Bayesian Semiparametric Models for Spatially Correlated and Arbitrarily Censored Data

Haiming Zhou

Division of Statistics
Northern Illinois University

Latent Variables 2016 Conference
October 14, 2016

(Joint work with Timothy Hanson, University of South Carolina)

Outline

Outline

Spatially Correlated Survival Data

- Spatial survival data are commonly observed in diverse areas such as epidemiology, environmental health, ecology, etc.
- Data structure: $\{(t_{ij}, \mathbf{x}_{ij}, \mathbf{s}_i) : i = 1, \ldots, m; j = 1, \ldots, n_i\}$, where
  - t_{ij} is a random survival time for individual j within region/location $\mathbf{s}_i$,
  - $\mathbf{x}_{ij}$ is a related p-vector of covariates, and
  - $\{\mathbf{s}_i\}_{i=1}^m$ is a set of *distinct* regions/locations.
- Spatial survival data are often classified into two types:
  - georeferenced data, where $\mathbf{s}_i$ is recorded as longitude and latitude;
  - areal data, where $\mathbf{s}_i$ represents a geographic region, e.g. county, state.

Arbitrary Censoring

- The survival time t_{ij} is said to be *arbitrarily censored* if we only observe an interval (a_{ij}, b_{ij}) in which t_{ij} lies, where $0 \leq a_{ij} \leq b_{ij} \leq \infty$.
- The arbitrary censoring is a mixture of
  - right censoring with $b_{ij} = \infty$,
  - left censoring with $a_{ij} = 0$,
  - interval censoring with $0 < a_{ij} < b_{ij} < \infty$,
  - and noncensoring with $a_{ij} = b_{ij}$, i.e., we define $(x, x) = \{x\}$.
- The observed data is $\{(a_{ij}, b_{ij}, \mathbf{x}_{ij}, \mathbf{s}_i) : i = 1, \ldots, m; j = 1, \ldots, n_i\}$.
- The goal is to model $S_{\mathbf{x}_{ij}}(t) = P(t_{ij} > t | \mathbf{x}_{ij})$ semiparametrically in the presence of arbitrary censoring and spatial dependence.

Popular Semiparametric Models

- Three commonly used models:
  - proportional hazards (PH) model

$$S_{\mathbf{x}_{ij}}(t) = S_0(t)^{e^{\mathbf{x}'_{ij}\beta + v_i}}$$

  - accelerated failure time (AFT) model

$$S_{\mathbf{x}_{ij}}(t) = S_0(e^{\mathbf{x}'_{ij}\beta + v_i}\, t)$$

  - proportional odds (PO) model

$$\frac{S_{\mathbf{x}_{ij}}(t)}{1 - S_{\mathbf{x}_{ij}}(t)} = e^{-\mathbf{x}'_{ij}\beta - v_i} \frac{S_0(t)}{1 - S_0(t)}.$$

- Here v_i is an unobserved frailty associated with $\mathbf{s}_i$, and $S_0(t)$ is the baseline survival function corresponding to $\mathbf{x}_{ij} = \mathbf{0}$ and $v_i = 0$.
- $e^{\mathbf{x}'_{ij}\beta}$ can be interpreted as the relative risk under PH, acceleration factor under AFT, and odds factor under PO of subject $\mathbf{x}_{ij}$ relative to $\mathbf{x}_{ij} = \mathbf{0}$.

Related Literature

- Zhang and Davidian (2008, Biometrics) model the baseline $f_0(t)$ by a polynomial-based seminonparametric density estimator under all three models for arbitrarily censored data, but not for spatial data.
- Zhao, Hanson and Carlin (2009, Biometrika) consider a mixture of Polya trees prior on $f_0(t)$ under all three models for right censored areal data. The mixing is not very good under AFT.
- Pan et al. (2014, CSDA), Lin et al. (2015, LiDA) and Wang et al. (2016, Biometrics) use monotone splines to approximate the baseline hazard $H_0(t)$ under PH for interval censored data. With appropriate data argumentations, the parameter estimates can be found via simple Gibbs sampler or EM algorithm. But their method has not been extended to fit the AFT model.

Available Packages

- `BayesX` (Belitz et al. 2015) uses penalized B-splines to model log baseline hazard under the PH. It allows for arbitrary censoring and spatial frailties (for both georeferenced and areal data).
- `ICBayes` (Pan et al. 2014) can be used to fit the PH and PO for interval-censored data, but not for spatial data yet.
- `bayesSurv` (Komárek and Lessffre, 2007) fits the AFT based on finite mixtures of normal and approximating B-splines.
- However, there is no package that can fit all three models using the same treatment on the baseline function, and allowing for arbitrary censoring and spatial dependence simultaneously.

Outline

Bernstein Polynomial Prior on $S_0(t)$

- Consider a Bernstein polynomial (BP) prior (Petrone 1999),

$$b(x|J, \mathbf{w}_J) = \sum_{j=1}^{J} w_{Jj} \beta(x|j, J-j+1),$$

where $\mathbf{w}_J = (w_{J1}, \ldots, w_{JJ})' \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$ and $\beta(\cdot|a, b)$ is the density of $\text{Beta}(a, b)$.

- Under mild conditions, for any density f with support $(0, 1)$,

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

- The corresponding cumulative distribution function (cdf) is

$$B(x|J, \mathbf{w}_J) = \sum_{j=1}^{J} w_{Jj} I_x(j, J-j+1),$$

where $I_x(a, b)$ is the cdf associated with $\beta(x|a, b)$.

- Note $E[b(x|J, \mathbf{w}_J)] = 1$ and $E[B(x|J, \mathbf{w}_J)] = x$ for $x \in (0, 1)$.

Bernstein Polynomial Prior on $S_0(t)$

- Let $\{S_\theta : \theta \in \mathbf{\Theta}\}$ denote a parametric family of survival functions with support on $\mathbb{R}^+$, e.g., log-logistic, lognormal, or Weibull.
- Note $S_\theta(t)$ always lies in the interval $(0, 1)$ for $0 < t < \infty$, so for a relatively large J, $S_0(t)$ and $f_0(t)$ can be well approximated by

$$S_0(t|J, \mathbf{w}_J, \theta) = B(S_\theta(t)|J, \mathbf{w}_J),$$

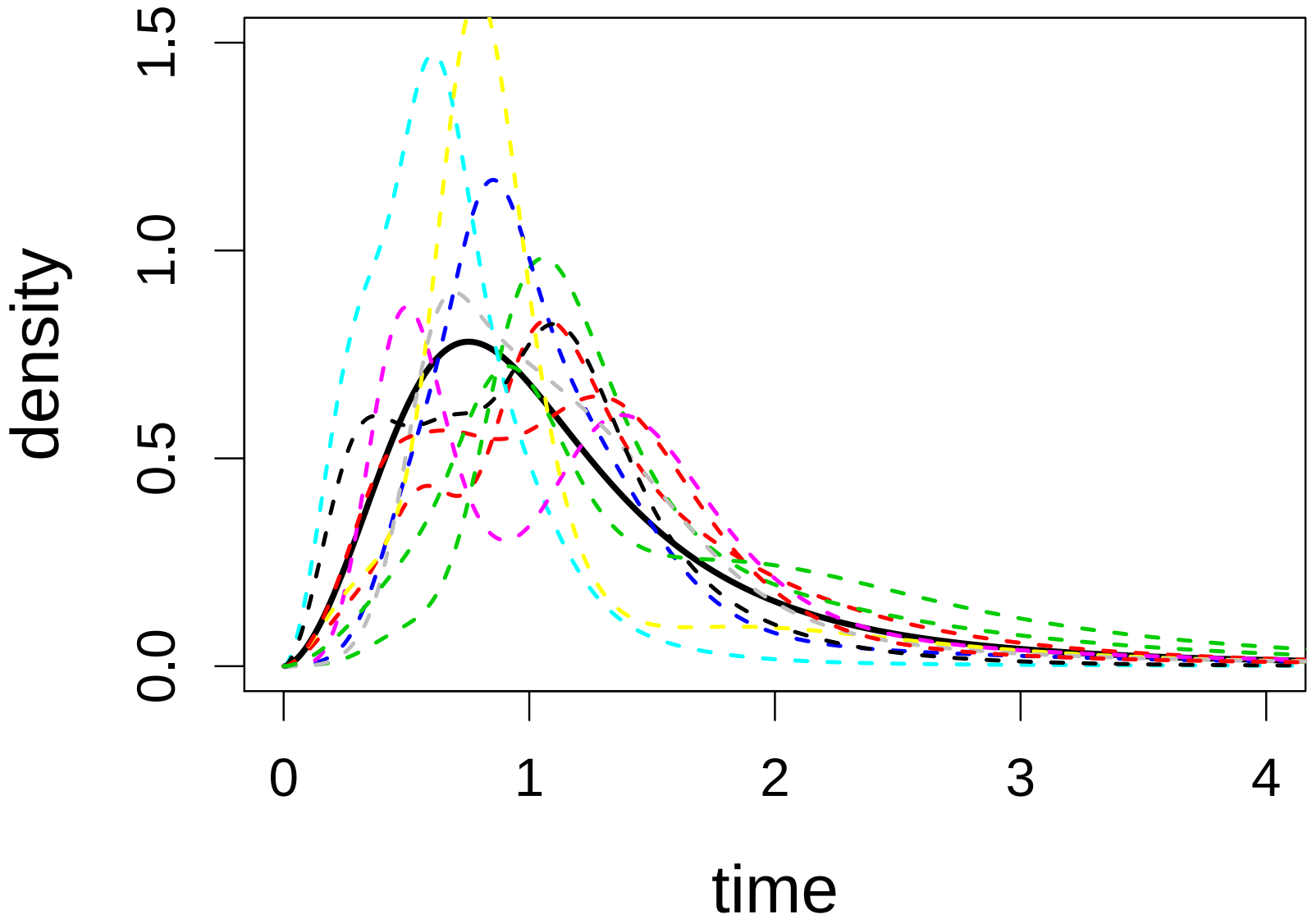
and

$$f_0(t|J, \mathbf{w}_J, \theta) = b(S_\theta(t)|J, \mathbf{w}_J) f_\theta(t),$$

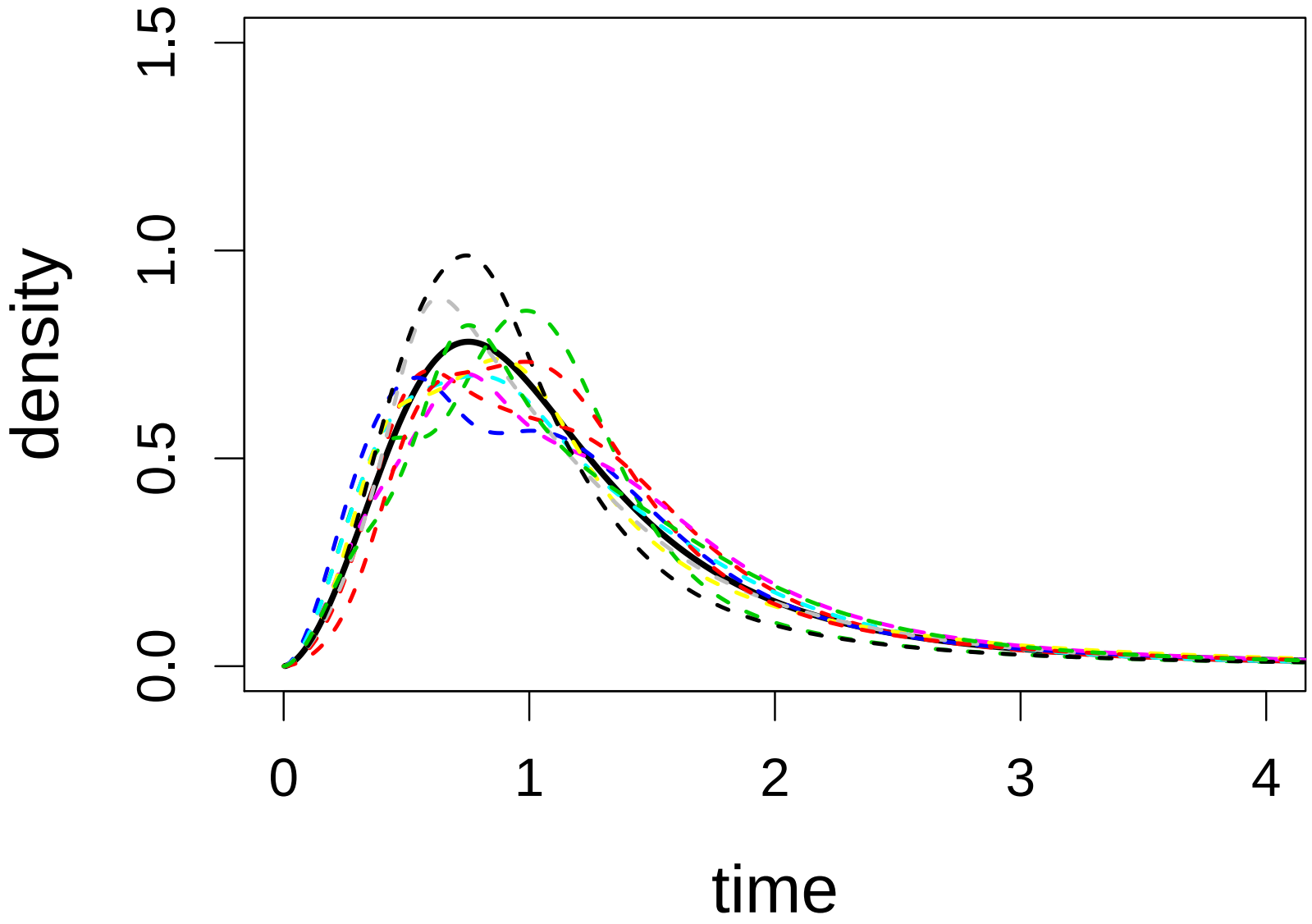
where f_θ is the density associated with S_θ.

- Clearly, $E[S_0(t|J, \mathbf{w}_J, \theta)] = S_\theta(t)$ and $E[f_0(t|J, \mathbf{w}_J, \theta)] = f_\theta(t)$.
- The weights $\mathbf{w}_J$ "adjust" the shape of S_0 relative to S_θ.

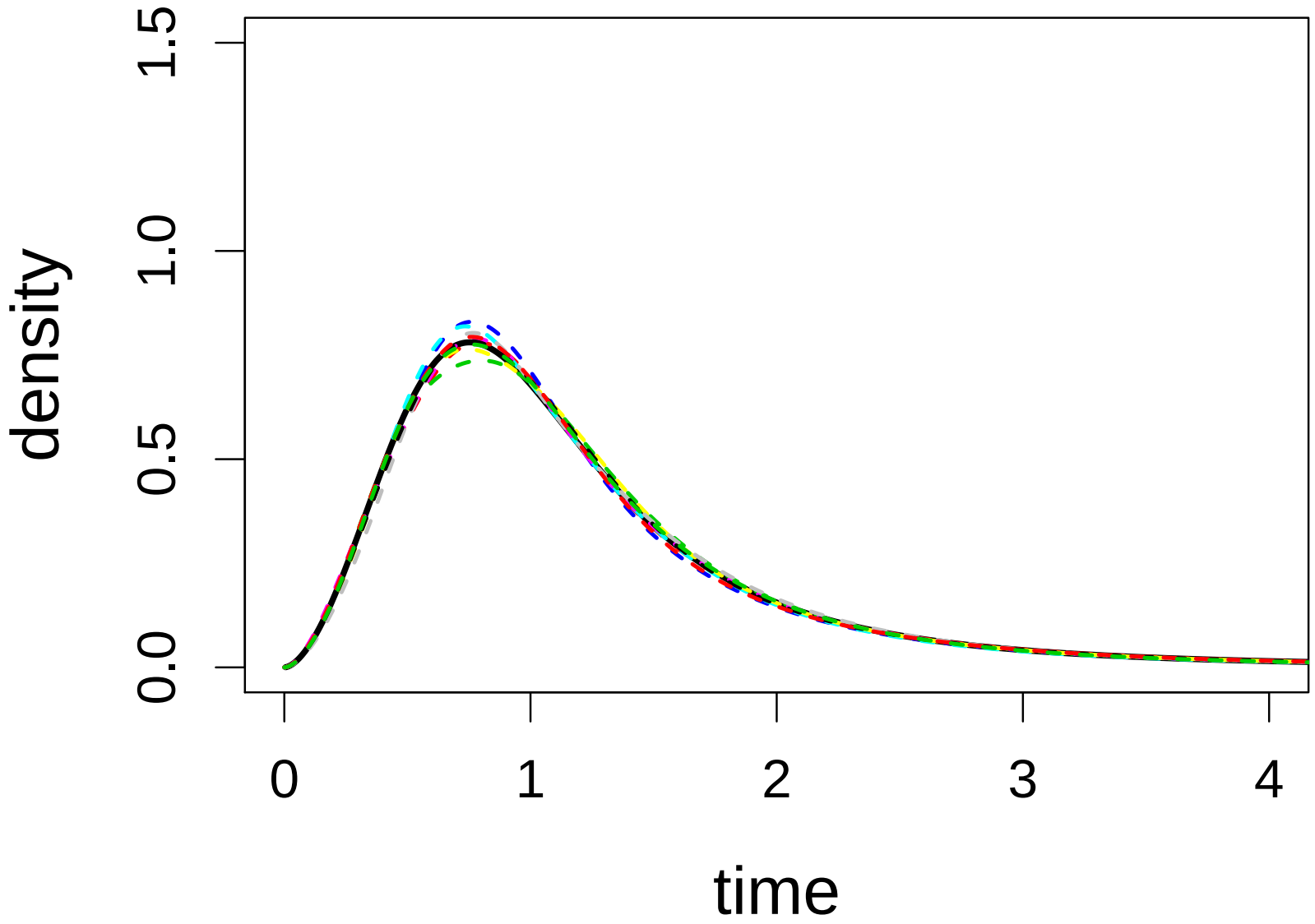
Bernstein Polynomial Prior with $J = 15$ and $\alpha = 0.5$



Bernstein Polynomial Prior with $J = 15$ and $\alpha = 5$



Bernstein Polynomial Prior with $J = 15$ and $\alpha = 100$



Prior on Frailties $\mathbf{v} = (v_1, \ldots, v_m)'$

- Areal data: conditionally autoregressive (CAR)
  - Let $e_{ij} = 1$ if i and j are adjacent and $e_{ij} = 0$ otherwise; set $e_{ii} = 0$.
  - The ICAR prior is defined through a set of conditional distributions

$$v_i | \{v_j\}_{j \neq i} \sim N\left(\sum_{\{j: j \neq i\}} e_{ij} v_j / e_{i+},\ \tau^2 / e_{i+} \right),\ i = 1, \ldots, m,$$

  where $e_{i+} = \sum_{\{j: j \neq i\}} e_{ij}$.

Prior on Frailties $\mathbf{v} = (v_1, \ldots, v_m)'$

- Areal data: conditionally autoregressive (CAR)
  - Let $e_{ij} = 1$ if i and j are adjacent and $e_{ij} = 0$ otherwise; set $e_{ii} = 0$.
  - The ICAR prior is defined through a set of conditional distributions

$$v_i | \{v_j\}_{j \neq i} \sim N\left(\sum_{\{j: j \neq i\}} e_{ij} v_j / e_{i+},\ \tau^2 / e_{i+} \right),\ i = 1, \ldots, m,$$

    where $e_{i+} = \sum_{\{j: j \neq i\}} e_{ij}$.

- Georeferenced data: Gaussian random field (GRF)
  - Assume $\mathbf{v} \sim N_m(\mathbf{0}, \tau^2 \mathbf{R})$, where $\mathbf{R}[i,j] = e^{-(\phi \|\mathbf{s}_i - \mathbf{s}_j\|)^\nu}$. Here $\phi > 0$ measures the spatial decay over distance, and $\nu \in (0, 2]$ is pre-specified.
  - The GRF prior is also a set of conditional distributions

$$v_i | \{v_j\}_{j \neq i} \sim N\left(- \sum_{\{j: j \neq i\}} p_{ij} v_j / p_{ii},\ \tau^2 / p_{ii} \right),\ i = 1, \ldots, m,$$

    where $p_{ij} = (\mathbf{R}^{-1})[i,j]$.

The Likelihood

- Observed data $\{(a_{ij}, b_{ij}, \mathbf{x}_{ij}, \mathbf{s}_i) : i = 1, \ldots, m; j = 1, \ldots, n_i\}$.
- The likelihood for $(\mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v})$ is given by

$$L(\mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} [S_{\mathbf{x}_{ij}}(a_{ij}) - S_{\mathbf{x}_{ij}}(b_{ij})]^{I\{a_{ij} < b_{ij}\}} f_{\mathbf{x}_{ij}}(a_{ij})^{I\{a_{ij} = b_{ij}\}},$$

  where $f_{\mathbf{x}_{ij}}$ is the density associated with $S_{\mathbf{x}_{ij}}$.
- The posterior density given the data $\mathcal{D}$ is

$$p(\mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}|\mathcal{D}) \propto L(\mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v}) p(\mathbf{w}|\alpha) p(\alpha) p(\boldsymbol{\theta}) p(\boldsymbol{\beta}) p(\mathbf{v}|\tau^2, \phi) p(\tau^2) p(\phi)$$

  where each $p(\cdot)$ represents a prior density, and $p(\phi)$ is needed only for georeferenced data.

Prior Specifications

- Assume $\alpha \sim \Gamma(a_0, b_0)$, $\boldsymbol{\theta} \sim N_2(\boldsymbol{\theta}_0, \mathbf{V}_0)$, $\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, \mathbf{S}_0)$, $\tau^{-2} \sim \Gamma(a_\tau, b_\tau)$, and $\phi \sim \Gamma(a_\phi, b_\phi)$.
- Note that when $\mathbf{w}_{Jj} = 1/J$ the underlying parametric model with $S_0(t) = S_\theta(t)$ is obtained and $\mathcal{L}(\mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v})$ is equal to the corresponding parametric likelihood function.
- A fit from a standard parametric survival model can provide good starting values and proposals for MCMC.
- Default hyperprior values: $a_0 = b_0 = 1$, $a_\tau = b_\tau = 0.001$, $\boldsymbol{\beta}_0 = \mathbf{0}$, $\mathbf{S}_0 = 10^{10}\mathbf{I}_p$, $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$, and $\mathbf{V}_0 = 10\hat{\mathbf{V}}$, where $\hat{\boldsymbol{\theta}}$ is the parametric point estimates of $\boldsymbol{\theta}$ and $\hat{\mathbf{V}}$ is its estimated covariance.
- For georeferenced data, set $a_\phi = 1$ and choose b_ϕ so that $\Pr(\phi > \phi_0) = 0.95$, where ϕ_0 satisfies $e^{-(\phi_0 \cdot \max \|\mathbf{s}_i - \mathbf{s}_j\|)^\nu} = 0.001$.

MCMC Overview

- Set $\mathbf{z}_{J-1} = (z_1, \ldots, z_{J-1})'$ with $z_j = \log(w_j) - \log(w_J)$.
- The β, $\boldsymbol{\theta}$, $\mathbf{z}_{J-1}$, α and ϕ are all updated using adaptive Metropolis samplers (Haario et al., 2001, Bernoulli), where the initial proposal variance is from the underlying parametric fit for β and $\boldsymbol{\theta}$, is $0.16\mathbf{I}_{J-1}$ for $\mathbf{z}_{J-1}$, and 0.16 for α and ϕ.
- The frailty term v_i is updated individually via Metropolis-Hastings, where the proposal is the conditional prior variance of $v_i|\{v_j\}_{j\neq i}$.
- The τ^{-2} is updated via Gibbs sampler from its full conditional.
- For large m, we use the full scale approximation (FSA) (Sang and Huang, 2012, JRSSB) to inverse $\mathbf{R}_{m\times m}$.

Variable Selection via Spike and Slab

- Multiply β_k by a latent indicator γ_k, where $\gamma_k = 1$ ($\gamma_k = 0$) indicates presence (absence) of covariate x_k in the model, $k = 1, \ldots, p$.
- Consider the prior

$$\beta \sim N_p(\mathbf{0}, gn(\mathbf{X}'\mathbf{X})^{-1}), \quad \gamma_k \overset{iid}{\sim} \text{Bern}(0.5),$$

  where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is the usual design matrix with mean-centred covariates, i.e. $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$.
- Hanson, Branscum and Johnson (2014) note that it is reasonable to assume $e^{\mathbf{x}_{ij}'\beta} \dot{\sim} \log N(0, gp)$ in many settings.
- The constant g is chosen so that $\Pr(e^{\mathbf{x}_{ij}'\beta} < 10) = .9$. It follows that $g = 3.228/p$.

Left-Truncation with Time-Dependent Covariates

- The survival time t_{ij} may be *left-truncated* at $u_{ij} \geq 0$, if u_{ij} is the time when subject ij is first observed.
- Given the observed left-truncated data $\{(u_{ij}, a_{ij}, b_{ij}, \mathbf{x}_{ij}, \mathbf{s}_i)\}$, the likelihood function becomes

$$L(\mathbf{w}, \boldsymbol{\theta}, \beta, \mathbf{v}) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} [S_{\mathbf{x}_{ij}}(a_{ij}) - S_{\mathbf{x}_{ij}}(b_{ij})]^{I\{a_{ij} < b_{ij}\}} f_{\mathbf{x}_{ij}}(a_{ij})^{I\{a_{ij} = b_{ij}\}} / S_{\mathbf{x}_{ij}}(u_{ij}).$$

Left-Truncation with Time-Dependent Covariates

- The survival time t_{ij} may be *left-truncated* at $u_{ij} \geq 0$, if u_{ij} is the time when subject ij is first observed.
- Given the observed left-truncated data $\{(u_{ij}, a_{ij}, b_{ij}, \mathbf{x}_{ij}, \mathbf{s}_i)\}$, the likelihood function becomes

$$L(\mathbf{w}, \theta, \beta, \mathbf{v}) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} [S_{\mathbf{x}_{ij}}(a_{ij}) - S_{\mathbf{x}_{ij}}(b_{ij})]^{I\{a_{ij} < b_{ij}\}} f_{\mathbf{x}_{ij}}(a_{ij})^{I\{a_{ij} = b_{ij}\}} / S_{\mathbf{x}_{ij}}(u_{ij}).$$

- Assume $\mathbf{x}_{ij}(t)$ is a step function:

$$\mathbf{x}_{ij}(t) = \sum_{k=1}^{o_{ij}} \mathbf{x}_{ij,k} I(t_{ij,k} \leq t < t_{ij,k+1}), \text{ where } t_{ij,1} = u_{ij}, t_{ij,o_{ij}+1} = \infty.$$

- Replace the observation $(u_{ij}, a_{ij}, b_{ij}, \mathbf{x}_{ij}(t), \mathbf{s}_i)$ by new o_{ij} observations $(t_{ij,1}, t_{ij,2}, \infty, \mathbf{x}_{ij,1}, \mathbf{s}_i), (t_{ij,2}, t_{ij,3}, \infty, \mathbf{x}_{ij,2}, \mathbf{s}_i), \ldots, (t_{ij,o_{ij}}, a_{ij}, b_{ij}, \mathbf{x}_{ij,o_{ij}}, \mathbf{s}_i)$, yielding a new left truncated data set of size $N = \sum_{i=1}^{m} \sum_{j=1}^{n_i} o_{ij}$.

Outline

Application to Childhood Mortality Data in Nigeria

- Data are from the 2003 Nigeria Demographic and Health Survey.
- The state of residence is available for each child, so the data type is areal. There are 37 states, and the sample size is $n = 4,363$.
- The survival time is *age at death* of the child. It was reported in days if it was less than one month, in months if it was less than two years and otherwise in years. If the child was still alive by the date of interview, the right censoring time can be calculated in days.
- To incorporate the inconsistency of time units, we treat all survival times recorded in months or years as interval censored (details in Appendix), yielding arbitrarily censored data.
- Kneib (2006, CSDA) fit a proportional hazards model with CAR frailties.

Application to Childhood Mortality Data in Nigeria

Continuous variables	**Mean**	**Std. Dev.**
Age at birth (yr.)	28.49	6.48
BMI	22.62	4.21
Breastfeed duration (mo.)	14.48	7.31
Preceding interval (mo.)	36.46	21.24
Categorical variables	**Level**	**Proportion (%)**
Censoring status	uncensored	1.67
	interval censored	7.54
	right censored	90.79
Place of delivery	hospital	32.78
	home/other	67.22
Gender of child	male	49.48
	female	50.52
Education	at least primary	47.26
	no education	52.74
place of residence	urban	34.82
	rural	65.18

Fit the Model using survregbayes

```
library(spBayesSurv);
### data preparation is omitted here ###
mcmc = list(nburn=50000, nsave=5000, nskip=9, ndisplay=1000);
res = survregbayes(formula=Surv(SurvLeft,SurvRight,type="interval2")~
                   AgeBirth+BMI+BreastfeedMonth+PrecedingInterval+
                   HospitalDelivery+Male+MotherEducation+Urban+
                   frailtyprior("car",State),data=d,survmodel="AFT",
                   mcmc=mcmc,Proximity=W,selection=FALSE);
summary(res);
```

- Fit PH via `survmodel="PH"` and PO via `survmodel="PO"`.
- Set `selection=TRUE` to perform the spike and slab variable selection.
- Set `frailtyprior("grf",State)` to fit Gaussian random field frailty models and `frailtyprior("iid",State)` to fit exchangeable Gaussian frailty models.
- Remove `frailtyprior()` to fit non-frailty models.

Output of the PO Model

Posterior inference of regression coefficients
(Adaptive M-H acceptance rate: 0.18116):

	Mean	Median	Std. Dev.	95%CI-Low	95%CI-Upp
AgeBirth	0.013442	0.013473	0.009282	-0.004765	0.031530
BMI	0.005937	0.005889	0.016905	-0.027046	0.038724
BreastfeedMonth	-0.378559	-0.378286	0.017017	-0.412091	-0.347309
PrecedingInterval	-0.016541	-0.016465	0.003913	-0.024405	-0.008966
HospitalDelivery	-0.553409	-0.549641	0.181878	-0.917547	-0.203444
Male	-0.081336	-0.080647	0.120485	-0.316681	0.152651
MotherEducation	-0.701258	-0.701159	0.161873	-1.014701	-0.378442
Urban	-0.362983	-0.362667	0.148649	-0.661083	-0.075890

Posterior inference of conditional CAR frailty variance

	Mean	Median	Std. Dev.	95%CI-Low	95%CI-Upp
variance	0.7904	0.7117	0.4062	0.2543	1.7858

Log pseudo marginal likelihood: LPML=-2079.558
Deviance Information Criterion: DIC=4153.352
Number of subjects: n=4363

Variable Selection

Table : Childhood mortality data. Selected models with high frequency.

Model	Proportions	Selected covariates
PH	0.402	Breastfeed, Preceding, Delivery, Education
	0.138	Breastfeed, Preceding, Delivery, Education, Residence
	0.124	Age, Breastfeed, Preceding, Delivery, Education
AFT	0.401	Breastfeed, Preceding, Delivery, Education
	0.244	Breastfeed, Preceding, Delivery, Education, Residence
	0.061	Age, Breastfeed, Preceding, Delivery, Education
PO	0.346	Breastfeed, Preceding,Delivery, Education, Residence
	0.256	Breastfeed, Preceding, Delivery, Education
	0.103	Age, Breastfeed, Preceding, Delivery, Education, Residence

Model Comparison and Results

Table : Model comparison.

Model	Covariates	LPML
PH	full	-2126
	selected	-2125
AFT	full	-2127
	selected	-2125
PO	full	-2080
	selected	-2077

Table : Covariate effects from fitting the PO model with selected covariates.

Breastfeed duration (mo.)	-0.376(-0.408, -0.347)
Preceding interval (mo.)	-0.015(-0.023, -0.008)
Delivery–hospital	-0.519(-0.876, -0.171)
Education–at least primary	-0.710(-1.024, -0.402)
Residence–urban	-0.338(-0.634, -0.047)

Survival Analysis of Loblolly Pine Trees

- Loblolly pine is the most commercially important timber species in Southeastern United States. Estimating its survival rate is a crucial task in forestry research.
- The dataset consists of 45,525 loblolly pine trees at 168 distinct sites, which were established in 1980-1981, and monitored annually until 2001-2002. The data type is georeferenced.
- During the 21-year follow-up, 5,379 trees experienced the death, and the rest which survived until the last follow-up are treated as right censored.
- It is of interest to investigate the association between the loblolly pine survival and several important risk factors after adjusting for spatial dependence among different sites.

Loblolly Pine Trees: Risk Factors

- *Time-independent* variables:
  - treatment (treat): 1–control, 2–light thinning, 3–heavy thinning
  - physiographic region (PhyReg): 1–coastal, 2-piedmont, 3-other.
- *Time-dependent* variables (repeatedly measured every 3 years):
  - total height of tree in meters (TH)
  - diameter at breast height in cm (DBH)
  - crown class (C): 1–dominant, 2–codominant, 3–intermediate, 4–suppressed.

Loblolly Pine Trees: Risk Factors

- *Time-independent* variables:
  - treatment (treat): 1–control, 2–light thinning, 3–heavy thinning
  - physiographic region (PhyReg): 1–coastal, 2-piedmont, 3-other.
- *Time-dependent* variables (repeatedly measured every 3 years):
  - total height of tree in meters (TH)
  - diameter at breast height in cm (DBH)
  - crown class (C): 1–dominant, 2–codominant, 3–intermediate, 4–suppressed.
- After incorporating the time-dependent variables, the final dataset contains $N = 180,676$ observations.
- Li et al. (2015, JASA) used a semiparametric PH model with several spatial frailty specifications to model the data. However, they showed that the PH assumption does not hold very well.

Loblolly Pine Trees: AFT, PH and PO

Table : *Model comparison.*

		PH	PO	AFT
GRF frailty	LPML	-23,991	-23,882	-23,812
IID frailty	LPML	-23,966	-23,865	-23,832
Non-frailty	LPML	-25,508	-25,549	-25,447

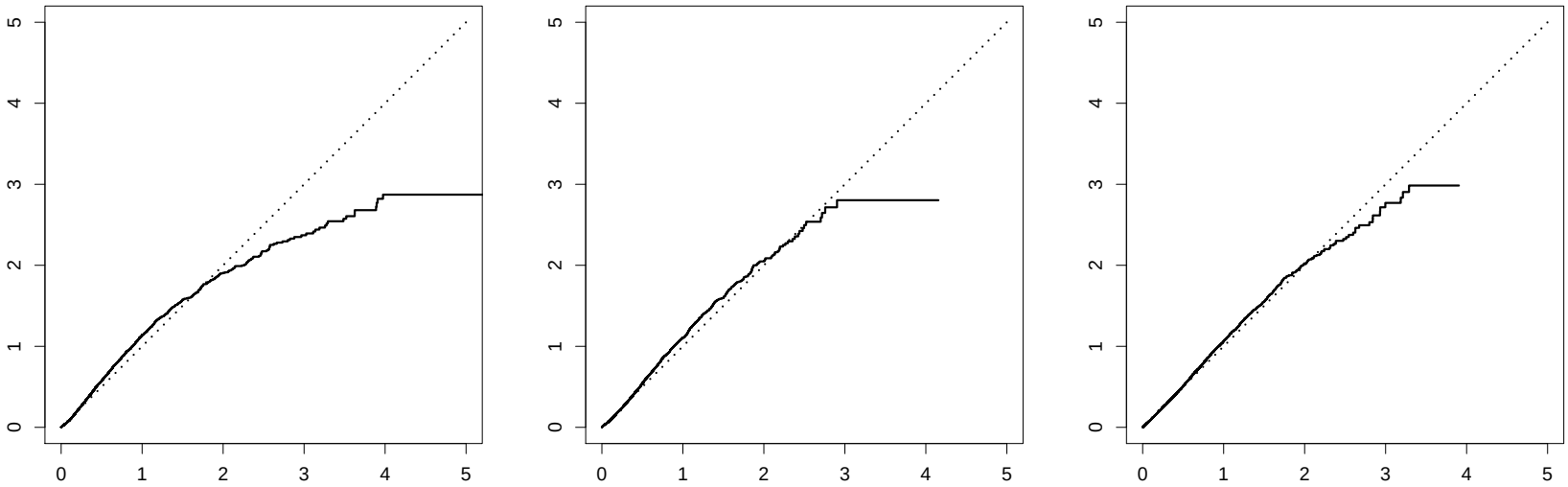
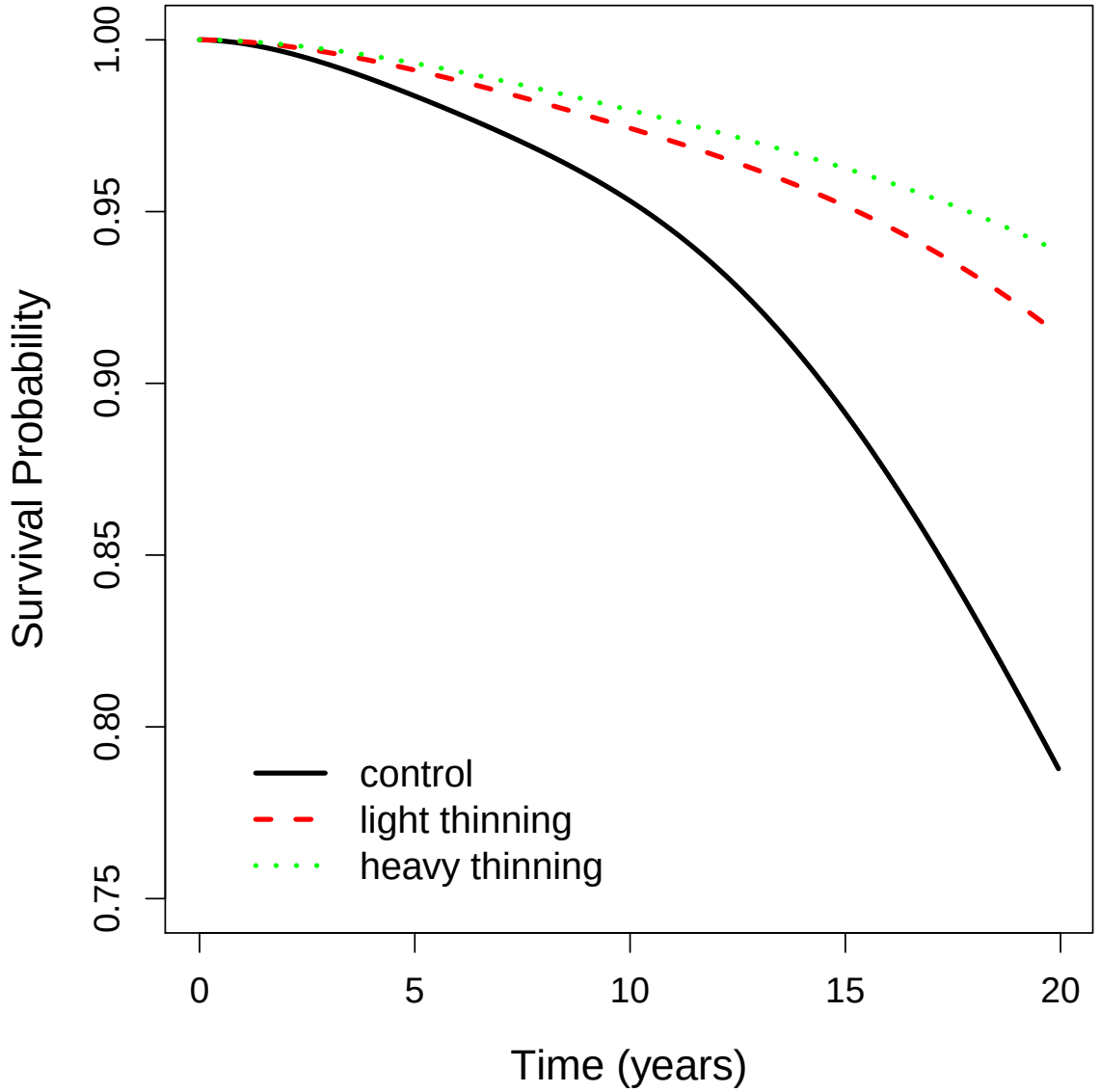


Figure : *Cox-Snell residual plot for GRF frailty PH, PO and AFT*

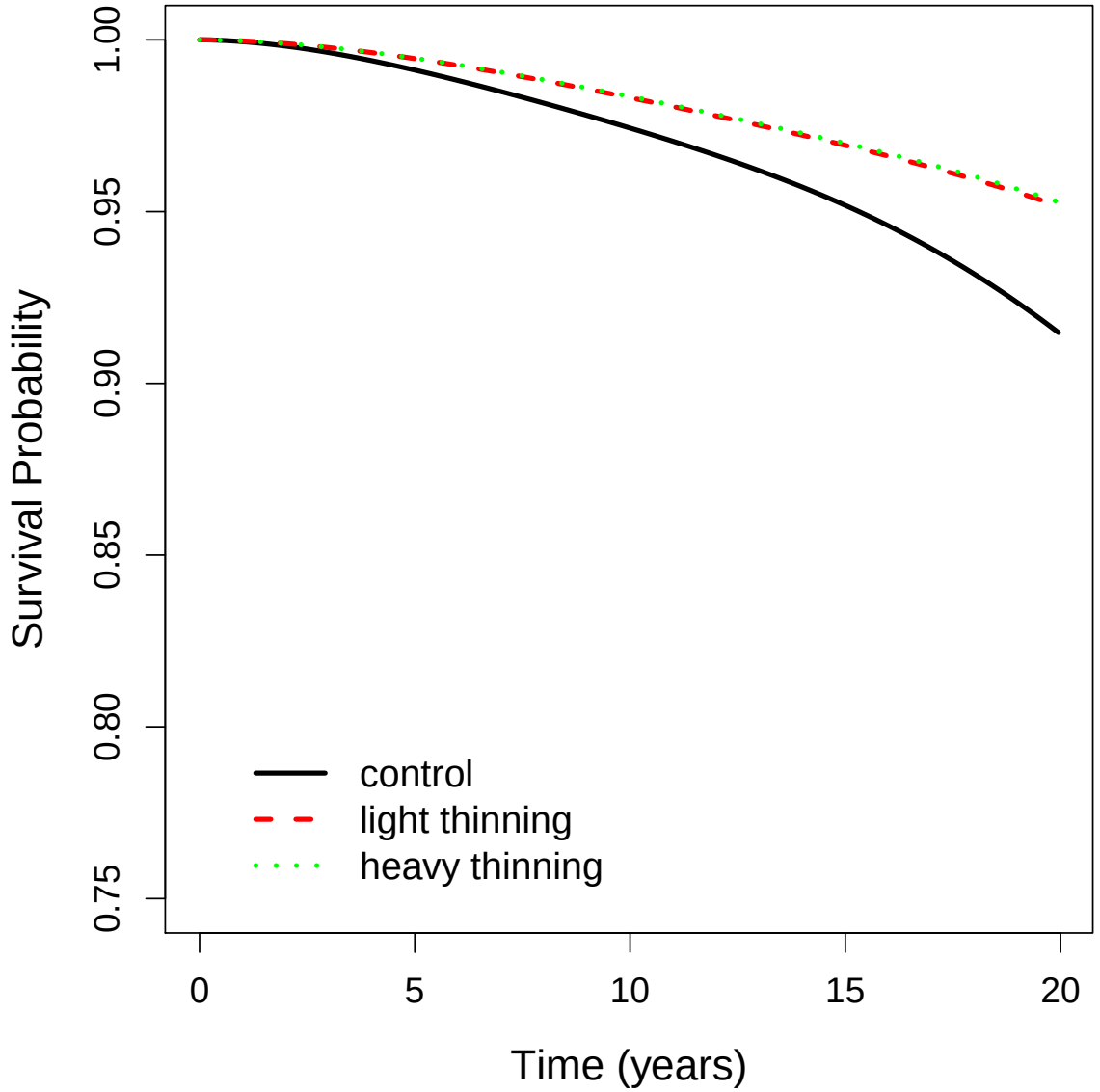
Loblolly Pine Trees: GRF-AFT

	Mean	Median	Std. Dev.	95%CI-Low	95%CI-Upp
DBH	-0.126270	-0.126519	0.008354	-0.141792	-0.109738
TH	-0.011462	-0.011488	0.001342	-0.014014	-0.008826
treat2	-0.388399	-0.387577	0.020644	-0.430511	-0.349127
treat3	-0.544378	-0.543409	0.027292	-0.601009	-0.495238
PhyReg2	-0.389881	-0.386379	0.106980	-0.593728	-0.200604
PhyReg3	-0.259512	-0.258088	0.132703	-0.510584	0.013621
C2	0.043812	0.043210	0.025837	-0.002139	0.097142
C3	0.429512	0.427719	0.031195	0.375179	0.491249
C4	1.101149	1.099480	0.046046	1.017613	1.194449
treat2:PhyReg2	0.105225	0.106106	0.031557	0.045876	0.167650
treat3:PhyReg2	0.246436	0.245954	0.042714	0.162279	0.331992
treat2:PhyReg3	-0.216354	-0.213024	0.079511	-0.367900	-0.063942
treat3:PhyReg3	0.125298	0.126770	0.084076	-0.036644	0.285920
variance	0.34961	0.34475	0.04802	0.26954	0.45747
range	0.2735	0.2643	0.0700	0.1651	0.4342

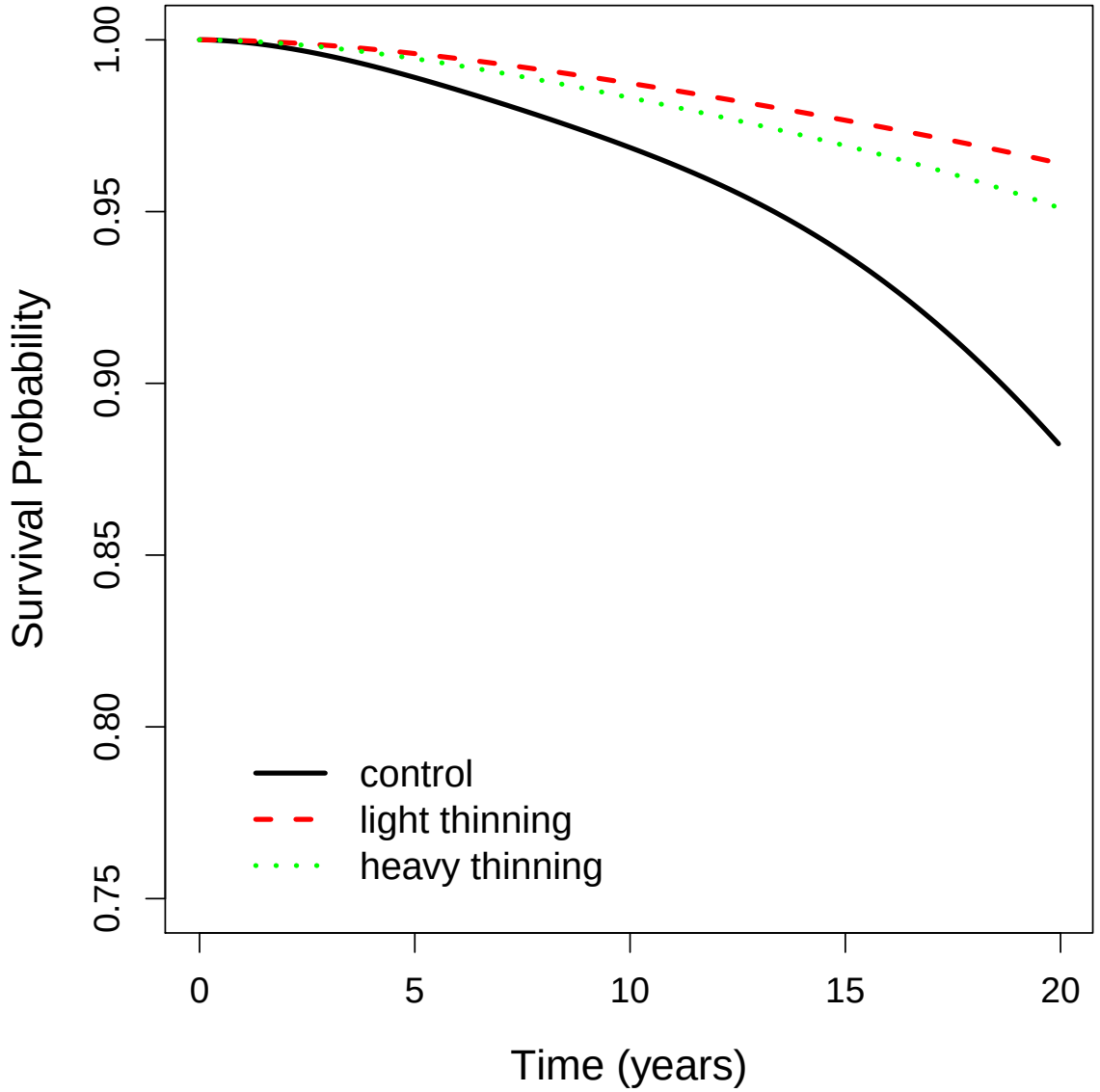
Survival Plots for Coastal Region under GRF-AFT



Survival Plots for Piedmont Region under GRF-AFT

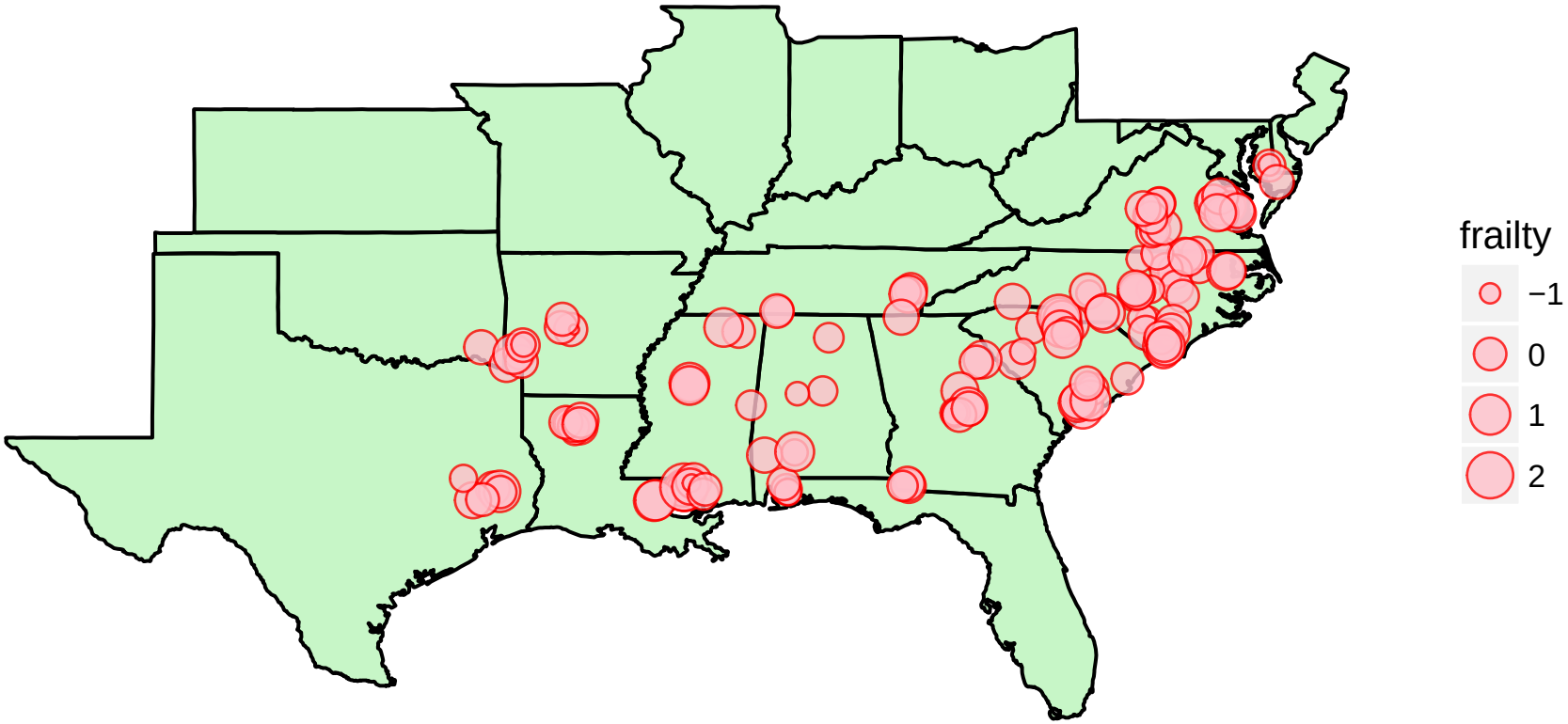


Survival Plots for Other Region under GRF-AFT



Loblolly Pine Trees: Spatial Dependence

Under the exponential correlation $\rho(\mathbf{s}_i, \mathbf{s}_j) = e^{-\phi\|\mathbf{s}_i - \mathbf{s}_j\|}$, the posterior mean is $\hat{\phi} = 0.2735$, indicating that the correlation decays by $1 - e^{-0.2735} = 24\%$ for every 1-km increase in distance.



Outline

Summary

- Studied semiparametric AFT, PH and PO frailty models for survival data subject to arbitrary censoring and spatial dependence.
- The baseline is modeled via the Bernstein polynomial prior, which is centered at a parametric family and always selects quite smooth densities, leading to more efficient posterior samplings.
- Developed a function `survregbayes` within the R package `spBayesSurv` for implementing the MCMC algorithms.
- Future work: marginal semiparametric models with spatial dependence modeled through copulas.

Determine the Endpoints of the Interval-Censoring

- Note the day of birth is not available for all the child who experienced the death before the interview.
- Suppose the age at death for child ij is recorded as 5 months. His birth date (month+year, but not day) is also available, say, May 1999. Then the date of death can be calculated, which is Oct 1999.
- The smallest possible survival time is the days between the *last day* of May 1999 and the *first day* of Oct 1999, denoted as a_{ij}.
- The largest possible survival time is the days between the *first day* of May 1999 and the *last day* of Oct 1999, denoted as b_{ij}.
- Therefore, the age at death in days for child ij lies in (a_{ij}, b_{ij}).

Simulation Settings

- Set $\mathbf{x}_{ij} = (x_{i1}, x_{i2})'$ with $x_{i1} \overset{iid}{\sim} \text{Bernoulli}(0.5)$ and $x_{i2} \overset{iid}{\sim} N(0,1)$.
- $S_0(t) = 0.5\Phi\left((\log t + 1)/0.5\right) + 0.5\Phi\left((\log t - 1)/0.5\right)$.
- v_i follows the ICAR based on the Nigeria data.
- Sample size is $n = 740$.
- The final data set yields around 20% right-censored, 40% uncensored, 25% left-censored and 15% interval-censored.

Simulation Results

Table : Simulation results based on 500 replicates for areal-referenced data. For each MCMC algorithm, 10,000 scans were thinned from 50,000 after a burn-in period of 50,000 iterations.

Model	Parameter	BIAS	PSD	SD-Est	CP	Effective size
AFT	$\beta_1 = 1$	0.003	0.066	0.062	0.960	3541
	$\beta_2 = 1$	0.001	0.035	0.034	0.954	3220
	$\tau^2 = 1$	0.010	0.296	0.245	0.968	7563
PH	$\beta_1 = 1$	-0.008	0.100	0.101	0.956	3299
	$\beta_2 = 1$	-0.004	0.061	0.060	0.952	2025
	$\tau^2 = 1$	-0.001	0.341	0.321	0.958	4862
PO	$\beta_1 = 1$	0.005	0.151	0.153	0.946	4397
	$\beta_2 = 1$	0.007	0.083	0.080	0.956	2925
	$\tau^2 = 1$	0.038	0.438	0.371	0.972	3497

Compare with ICBayes

Table : Simulation results based on 500 replicates of size $n = 500$ under the non-frailty PH model for pure interval-censored data. For each MCMC, we retain 10, 000 scans without thinning after a burn-in period of 10, 000 iterations.

Method	Time	Parameter	BIAS	PSD	SD-Est	CP	Effective size
spBayesSurv	4.14	$\beta_1 = 1$	0.032	0.136	0.132	0.962	1028
		$\beta_2 = 1$	0.020	0.088	0.089	0.940	745
ICBayes	8.81	$\beta_1 = 1$	-0.022	0.133	0.124	0.958	342
		$\beta_2 = 1$	-0.018	0.084	0.081	0.954	291

Variable Selection Results

- Example 1: $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ij5})$ with $x_{ij1} \sim \text{Bernoulli}(0.5)$ and $x_{ij2}, \ldots, x_{ij5} \overset{iid}{\sim} N(0,1)$, and $\boldsymbol{\beta} = (1,1,0,0,0)'$.
- Example 2: $x_{ij3} = x_{ij2} + 0.15z$ where $z \sim N(0,1)$, yielding a 0.989 correlation between x_2 and x_3.
- Example 3: $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ij10})$ with $\boldsymbol{\beta} = (1,1,1,1,1,0,0,0,0,0)$ and $x_{ijk}|z \overset{iid}{\sim} N(z,1)$ where $z \sim N(0,1)$, which induces pairwise correlations of about 0.5.

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