Marginal Survival Modeling through Spatial Copulas

Tim Hanson

Department of Statistics
University of South Carolina, U.S.A.

University of Michigan
Department of Biostatistics

March 31, 2016

Outline

1. Fundamental concepts
2. Semiparametric models
3. Spatial copula models

Survival data

- Can be time to any event of interest, e.g. death, leukemia remission, bankruptcy, electrical component failure, etc.
- Data $T_1, T_2, \ldots, T_n$ live in $\mathbb{R}^+$.
- Called: survival data, reliability data, time to event data.
- Interest often focuses on relating aspects of the distribution on T_i to covariates or risk factors $\mathbf{x}_i$.

Survival data: covariates and censoring

- Uncensored data: $(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_n, t_n)$. Observe $T_i = t_i$.
- Right censored data: $(\mathbf{x}_1, t_1, \delta_1), \ldots, (\mathbf{x}_n, t_n, \delta_n)$. Observe

$$\left\{ \begin{array}{ll} T_i = t_i & \delta_i = 1 \\ T_i > t_i & \delta_i = 0 \end{array} \right\}.$$

- Interval censored data: $(\mathbf{x}_1, a_1, b_1), \ldots, (\mathbf{x}_n, a_n, b_n)$. Observe $T_i \in [a_i, b_i]$.

Density and survival

- Continuous T has density $f(t)$.
- Survival function is

$$S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(s)ds.$$

- Regression model: proportional odds.

Quantiles

- p^{th} quantile q_p for T solves $P(T \leq q_p) = p$.
- Continuous $T \Rightarrow q_p = F^{-1}(p)$.
- Median lifetime is $q_{0.5} = F^{-1}(0.5)$.
- Quantile regression active area of research from frequentist & Bayesian perspectives, e.g. Koenker's excellent `quantreg` package for `R`.

Residual life

- Mean residual life

$$m(t) = E\{T - t | T > t\} = \frac{\int_t^\infty S(s) ds}{S(t)}.$$

- Regression model: proportional mean residual life.

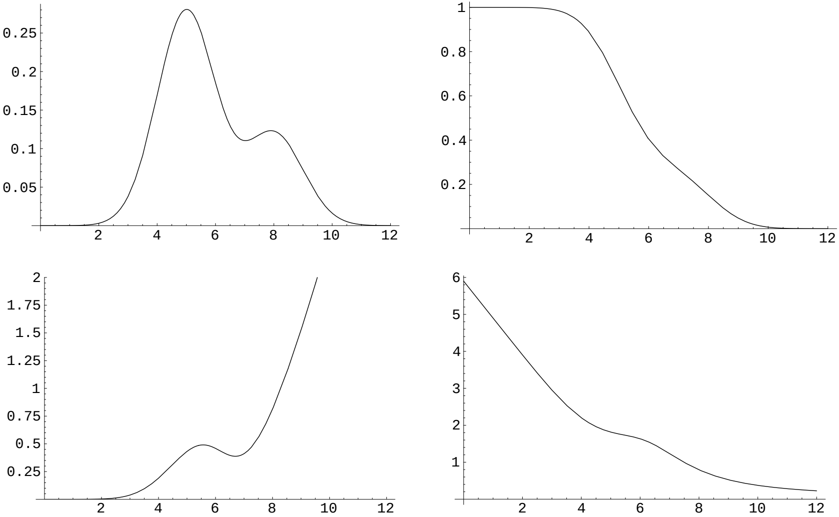
Hazard function

- Hazard at t:

$$h(t) = \lim_{dt \to 0^+} \frac{P(t \leq T < t + dt | T \geq t)}{dt} = \frac{f(t)}{S(t)}.$$

- Regression models: proportional hazards (Cox), additive hazards (Aalen), accelerated hazards, & extended hazards.

Density, survival, hazard, and MRL



Nonparametric survival priors

- Infinite-dimensional process defined on one of $h(t)$, $H(t)$, $f(t)$, or $S(t)$.
- Priors on $h(t)$ include extended gamma, piecewise exponential, B-splines, etc.
- Priors on $H(t) = -\log S(t)$ include gamma, beta, etc.
- Priors on $S(t)$ include Dirichlet process (DP).
- Priors on $f(t)$ include DP mixtures, transformed Bernstein polynomials, Polya trees, B-splines, etc.
- We'll consider MPT, B-spline, and DPM.

Semiparametric models

Work covariates $\mathbf{x}_i$ into model for T_i. Most common: semiparametric model. Why?

- Splits inference into two pieces: $\boldsymbol{\beta}$ and $S_0(t)$.
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ succinctly summarizes effects of risk factors **x** on aspects of survival.
- Make $S_0(t)$ as flexible as possible.
- Can make easily digestible statements concerning the population, e.g. “Median life on those receiving treatment A is 1.7 times those receiving B, adjusting for other factors.”

Some semiparametric models

- PH: $h_{\mathbf{x}}(t) = \exp(\mathbf{x}'\beta)h_0(t)$.
- AddH: $h_{\mathbf{x}}(t) = h_0(t) + \beta'\mathbf{x}$.
- AFT: $S_{\mathbf{x}}(t) = S_0\{e^{\beta'\mathbf{x}}t\}$.
- PO: $F_{\mathbf{x}}(t)/S_{\mathbf{x}}(t) = e^{\beta'\mathbf{x}}F_0(t)/S_0(t)$.
- PMRL: $m_{\mathbf{x}}(t) = e^{\beta'\mathbf{x}}m_0(t)$.
- AccH: $h_{\mathbf{x}}(t) = h_0\{e^{\beta'\mathbf{x}}t\}$.
- ExtH: $h_{\mathbf{x}}(t) = h_0\{e^{\beta'\mathbf{x}}t\}e^{\gamma'\mathbf{x}}$.
- Others, but this covers 99%.

Proportional hazards (PH)

- Model is:

$$h_{\mathbf{x}}(t) = \exp(\mathbf{x}'\beta)h_0(t) \text{ or } S_{\mathbf{x}}(t) = S_0(t)^{\exp(\mathbf{x}'\beta)}.$$

- `BayesX` assigns penalized B-spline prior on log $h_0(t)$ and allows for additive predictors, structured frailties, time-varying coefficients, etc. Free: http://www.statistik.lmu.de/∼bayesx/bayesx.html. Also R package to call BayesX.
- `BAYES` in SAS `PROC PHREG` gives p.w. exponential.
- Haiming Zhou's `spBayesSurv` has S_0 modeled as MPT in `survregbayes`.

Accelerated failure time (AFT)

- Model is

$$S_{\mathbf{x}}(t) = S_0\left(e^{-\mathbf{x}'\boldsymbol{\beta}}t\right), \quad \text{or} \quad \log T_{\mathbf{x}} = \mathbf{x}'\boldsymbol{\beta} + e_0.$$

- Implies $q_p(\mathbf{x}) = e^{\mathbf{x}'\boldsymbol{\beta}}q_p(0)$.
- Komarek's `bayesSurv` for AFT models; spline and discrete normal mixture on error.
- `bj()` in Harrell's `Design` library fits Buckley-James version.
- `spBayesSurv` has S_0 modeled as MPT.

Proportional odds (PO)

- Model is
$$\frac{1 - S_{\mathbf{x}}(t)}{S_{\mathbf{x}}(t)} = \exp(\mathbf{x}'\beta)\frac{1 - S_0(t)}{S_0(t)}.$$
- Attenuation of risk:
$$\lim_{t\to\infty} \frac{h_{\mathbf{x}_1}(t)}{h_{\mathbf{x}_2}(t)} = 1.$$
- Haiming Zhou's `spBayesSurv` has S_0 modeled as MPT. `timereg` has frequentist version.

Spatial frailty survival models

- Survival data often collected over region.
- Georeferenced includes $\boldsymbol{s}_i = (x_i, y_i)$, e.g. latitude & longitude.
- Areal includes $c_i \in \{1, \ldots, C\}$, e.g. the county of residence (there are C counties).
- Traditionally, spatial dependence induced by adding frailty (random effect) to linear predictor in semiparametric model.

Georeferenced spatial frailty

- Replace $\mathbf{x}_i'\beta$ by $\mathbf{x}_i'\beta + g_i$.
- Take $g_i = g(x_i, y_i)$ where $\{g(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{S}\}$ is mean-zero stationary Gaussian process.
- Yields $\mathbf{g} = (g_1, \ldots, g_n) \sim N_n(\mathbf{0}, \mathbf{C}_\theta)$; $\mathbf{C}_\theta$ e.g. Matérn.

Mean-zero, smoothed spatial surface $g(\boldsymbol{s})$ for $\boldsymbol{s} \in \mathcal{S}$.

Areal spatial frailty

- Replace $\mathbf{x}_i'\beta$ by $\mathbf{x}_i'\beta + g_{c_i}$.
- Define $\mathbf{W}$ to be adjacency matrix: $w_{ij} = 1$ if counties i and j share a border, otherwise $w_{ij} = 0$ (assume $w_{ii} = 0$).
- CAR model assumes $g_j|\mathbf{g}_{-j} \sim N(\rho\tilde{g}_j, \frac{\lambda}{w_{j+}})$ where $\rho \in (0,1)$ and $\tilde{g}_j = \frac{1}{w_{j+}}\sum_{i=1}^{C} w_{ij}g_i$.
- Limiting case $\rho \to 1$ called ICAR, requires $\sum_{j=1}^{C} g_j = 0$.

Mean-zero, smoothed spatial surface g_j for $j \in \mathcal{S}$.

Choosing among survival models with spatial frailties

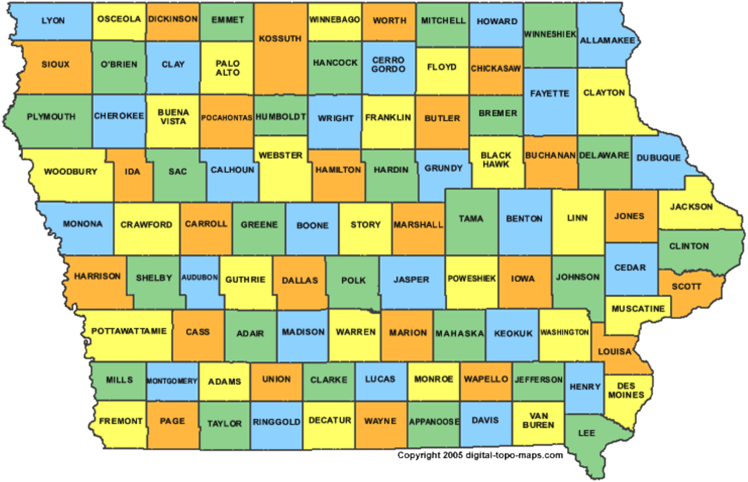
For SEER data look at survival of women in 99 counties from Iowa. Examined 3 models:

- Proportional hazards (PH)
- Accelerated failure time (AFT)
- Proportional odds (PO)

In each case simply use $\mathbf{x}_i'\beta + g_{C_i}$ instead of $\mathbf{x}_i'\beta$.

Analysis of the 1995-1998 Iowa SEER data

- Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute provides county-level cancer data on annual basis for public use.
- 488 events; 585 censorings.
- Covariates: race (white or other), age in years at diagnosis, number of primaries, and the stage of the disease: local (baseline, confined to the breast), regional (spread beyond the breast tissue), or distant (metastatis).



Copyright 2005 digital-topo-maps.com

Figure: $n = 99$ Iowa counties.

- **Question**: which is predictively most important?
  - (a) Parametric versus nonparametric assumptions on baseline survival S_0
  - (b) assumptions on frailty terms
  - (c) assumptions built into survival model (PH, AFT, PO) itself?
- Frailties enter into linear predictor; if model grossly invalid then no way to “fix” frailty distribution or assumptions on S_0 to make model fit adequate. *Need to consider alternative models*.
- Assume $S_0 \sim PT_5(c, \rho, G_\theta)$ where G_θ Weibull or log-logistic. Different priors on c and $c \to \infty$.

		PH		AFT		PO	
Model	c prior	Weibull	Log-logistic	Weibull	Log-logistic	Weibull	Log-logistic
CAR frailty	$\Gamma(5,1)$	–25.8	–24.5	–31.1	–25.2	–9.2	–8.7
	$\Gamma(20,2)$	–26.1	–28.2	–33.8	–26.3	–12.7	–12.0
	$c \to \infty$	–33.0	–40.6	–33.1	–29.6	–20.9	–29.5
iid frailty	$\Gamma(5,1)$	–28.2	–25.8	–31.7	–26.2	–12.5	–11.9
	$\Gamma(20,2)$	–27.7	–29.1	–37.6	–27.9	–15.9	–15.2
	$c \to \infty$	–34.8	–42.3	–34.9	–32.5	–23.2	–32.4
Non-frailty	$\Gamma(5,1)$	–44.2	–40.1	–40.7	–34.7	–23.6	–22.7
	$\Gamma(20,2)$	–44.3	–41.5	–43.0	–35.9	–24.9	–24.5
	$c \to \infty$	–47.7	–54.8	–47.9	–39.5	–30.8	–39.2

LPML ($+2200$) Parametric model obtains when $c \to \infty$.

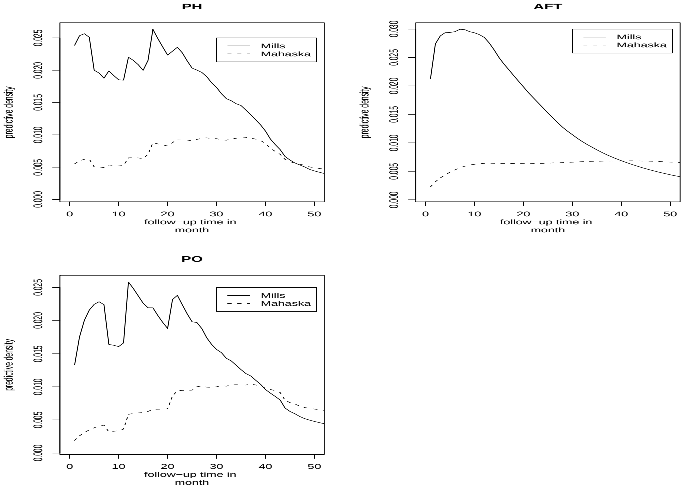
- Among MPT survival models, overall PO$>$PH$>$AFT. For every model, PO best. PBF$\approx 3,000,000$ of PO over PH.
- Overall, MPT$>$log-logistic or Weibull.
- For PO and PH models, CAR$>$i.i.d.$>$none.
- Overall, survival model most important, followed by assumptions on baseline, *followed by frailty model*.
- Focus in literature is on development of complex frailty models within context of PH; alternative survival models often not considered.

Regression effects across models

Model	Centered age	Regional stage	Distant stage
MPT CAR frailty PH	0.018 (0.012, 0.025)	0.22 (0.01, 0.49)	1.65 (1.40, 1.93)
Standard iid frailty PH	0.019 (0.013, 0.025)	0.26 (0.04, 0.49)	1.68 (1.45, 1.92)
Standard non-frailty PH	0.019 (0.013, 0.025)	0.30 (0.08, 0.52)	1.64 (1.42, 1.87)
MPT CAR AFT	0.017 (0.012, 0.022)	0.18 (0.00, 0.38)	1.49 (1.26, 1.74)
Standard iid frailty AFT	0.017 (0.012, 0.022)	0.20 (0.03, 0.38)	1.45 (1.27, 1.64)
Standard non-frailty AFT	0.017 (0.012, 0.021)	0.21 (0.04, 0.38)	1.42 (1.24, 1.61)
MPT CAR frailty PO	-0.030 (-0.038, -0.022)	-0.47 (-0.77, -0.22)	-2.68 (-3.00, -2.36)
Standard iid frailty PO	-0.028 (-0.036, -0.020)	-0.37 (-0.66, -0.08)	-2.58 (-2.92, -2.24)
Standard non-frailty PO	-0.029 (-0.037, -0.020)	-0.40 (-0.68, -0.12)	-2.53 (-2.86, -2.21)

- Regression effects fairly stable.
- Well identified regardless of frailty assumptions.

PDF's: two counties, mean age and local stage.



Discussion

- Three models fit using same nonparametric prior on S_0.
- MCMC scheme based on initial fits of corresponding parametric models.
- Implemented in `spBayesSurv` for interval censored data incorporating variable selection; paper w/ Haiming Zhou in progress.

Spatial copula in a nutshell

- Let $T_i \sim F_{\mathbf{x}_i}(\cdot)$ where $F_{\mathbf{x}}$ c.d.f. from any survival model: parametric, semiparametric, nonparametric.
- $U_i = F_{\mathbf{x}_i}(T_i) \sim U(0,1)$ and $Y_i = \Phi^{-1}(U_i) \sim N(0,1)$. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$.
- No spatial correlation $\Rightarrow \mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I}_n)$.
- Spatial correlation $\Rightarrow \mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{\Gamma})$. Here $\mathbf{\Gamma}_{n\times n} = [\gamma_{ij}]$ with pairwise correlations γ_{ij}.
- Li and Lin (2006) use this in PH model, term it “normal transformation model.”
- Gives marginal (population-averaged) model.
- Unlike frailties, can be used in models *without* a linear predictor.

SCCCR data set on prostate cancer survival

- Large dataset on prostate cancer survival that does not follow proportional hazards.
- $n = 20599$ patients from South Carolina Central Cancer Registry (SCCCR) for the period 1996–2004; each recorded with county, race, marital status, grade of tumor, and SEER summary stage; 72.3% are censored.
- Need to allow for non-proportional hazards and accommodate correlation of survival times within county.

Extended hazards model

- Etezadi-Amoli and Ciampi (1987) propose ExtH model

$$h_{\mathbf{x}}(t) = h_0(te^{\mathbf{x}'\beta})e^{\mathbf{x}'\gamma}.$$

- Say $\mathbf{x} = (x_1, x_2)$, then ExtH is

$$h_{\mathbf{x}}(t) = h_0(te^{\beta_1 x_1 + \beta_2 x_2})e^{\gamma_1 x_1 + \gamma_2 x_2}.$$

- $\gamma_1 = \beta_1 \Rightarrow x_1$ has AFT interpretation; $\beta_1 = 0 \Rightarrow x_1$ has PH interpretation; $\gamma_1 = 0 \Rightarrow x_1$ has AccH interpretation.
- B-spline baseline hazard $h(t)$ shrunk toward parametric target h_θ. Posterior updating through clever McMC w/ augmented likelihood.

Spatial dependence via frailties impractical

- PH with frailties:

$$h(t_i|\mathbf{x}) = h_0(t_i)e^{\gamma'\mathbf{x}_i + g_{c_i}},$$

  where g_{c_i} are county-level frailties, c_i is county subject i in.
- EH with frailties:

$$h(t_i|\mathbf{x}) = h_0\{t_i e^{\beta'\mathbf{x}_i + b_{c_i}}\}e^{\gamma'\mathbf{x}_i + g_{c_i}},$$

  where, for our data, $b_1, \ldots, b_{46}$ and $g_1, \ldots, g_{46}$ are county-level frailties.
- Possible but impractical, and hard to interpret.

Spatial dependence via copula works great

- Define $Y_i = \Phi^{-1}\left\{F_{\mathbf{x}_i}(T_i)\right\}$.
- Under Li and Lin (2006) $\mathbf{Y} \sim N(\mathbf{0}, \boldsymbol{\Gamma})$.
- Likelihood from data $\{(t_i, \mathbf{x}_i, \delta_i)\}_{i=1}^n$ is

$$L(\beta, \gamma, \mathbf{b}, \theta, \boldsymbol{\Gamma}) = \left[\prod_{i \in S} \frac{f_i(t_i)}{\phi(y_i)}\right] \int \left[\prod_{i \in S^c} \frac{f_i(z_i)}{\phi(y_i)} I(z_i > t_i)\right] \phi(\mathbf{y}; \mathbf{0}, \boldsymbol{\Gamma}) \prod_{i \in S^c} dz_i$$

- $\boldsymbol{\Gamma}$ defined through ICAR correlation matrix; details in paper but not straightforward. SVD saves the day.

Savage-Dickey ratio for global and per-variable tests

- Example of global test of PH vs. EH

$$BF_{12} = \frac{\pi(\boldsymbol{\beta} = \mathbf{0} | \mathcal{D}, EH)}{\pi(\boldsymbol{\beta} = \mathbf{0} | EH)}.$$

- Example of per-variable of PH for x_j vs. EH

$$BF_{12} = \frac{\pi(\beta_j = 0 | \mathcal{D}, EH)}{\pi(\beta_j = 0 | EH)}.$$

SCCCR data

- SCCCR prostate cancer data for the period 1996–2004.
- Baseline covariates are county of residence, age, race, marital status, grade of tumor differentiation, and SEER summary stage.
- $n = 20599$ patients in the dataset after excluding subjects with missing information.
- 72.3% of the survival times are right-censored.

Goal: assess racial disparity in prostate cancer survival, adjusting for the remaining risk factors and accounting for the county the subject lives in.

SCCCR data

Table: Summary characteristics of prostate cancer patients in SC from 1996-2004.

Covariate		n	Sample percentage
Race	Black	6483	0.32
	White	14116	0.68
Marital status	Non-married	4525	0.22
	Married	16074	0.78
Grade	well or moderately differentiated	15309	0.74
	poorly differentiated or undifferentiated	5290	0.26
SEER summary stage	Localized or regional	19792	0.96
	Distant	807	0.04

Non-spatial EH and reduced models

Table: Summary of fitting the extended hazard model EH, the reduced model, AFT, and PH; $*$ indicates *LPML* $-$ 21000 and *DIC* $-$ 42000.

Covar		EH	Reduced	AFT	PH	PH+additive age
				$\boldsymbol{\beta} = \boldsymbol{\gamma}$	$\boldsymbol{\beta} = 0$	$\boldsymbol{\beta} = 0$
Age	β_1	0.50(0.48,0.52)	0.48(0.46,0.50)	0.48(0.45,0.51)	–	
	γ_1	0.45(0.42,0.49)	$\gamma_1 = \beta_1$	–	0.65(0.62,0.68)	–
Race	β_2	0.18(0.15,0.21)	0.20(0.16,0.21)	0.18(0.15,0.22)	–	–
	γ_2	0.18(0.12,0.24)	$\gamma_2 = \beta_2$	–	0.26(0.21,0.32)	0.26(0.20,0.31)
Marital	β_3	-0.06(-0.11,-0.02)	-0.05(-0.09,-0.00)	0.26(0.21,0.30)	–	–
status	γ_3	0.35(0.29,0.40)	0.33(0.28,0.40)	–	0.33(0.27,0.39)	0.31(0.26,0.37)
Grade	β_4	0.03(-0.02,0.08)	$\beta_4 = 0$	0.27(0.22,0.32)	–	–
	γ_4	0.36(0.29,0.41)	0.37(0.31,0.43)	–	0.38(0.32,0.44)	0.37(0.33,0.43)
SEER	β_5	3.19(2.80,3.53)	3.27(2.79,3.57)	1.50(1.41,1.59)	–	–
stage	γ_5	1.02(0.83,1.20)	1.00(0.82,1.19)	–	1.56(1.47,1.64)	1.57(1.19,1.65)
*LPML**		-161.0	-162.0	-206.5	-242.5	-231.9
*DIC**		267.7	270.7	366.0	443.0	412.8

Non-spatial EH and reduced models

Table: Bayes factors for comparing EH to PH, AFT, and AH with and without spatial correlation.

		EH			Spatial+EH	
Covariate	PH	AFT	AH	PH	AFT	AH
Age	> 1000	0.08	> 1000	> 1000	0.01	> 1000
Race	> 1000	0.01	> 1000	> 1000	< 0.01	> 1000
Marital status	1.79	> 1000	> 1000	1.18	> 1000	> 1000
Grade	0.14	> 1000	> 1000	0.08	> 1000	> 1000
SEER stage	> 1000	> 1000	> 1000	> 1000	> 1000	> 1000

Spatial EH and reduced models

Table: Summary of spatial models; $*$ indicates $LPML - 21000$ and $DIC - 42000$.

Covariates		Marginal EH	Marginal reduced	PH+ICAR+additive age
				$\boldsymbol{\beta} = 0$
Age	β_1	0.50(0.47,0.52)	0.47(0.46,0.49)	–
	γ_1	0.46(0.43,0.49)	$\gamma_1 = \beta_1$	–
Race	β_2	0.18(0.15,0.21)	0.20(0.17,0.22)	–
	γ_2	0.17(0.11,0.23)	$\gamma_2 = \beta_2$	0.24(0.18,0.30)
Marital status	β_3	-0.06(-0.10,-0.02)	-0.02(-0.05,-0.00)	–
	γ_3	0.34(0.28,0.41)	0.33(0.27,0.39)	0.32(0.25,0.38)
Grade	β_4	0.03(-0.01,0.07)	$\beta_4 = 0$	–
	γ_4	0.36(0.30,0.42)	0.38(0.32,0.43)	0.37(0.32,0.44)
SEER stage	β_5	3.16(2.86,3.34)	2.77(2.72,2.82)	–
	γ_5	1.10(0.94,1.26)	1.21(1.01,1.33)	1.55(1.46,1.64)
φ^*		50.1(19.9,113.7)	54.6(22.7,120.8)	33.08(9.2,100.1)
$LPML^*$		-142.7	-143.2	-215.7
DIC^*		192.4	164.0	332.5

Spatial EH and reduced models

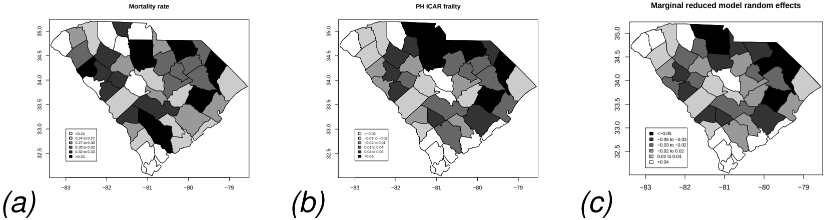


Figure: Map of (a) Mortality rate, (b) ICAR frailties in the PH model and (c) random effects in the marginal reduced model for SC counties.

Spatial EH and reduced models

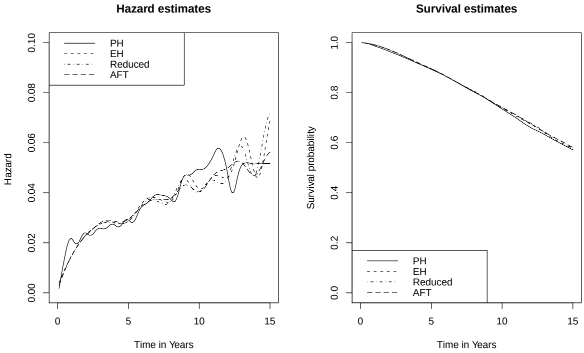


Figure: Baseline hazard (left) and survival probabilities (right) estimates.

Spatial EH and reduced models

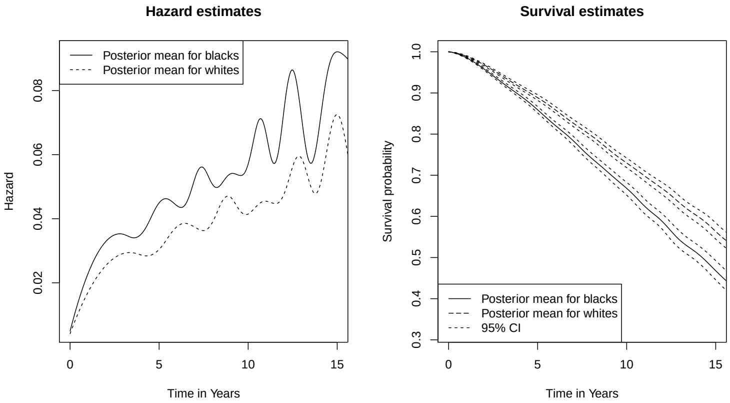


Figure: Hazard and survival for black patients (solid line) and white patients; baseline covariates.

Interpretation for race effect

- Reduced models, white South Carolina subjects diagnosed with prostrate cancer in live 22% longer ($e^{0.20} \approx 1.22$) than black patients (95% CI is 18% to 25%) adjusting for rest.
- Cox: "*...the physical or substantive basis for...proportional hazards models...is one of its weaknesses...*" and goes on to suggest that "*...accelerated failure time models are in many ways more appealing because of their quite direct physical interpretation.*"
- Main covariate of interest, race, best modeled as AFT effect.

More interpretation

- Decreasing age one year increases survival 5.4%.
- Hazard of dying increases 46% for poorly or undifferentiated grades vs. well or moderately differentiated, holding all else constant.
- SEER stage has general ExtH effects, $e^{2.77} \approx 16$ (AH) and $e^{1.21} \approx 3.4$ (PH). Those with distant stage are at least three times worse in one-sixteenth of the time as those with localized or regional.
- Marital status essentially has PH interpretation; single (incl. widowed & separated) subjects $e^{0.33} \approx 1.39$ times more likely to die at any instant than married.

Extinction of mountain yellow-legged frog

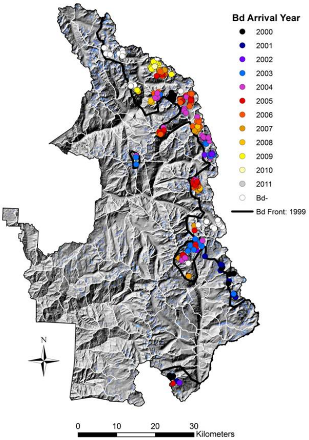
- Frogs and other amphibians have been dying off in large numbers since the 1980s because of a deadly fungus called *Batrachochytrium dendrobatidis*, or Bd.
- Dr. Knapp has been studying the amphibian declines for the past decade at Sierra Nevada Aquatic Research Laboratory; he has hiked thousands of miles and surveyed hundreds of frog populations in Sequoia-Kings Canyon National Park collecting the data by hand.
- As with the SCCCR data, proportional hazards grossly violated.
- Instead of semiparametric, pursue nonparametric $F_{\mathbf{x}_i}$; not able to use frailties.

The Frog Data (2000-2011)

- Contains 309 frog populations. Each was followed up until infection or being censored (10% censoring).
- Response T_i is time to Bd infection. (i.e. Bd arrival year $-$ baseline year).
- Main covariates:

  $x_{i1} \in \{0, 1\}$ is whether or not Bd has been found in the watershed.

  x_{i2} is straight-line distance to the nearest Bd location.
- Populations near each other tend to become infected at about the same time.



LDDPM model and Spatial Extension

- LDDPM (De Iorio et al., 2009; Jara et al., 2010): $Z_i = \log T_i$ given $\mathbf{x}_i$ follows mixture model

  $$F_{\mathbf{x}_i}(z) = \int \Phi\left(\frac{z - \mathbf{x}_i'\beta}{\sigma}\right) dG(\beta, \sigma^2),$$

  where G follows Dirichlet Process (DP) prior: $G \sim DP(\alpha, G_0)$.
- Countable mixture of parametric linear models $F_{\mathbf{x}_i} = \sum_{j=1}^{\infty} w_j N(\mathbf{x}_i'\boldsymbol{\beta}_j, \boldsymbol{\sigma}_j^2)$.
- As before, take $Y_i = \Phi^{-1}\{F_{\mathbf{x}_i}(\log T_i)\}$ and $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{\Gamma})$.
- $\mathbf{\Gamma}_{\boldsymbol{\theta}}$ used for capturing spatial dependence; $\gamma_{ij} = \theta_1 \exp\{-\theta_2 ||\boldsymbol{s}_i - \boldsymbol{s}_j||\} + (1 - \theta_1) I\{\boldsymbol{s}_i = \boldsymbol{s}_j\}$.

MCMC Overview

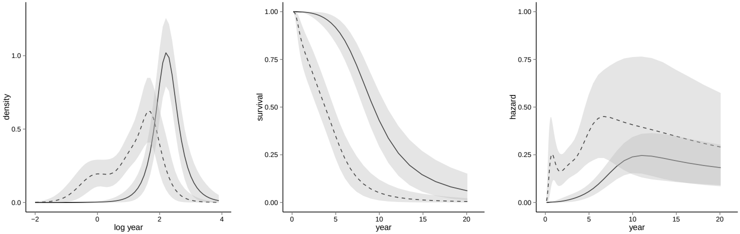
- Truncated stick-breaking representation $G = \sum_{i=1}^{N}[v_i \prod_{j<i}(1-v_j)]\delta_{\beta_j,\sigma_j^2}$ where $v_1,\ldots,v_{N-1} \overset{iid}{\sim} \text{beta}(1,\alpha)$, $v_N = 1$, and $(\beta_j,\sigma_j^2) \overset{iid}{\sim} G_0$.
- G parameters updated based on a M-H proposal from blocked Gibbs sampler (Ishwaran and James, 2001).
- The latent censored t_i updated via M-H sampler.
- Delayed rejection (Tierney and Mira, 1999) used for several parameters; helps sampler not get "stuck."
- Correlation parameters $\boldsymbol{\theta}$ are updated using adaptive M-H (Haario et al., 2001).
- For large n, the inversion of the $n \times n$ matrix $\boldsymbol{C}$ substantially sped up using a full scale approximation (FSA) (Sang and Huang, 2012).

Frog Data: Inference on Spatial Correlation

- Posterior mean $\hat{\theta}_1 = 0.9937$.
- Posterior mean $\hat{\theta}_2 = 0.0866$, indicating the correlation decays by $1 - \exp\{-0.0866(1)\} = 8\%$ for every 1-km increase in distance and $1 - \exp\{-0.0866(10)\} = 58\%$ for every 10-km increase in distance.

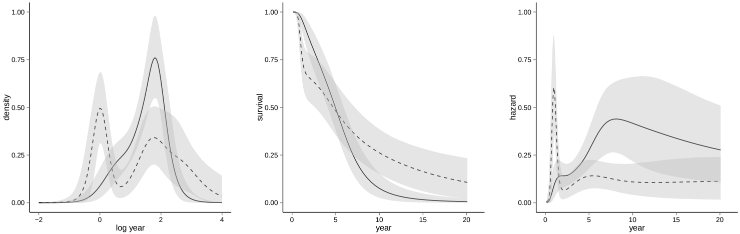
Table: Posterior summary statistics for the spatial correlation parameters

Par.	Mean	Median	Std. dev.	95% HPD Interval
θ_1	0.9937	0.9941	0.0029	(0.9879, 0.9988)
θ_2	0.0866	0.0841	0.0211	(0.0493, 0.1297)



(a) (b) (c)

Figure: Fitted marginal densities, survival curves, and hazard curves w/ 90% CI for high versus low value of bddist when bdwater is equal to 0; bddist=95% and bddist=5% quantiles are solid and dashed lines.



(a) (b) (c)

Figure: Fitted marginal densities, survival curves, and hazard curves w/ 90% CI for bdwater=0 versus bdwater=1 when bddist is equal to population mean of 2.7 km; results for bdwater=0 and bdwater=1 are solid and dashed lines.

Frog Data: Spatial Prediction

Spatial map for the transformed process $z(\boldsymbol{s}) = \Phi^{-1}\left\{F_{\mathbf{x}(\boldsymbol{s})}(\log T(\boldsymbol{s})|G)\right\}$.

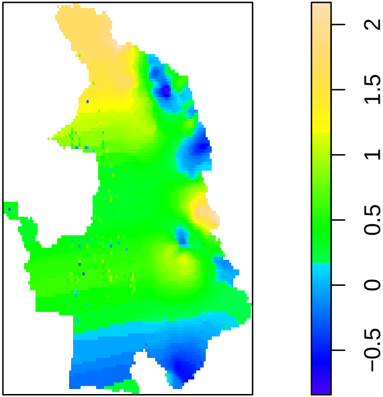


Figure: Predictive spatial map across $\mathcal{D}$.

Which is better, copula or frailty?

LPML	Model
-276	LDDPM-copula
-304	PH-copula
-632	LDDPM-independent
-705	PH-independent
-703	PH-frailty

LDDPM copula model better than PH copula model. However, PH copula better than LDDPM without copula. Modeling via copula grossly improves predictive performance of the models. Frailty improves PH model only slightly.

Remarks

- PH, PO, AFT frailty models developed w/ iid or ICAR.
- Bayesian spatial copula semiparametric (ExtH model) and nonparametric (LDDPM); wins over frailty.
- Implementation of semiparametric models focus of current research, both frailty and copula.
- Thanks to my co-authors Haiming Zhou, Li Li, Roland Knapp, Luping Zhao, and Jiajia Zhang.
- Papers based on this work are available; email if interested.
- Thanks for invitation!