A unified framework for fitting Bayesian semiparametric models to arbitrarily censored spatial survival data

Tim Hanson

Department of Statistics
University of South Carolina

hansont@stat.sc.edu

Statistics Graduate Interdisciplinary Program
University of Arizona at Tucson December 7, 2016

Joint work with Haiming Zhou, University of Northern Illinois

Outline

Outline

Spatially correlated survival data

- Spatial survival data commonly seen in 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 typically classified into two types:
  - georeferenced data, where $\mathbf{s}_i \in \mathbb{R}^2$ is recorded as longitude and latitude;
  - areal data, where $\mathbf{s}_i \in \{1, \ldots, m\}$ represents a geographic region, e.g. county, state.

Arbitrary censoring

- 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 $0 \leq a_{ij} \leq b_{ij} \leq \infty$.
- Arbitrary censoring is 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 $a_{ij} = b_{ij}$; define $(x, x) = \{x\}$.
- The observed data are $\{(a_{ij}, b_{ij}, \mathbf{x}_{ij}, \mathbf{s}_i) : i = 1, \ldots, m; j = 1, \ldots, n_i\}$.
- Goal: 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

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

- v_i is unobserved "frailty" associated with $\mathbf{s}_i$; $S_0(t)$ is baseline survival function corresponding to $\mathbf{x}_{ij} = \mathbf{0}$ and $v_i = 0$.
- $e^{\mathbf{x}_{ij}'\beta}$ interpreted as relative risk under PH, acceleration factor under AFT, or relative odds of surviving past any time t under PO for those w/ $\mathbf{x}_{ij}$ relative to $\mathbf{x}_{ij} = \mathbf{0}$.

15 years of spatial survival modeling...

- Human health: data on leukemia survival (Henderson et al., 2002), infant/childhood mortality (Banerjee et al., 2003; Kneib, 2006), coronary artery bypass grafting (Hennerfeind et al., 2006), asthma (Li and Ryan, 2002; Li and Lin, 2006), breast cancer (Zhao and Hanson, 2011; Hanson et al., 2012; Zhou et al., 2015), mortality due to air pollution (Jerrett et al., 2013), colorectal cancer survival (Liu et al., 2014), smoking cessation (Pan et al., 2014), HIV/AIDS patients (Martins et al., 2016), time to tooth loss (Schnell et al., 2015).
- Other: political event processes (Darmofal, 2009), gourd mildew outbreaks (Ojiambo and Kang, 2013), forest fires (Morin, 2014), pine trees (Li et al., 2015 JASA), health and pharmaceutical firms (Arbia et al., 2016), emergency service response times (Taylor, 2016).
- All twenty of these use proportional hazards; other semiparametric models not considered or compared to.

Alternative models do exist...

- e.g. Diva et al. (2008), Zhao et al. (2009), Wang et al. (2012), Li et al. (2015 Bcs), a few others.
- These only consider areal (e.g. county-level) data; all right-censored; time-dependent covariates not considered nor is variable selection; diagnostics limited.
- Our goal is to provide *broadly comprehensive* approach to modeling spatial survival data semiparametrically, including AFT and PO as well as PH. Bring together many ideas in literature and provide easy-to-use R package.

More 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), etc. use monotone splines to approximate the baseline hazard $H_0(t)$ under PH for interval censored data. With clever data augmentation, inference obtained via simple Gibbs sampler or EM algorithm. But their method has not been extended to fit the AFT model, georeferenced data, etc. Also *requires* each survival time interval censored – cannot handle times that are actually observed.

Some available R 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). Also `R2BayesX`. No interval censored 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. Frailties, but not spatial.
- However, there is no approach/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

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

$$b(x) = \sum_{j=1}^{J} w_j \beta(x|j, J-j+1),$$

where $\mathbf{w} = (w_1, \ldots, w_J)' \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$ and $\beta(\cdot|a, b)$ is the density of $\text{Beta}(a, b)$.

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

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

- Corresponding CDF is

$$B(x) = \sum_{j=1}^{J} w_j I_x(j, J-j+1),$$

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

- Note $E\{b(x)\} = 1$ for $x \in (0, 1)$.

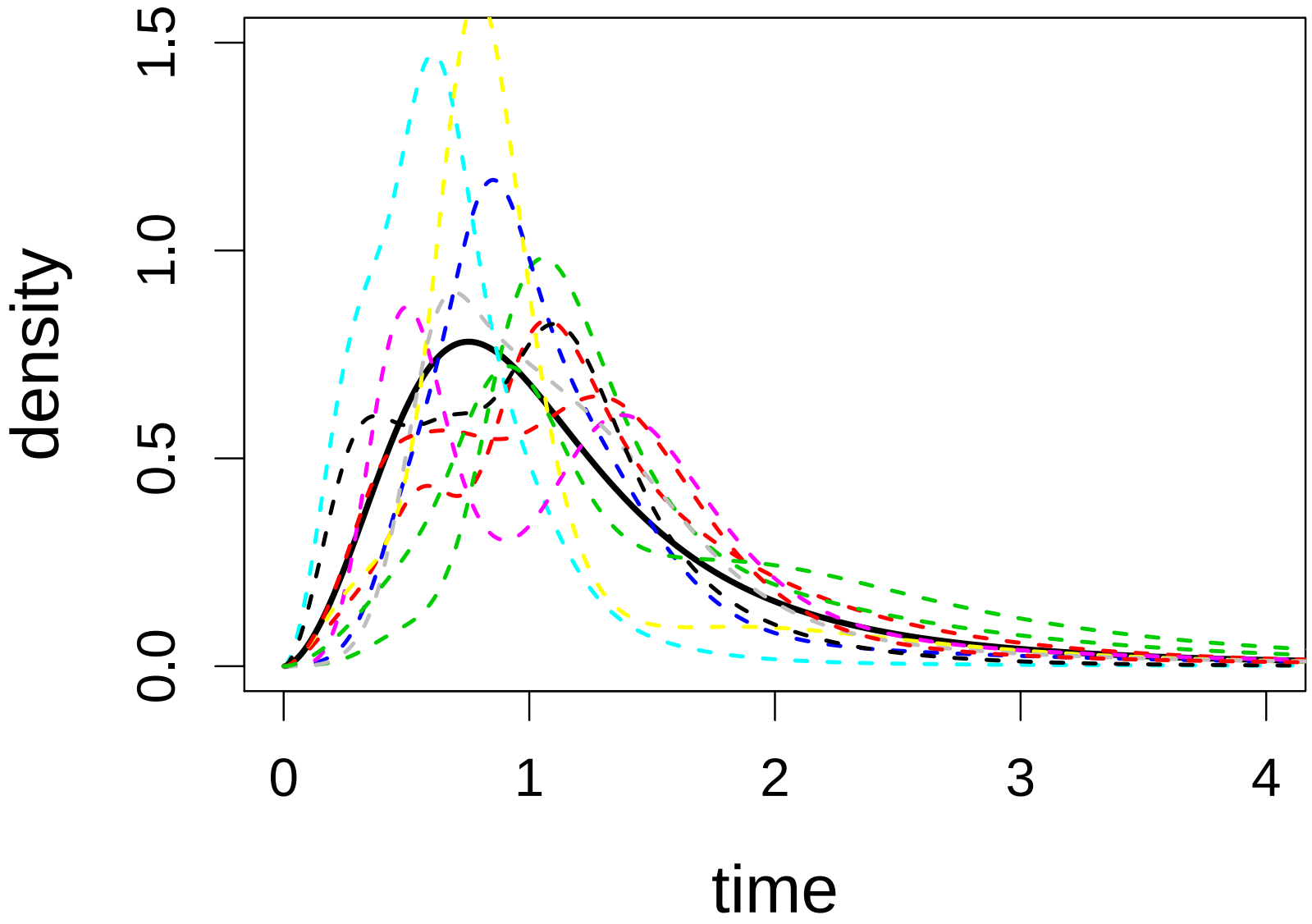
Transformed Bernstein Polynomial Prior (TBPP) on $S_0(t)$

- Let $\{S_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ denote parametric family of survival functions with support on $\mathbb{R}^+$; e.g. log-logistic, lognormal, or Weibull.
- Note $S_{\boldsymbol{\theta}}(t)$ 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

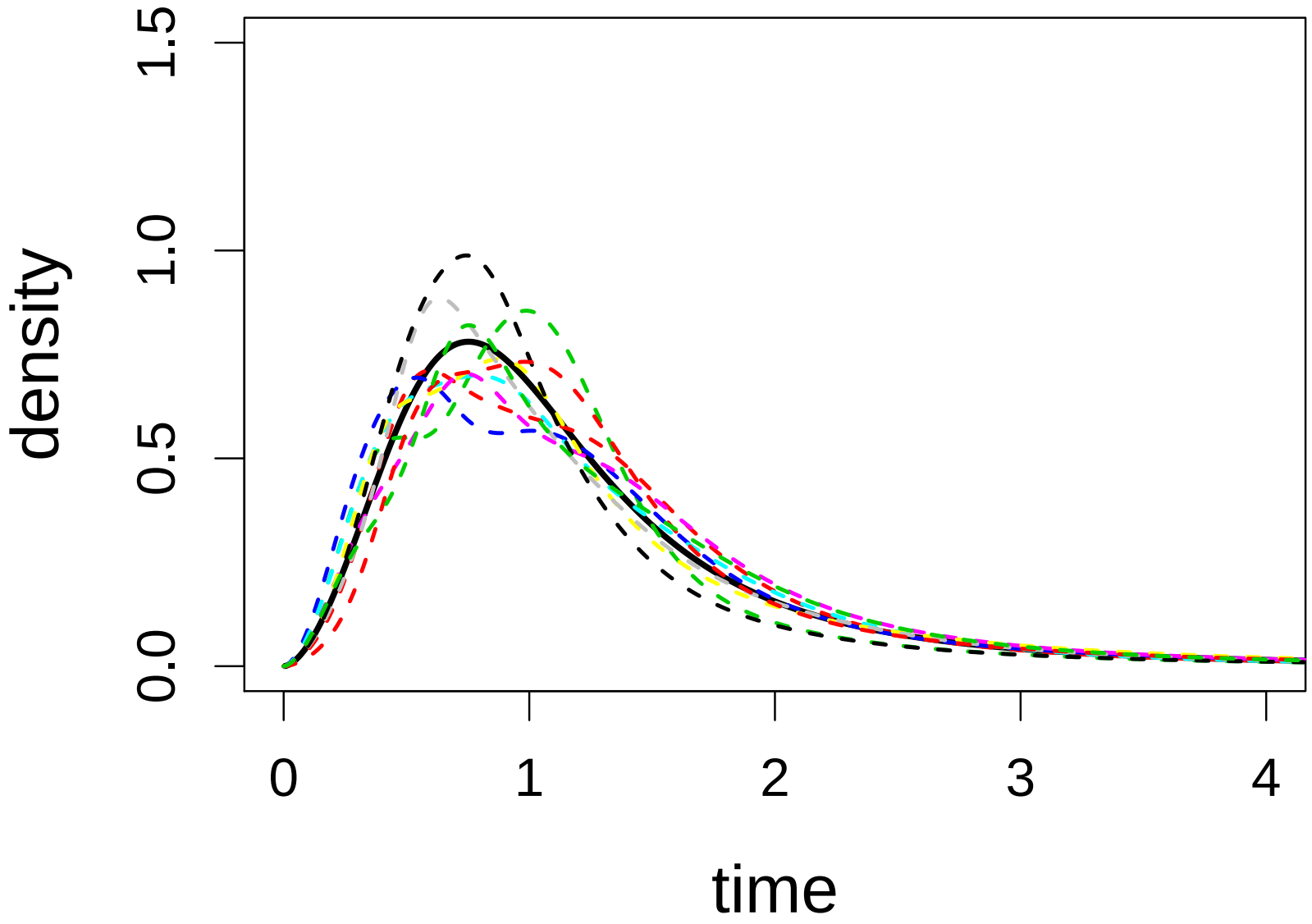
$$S_0(t) = B(S_{\boldsymbol{\theta}}(t)), \quad f_0(t) = b(S_{\boldsymbol{\theta}}(t)) f_{\boldsymbol{\theta}}(t)$$

  where $f_{\boldsymbol{\theta}}$ is density associated with $S_{\boldsymbol{\theta}}$; Chen et al. (2014).
- Then $E\{S_0(t)\} = S_{\boldsymbol{\theta}}(t)$ and $E\{f_0(t)\} = f_{\boldsymbol{\theta}}(t)$.
- The weights $\mathbf{w}$ "adjust" the shape of S_0 relative to $S_{\boldsymbol{\theta}}$. Increasing J gives greater flexibility.

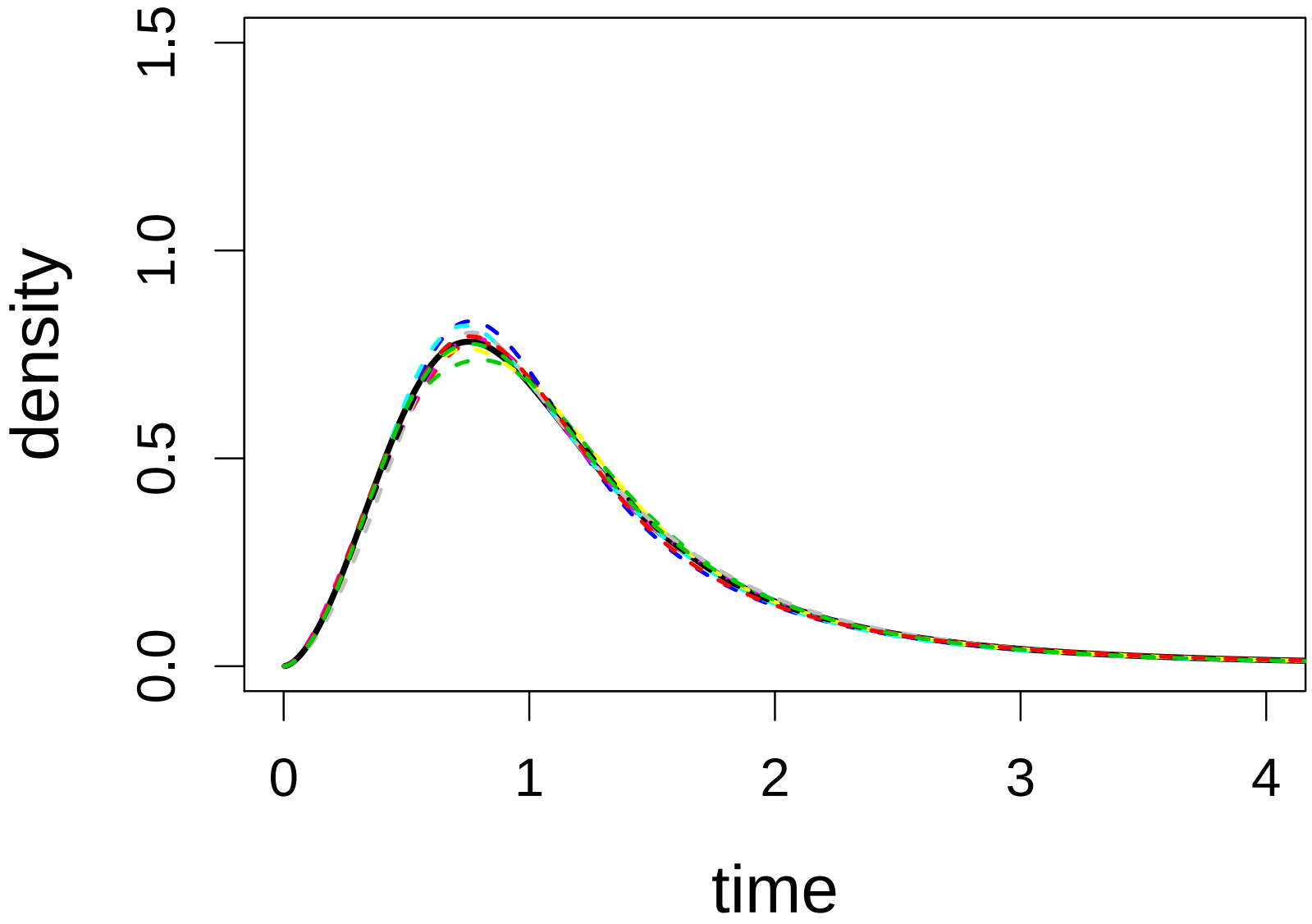
TBPP with $J = 15$ and $\alpha = 0.5$



TBPP with $J = 15$ and $\alpha = 5$



TBPP with $J = 15$ and $\alpha = 100$



Model for frailties $\mathbf{v} = (v_1, \ldots, v_m)'$

- Areal data: intrinsic conditionally autoregressive (CAR)
  - Let $e_{ij} = 1$ if i and j are adjacent and $e_{ij} = 0$ otherwise; set $e_{ii} = 0$.
  - The CAR 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 $p_{ij} = (\mathbf{R}^{-1})[i,j]$.

Likelihood & posterior

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

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

  $p(\phi)$ needed only for georeferenced data.

Prior specification

- 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 $\phi \sim \Gamma(a_\phi, b_\phi)$.
- When $\mathbf{w}_j = 1/J$ the underlying parametric model $S_0(t) = S_\theta(t)$ is obtained and $\mathcal{L}(\mathbf{w}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{v})$ is same as parametric likelihood function.
- Fit from 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 parametric MLE of $\boldsymbol{\theta}$ and $\hat{\mathbf{V}}$ is 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 $e^{-(\phi_0 \max \|\mathbf{s}_i - \mathbf{s}_j\|)^\nu} = 0.001$.

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 all block-adaptive Metropolis samplers (Haario et al., 2001); initial proposal covariance from underlying parametric fit $\hat{\mathbf{V}}$ & $\hat{\mathbf{V}}_{\boldsymbol{\theta}}$ for β & $\boldsymbol{\theta}$; $0.16\mathbf{I}_{J-1}$ for $\mathbf{z}_{J-1}$; and 0.16 for α and ϕ.
- Frailty v_i updated individually via Metropolis-Hastings; proposal uses conditional variance of $v_i | \{v_j\}_{j \neq i}$.
- τ^{-2} updated from full conditional.
- For large m, full scale approximation (FSA) (Sang and Huang, 2012, JRSSB) used to invert $\mathbf{R}_{m \times m}$.

spBayesSurv compared to ICBayes

500 replicates size $n = 500$ under non-frailty PH model for pure interval-censored data; 10,000 MCMC scans kept after burn-in of 10,000 iterations.

Method	Time	Parameter	BIAS	PSD	SD-Est	CP	Effective size
survregbayes	63	$\beta_1 = 1$	-0.018	0.134	0.134	0.940	1139
		$\beta_2 = 1$	-0.015	0.086	0.087	0.940	934
ICBayes	310	$\beta_1 = 1$	-0.036	0.133	0.132	0.938	346
		$\beta_2 = 1$	-0.019	0.084	0.085	0.938	292

Authors of ICBayes claim their method is efficient and "...*does not require imputing any unobserved failure times or contain any complicated Metropolis- Hastings steps...*" In fact, their approach augments every interval censored time with as many latent variables as there are spline basis functions, e.g. nJ additional parameters. Their approach cannot be used with uncensored data, nor can it be generalized to AFT.

Spike and slab variable selection (Kuo and Mallick, 1998)

- Multiply β_k by a latent γ_k; $\gamma_k = 0/1$ indicates absence/presence of x_k in model, $k = 1, \ldots, p$.
- Prior is

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

  where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ is the usual design matrix with mean-centered covariates, i.e. $\mathbf{1}_n'\mathbf{X} = \mathbf{0}_p'$.
- Hanson, Branscum and Johnson (2014) note that $e^{\mathbf{x}_{ij}'\boldsymbol{\beta}} \overset{\bullet}{\sim} \log N(0, gp)$ in many settings.
- Constant g is chosen so that $\Pr(e^{\mathbf{x}_{ij}'\boldsymbol{\beta}} < 10) = 0.9$: $g = 3.228/p$.

Left-truncation

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

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

Time-dependent covariates

- With left-truncation AFT, PH and PO models can be extended to time-dependent covariates (Hanson et al., 2009, CJS).
- 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 $N = \sum_{i=1}^{m} \sum_{j=1}^{n_i} o_{ij}$.

Cox-Snell (1968) variable plots

- $r(t_{ij}) = -\log\{S_{\mathbf{x}_{ij}}(t_{ij})|\mathcal{D}\}$, depends on posterior $[\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}, v_i|\mathcal{D}]$.
- Given $S_{\mathbf{x}_{ij}}(\cdot)$, $-\log S_{\mathbf{x}_{ij}}(t_{ij})$ has standard exponential distribution.
- If model is "correct" pairs $\{r(a_{ij}), r(b_{ij})\}$ are approximately random arbitrarily censored sample from exp(1).
- Estimated integrated hazard plot (using Turnbull, 1974) should be approximately straight with slope 1.
- Uncertainty visualized by plotting several from $[\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}, v_i|\mathcal{D}]$.
- Problem: AFT model typically "fits" regardless, e.g. (Baltazar-Aban and Peña 1995).

LPML and DIC model selection criteria

- DIC is Bayesian version of AIC. Ω all model parameters and $f(\mathcal{D}|\Omega)$ is likelihood function based on observed data $\mathcal{D}$.

$$\text{DIC} = E_{\Omega|\mathcal{D}}[D(\Omega)] + p_D$$

  where $D(\Omega) = -2\log f(\mathcal{D}|\Omega)$ and $p_D = E_{\Omega|\mathcal{D}}[D(\Omega)] - D(E_{\Omega|\mathcal{D}}[\Omega])$.
- The conditional predictive ordinate (Geisser and Eddy, 1979) for observation ij is

$$\text{CPO}_{ij} = f(\mathcal{D}_{ij}|\mathcal{D}_{-ij}),$$

  where $\mathcal{D}_{-ij} = \{(\mathbf{x}_{st}, a_{st}, b_{st}) : (s,t) \neq (i,j)\}$.
- $\text{LPML} = \log \prod_{i=1}^{m} \prod_{j=1}^{n_i} \text{CPO}_{ij}$.
- Over 100's of data analyses DIC & LPML typically pick same model. Differences occur in richly parameterized models and random effects models (we have both).
- DIC requires thought about what goes into Ω, e.g. including frailties $(v_1, \ldots, v_m)'$ is only easy way to compute DIC. LPML does not require such thought; purely a cross-validated predictive measure.

Partially linear (additive) models

- Additive PH first considered by Gray (1992 JASA) as

$$h_{\mathbf{x}_{ij}}(t) = h_0(t) \exp\left\{ \mathbf{x}_{ij}'\beta + \sum_{\ell=1}^{p} b_\ell(x_{ij\ell}) \right\}.$$

- $b_1(\cdot), \ldots, b_p(\cdot)$ penalized B-splines w/ linear portion removed.
- Setting some $b_\ell(\cdot) \equiv 0$ gives the "partially linear PH model."
- Spatial versions for PH (Kneib, 2006; Hennerfeind et al., 2006) can be fit in `R2BayesX`.

Partially linear (additive) models

- Want additive PH, PO, and AFT models for arbitrarily censored spatial data.
- Take

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

where $\{B_{\ell k}(\cdot) : k = 0, \ldots, K+1\}$ are cubic B-spline basis functions.

- Priors for β and $\boldsymbol{\xi}_\ell = (\xi_{\ell 1}, \ldots, \xi_{\ell K})$ are

$$\beta \sim N(\mathbf{0}, \mathbf{S}_0), \quad \boldsymbol{\xi}_\ell \sim N(\mathbf{0}, gn(\mathbf{X}_\ell' \mathbf{X}_\ell)^{-1}), \quad \ell = 1, \ldots, p$$

where $\mathbf{S}_0 = 10^{10}\mathbf{I}_p$, $\mathbf{X}_\ell$ is design for the $b_\ell(\cdot)$ term, and $g = \left[\log 10 / \Phi^{-1}(0.9)\right]^2 / K$.

Test linearity of $x_{ij\ell}$

- Formally $H_0 : \boldsymbol{\xi}_\ell = \mathbf{0}$ vs. $H_1 : \boldsymbol{\xi}_\ell \neq \mathbf{0}$.
- Let BF_{10} be Bayes factor between H_1 and H_0. BF_{10} estimated large-sample approximation to the Savage-Dickey density ratio (Verdinelli and Wasserman, 1995):

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

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

Outline

Application to Nigerian childhood mortality data

- 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 paper), yielding arbitrarily censored data.
- Kneib (2006, CSDA) fit a PH model with CAR frailties.

Application to Nigerian childhood mortality data

Continuous variables	**Mean**	**Std. Dev.**
Age at birth (yr.)	28.49	6.48
BMI	22.62	4.21
Breastfeeding 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

Model fit 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

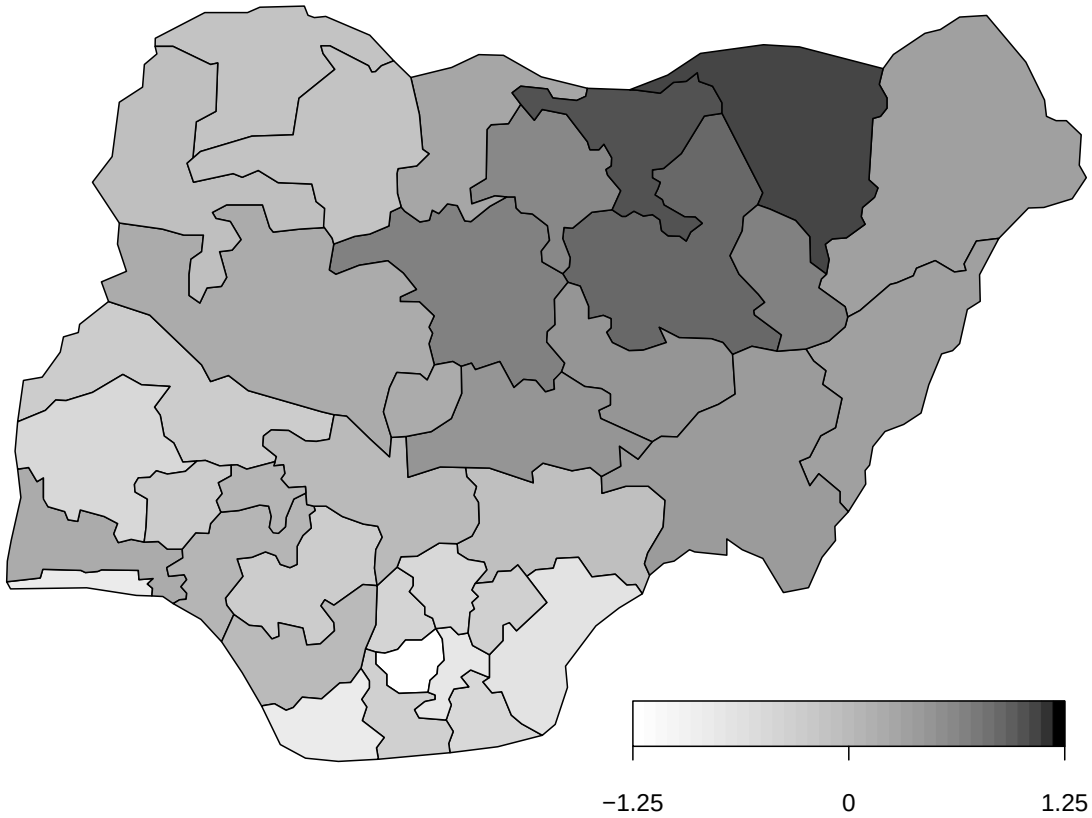
Table: Model comparison.

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.

Breastfeeding 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)

Childhood mortality: posterior mean frailties PO CAR frailty



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 (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 but noted there are no alternatives.

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

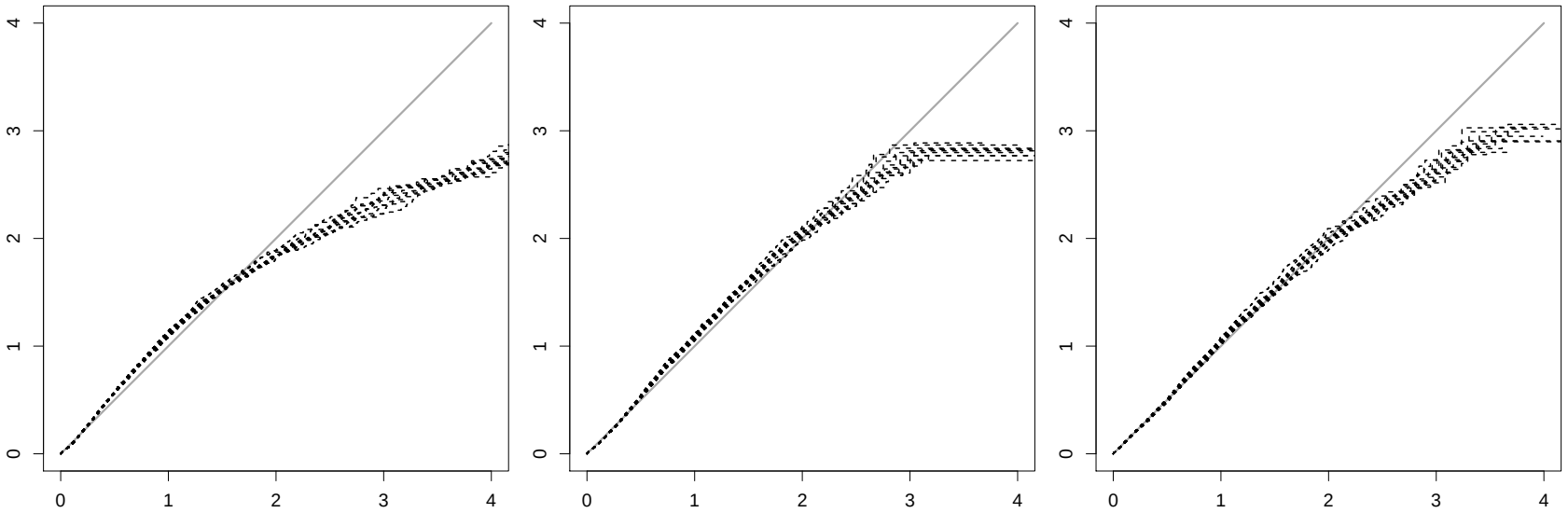
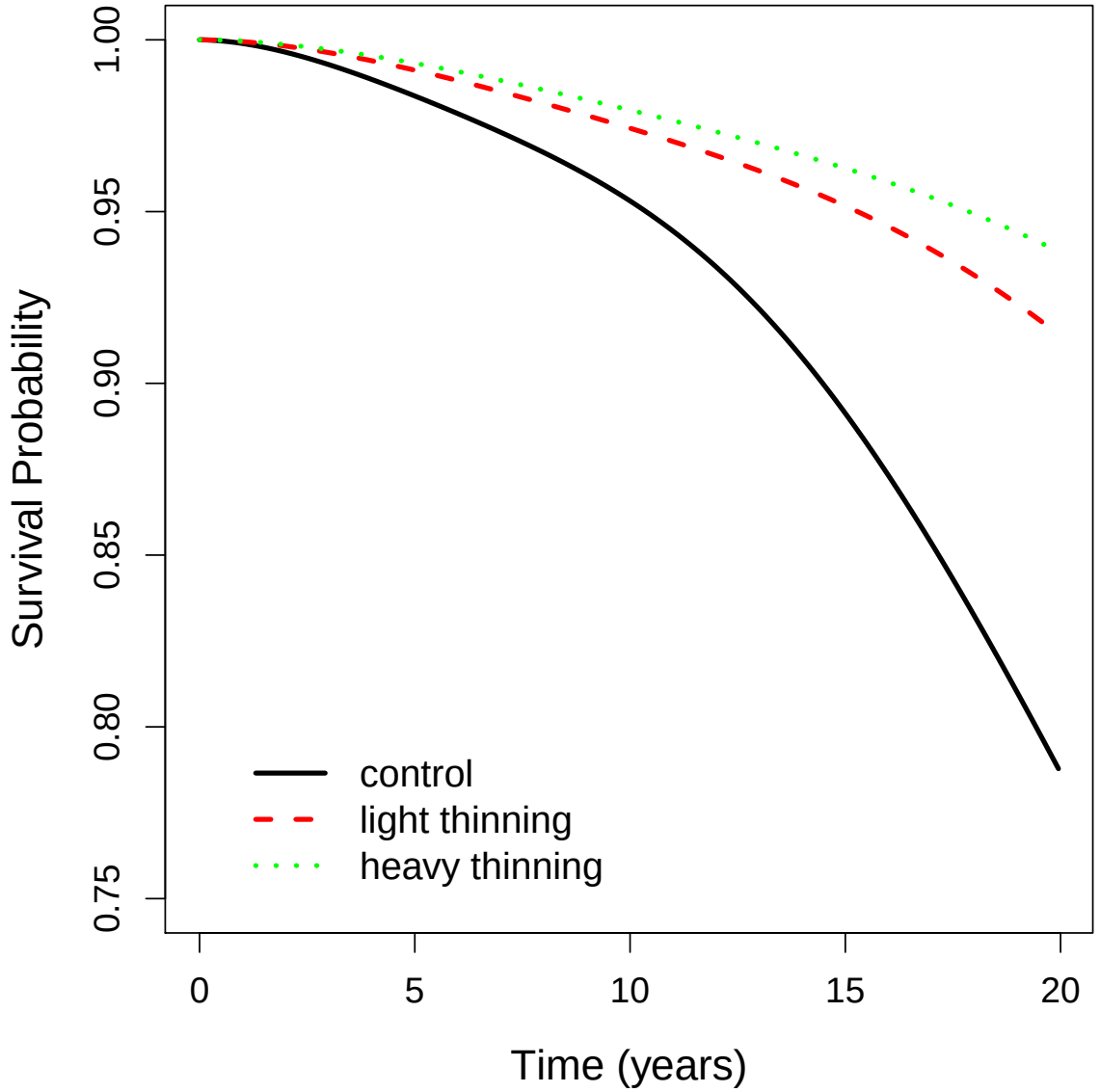


Figure: *Cox-Snell residual plot for GRF frailty PH, PO and AFT*

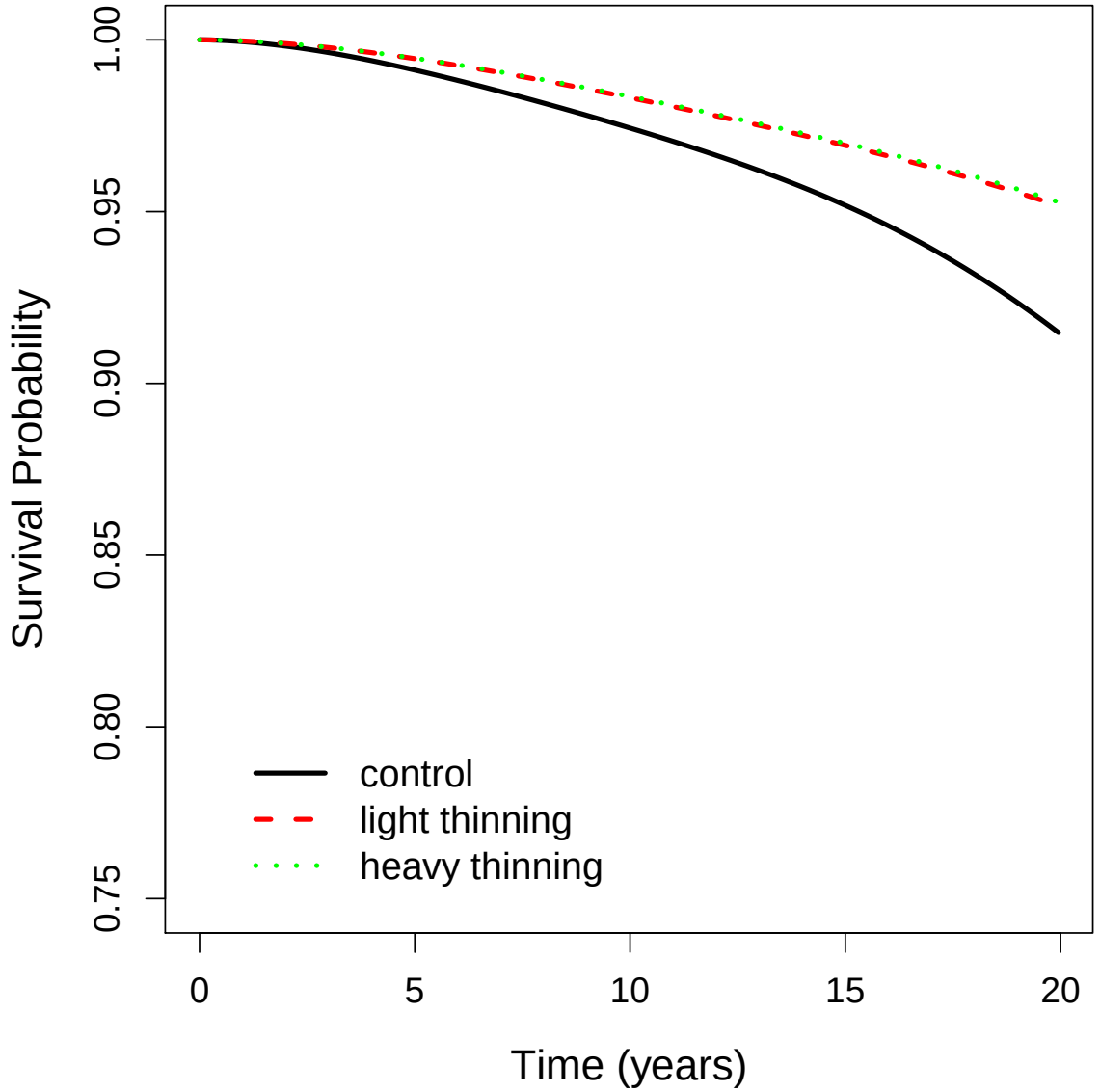
Loblolly pine trees: GRF-AFT

	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

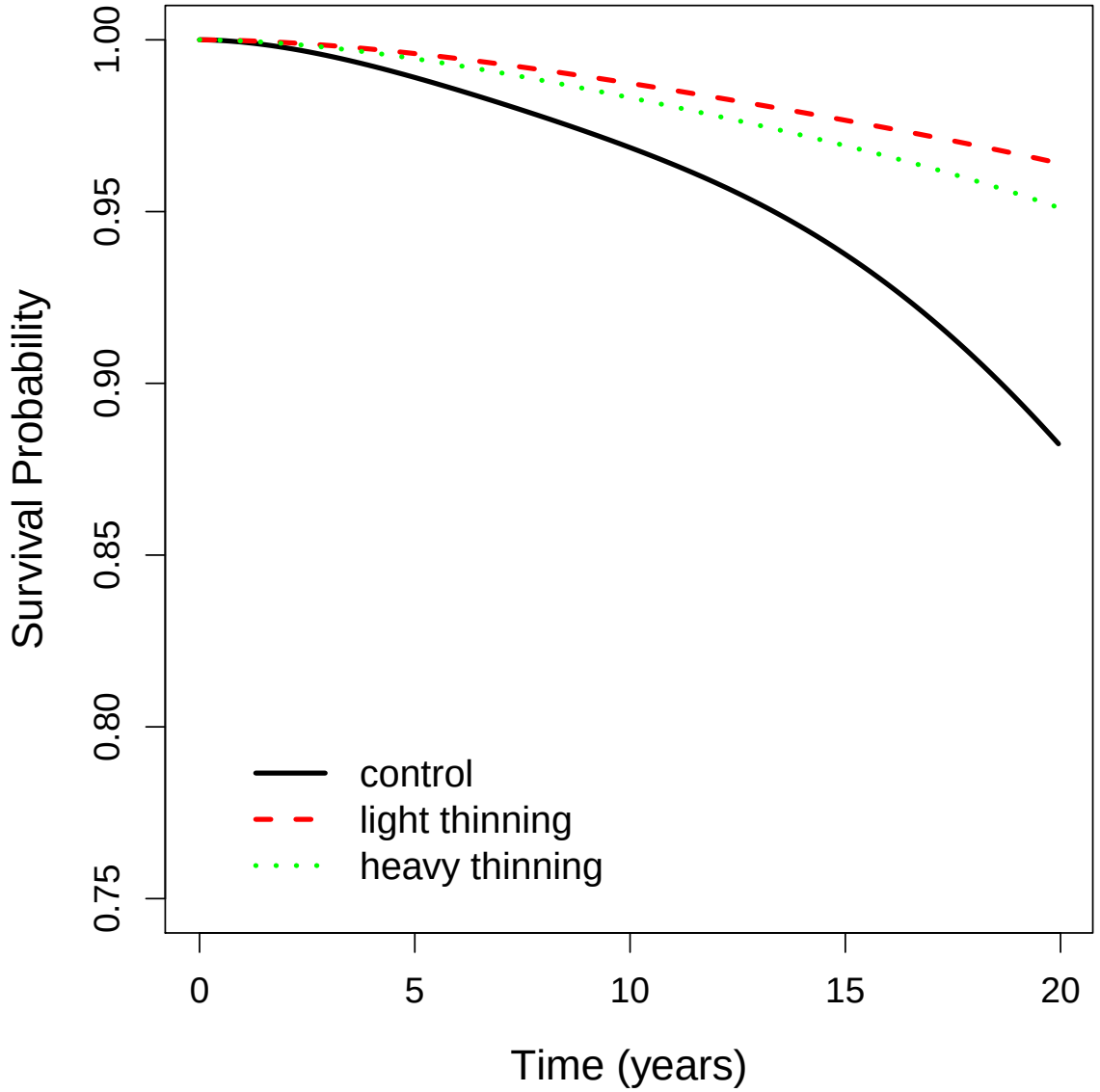
Survival plots for coastal region under GRF-AFT



Survival plots for Piedmont region under GRF-AFT

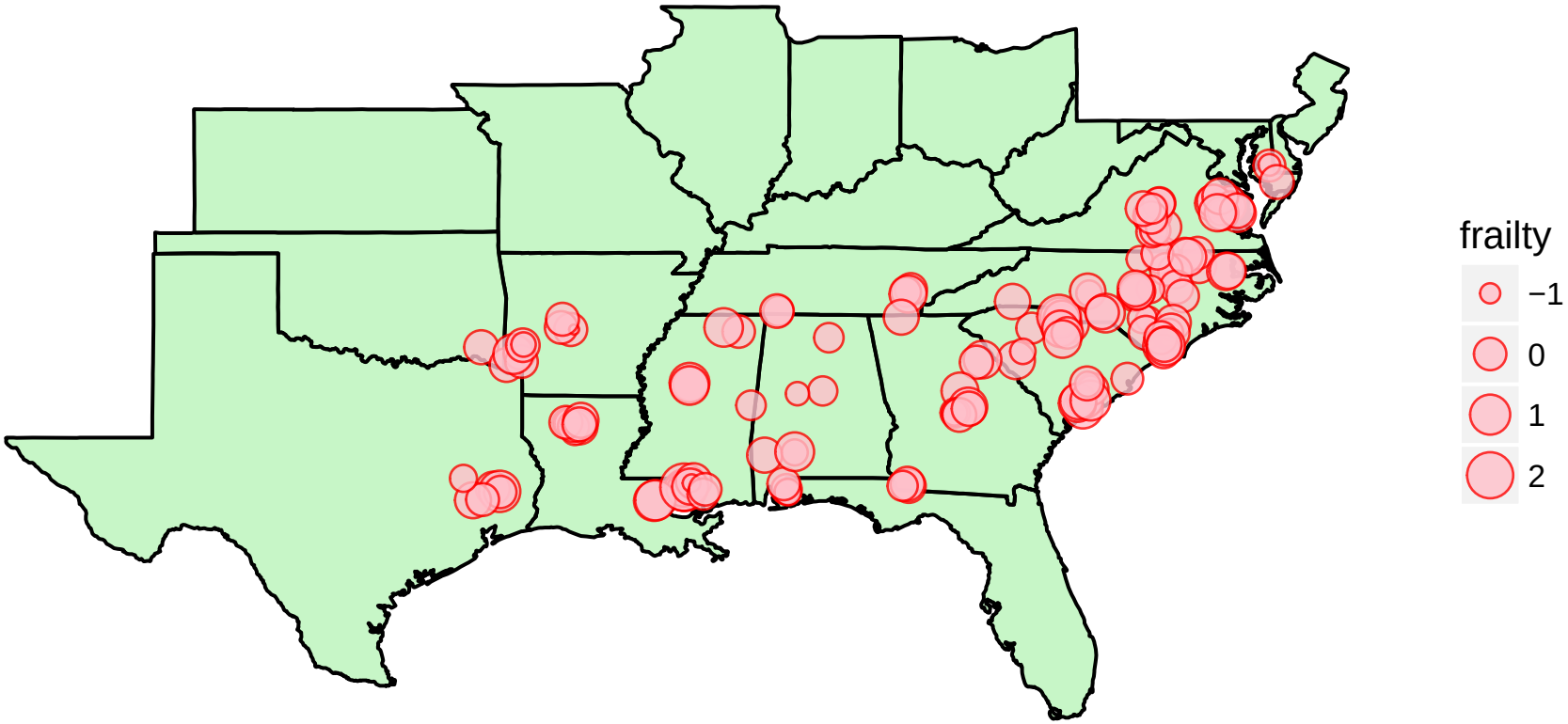


Survival plots for other region under GRF-AFT



Loblolly pine trees: spatial dependence

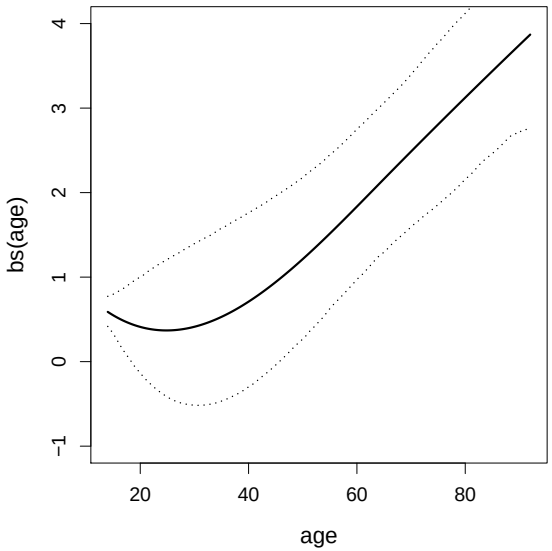
Under the exponential correlation $\rho(\mathbf{s}_i, \mathbf{s}_j) = e^{-\phi\|\mathbf{s}_i - \mathbf{s}_j\|}$, the posterior mean is $\hat{\phi} = 0.2735$, indicating that the correlation decays by $1 - e^{-0.2735} = 24\%$ for every 1-km increase in distance.



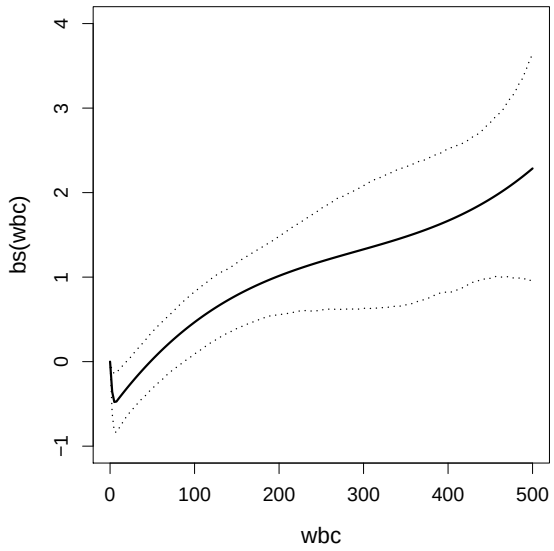
Leukemia data

- Survival of acute myeloid leukemia in $n = 1043$ patients
- Of interest to investigate possible spatial variation in survival after accounting for age, sex, log white blood cell count (wbc) at diagnosis, and Townsend score (tpi, higher = less affluent).
- $m = 24$ administrative districts.
- Henderson et al. (2002) fitted PH CAR model w/ linear predictors.
- We fit additive PH, AFT and PO models with CAR frailties: LPML for PH, AFT and PO are -5946, -5945, and -5919, respectively.
- BF for testing linearity of age, wbc and tpi are 0.13, 0.04 and 0.01; linear effects fine.

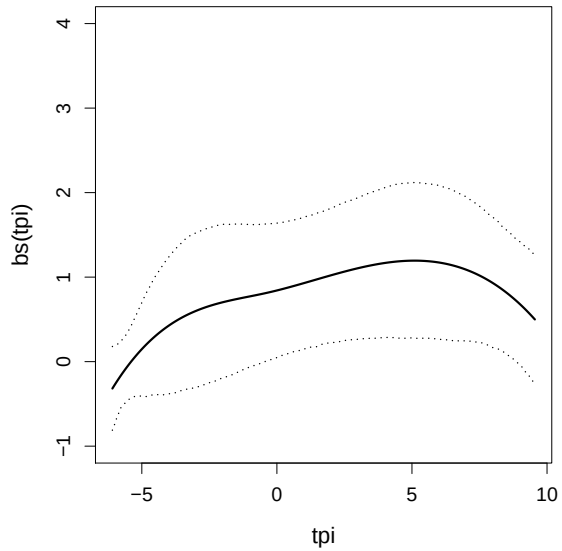
Leukemia data: nonlinear age, log-wbc, and tpi effects



(a)

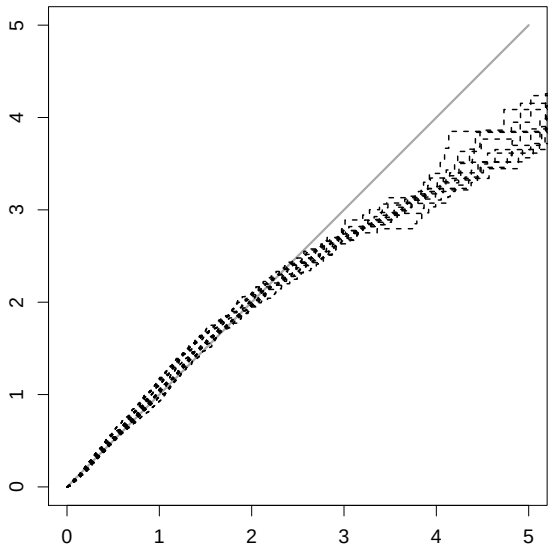


(b)

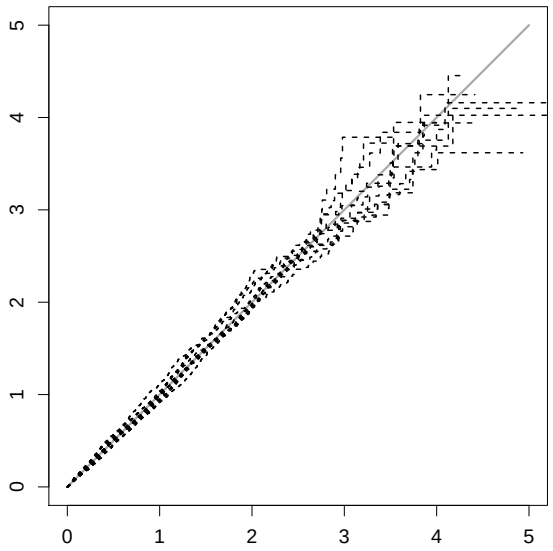


(c)

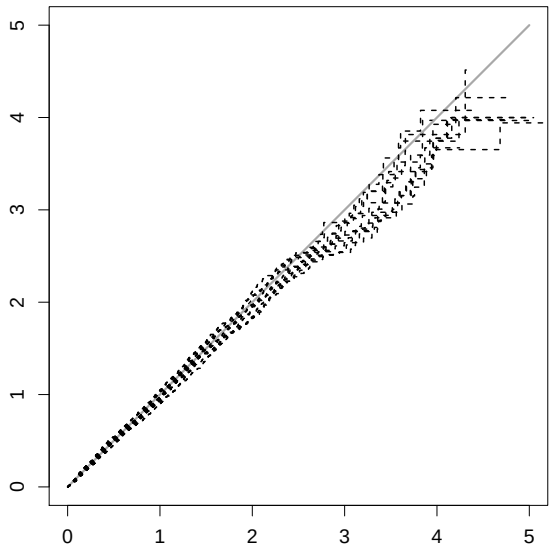
Leukemia data: Cox-Snell plots for PH, AFT, and PO



(d)



(e)



(f)

Outline

Summary

- Proposed new AFT, PH and PO frailty models for survival data subject to arbitrary censoring and spatial dependence.
- All three data analyses did not choose PH despite this being how data initially analyzed.
- Baseline modeled via Bernstein polynomial centered at parametric family; smooth densities leads to efficient posterior updating.
- Developed a function `survregbayes` within the R package `spBayesSurv` for implementing the MCMC algorithms.
- Joint work w/ Haiming Zhou at U. Northern Illinois.
- Future work: marginal semiparametric models with spatial dependence modeled through copulas; specialized MCMC for additive models w/ penalized B-splines and inclusion of pairwise interaction surfaces.
- Thanks for the invitation!