ set to be $\hat{\mathbf{V}}$.

Step 3: Update $\boldsymbol{\beta}$.
The full conditional distribution for $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\theta}|\text{else}) \propto L(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}) \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{W}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}.$$

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

$$\min\left\{1, \frac{p(\boldsymbol{\beta}^*|\text{else})}{p(\boldsymbol{\beta}^{(l-1)}|\text{else})}\right\}.$$

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

Step 4: Update α.
The full conditional distribution for α is

$$p(\alpha|\text{else}) \propto \frac{\Gamma(\alpha J)}{\Gamma(\alpha)^J}\prod_{j=1}^{J}(w_j)^{\alpha-1} \times \alpha^{a_\alpha-1}\exp\{-b_\alpha\alpha\}.$$

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

$$\min\left\{1, \frac{p(\alpha^*|\text{else})}{p(\alpha^{(l-1)}|\text{else})}\right\}.$$

Step 5: Update $\mathbf{v}$.
Let $L(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}) = \prod_{i=1}^{m}\prod_{j=1}^{n_i} L_{ij}(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v})$. For the ICAR prior, the full conditional distribution for v_i, $i = 1, \ldots, m$, is

$$p(v_i|\text{else}) \propto \prod_{j=1}^{n_i} L_{ij}(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}) \exp\left\{-\frac{e_{i+}}{2\tau^2}\left(v_i - \sum_{j=1}^{m} e_{ij}v_j/e_{i+}\right)^2\right\}.$$

The v_j is updated via Metropolis-Hastings sampling steps with proposal $v_j^* \sim N(v_j^{(l-1)}, \tau^2/e_{j+})$. The candidate v_j^* is accepted with probability

$$\min\left\{1, \frac{p(v_j^*|\text{else})}{p(v_j^{(l-1)}|\text{else})}\right\}.$$

After each individual frailty update, the vector of $\mathbf{v}$ is updated to have sample mean zero through $\mathbf{v} \leftarrow \mathbf{v} - \frac{1}{m}\mathbf{1}'_m\mathbf{v}$. Although *ad hoc*, this approach to enforcing the sum-to-zero constraint on $v_1, \ldots, v_m$ has negligible effect on the posterior and has been advocated by many authors, e.g. Banerjee et al. (2014) and Lang and Brezger (2004).

For the IID prior, the full conditional distribution for v_i, $i = 1, \ldots, m$, is

$$p(v_i|\text{else}) \propto \prod_{j=1}^{n_i} L_{ij}(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}) \exp\left\{-\frac{1}{2\tau^2}v_i^2\right\}.$$

The v_j is updated via Metropolis-Hastings sampling steps with proposal $v_j^* \sim N(v_j^{(l-1)}, \tau^2)$. The candidate v_j^* is accepted with probability

$$\min\left\{1, \frac{p(v_j^*|\text{else})}{p(v_j^{(l-1)}|\text{else})}\right\}.$$

For the GRF prior, the full conditional distribution for v_i, $i = 1, \ldots, m$, is

$$p(v_i|\text{else}) \propto \prod_{j=1}^{n_i} L_{ij}(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}) \exp\left\{-\frac{p_{ii}}{2\tau^2}\left(v_i + \sum_{\{j:j\neq i\}} p_{ij}v_j/p_{ii}\right)^2\right\},$$

where p_{ij} is the (i, j) element of $\mathbf{R}^{-1}$. The v_j is updated via Metropolis-Hastings sampling steps with proposal $v_j^* \sim N(v_j^{(l-1)}, \tau^2/p_{ii})$. The candidate v_j^* is accepted with probability

$$\min\left\{1, \frac{p(v_j^*|\text{else})}{p(v_j^{(l-1)}|\text{else})}\right\}.$$

Step 6: Update τ^2.
The full conditional distribution for τ^{-2} is

$$p(\tau^{-2}|\text{else}) \propto \left(\tau^{-2}\right)^{a_\tau + \frac{\text{rank}(\mathbf{C})}{2} - 1} \exp\left\{-\left[b_\tau + \frac{1}{2}\mathbf{v}'\mathbf{C}\mathbf{v}\right]\tau^{-2}\right\}.$$

Thus τ^{-2} is sampled from $\Gamma(a_\tau^*, b_\tau^*)$, where $a_\tau^* = a_\tau + \frac{\text{rank}(\mathbf{C})}{2} - 1$ and $b_\tau^* = b_\tau + \frac{1}{2}\mathbf{v}'\mathbf{C}\mathbf{v}$.

Step 7: Update ϕ for georeferenced data.
The full conditional distribution for ϕ is

$$p(\phi|\text{else}) \propto |\mathbf{R}|^{-1/2}\text{exp}\left\{-\frac{1}{2\tau^2}\mathbf{v}'\mathbf{R}^{-1}\mathbf{v}\right\}\phi^{a_\tau - 1}\exp\left\{-b_\phi\phi\right\}$$

The range parameter ϕ is updated via adaptive Metropolis samplers with normal proposal $\phi^* \sim N_1(\phi^{(l-1)}, \boldsymbol{\Sigma}_l)$ with $\boldsymbol{\Sigma}_l$ is defined similarly as above with $\boldsymbol{\Sigma}_0 = 0.16$, and the acceptance probability is

$$\min\left\{1, \frac{p(\phi^*|\text{else})}{p(\phi^{(l-1)}|\text{else})}\right\}.$$

Step 8: Update $\boldsymbol{\gamma}$ when variable selection is performed.
When variable selection is performed, all $\boldsymbol{\beta}$s in steps 1-7 need to be replaced by $\boldsymbol{\gamma} \odot \boldsymbol{\beta}$, where $\odot$

denotes componentwise multiplication. Then each γ_j is generated from its full conditional, i.e. a Bernoulli distribution with the success probability

$$\frac{q_j}{q_j + (1-q_j)L(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\gamma}_{j0} \odot \boldsymbol{\beta}, \mathbf{v})/L(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\gamma}_{j1} \odot \boldsymbol{\beta}, \mathbf{v})},$$

where the vector $\boldsymbol{\gamma}_{j0}$ ($\boldsymbol{\gamma}_{j1}$) is obtained from $\boldsymbol{\gamma}$ with the jth element replaced by 0 (1).

Appendix B The Full Scale Approximation

For georeferenced data, a computational bottleneck of the MCMC sampling scheme is inverting the $m \times m$ matrix $\mathbf{R}$, which typically has computational cost $O(m^3)$. In this section, we introduce a full scale approximation (FSA) approach proposed by Sang and Huang (2012), which provides a high quality approximation to the correlation function ρ at both the large and the small spatial scales, such that the inverse of $\mathbf{R}$ can be substantially sped up for large value of m, e.g., $m \geq 500$.

Consider a fixed set of "knots" $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_A^*\}$ chosen from the study region. These knots can be chosen using the function `cover.design` within the R package `fields`, which computes space-filling coverage designs using the swapping algorithm (Johnson et al., 1990). Let $\rho(\mathbf{s}, \mathbf{s}')$ be the correlation between locations $\mathbf{s}$ and $\mathbf{s}'$. The FSA approach approximates the correlation function $\rho(\mathbf{s}, \mathbf{s}')$ with

$$\rho^\dagger(\mathbf{s}, \mathbf{s}') = \rho_l(\mathbf{s}, \mathbf{s}') + \rho_s(\mathbf{s}, \mathbf{s}'). \tag{B.2}$$

The $\rho_l(\mathbf{s}, \mathbf{s}')$ in (B.2) is the reduced-rank part capturing the long-scale spatial dependence, defined as $\rho_l(\mathbf{s}, \mathbf{s}') = \rho'(\mathbf{s}, \mathcal{S}^*)\rho_{AA}^{-1}(\mathcal{S}^*, \mathcal{S}^*)\rho(\mathbf{s}', \mathcal{S}^*)$, where $\rho(\mathbf{s}, \mathcal{S}^*) = [\rho(\mathbf{s}, \mathbf{s}_i^*)]_{i=1}^A$ is an $A \times 1$ vector, and $\rho_{AA}(\mathcal{S}^*, \mathcal{S}^*) = [\rho(\mathbf{s}_i^*, \mathbf{s}_j^*)]_{i,j=1}^A$ is an $A \times A$ correlation matrix at knots $\mathcal{S}^*$. However, $\rho_l(\mathbf{s}, \mathbf{s}')$ cannot well capture the short-scale dependence due to the fact that it discards entirely the residual part $\rho(\mathbf{s}, \mathbf{s}') - \rho_l(\mathbf{s}, \mathbf{s}')$. The idea of FSA is to add a small-scale part $\rho_s(\mathbf{s}, \mathbf{s}')$ as a sparse approximate of the residual part, defined by $\rho_s(\mathbf{s}, \mathbf{s}') = \{\rho(\mathbf{s}, \mathbf{s}') - \rho_l(\mathbf{s}, \mathbf{s}')\}\,\Delta(\mathbf{s}, \mathbf{s}')$, where $\Delta(\mathbf{s}, \mathbf{s}')$ is a modulating function, which is specified so that the $\rho_s(\mathbf{s}, \mathbf{s}')$ can well capture the local residual spatial dependence while still permits efficient computations. Motivated by Konomi et al. (2014), we first partition the total input space into B disjoint blocks, and then specify $\Delta(\mathbf{s}, \mathbf{s}')$ in a way such that the residuals are independent across input blocks, but the original residual dependence structure within each block is retained. Specifically, the function $\Delta(\mathbf{s}, \mathbf{s}')$ is taken to be 1 if $\mathbf{s}$ and $\mathbf{s}'$ belong to the same block and 0 otherwise. The approximated correlation function $\rho^\dagger(\mathbf{s}, \mathbf{s}')$ in (B.2) provides an exact recovery of the true correlation within each block, and the approximation errors are $\rho(\mathbf{s}, \mathbf{s}') - \rho_l(\mathbf{s}, \mathbf{s}')$ for locations $\mathbf{s}$ and $\mathbf{s}'$ in different blocks. Those errors are expected to be small for most entries because most of these location pairs are farther apart. To determine the blocks, we first use the R function `cover.design` to choose $B \leq m$ locations among the m locations forming B blocks, then assign each $\mathbf{s}_i$ to the block that is closest to $\mathbf{s}_i$. Here B does not need to be equal to A. When $B = 1$, no approximation is applied to the correlation ρ. When $B = m$, it reduces to the approach of Finley et al. (2009), so the local residual spatial dependence may not be well captured.

Applying the above FSA approach to approximate the correlation function $\rho(\mathbf{s}, \mathbf{s}')$, we can approximate the correlation matrix $\mathbf{R}$ with

$$\boldsymbol{\rho}_{mm}^\dagger = \boldsymbol{\rho}_l + \boldsymbol{\rho}_s = \boldsymbol{\rho}_{mA}\boldsymbol{\rho}_{AA}^{-1}\boldsymbol{\rho}_{mA}' + \left(\boldsymbol{\rho}_{mm} - \boldsymbol{\rho}_{mA}\boldsymbol{\rho}_{AA}^{-1}\boldsymbol{\rho}_{mA}'\right) \circ \boldsymbol{\Delta}, \tag{B.3}$$

where $\boldsymbol{\rho}_{mA} = [\rho(\mathbf{s}_i, \mathbf{s}_j^*)]_{i=1:m, j=1:A}$, $\boldsymbol{\rho}_{AA} = [\rho(\mathbf{s}_i^*, \mathbf{s}_j^*)]_{i,j=1}^A$, and $\boldsymbol{\Delta} = [\Delta(\mathbf{s}_i, \mathbf{s}_j)]_{i,j=1}^m$. Here, the notation "$\circ$" represents the element-wise matrix multiplication. To avoid numerical instability, we

add a small nugget effect $\epsilon = 10^{-10}$ when define $\mathbf{R}$, that is, $\mathbf{R} = (1-\epsilon)\boldsymbol{\rho}_{mm} + \epsilon\mathbf{I}_m$. It follows from equation (B.3) that $\mathbf{R}$ can be approximated by

$$\mathbf{R}^\dagger = (1-\epsilon)\boldsymbol{\rho}^\dagger_{mm} + \epsilon\mathbf{I}_m = (1-\epsilon)\boldsymbol{\rho}_{mA}\boldsymbol{\rho}^{-1}_{AA}\boldsymbol{\rho}'_{mA} + \mathbf{R}_s,$$

where $\mathbf{R}_s = (1-\epsilon)\left(\boldsymbol{\rho}_{mm} - \boldsymbol{\rho}_{mA}\boldsymbol{\rho}^{-1}_{AA}\boldsymbol{\rho}'_{mA}\right) \circ \boldsymbol{\Delta} + \epsilon\mathbf{I}_m$. Applying the Sherman-Woodbury-Morrison formula for inverse matrices, we can approximate $\mathbf{R}^{-1}$ by

$$\left(\mathbf{R}^\dagger\right)^{-1} = \mathbf{R}_s^{-1} - (1-\epsilon)\mathbf{R}_s^{-1}\boldsymbol{\rho}_{mA}\left[\boldsymbol{\rho}_{AA} + (1-\epsilon)\boldsymbol{\rho}'_{mA}\mathbf{R}_s^{-1}\boldsymbol{\rho}_{mA}\right]^{-1}\boldsymbol{\rho}'_{mA}\mathbf{R}_s^{-1}. \tag{B.4}$$

In addition, the determinant of $\mathbf{R}$ can be approximated by

$$\det\left(\mathbf{R}^\dagger\right) = \det\left\{\boldsymbol{\rho}_{AA} + (1-\epsilon)\boldsymbol{\rho}'_{mA}\mathbf{R}_s^{-1}\boldsymbol{\rho}_{mA}\right\}\det(\boldsymbol{\rho}_{AA})^{-1}\det(\mathbf{R}_s). \tag{B.5}$$

Since the $m \times m$ matrix $\mathbf{R}_s$ is a block matrix, the right-hand sides of equations (B.4) and (B.5) involve only inverses and determinants of $A \times A$ low-rank matrices and $m \times m$ block diagonal matrices. Thus the computational complexity can be greatly reduced relative to the expensive computational cost of using original correlation function for large value of m.

Appendix C The DIC and LPML Criteria

To set notation, denote by $\mathcal{D}$ the observed dataset, by $\mathcal{D}_i$ the ith data point, and by $\mathcal{D}_{-i}$ the dataset with $\mathcal{D}_i$ removed, $i = 1, \ldots, n$. Let Ω denote the entire collection of model parameters under a particular model, $L(\mathcal{D}|\Omega)$ be the likelihood function based on observed data $\mathcal{D}$, and $L_i(\cdot|\Omega)$ be the likelihood contribution based on $\mathcal{D}_i$. Suppose $\{\Omega^{(1)}, \ldots, \Omega^{(\mathcal{L})}\}$ are random draws from the full posterior $p_{post}(\Omega|\mathcal{D})$. Let $\hat{\Omega} = \sum_{l=1}^{\mathcal{L}} \Omega^{(l)}/\mathcal{L}$ be the posterior mean estimate for Ω.

The DIC, a generalization of the Akaike information criterion (AIC), is commonly used for comparing complex hierarchical models for which the asymptotic justification of AIC is not appropriate. The DIC is defined as

$$\text{DIC} = -2\log L(\mathcal{D}|\hat{\Omega}) + 2p_D, \tag{C.6}$$

where

$$p_D = 2\left(\log L(\mathcal{D}|\hat{\Omega}) - \frac{1}{\mathcal{L}}\sum_{l=1}^{\mathcal{L}}\log L(\mathcal{D}|\Omega^{(l)})\right)$$

is referred to as the effective number of parameters measuring the model complexity. Similar to AIC, a smaller value of DIC indicates a better fit model.

The definition of LPML is based on the conditional predictive ordinate (CPO) statistic. The CPO for data point $\mathcal{D}_i$ is given by

$$\text{CPO}_i = f(\mathcal{D}_i|\mathcal{D}_{-i}) = \int L_i(\mathcal{D}_i|\Omega)p_{post}(\Omega|\mathcal{D}_{-i})d\Omega,$$

where $p_{post}(\cdot|\mathcal{D}_{-i})$ is the posterior density of Ω give $\mathcal{D}_{-i}$. Let $\text{CPO}_{i,1}$ and $\text{CPO}_{i,2}$ denote the CPO for the ith data point under models 1 and 2, respectively. The ratio $\text{CPO}_{i,1}/\text{CPO}_{i,2}$ measures how well model 1 supports the data point $\mathcal{D}_i$ relative to model 2, based on the remaining data $\mathcal{D}_{-i}$. The product of the CPO ratios gives an overall aggregate summary of how well supported the data are by model 1 relative to model 2 and is called the pseudo Bayes factor (PBF),

$$B_{12} = \prod_{i=1}^{n}\frac{\text{CPO}_{i,1}}{\text{CPO}_{i,2}}.$$

It is well known that Bayes factors (Kass and Raftery, 1995; Han and Carlin, 2001) are usually difficult to obtain in practice. The PBF is a surrogate for the more traditional Bayes factor and can be interpreted similarly, but is more analytically tractable, much less sensitive to prior assumptions, and does not suffer from Lindley's paradox.

As noted by Gelfand and Dey (1994), one can use importance sampling to estimate CPO_i by

$$\left\{ \frac{1}{\mathcal{L}} \sum_{l=1}^{\mathcal{L}} \frac{1}{L_i(\mathcal{D}_i|\Omega^{(l)})} \right\}^{-1}.$$

However, these estimates may be unstable since the weights $\omega_{i,l} = 1/L_i(\mathcal{D}_i|\Omega^{(l)})$ can have infinite variance (Epifani et al., 2008), depending on the tail behavior of $p_{post}(\Omega|\mathcal{D}_{-i})$ relative to $L_i(\mathcal{D}_i|\Omega)$ as a function of Ω. To stabilize the weights, Vehtari and Gelman (2014) suggest replacing $\omega_{i,l}$ with $\tilde{\omega}_{i,l} = \min\{\omega_{i,l}, \sqrt{\mathcal{L}}\bar{\omega}_i\}$, where $\bar{\omega}_i = \sum_{l=1}^{\mathcal{L}} \omega_{i,l}/\mathcal{L}$. Therefore, the stabilized estimate of the CPO statistic is

$$\widehat{\text{CPO}}_i = \frac{\sum_{l=1}^{\mathcal{L}} L_i(\mathcal{D}_i|\Omega^{(l)})\tilde{\omega}_{i,l}}{\sum_{l=1}^{\mathcal{L}} \tilde{\omega}_{i,l}}.$$

Finally, the LPML is defined as

$$\text{LPML} = \sum_{i=1}^{n} \log \widehat{\text{CPO}}_i. \tag{C.7}$$

A further improved estimate was recently proposed by Vehtari et al. (2017) using Pareto-smoothed importance sampling; this version will be implemented in later versions of the R package.

The LPML can be viewed as a predictive measure that generalizes leave-one-out cross-validated prediction error to more heavily penalize "bad predictions." Consider the frequentist LPML proposed by Geisser and Eddy (1979) for normal-errors regression data. Let $y_i \overset{ind.}{\sim} N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$; then

$$\text{CPO}_i = \frac{1}{\sqrt{2\pi}\hat{\sigma}_i} \exp\left\{ -\frac{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_i)^2}{2\hat{\sigma}_i^2} \right\},$$

where $(\hat{\boldsymbol{\beta}}_i, \hat{\sigma}_i)$ is the MLE of $(\boldsymbol{\beta}, \sigma)$ leaving out $(\mathbf{x}_i, y_i)$. Then

$$-\text{LPML} = \underbrace{\sum_{i=1}^{n} \frac{1}{2\hat{\sigma}_i}(y_i - \hat{y}_{-i})^2}_{\text{squared bias}} + \underbrace{\sum_{i=1}^{n} \log \hat{\sigma}_i}_{\text{variance}} + \text{constant}.$$

This generalizes to any location-scale family, e.g. parametric survival models $\log t_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$, where ϵ_i has a scaled standard extreme value distribution, scaled log-logistic distribution, or scaled normal distribution yielding common Weibull, log-logistic, and log-normal regression models. Note that unlike the usual predicted residual error sum of squares (PRESS) statistic the bias terms are weighted by the variability of the prediction: "bad" predictions with less variability (more precision) provide much more discrepancy than "bad" predictions with large variability. Having both bias and variance pieces, the LPML is of similar form to the L-measure (Ibrahim et al., 2001), but more naturally generalizes to survival data; note that Ibrahim et al. (2001) advocating taking the log of the survival time and require a different L-measure for each family of distributions.

A Bayesian might view the frequentist CPO_i using the MLE above as overoptimistic. The MLE is the posterior mode under a flat prior and sampling variability is not taken into account. Instead, one might want to average the CPO_i statistic over the (perhaps asymptotic) estimated

sampling distribution of Ω_i, e.g. $\Omega_i \overset{\bullet}{\sim} N(\hat{\Omega}_i, \mathbf{V}_i)$. Equivalently, and more precisely, the Bayesian approach averages the predictive density for a new observation with covariates $\mathbf{x}_i$ over the leave-i-out posterior $[\Omega|\mathcal{D}_{-i}]$. Thus the Bayesian LPML used here can be viewed as a measure similar to PRESS or prediction error, but a properly pessimistic one that averages over the (non-asymptotic) sampling distribution of the parameters. The more sampling variability there is (reflecting smaller sampler sizes n), the more heavily each CPO_i is penalized.

In addition to DIC and LPML, the Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010) has also gained popularity in recent years due to its stability compared to DIC (Gelman et al., 2014; Vehtari and Gelman, 2014). The WAIC is defined as

$$\text{WAIC} = -2\sum_{i=1}^{n} \log\left(\frac{1}{\mathcal{L}}\sum_{l=1}^{\mathcal{L}} L_i(\mathcal{D}_i|\Omega^{(l)})\right) + 2p_W, \tag{C.8}$$

where

$$p_W = \sum_{i=1}^{n}\left[\frac{1}{\mathcal{L}-1}\sum_{l=1}^{\mathcal{L}}\left\{\log L_i(\mathcal{D}_i|\Omega^{(l)}) - \frac{1}{\mathcal{L}}\sum_{k=1}^{\mathcal{L}}\log L_i(\mathcal{D}_i|\Omega^{(k)})\right\}^2\right]$$

is the effective number of parameters. A smaller value of WAIC indicates a better predictive model. WAIC can be viewed as an approximation to $-2\sum_{i=1}^{n}\log \text{CPO}_i$ (Gelman et al., 2014), so WAIC is also used to compare models' predictive performance. The WAIC has been implemented in the function `survregbayes` and saved in its returned object.

Appendix D Parametric vs. Nonparametric $S_0(\cdot)$

Many authors have found parametric models to fit as well or better than competing semiparametric models (Cox and Oakes, 1984, p. 123; Nardi and Schemper, 2003). Here, testing for the adequacy of the simpler underlying parametric model is developed. The proposed semiparametric models have their baseline survival functions centered at a parametric family $S_{\boldsymbol{\theta}}(t)$. Note that $\mathbf{z}_{J-1} = \mathbf{0}$ implies $S_0(t) = S_{\boldsymbol{\theta}}(t)$. Therefore, testing $H_0 : \mathbf{z}_{J-1} = \mathbf{0}$ versus $H_1 : \mathbf{z}_{J-1} \neq \mathbf{0}$ leads to the comparison of the semiparametric model with the underlying parametric model. Let BF_{10} be the Bayes factor between H_1 and H_0. Zhou et al. (2017) proposed to estimate BF_{10} by a large-sample approximation to the generalized Savage-Dickey density ratio (Verdinelli and Wasserman, 1995). Adapting their approach BF_{10} is estimated

$$\widehat{BF}_{10} = \frac{p(\mathbf{0}|\hat{\alpha})}{N_{J-1}(\mathbf{0}; \hat{\mathbf{m}}, \hat{\boldsymbol{\Sigma}})}, \tag{D.9}$$

where $p(\mathbf{0}|\alpha) = \Gamma(\alpha J)/[J^{\alpha}\Gamma(\alpha)]^J$ is the prior density of $\mathbf{z}_{J-1}$ evaluated at $\mathbf{z}_{J-1} = \mathbf{0}$, $\hat{\alpha}$ is the posterior mean of α, $N_p(\cdot; \mathbf{m}, \boldsymbol{\Sigma})$ denotes a p-variable normal density with mean $\mathbf{m}$ and covariance $\boldsymbol{\Sigma}$, and $\hat{\mathbf{m}}$ and $\hat{\boldsymbol{\Sigma}}$ are posterior mean and covariance of $\mathbf{z}_{J-1}$.

Appendix E Variable Selection

There is a large amount of literature on Bayesian variable selection methods; see O'Hara and Sillanpää (2009) for a comprehensive review. Let $\mathbf{x} = (x_1, \ldots, x_p)'$ denote the p-vector of covariates in general. The most direct approach is to multiply β_ℓ by a latent Bernoulli variable γ_ℓ for $\ell = 1, \ldots, p$, where $\gamma_\ell = 1$ indicates the presence of x_ℓ in the model, and then assume an appropriate prior on $(\boldsymbol{\beta}, \boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$. Kuo and Mallick (1998) considered an independent prior

$p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = N_p(\mathbf{0}, \mathbf{W}_0) \times \prod_{\ell=1}^{p} \text{Bern}(q_\ell)$, where $\mathbf{W}_0$ was taken as a diagonal matrix yielding a diffuse prior on $\boldsymbol{\beta}$, and q_ℓ is a prior probability of including x_ℓ in the model. The resulting MCMC algorithm does not require any tuning, but mixing can be poor if the prior on $\boldsymbol{\beta}$ is too diffuse (O'Hara and Sillanpää, 2009). The g-prior of Zellner (1983) and its various extensions (Bové et al., 2011; Hanson et al., 2014) have been widely used for variable selection. We consider one such prior adapted for use in the semiparametric survival models considered here. Specifically, the same prior as Kuo and Mallick (1998) is considered, but with

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

where $\mathbf{X}$ is the usual design matrix, but with mean-centered covariates, i.e. $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$. Assume that the covariate vectors $\mathbf{x}_{ij}$ arise from a distribution G with support on $\mathcal{X} \subseteq \mathbb{R}^p$, and are independent of $\boldsymbol{\beta}$. Following Hanson et al. (2014) g is set equal to a constant based on prior information on $e^{\mathbf{x}'\boldsymbol{\beta}}$, i.e. the relative risks (PH), acceleration factors (AFT), or odds factors (PO) of random subjects $\mathbf{x}$ relative to their mean $\int_{\mathcal{X}} \mathbf{x} G(d\mathbf{x})$. Under the prior (E.10), Hanson et al. (2014) showed that $\mathbf{x}'\boldsymbol{\beta}$ has an approximately normal distribution with mean 0 and variance ng. Thus, a simple method of choosing g is to pick a number M such that a random $e^{\mathbf{x}'\boldsymbol{\beta}}$ is less than M with probability q. It follows that $g = \left[\log M / \Phi^{-1}(q)\right]^2 / p$. Here, $M = 10$ and $q = 0.9$ are fixed. The MCMC procedure is described in supplementary Appendix A. Posterior output includes a list of sub-models with their posterior probabilities, i.e. a ranking of models much like the best subsets C_p statistic.

This variable selection method was originally termed "stochastic search variable selection" (SSVS) by George and McCulloch (1993) who instead of using Bernoulli point masses for each regression effect used highly concentrated normal distributions centered at zero. This approach has also been called "spike and slab" variable selection by many authors. A recent review and extensive simulation study by Pavlou et al. (2016) suggests that SVSS can routinely outperform other variable selection approaches. They found that SSVS performed overall the best across many realistic data scenarios for variable selection among methods that also include versions of the LASSO (regular, adaptive, and Bayesian), SCAD, and the elastic net. All methods grossly outperformed backwards elimination; see Table 4 in Pavlou et al. (2016).

Appendix F Left-Truncation and Time-Dependent Covariates

To avoid an explosion of subscripts, drop the ij from t_{ij}, etc. The survival time t is left-truncated at $u \geq 0$ if u is the time when the subject under consideration is first observed. Left-truncation often occurs when age is used as the time scale. Given the observed left-truncated data $\{(u, a, b, \mathbf{x}, \mathbf{s})\}$, where $a \geq u$, the likelihood contribution is

$$L = [S_{\mathbf{x}}(a) - S_{\mathbf{x}}(b)]^{I\{a<b\}} f_{\mathbf{x}}(a)^{I\{a=b\}} / S_{\mathbf{x}}(u).$$

Note that the left censored data under left-truncation are of the form (u, b).

We next discuss how to extend the semiparametric AFT, PH and PO models to handle time-dependent covariates. Let $\{(u, a, b, \mathbf{x}(t), \mathbf{s}) : u \leq t \leq a\}$ be the observed data with time-dependent covariates and possible left-truncation. Suppose we observe $\mathbf{x}(t)$ at o ordered times $t = t_1, \ldots, t_o$, denoted as $\mathbf{x}_1, \ldots, \mathbf{x}_o$, respectively, where $t_1 = u$ and $t_o \leq a$. Following Kneib (2006) and Hanson et al. (2009), we assume that $\mathbf{x}(t)$ is a step function given by

$$\mathbf{x}(t) = \sum_{k=1}^{o} \mathbf{x}_k I(t_k \leq t < t_{k+1}),$$

where $t_{o+1} = \infty$. Assuming the AFT, PH or PO holds conditionally on each interval, the survival function at time a is

$$\begin{aligned} P(t > a) &= P(t > a | t > t_o) \prod_{k=1}^{o-1} P(t > t_{k+1} | t > t_k) \\ &= \frac{S_{\mathbf{x}_o}(a)}{S_{\mathbf{x}_o}(t_o)} \prod_{k=1}^{o-1} \frac{S_{\mathbf{x}_k}(t_{k+1})}{S_{\mathbf{x}_k}(t_k)}. \end{aligned}$$

This leads to the usual PH model for time-dependent covariates (Cox, 1972), the AFT model first proposed by Prentice and Kalbfleisch (1979), and a particular piecewise PO model.

Returning to the use of subscripts for the ijth subject, for time-dependent covariates replace $(u_{ij}, a_{ij}, b_{ij}, \mathbf{x}_{ij}(t), \mathbf{s}_i)$ by a set of new o_{ij} observations $(t_{ij,1}, t_{ij,2}, \infty, \mathbf{x}_{ij,1}, \mathbf{s}_i)$, $(t_{ij,2}, t_{ij,3}, \infty, \mathbf{x}_{ij,2}, \mathbf{s}_i)$, ..., $(t_{ij,o_{ij}}, a_{ij}, b_{ij}, \mathbf{x}_{ij,o_{ij}}, \mathbf{s}_i)$ yielding an augmented left-truncated data set of size $\sum_{i=1}^{m} \sum_{j=1}^{n_i} o_{ij}$. Then the likelihood function becomes

$$\begin{aligned} L(\mathbf{w}_J, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \Bigg\{ & \left[S_{\mathbf{x}_{ij,o_{ij}}}(a_{ij}) - S_{\mathbf{x}_{ij,o_{ij}}}(b_{ij}) \right]^{I\{a_{ij} < b_{ij}\}} f_{\mathbf{x}_{ij,o_{ij}}}(a_{ij})^{I\{a_{ij} = b_{ij}\}} / S_{\mathbf{x}_{ij,o_{ij}}}(t_{ij,o_{ij}}) \\ & \times \prod_{k=1}^{o_{ij}-1} \frac{S_{\mathbf{x}_{ij,k}}(t_{ij,k+1})}{S_{\mathbf{x}_{ij,k}}(t_{ij,k})} \Bigg\}. \end{aligned}$$

Note that the derivations above still hold for time-dependent covariates without left-truncation (i.e. $u_{ij} = 0$ for all i and j).

Appendix G Partially linear predictors

An additive PH model was first considered by Gray (1992) as

$$h_{\mathbf{x}_{ij}}(t) = h_0(t) \exp\{\mathbf{x}_{ij}' \boldsymbol{\beta} + \sum_{\ell=1}^{p} u_\ell(x_{ij\ell})\},$$

where the nonlinear functions $u_1(\cdot), \ldots, u_p(\cdot)$ are modeled via penalized B-splines with the linear portion removed. Setting some of the $u_\ell(\cdot) \equiv 0$ gives the so-called "partially linear PH model" that has been given a great deal of attention in recent literature. This model has been extended to spatial versions by Kneib (2006) and Hennerfeind et al. (2006) for PH and can be easily fit in `R2BayesX`.

Additive partially linear predictors can be implemented in the proposed AFT, PH and PO models by simply adding a linear basis expansion for any continuous covariate; cubic B-splines are considered here and illustrated in Section 3.3. Specifically, $u_\ell(\cdot)$ is parameterized as

$$u_\ell(\cdot) = \sum_{k=1}^{K} \xi_{\ell k} B_{\ell k}(\cdot),$$

where $\{B_{\ell k}(\cdot) : k = 0, \ldots, K+1\}$ are the standard cubic B-spline basis functions with knots determined by the data; the first and last basis functions have been dropped to ensure a full-rank model (the linear term is already included). Independent normal priors are considered for $\boldsymbol{\beta}$ and $\boldsymbol{\xi}_\ell = (\xi_{\ell 1}, \ldots, \xi_{\ell K})'$:

$$\boldsymbol{\beta} \sim N_p(\mathbf{0}, \mathbf{W}_0), \quad \boldsymbol{\xi}_\ell \sim N_K(\mathbf{0}, g n (\mathbf{X}_\ell' \mathbf{X}_\ell)^{-1}), \ell = 1, \ldots, p$$

where $\mathbf{W}_0 = 10^{10}\mathbf{I}_p$, $\mathbf{X}_\ell$ is the design matrix for the $u_\ell(\cdot)$ term, and $g = \left[\log 10/\Phi^{-1}(0.9)\right]^2/K$. This approach can be viewed as a simplified version of the Bayesian P-splines (Lang and Brezger, 2004) with fewer basis functions and a g-prior "penalty" instead of a random-walk penalty. Note that posterior updating could be inefficient if a large number of basis functions is considered, as $(\boldsymbol{\beta}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_p)$ is currently updated in one large block via adaptive Metropolis. For this reason, the full Bayesian P-spline approach may be a better choice, but requires updating high-dimensional vectors of spline coefficient parameters, and their suggested iteratively weighted least squares proposals would need to be modified to handle our survival models. We hope to include this in future updates of the R package

Bayes factors can be used to test the linearity of $x_{ij\ell}$ through the hypothesis $H_0 : \boldsymbol{\xi}_\ell = \mathbf{0}$ versus $H_1 : \boldsymbol{\xi}_\ell \neq \mathbf{0}$. Let BF_{10} be the Bayes factor between H_1 and H_0. We estimate BF_{10} by a large-sample approximation to the Savage-Dickey density ratio (Dickey, 1971)

$$\widehat{BF}_{10} = \frac{N_K(\mathbf{0}; \mathbf{0}, gn(\mathbf{X}_\ell'\mathbf{X}_\ell)^{-1})}{N_K(\mathbf{0}; \hat{\mathbf{m}}_\ell, \hat{\boldsymbol{\Sigma}}_\ell)}, \tag{G.11}$$

where $\hat{\mathbf{m}}_\ell$ and $\hat{\boldsymbol{\Sigma}}_\ell$ are posterior mean and covariance of $\boldsymbol{\xi}_\ell$.

Appendix H Implementation Using R

An illustrative use of the `R` function `survregbayes` in the package `spBayesSurv` is presented to fit AFT, PH and PO frailty models with the TBP prior on baseline survival functions using simulated data. We take Example 2 of the variable selection simulation (see **Simulation IV** below) as an example. The following code is used to generate data:

```
##-------------Load libraries-------------------##
rm(list=ls())
library(coda)
library(survival)
library(spBayesSurv)
library(BayesX)

##-------------Set the true models--------------##
betaT = c(1,1,0,0,0);
## Baseline Survival
f0oft = function(t) 0.5*dlnorm(t, -1, 0.5)+0.5*dlnorm(t,1,0.5);
S0oft = function(t) 0.5*plnorm(t, -1, 0.5, lower.tail=FALSE)+
0.5*plnorm(t, 1, 0.5, lower.tail=FALSE)
## The Survival function:
Sioft = function(t,x,v=0)  exp( log(S0oft(t))*exp(sum(x*betaT)+v) ) ;
Fioft = function(t,x,v=0) 1-Sioft(t,x,v);
## The inverse for Fioft
Finv = function(u, x,v=0) uniroot(function (t) Fioft(t,x,v)-u, lower=1e-100, upper=1e100,
extendInt ="yes")$root

##-------------Generate data-------------------##
## read the adjacency matrix of Nigeria for the 37 states
nigeria=read.bnd(system.file("otherdata/nigeria.bnd",
package="spBayesSurv"));
adj.mat=bnd2gra(nigeria)
W = diag(diag(adj.mat)) - as.matrix(adj.mat); m=nrow(W);
tau2T = 1;
covT = tau2T*solve(diag(rowSums(W))-W+diag(rep(1e-10, m)));
```

```
v0 = MASS::mvrnorm(n=1, mu=rep(0,m), Sigma=covT);
v = v0-mean(v0);
mis = rep(20, m); n = sum(mis);
vn = rep(v, mis);
id = rep(1:m, mis);
## generate x
x1 = rbinom(n, 1, 0.5); x2 = rnorm(n, 0, 1);
x3 = x2+0.15*rnorm(n); x4 = rnorm(n, 0, 1); x5 = rnorm(n, 0, 1);
X = cbind(x1, x2, x3, x4, x5);
colnames(X) = c("x1", "x2", "x3", "x4", "x5");
## generate survival times
u = runif(n);
tT = rep(0, n);
for (i in 1:n){
tT[i] = Finv(u[i], X[i,], vn[i]);
}
## generate partly interval-censored data
t1=rep(NA, n);t2=rep(NA, n); delta=rep(NA, n);
n1 = floor(0.5*n); ## right-censored part
n2 = n-n1; ## interval-censored part
# right-censored part
rcen = sample(1:n, n1);
t1_r=tT[rcen];t2_r=tT[rcen];
Centime = runif(n1, 2, 6);
delta_r = (tT[rcen]<=Centime) +0 ; length(which(delta_r==0))/n1;
t1_r[which(delta_r==0)] = Centime[which(delta_r==0)];
t2_r[which(delta_r==0)] = NA;
t1[rcen]=t1_r; t2[rcen]=t2_r; delta[rcen] = delta_r;
# interval-censored part
intcen = (1:n)[-rcen];
t1_int=rep(NA, n2);t2_int=rep(NA, n2); delta_int=rep(NA, n2);
npois = rpois(n2, 2)+1;
for(i in 1:n2){
gaptime = cumsum(rexp(npois[i], 1));
pp = Fioft(gaptime, X[intcen[i],], vn[intcen[i]]);
ind = sum(u[intcen[i]]>pp);
if (ind==0){
delta_int[i] = 2;
t2_int[i] = gaptime[1];
}else if (ind==npois[i]){
delta_int[i] = 0;
t1_int[i] = gaptime[ind];
}else{
delta_int[i] = 3;
t1_int[i] = gaptime[ind];
t2_int[i] = gaptime[ind+1];
}
}
t1[intcen]=t1_int; t2[intcen]=t2_int; delta[intcen] = delta_int;
## make a data frame
d = data.frame(t1=t1, t2=t2, X, delta=delta, tT=tT, ID=id, frail=vn); table(d$delta)/n;

##------- Fit the PH model with variable selection -----------##
# MCMC parameters
nburn=10000; nsave=2000; nskip=4; niter = nburn+nsave
mcmc=list(nburn=nburn, nsave=nsave, nskip=nskip, ndisplay=500);
prior = list(maxL=15, a0=1, b0=1);
state <- list(cpar=1);
ptm<-proc.time()
```

```
res2 = survregbayes(formula = Surv(t1, t2, type="interval2")~x1+x2+x3+
x4+x5+frailtyprior("car", ID),
data=d, survmodel="PH", selection=TRUE, prior=prior, mcmc=mcmc, state=state,
dist="loglogistic", Proximity = W);
sfit2=summary(res2); sfit2;
systime2=proc.time()-ptm; systime2;
```

Note that the data have to be sorted by region ID before model fitting. The argument `mcmc` above specifies that the chain is subsampled every 5 iterates to get a total of $2,000$ scans after a burn-in period of $10,000$ iterations. The argument `prior` is used set all the priors; if nothing is specified, the default priors in the paper are used. The output is given below:

```
Posterior inference of regression coefficients
(Adaptive M-H acceptance rate: 0.105):
     Mean      Median    Std. Dev.  95%CI-Low  95%CI-Upp
x1   1.00002   1.00201   0.09491    0.82401    1.18337
x2   0.93568   0.97427   0.16605    0.44710    1.15790
x3  -0.68349  -0.66875   0.83818   -2.26155    0.56054
x4   0.03566   0.06164   0.75003   -1.42845    1.47316
x5  -0.02822   0.01343   0.72955   -1.43050    1.28246

Posterior inference of precision parameter
(Adaptive M-H acceptance rate: 0.2652):
       Mean     Median   Std. Dev.  95%CI-Low  95%CI-Upp
alpha  0.3843  0.3642  0.1509     0.1541     0.7288

Posterior inference of conditional CAR frailty variance
          Mean    Median  Std. Dev.  95%CI-Low  95%CI-Upp
variance  0.6576  0.6162  0.2504     0.2994     1.2456

Variable selection:
       x1,x2   x1,x2,x3  x1,x2,x4  x1,x2,x5  x1,x2,x3,x5  x1,x2,x3,x4  x1,x2,x4,x5
prop.  0.6490  0.2245    0.0505    0.0485    0.0155       0.0075       0.0045

Log pseudo marginal likelihood: LPML=-417.0232
Deviance Information Criterion: DIC=833.0498
Number of subjects: n=740
```

Remarks: The function `survregbayes` can also fit a semiparametric survival model (AFT, PH, or PO) with independent Gaussian frailties by setting `frailtyprior("iid", ID)`, with Gaussian random field frailties by setting `frailtyprior("grf", ID)`, a model without frailties by removing `frailtyprior()` in the formula, and a parametric (loglogistic, lognormal or weibull) survival model by specifying `a0` at a negative value and adding an argument `state=list(cpar=Inf)`. If FSA is used for GRF frailty models, the number of knots A and the number of blocks B are specified via `prior=list(nknots=`A`, nblock=`B`)`.

Appendix I Additional Results for Real Data Applications

I.1 Loblolly Pine Survival Data

Table S3 presents some baseline characteristics for the trees.

Table S3: Loblolly pine data. Baseline characteristics of the 45,525 trees.

Categorical variables	**Level**	**Proportion (%)**
Censoring status	uncensored	12.65
	right censored	87.35
Treatment (treat)	1–control	24.78
	2–light thinning	40.32
	3–heavy thinning	34.90
Physiographic region (PhyReg)	1–coastal	55.53
	2–piedmont	37.01
	3–other	7.46
Crown class (C)	1–dominant	28.21
	2–codominant	52.22
	3–intermediate	15.50
	4–suppressed	4.07
Continuous variables	**Mean**	**Std. Dev.**
Total height of tree in meters (TH)	38.47	11.77
Diameter at breast height in cm (DBH)	5.88	1.77

Appendix J Additional Results for Simulations

J.1 Simulation I: Areal Data

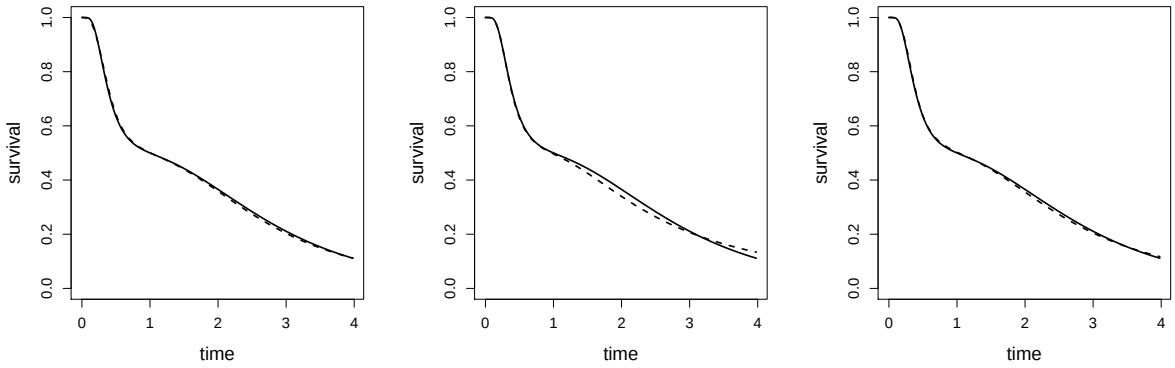
Figure S1 presents the average, across the 500 MC replicates, of fitted (posterior means over a grid of time points) baseline survival functions; the proposed method capably captures complex (here bimodal) baseline survival curves.

J.2 Simulation III: Georeferenced Data

We generated the data using the same settings as **Simulation I** except that $i = 1, \ldots, 150, j = 1, \ldots, 5$, and v_i follows the GRF prior with $\tau^2 = 1$, $\nu = 1$ and $\phi = 1$. The locations $\{\mathbf{s}_i\}_{i=1}^{150}$ were generated from $[0, 10]^2$ uniformly. Table S4 summaries the results, where we see that the point estimates of $\boldsymbol{\beta}$ are unbiased under all three models, SD-Est values are close to the corresponding PSDs, and the CP values are close to the nominal 95% level. We also observe that ϕ tends to be overestimated and the standard deviations for τ^2 and ϕ are underestimated (because SD-Est is smaller than PSD). Even though, the CP values are still close to 95%. The ESS values for $\boldsymbol{\beta}$ are much smaller than these obtained for areal data, indicating that the georeferenced spatial dependency makes the posterior samples more correlated.

J.3 Simulation IV: Variable Selection

We next assess the performance of our variable selection method via three simulated examples. For each example, one data set was generated from the PH model with $S_0(t)$ and ICAR as in **Simulation I**. Under Example 1, we set $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ij5})$ with $x_{ij1} \sim$Bernoulli(0.5) and $x_{ij2}, \ldots, x_{ij5} \overset{iid}{\sim} N(0, 1)$, and $\boldsymbol{\beta} = (1, 1, 0, 0, 0)'$. Example 2 is identical to Example 1 except that $x_{ij3} = x_{ij2} + 0.15z$ where $z \sim N(0, 1)$, yielding a 0.989 correlation between x_2 and x_3. For Example 3, we set $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ij10})$ with $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$ and $x_{ijk}|z \overset{iid}{\sim} N(z, 1)$ where



(a) (b) (c)

Figure S1: Simulated I. Mean, across the 500 MC replicates, of the posterior mean of the baseline survival functions under AFT (panel a), PH (panel b) and PO (panel c). The true curves are represented by continuous lines and the fitted curves are represented by dashed lines.

Table S4: Simulation III. Averaged bias (BIAS) and posterior standard deviation (PSD) of each point estimate, standard deviation (across 500 MC replicates) of the point estimate (SD-Est), coverage probability (CP) for the 95% credible interval, and effective sample size (ESS) for each point estimate.

Model	Parameter	BIAS	PSD	SD-Est	CP	ESS
AFT	$\beta_1 = 1$	-0.002	0.085	0.089	0.946	1933
	$\beta_2 = 1$	-0.000	0.045	0.042	0.964	1815
	$\tau^2 = 1$	0.000	0.329	0.220	0.948	548
	$\phi = 1$	0.082	0.388	0.357	0.962	471
PH	$\beta_1 = 1$	-0.016	0.112	0.116	0.934	1943
	$\beta_2 = 1$	-0.015	0.068	0.068	0.942	1110
	$\tau^2 = 1$	0.042	0.451	0.316	0.938	366
	$\phi = 1$	0.066	0.471	0.420	0.918	351
PO	$\beta_1 = 1$	-0.001	0.157	0.159	0.952	3006
	$\beta_2 = 1$	0.003	0.087	0.088	0.944	1960
	$\tau^2 = 1$	0.034	0.410	0.341	0.954	502
	$\phi = 1$	0.313	1.361	0.768	0.912	353

Table S5: Simulated IV. High frequency models with selected variables.

Example 1		Example 2		Example 3	
Variables	Proportions	Variables	Proportions	Variables	Proportions
1 2	0.80	1 2	0.49	1-5	0.63
1 2 3	0.08	1 2 3	0.22	1-5, 10	0.15
1 2 5	0.05	1 3	0.17	1-5, 7	0.09
1 2 4	0.05	1 2 5	0.04	1-5, 8	0.05

$z \sim N(0,1)$, which induces pairwise correlations of about 0.5. We applied our method to the three simulated datasets using all default priors designed for variable selection. A sample of $10,000$ scans was thinned from $50,000$ after a burn-in period of $10,000$ iterations. Table S5 lists the proportions for the four highest frequency models under each example. The results reveal that our method predicts the right model very well even in the presence of extreme collinearity.

J.4 Comparing with Polya Trees

Zhao et al. (2009) considered the AFT, PH and PO models for right censored areal data, and used the mixture of Polya trees (MPT) prior on the baseline survival function. In their MCMC scheme, most parameters were updated using simple random walk Metropolis-Hastings steps, so a careful tuning of the proposal distribution was required to achieve desirable acceptance rate. We instead used adaptive Metropolis samplers (Haario et al., 2001) on most parameters and implemented the three MPT models into an R function `survregbayes2`; this function can also fit arbitrarily censored data. We generated data using the same settings as **Simulation I**, then fitted each model with finite Polya tree level equal to 4, a $\Gamma(5,1)$ prior on the Polya tree precision parameter, and priors on other parameters similar to Section 2.3 in the main paper. For each MCMC algorithm, 5,000 scans were thinned from 50,000 after a burn-in period of 10,000 iterations.

Table S6 summarizes the results for regression parameters $\boldsymbol{\beta}$ and the ICAR variance τ^2, including the averaged bias (BIAS) and posterior standard deviation (PSD) of each point estimate (posterior mean for $\boldsymbol{\beta}$ and median for τ^2), the standard deviation (across 500 MC replicates) of the point estimate (SD-Est), the coverage probability (CP) of the 95% creditable interval, and effective sample size (ESS) out of 5,000 (Sargent et al., 2000) for each point estimate. We can see that effective sample sizes for β_1 and β_2 under the MPT AFT are 2 times smaller than those under the TBP AFT. In addition, the MPT PH model provides more biased estimates than the TBP PH.

Due to the non-smoothness of Polya tree densities, the MPT AFT often suffers poor mixing when the true baseline survival function is far away from the centering parametric distribution family $S_{\boldsymbol{\theta}}$ and uncensored survival times are available. For example, for right censored data, the likelihood will involve $f_{\mathbf{x}_{ij}}(t) = e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+v_i} f_0(e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+v_i}t)$, where $f_0(\cdot)$ is the density of a Polya tree. Note that $f_0(\cdot)$ consists of many big jumps when the precision parameter of the Polya trees is small, and hence a tiny change in $\boldsymbol{\beta}$ imply a big jump in the likelihood value, leading to poor mixing. However, MCMC mixing issues are mitigated for interval censored data, since only the survival function $S_0(e^{\mathbf{x}_{ij}'\boldsymbol{\beta}+v_i}t)$ is involved in the likelihood and $S_0(t)$ is continuous.

J.5 Model Selection via LPML and DIC

We next demonstrate via simulations that the LPML and DIC are reasonable criteria for model selection among AFT, PH and PO models. Arbitrarily censored survival data of size $n = 740$ and $n = 1850$ were generated from each of the three models with ICAR frailties using the same settings

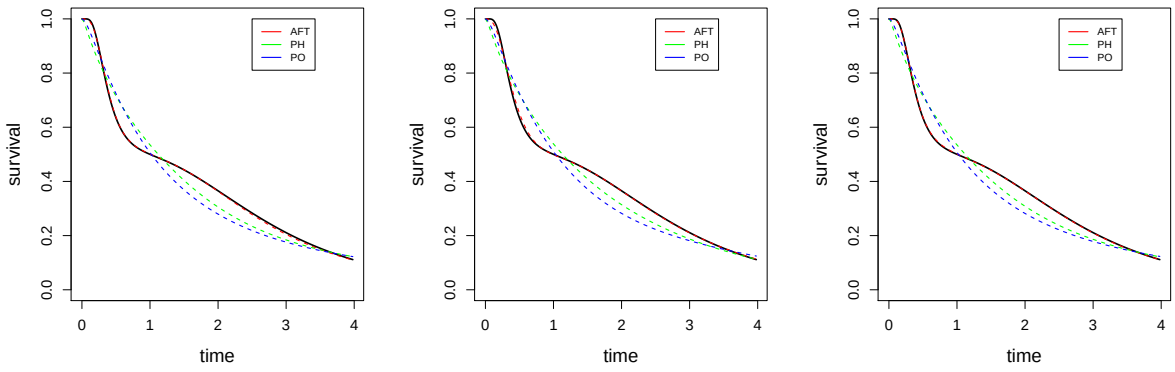
Table S6: Simulation under MPT. Averaged bias (BIAS) and posterior standard deviation (PSD) of each point estimate, standard deviation (across 500 MC replicates) of the point estimate (SD-Est), coverage probability (CP) for the 95% credible interval, and effective sample size (ESS) out of 5,000 with thinning=10 for each point estimate.

Model	Parameter	BIAS	PSD	SD-Est	CP	ESS
AFT	$\beta_1 = 1$	0.002	0.094	0.071	0.986	1009
	$\beta_2 = 1$	0.002	0.050	0.038	0.988	1079
	$\tau^2 = 1$	0.013	0.309	0.243	0.976	3760
PH	$\beta_1 = 1$	-0.045	0.099	0.098	0.932	2887
	$\beta_2 = 1$	-0.043	0.060	0.060	0.874	1794
	$\tau^2 = 1$	-0.084	0.318	0.280	0.954	3599
PO	$\beta_1 = 1$	-0.014	0.149	0.142	0.962	3579
	$\beta_2 = 1$	-0.029	0.082	0.078	0.938	2561
	$\tau^2 = 1$	-0.038	0.407	0.346	0.966	2903

as **Simulation I**. For each model, 200 MC replicates were generated. We fitted each dataset using all three models with the default priors and the same MCMC settings as **Simulation I**. Table S7 (under log-logistic $S_{\boldsymbol{\theta}}(\cdot)$) presents the proportion (out of 200 MC replicates) of times each model is picked. The model picked is the one with largest LPML or smallest DIC. DIC and LPML yield very similar proportions for $n = 740$ and identical results when $n = 1850$, indicating that the two criteria are consistent for model comparison. When the true model is PH, DIC has a 3% chance of picking PO under $n = 740$, but is reduced to zero for the larger sample size $n = 1850$.

Table S7: Simulation for model selection via LPML and DIC. Proportion of times DIC or LPML selects each model when truth is known out of 200 replicated datasets.

		log-logistic			Weibull			log-normal		
		Model picked			Model picked			Model picked		
True model	Criteria	AFT	PH	PO	AFT	PH	PO	AFT	PH	PO
					$n = 740$					
AFT	DIC	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000
	LPML	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000
PH	DIC	0.000	0.985	0.015	0.000	1.000	0.000	0.000	1.000	0.000
	LPML	0.000	0.970	0.030	0.000	1.000	0.000	0.000	0.995	0.005
PO	DIC	0.000	0.000	1.000	0.000	0.075	0.925	0.000	0.000	1.000
	LPML	0.000	0.000	1.000	0.000	0.025	0.975	0.000	0.000	1.000
					$n = 1850$					
AFT	DIC	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000
	LPML	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000
PH	DIC	0.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000
	LPML	0.000	1.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000
PO	DIC	0.000	0.000	1.000	0.000	0.010	0.990	0.000	0.000	1.000
	LPML	0.000	0.000	1.000	0.000	0.010	0.990	0.000	0.000	1.000



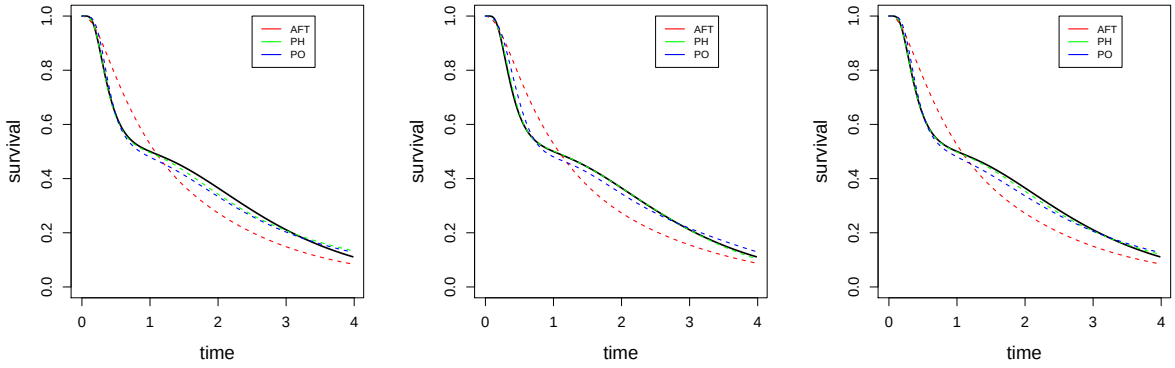
(a) (b) (c)

Figure S2: Simulation for sensitivity analysis of the TBP's centering distribution when AFT is the true model. Mean, across the 200 MC replicates, of the posterior mean of the baseline survival functions under log-logistic (panel a), Weibull (panel b) and log-normal (panel c). The true curves are represented by continuous lines and the fitted curves are represented by dashed lines (red is for AFT, green is for PH and blue is for PO).

J.6 Sensitivity Analysis of The TBP's Centering Distribution

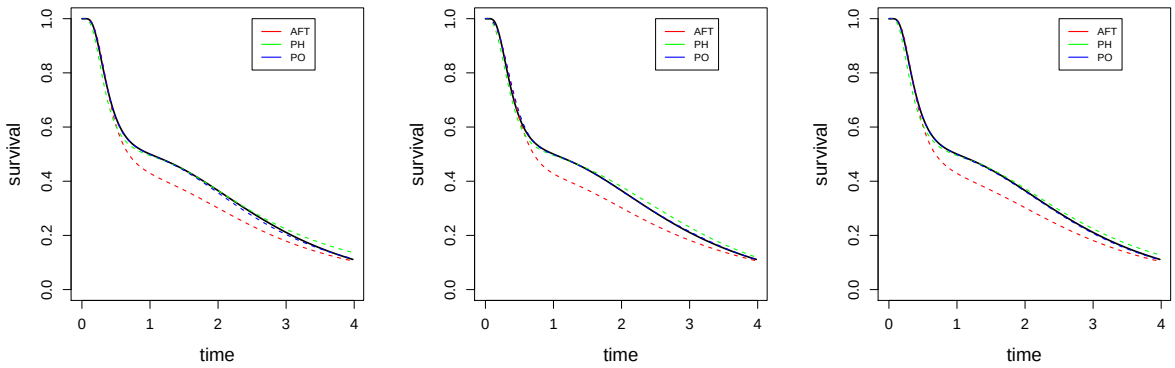
The TBP prior is centered at a parametric family of distributions. The log-logistic $S_{\boldsymbol{\theta}}(t) = \{1 + (e^{\theta_1} t)^{\exp(\theta_2)}\}^{-1}$, the log-normal $S_{\boldsymbol{\theta}}(t) = 1 - \Phi\{(\log t + \theta_1)\exp(\theta_2)\}$, and the Weibull $S_{\boldsymbol{\theta}}(t) = 1 - \exp\left\{-(e^{\theta_1} t)^{\exp(\theta_2)}\right\}$ families are implemented in the software. We next demonstrate via simulations that posterior inference and model selection is not very sensitive to the choice of centering parametric family. Table S7 presents the proportion (out of 200 MC replicates) of times each model is picked under all settings when data are generated as in the previous simulation J.5. When the true model is PH with $n = 740$, the Weibull centering distribution has a improved chance to pick the correct model than log-logistic, indicating that the Weibull slightly favors PH for this bimodal baseline S_0. As sample sizes increase, all three centering distributions give the same model selection results.

Figures S2, S3, and S4 present the averaged (across the 200 MC replicates) fitted baseline survival functions under three centering distribution families when the true model is AFT, PH and PO, respectively. Overall, the three families yield almost the same estimates regardless of what the true model is, although we do see that Weibull provides a slightly better estimate than the other two (Figure S3) when the true model is PH with the bimodal baseline S_0 and PH is used to fit the model. We also compared the inference results on the coefficient estimates (not shown), which resulted in very similar biases, coverage probabilities and effective sample sizes.



(a) (b) (c)

Figure S3: Simulation for sensitivity analysis of the TBP's centering distribution when PH is the true model. Mean, across the 200 MC replicates, of the posterior mean of the baseline survival functions under log-logistic (panel a), Weibull (panel b) and log-normal (panel c). The true curves are represented by continuous lines and the fitted curves are represented by dashed lines (red is for AFT, green is for PH and blue is for PO).



(a) (b) (c)

Figure S4: Simulation for sensitivity analysis of the TBP's centering distribution when PO is the true model. Mean, across the 200 MC replicates, of the posterior mean of the baseline survival functions under log-logistic (panel a), Weibull (panel b) and log-normal (panel c). The true curves are represented by continuous lines and the fitted curves are represented by dashed lines (red is for AFT, green is for PH and blue is for PO).

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. Chapman and Hall/CRC Press.

Bové, D. S., Held, L., et al. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall: London.

Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42(1):204–223.

Epifani, I., MacEachern, S. N., and Peruggia, M. (2008). Case-deletion importance sampling estimators: Central limit theorems and related results. *Electronic Journal of Statistics*, 2:774–806.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis*, 53(8):2873–2884.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, 74(365):153–160.

Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B*, 56(3):501–514.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420):942–951.

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

Han, C. and Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors. *Journal of the American Statistical Association*, 96(455):1122–1132.

Hanson, T., Johnson, W., and Laud, P. (2009). Semiparametric inference for survival models with step process covariates. *Canadian Journal of Statistics*, 37(1):60–79.

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

Hennerfeind, A., Brezger, A., and Fahrmeir, L. (2006). Geoadditive survival models. *Journal of the American Statistical Association*, 101(475):1065–1075.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). Criterion-based methods for Bayesian model assessment. *Statistica Sinica*, 11(2):419–443.

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26(2):131–148.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Kneib, T. (2006). Mixed model-based inference in geoadditive hazard regression for interval-censored survival times. *Computational Statistics & Data Analysis*, 51(2):777–792.

Konomi, B. A., Sang, H., and Mallick, B. K. (2014). Adaptive Bayesian nonstationary modeling for large spatial datasets using covariance approximations. *Journal of Computational and Graphical Statistics*, 23(3):802–929.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 60(1):65–81.

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212.

Nardi, A. and Schemper, M. (2003). Comparing Cox and parametric models in clinical studies. *Statistics in Medicine*, 22(23):3597–3610.

O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4(1):85–118.

Pavlou, M., Ambler, G., Seaman, S., De Iorio, M., and Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, 35(7):1159–1177.

Prentice, R. L. and Kalbfleisch, J. D. (1979). Hazard rate models with covariates. *Biometrics*, 35(1):25–39.

Sang, H. and Huang, J. Z. (2012). A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(1):111–132.

Sargent, D. J., Hodges, J. S., and Carlin, B. P. (2000). Structured Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 9(2):217–234.

Vehtari, A. and Gelman, A. (2014). *WAIC and cross-validation in Stan*. http://www.stat.columbia.edu/˜gelman/research/unpublished/waic_stan.pdf.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

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.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

Zellner, A. (1983). Applications of Bayesian analysis in econometrics. *The Statistician*, 32:23–34.

Zhao, L., Hanson, T. E., and Carlin, B. P. (2009). Mixtures of Polya trees for flexible spatial frailty survival modelling. *Biometrika*, 96(2):263–276.

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.