Semiparametric Spatial Model for Interval-censored Data with Time-varying Covariate Effects

Yue Zhang

Department of Bioinformatics and Biostatistics
SJTU-Yale Joint Biostatistics Center
Shanghai Jiao Tong University

August 30, 2018

Outline

- Introduction and motivating example
- Likelihood, prior and posterior
- Simulation study and application
- Summary and future work

Survival Analysis

Survival analysis is used to analyze data in which the time until the event is of interest. The response is often referred to as a failure time, survival time, or event time, e.g.,

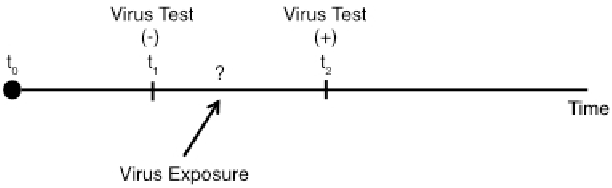
- Time until tumor recurrence
- Time until cardiovascular death after some treatment intervention
- Time until AIDS for HIV patients
- Time until a machine part fails

The survival time response

- Usually Continuous
- Incompletely observed responses are censored
- Let survival time T be a nonnegative random variable and

$$T \in (L, R] \quad \text{(Sun, 2007)}$$

  - Exact observation: $0 < L = R < \infty$
  - Right Censoring: $0 < L < R = \infty$
  - Interval Censoring: $0 < L < R < \infty$
- e.g., HIV infection time



- Ignore interval-censoring?

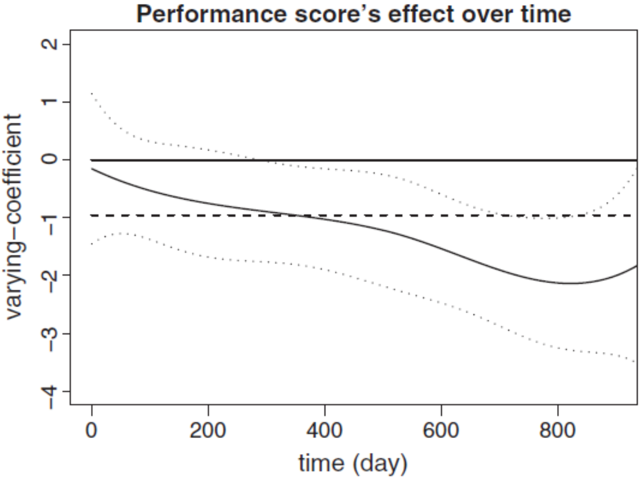
Statistical Methods

- Estimation of the survival distribution
  - Kaplan-Meier or Product Limit Estimator
  - Life-Table
- Comparison of survival curves
  - Log-Rank Test
- Regression Models with respect to hazard
  - Parametric regression models: exponential, Weibull, etc.
  - Semiparametric regression models: Cox, etc.

$$\lambda(t) = \lambda_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta})$$

- Cox model, or relative risk model = proportional hazards model? $\boldsymbol{\beta}$ or $\boldsymbol{\beta}(t)$.

Time-varying coefficient example



Time-varying effect of the performance score on stroke readmission: $\beta(t)$, solid; 95% point-wise confidence interval, dotted; performance score effect in constant coefficient model, dashed (Yu et al., 2013).

Motivating example

Smoking cessation data in southeastern Minnesota

- Lung health study (Murray et al., 1998)
- Event of interest: Time to smoking relapse
- Covariates: Gender (F/M), Treatment (Intervention/Usual)

ObsLHS	SexF	Relapse	Timept1	Timept2	Zip	Treatment
4266	0	1	0.997	2.146	55009	1
4213	0	0	4.895	Inf	55009	2

Challenge

- The effect of **Treatment** may vary over time
- **Cox** model with **interval-censored** data
- The **correlation** of subjects **within/between** zip code areas

Methods

Challenge

- The effect of **Treatment** may vary over time
  Solution: Time-varying coefficient $\beta(t)$
  - Gibbs sampling with piecewise constant coefficient assumption (Sinha et al., 1999)
  - Penalized splines (Cai and Betensky, 2003; Kneib, 2006)
  - ✓ Reversible jump Markov chain Monte Carlo (Green, 1995)
- **Cox** model with **interval-censored** data
  Solution: Piecewise constant baseline and augmented likelihood (Sinha et al., 1999)
- The **correlation** of subjects **within/between** zip code areas
  - Frailty model (Yu et al., 2013; Zhang et al., 2018)
  - ✓ Spatially correlated frailty (Carlin and Louis, 1997; Banerjee et al., 2003)

Model

Cox model with time-varying coefficient and frailty:

$$\lambda(t|\omega_i, \mathbf{x}_{i,j}) = \lambda_0(t)\exp(\mathbf{x}_{i,j}^T\beta(t) + \omega_i)$$

- $i = 1, 2, ..., n$, $j = 1, 2, ..., m_i$, $N = \sum_{i=1}^{n} m_i$
- $\lambda_0(.)$ is an unknown baseline hazard function common to all subjects
- $\mathbf{x}_{i,j}$ is the covariate vector for the j^{th} subject in the i^{th} cluster
- $\beta(t)$ is the time-varying coefficient of main interest
- ω_i is the frailty of the i^{th} cluster

likelihood

Observed data likelihood of subject i

- Interval censoring: $\ell_i = \mathcal{P}(T_i > L_i) - \mathcal{P}(T_i > R_i)$
- Right censoring: $\ell_i = \lambda(t_i)^{\delta_i} S(t_i)$

Latent variables

- $dN_{i,j,k} = \mathbb{1}(T_{i,j} \in (\tau_{k-1}, \tau_k]), k = 1, 2, \dots K$
- $Y_{i,j,\ell} = 1$ for $\ell < k$, $Y_{i,j,\ell} = 0$ for $\ell > k$, and $Y_{i,j,\ell} = (T_{i,j} - \tau_{\ell-1})/\Delta_\ell$ for $\ell = k$

Augmented likelihood
Set $\Theta = \{\log \lambda_0(t), \boldsymbol{\beta}(t)\}$, $D = \{dN_{i,j,k}, Y_{i,j,k}\}$, $W = \{\omega_i\}$, $\lambda_k = \lambda_0(\tau_k)$ and $\boldsymbol{\beta}_k = \boldsymbol{\beta}(\tau_k)$ for $k = 1, 2, \dots, K$,

$$\ell_{i,j}(\Theta | D, W, \mathbf{x}_{i,j}) = \prod_{k=1}^{K} \{\lambda_k \omega_i \exp(\mathbf{x}_{i,j}^T \boldsymbol{\beta}_k)\}^{dN_{i,j,k}} \exp\{-\Delta_k \lambda_k \omega_i \exp(\mathbf{x}_{i,j}^T \boldsymbol{\beta}_k) Y_{i,j,k}\}$$

Prior specification

- Three models
  - Model 1: Fixed $\boldsymbol{\beta}$ with spatially correlated ω_i's
  - Model 2: Time-varying $\boldsymbol{\beta}(t)$ with independent ω_i's
  - Model 3: Time-varying $\boldsymbol{\beta}(t)$ with spatially correlated ω_i's
- Prior of coefficient
  - Fixed $\boldsymbol{\beta}$: $\lambda_k \sim \mathcal{G}(c_k, d_k)$, $\beta \sim \mathcal{N}(\mu_0, \sigma_0^2)$
  - Time-varying $\boldsymbol{\beta}(t)$: $\theta(\tau_p)|\theta(\tau_{p-1}) \sim \mathcal{N}(\theta(\tau_{p-1}), \nu)$, where $\theta = \{\log \lambda_0, \boldsymbol{\beta}\}$
- Prior of frailty
  - Independent ω_i: $\omega_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.
  - Spatially correlated ω_i: $\omega_i|\omega_{-i} \sim \mathcal{N}(\bar{\omega}_{ii}, 1/(m_i\pi_\omega))$ Intrinsic conditional autoregressive (ICAR) model prior (Besag and Kooperberg, 1995)

Posterior computation of latent variables: $dN_{i,j,k}$ and $Y_{i,j,k}$

- Event indicator vector $(dN_{i,j,1}, dN_{i,j,2}, ..., dN_{i,j,k})$ follows a multinomial distribution with size 1 and probability vector $(e_{i,j,1}, e_{i,j,2}, ..., e_{i,j,k})$, where for $k = 1, 2, ..., K$,

$$e_{i,j,k} = \frac{p_{i,j,k}\mathbb{1}(s_k \in (L_{i,j}, R_{i,j}])}{\sum_{s_l \in (L_{i,j}, R_{i,j}]} p_{i,j,l}},$$

$$p_{i,j,k} = \begin{cases} \exp\left\{-\sum_{l=1}^{k-1} \Delta_l \lambda_l \exp(\mathbf{x}_{i,j}^T \boldsymbol{\beta}_k + \omega_i)\right\} - & \\ \exp\left\{-\sum_{l=1}^{k} \Delta_l \lambda_l \omega_i \exp(\mathbf{x}_{i,j}^T \boldsymbol{\beta}_k + \omega_i)\right\} & \text{if } k > 1 \\ 1 - \exp\left\{-\Delta_1 \lambda_1 \exp(\mathbf{x}_{i,j}^T \boldsymbol{\beta}_1 + \omega_i)\right\} & \text{if } k = 1 \end{cases}$$

- Sample failure time $T_{i,j}$, where $T_{i,j}$ follows a doubly truncated exponential distribution on $(\tau_{k-1}, \tau_k]$

$$F(u) = \frac{1 - \exp\{-\lambda_k (u - s_{k-1}) \exp(\mathbf{x}_{i,j}^T \boldsymbol{\beta}_k) + \omega_i\}}{1 - \exp\{-\lambda_k \Delta_k \exp(\mathbf{x}_{i,j}^T \boldsymbol{\beta}_k + \omega_i)\}}$$

- $Y_{i,j,k} = (T_{i,j} - \tau_{k-1})/\Delta_k$

Posterior computation: $\lambda_0(t)$, $\boldsymbol{\beta}(t)$ and ω_i

- $\theta(t) = \{\log(\lambda_0(t)), \beta(t)\}$. Reversible jump MCMC
  - Update move: Number of jumps P and jump times are fixed

$$\Pr(\theta(\tau_p)|\Theta/\{\theta(\tau_p)\}, D, \nu, W) \propto \exp\left\{-\frac{(\theta(\tau_p)-\mu_p)^2}{2\sigma_p^2}\right\}$$

$$\times \exp\left\{-\sum_{i=1}^{n}\sum_{j=1}^{m_i}\sum_{k=1}^{K} \mathbb{1}(\tau_k \in (\tau_{p-1}, \tau_p])\Delta_k \lambda_k \omega_i \exp(\mathbf{x}_{i,j}^T \boldsymbol{\beta}_k) Y_{i,j,k}\right\},$$

  - Birth move: A new jump time τ' is randomly selected from non-jump time grids
  - Death move: A current jump time τ' is randomly selected and deleted

- Sample ω_i with Metropolis-Hastings algorithm.

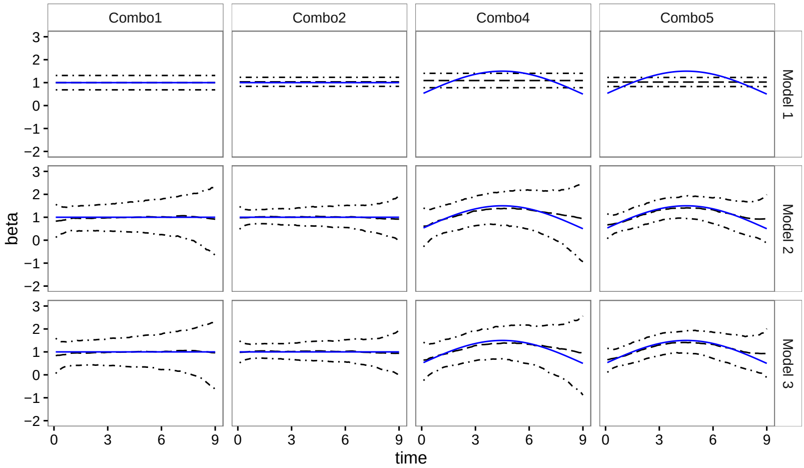
$$\Pr(\omega_i|\Theta, D, \pi_\omega, \omega_{-i}) \propto \prod_{j=1}^{J_i}\prod_{k=1}^{K}\{\lambda_k \exp(\mathbf{x}_{i,j}^T\boldsymbol{\beta}_k + \omega_i)\}^{dN_{i,j,k}}$$

$$\exp\{-\Delta_k \lambda_k \exp(\mathbf{x}_{i,j}^T\boldsymbol{\beta}_k + \omega_i) Y_{i,j,k}\}\Pr(\omega_i|\omega_{-i}).$$

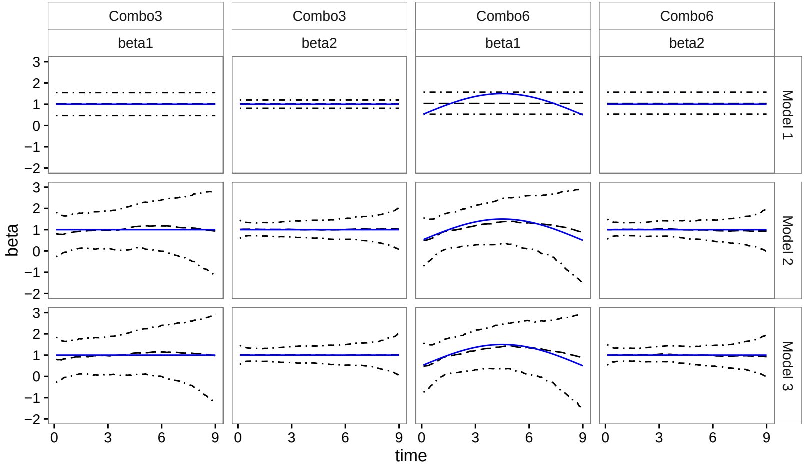
Simulation

- Six combinations
  Combo 1: $\beta_1 = 1$, $x_1 \sim \mathcal{B}(N, 0.5)$
  Combo 2: $\beta_1 = 1$, $x_1 \sim \mathcal{N}(0, 1)$
  Combo 3: $\beta_1 = 1$, $x_1 \sim \mathcal{B}(N, 0.5)$, $\beta_2 = 1$, $x_2 \sim \mathcal{N}(0, 1)$
  Combo 4: $\beta_1 = 0.5 + \sin(t\pi/9)$, $x_1 \sim \mathcal{B}(N, 0.5)$
  Combo 5: $\beta_1 = 0.5 + \sin(t\pi/9)$, $x_1 \sim \mathcal{N}(0, 1)$
  Combo 6: $\beta_1 = 0.5 + \sin(t\pi/9)$, $x_1 \sim \mathcal{B}(N, 0.5)$, $\beta_2 = 1$, $x_2 \sim \mathcal{N}(0, 1)$
- Spatial frailties are based on 45 zip code areas in Cincinnati, there are 15 subjects in each zip code area.
- Baseline hazard function: $\lambda_0(t) = 0.1\sqrt{t}$.
- Log-normal density function $\mathcal{LN}(x; 0, 0.4)$ is used to simulate follow-up times.

Coefficient estimates for combinations with one covariate



Coefficient estimates for combinations with two covariates



LPML comparison between Model 2 and Model 3

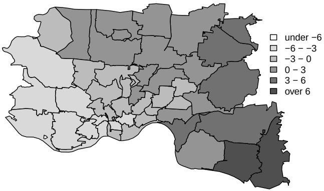
LPML comparison result between Model 2 and Model 3

	Combo 1	Combo 2	Combo 3
LPML Diff	117.6 (-139.0, 391.3)	145.8 (-287.1, 498.3)	125.5 (-92.9, 498.3)
% Diff $>$ 0	84%	79%	89%
	Combo 4	**Combo 5**	**Combo 6**
LPML Diff	116.4 (-129.5, 343.7)	115.1 (-159.9, 539.4)	100.6 (-117.9, 357.6)
% Diff $>$ 0	85%	70%	81%

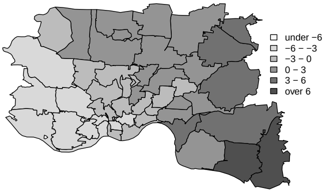
LPML Diff $=$ LPML of Model 3 $-$ LPML of Model 2. Mean and $(0.025, 0.975)$ quantile from 100 replicates are reported. % Diff $>$ 0 is calculated as percentage of LPML Diff $>$ 0 over 100 replicates.

Maps of posterior means for the 45 spatial frailties

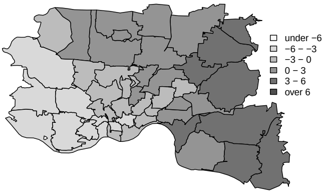
Simulated Frailties



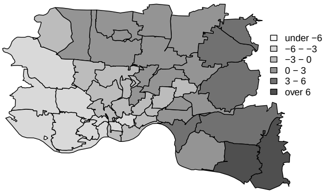
Model 1



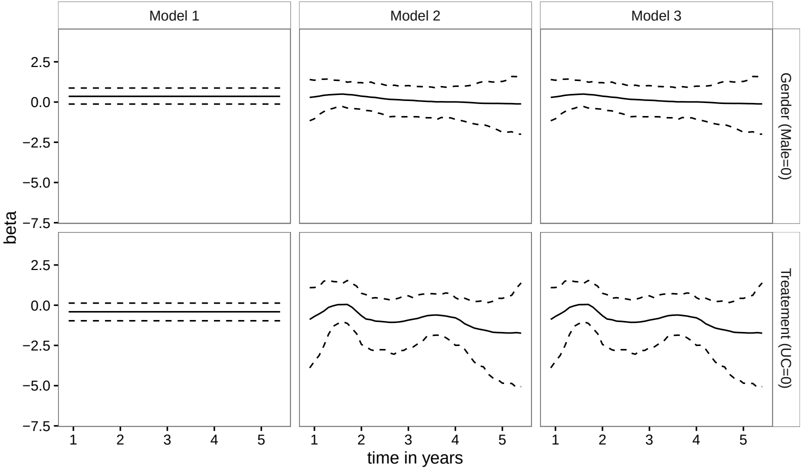
Model 2



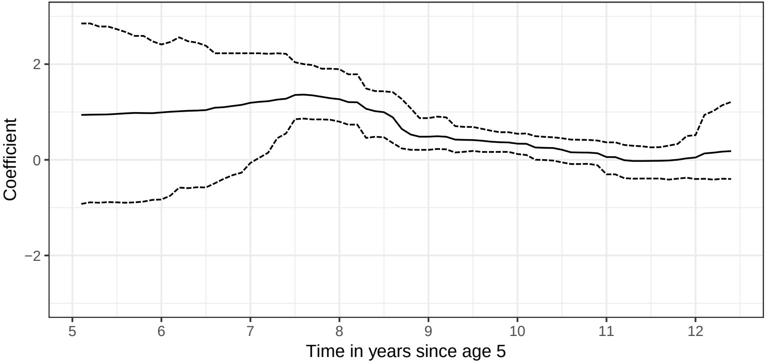
Model 3



Smoking data: Coefficient of *Gender* and *Treatment*

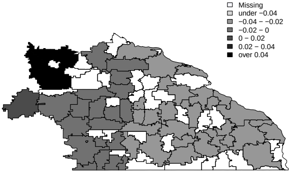


Tooth data (Zhang et al., 2018)

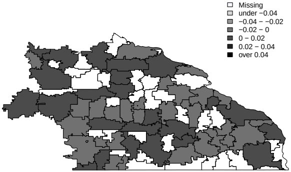


Posterior mean of frailties

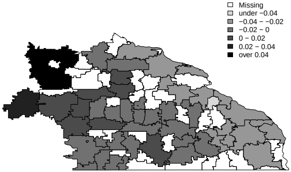
Model 1



Model 2



Model 3



Summary

- Reversible jump MCMC is a powerful tool to deal with model dimensionality, i.e., smooth the time-varying curve in this project.

- Spatial correlation needs to be considered.

- Current & future work
  - Improvement of reversible jump MCMC
  - Spline model?
  - Sample size calculation

Reference

Banerjee S, Wall MM, Carlin BP (2003) Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. Biostatistics 4(1):123–142

Besag J, Kooperberg C (1995) On conditional and intrinsic autoregressions. Biometrika 82(4):733–746

Cai T, Betensky RA (2003) Hazard regression for interval-censored data with penalized spline. Biometrics 59(3):570–579

Carlin BP, Louis TA (1997) Bayes and empirical Bayes methods for data analysis. Statistics and Computing 7(2):153–154

Green PJ (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82(4):711–732

Kneib T (2006) Mixed model based inference in structured additive regression. PhD thesis, lmu

Murray RP, Anthonisen NR, Connett JE, Wise RA, Lindgren PG, Greene PG, Nides MA, Group LHSR, et al. (1998) Effects of multiple attempts to quit smoking and relapses to smoking on pulmonary function. Journal of Clinical Epidemiology 51(12):1317–1326

Sinha D, Chen MH, Ghosh SK (1999) Bayesian analysis and model selection for interval-censored survival data. Biometrics 55(2):585–590

Sun J (2007) The Statistical Analysis of Interval-censored Failure Time Data. Springer Science & Business Media

Yu Z, Liu L, Bravata DM, Williams LS, Tepper RS (2013) A semiparametric recurrent events model with time-varying coefficients. Statistics in Medicine 32(6):1016–1026

Zhang Y, Wang X, Zhang B (2018) Bayesian approach for clustered interval-censored data with time-varying variate effects. Statistics and Its Interface