Frailty Probit model for multivariate and clustered interval-censored failure time data

Lianming Wang

University of South Carolina
Department of Statistics

June 4, 2013

Outline

- Introduction
- Proposed models
- Simulation studies
- Data analysis
- Discussion

Basic concepts

- Interval-censored data: The failure time of interest can not be observed exactly but is known to fall within some time interval.
- Clustered interval-censored data: One failure time event. The failure time observations are correlated because they are in the same cluster.
- Multivariate interval-censored data: Multiple failure time events. Subjects are independent.

Sexually transmitted infection (STI) data

- STIs are prevalent in the US population, especially among young people aged 15-24.
- STIs can cause many serious problems such as pelvic inflammatory disease, ectopic pregnancy, tubal infertility, preterm birth, and increased susceptibility to HIV.
- A longitudinal study referred to as the Young Women's Project (YWP) conducted between year 1999 and year 2008.
- Young women aged 14 to 17 years old participated the study, regardless of prior sexual experience.
- At enrollment, participants had face-to-face interviews and took STI tests.
- They were scheduled to visits and tests every 3 months.

Research objectives

- Study the times to infections and reinfections of each type of STIs.
- Focus on three types: Chlamydia trachomatis (CT), Neisseria gonorrhoeae (NG), and Trichomonas vaginalis (TV).
  - Estimate the risk incidence functions;
  - Estimate the covariate effects on infections. Potential covariates: race, age at enrollment, infection history, number of sexual partners, etc.
- Interval-censored data available for each time to infection or re-infection.

Two types of interval-censored data

- Focus on all the infections and reinfections for a specific type of STI:
  - Different women have different numbers of infections.
  - Each woman forms a cluster. Times to new infections from the same woman are correlated.
  - Times to new infections from different women are independent.
  - Clustered interval-censored data.
- Joint analysis of times to first infections of CT, GC, and TV:
  - Each woman has one observed interval for each infection.
  - The three types of infections are correlated.
  - Multivariate interval-censored data.

Existing approaches for **clustered** interval-censored data

- Weibull model
  - Bellamy et al. (2004) and Wong et al.(2005): normal frailty.
  - Goethals et al. (2005): gamma frailty.
  - Zhang and Sun (2010): informative cluster size.
  - Lam et al. (2010): multiple imputation method.
- Cox model
  - Wong et al. (2006): gamma frailty.
  - Kim (2010): joint modeling approach.
  - Kor, Cheng, and Chen (2013) frailty PH model.
- Additive hazards model: Li et al. (2012)

Proposed model

- The semiparametric normal frailty Probit model:

$$F(t|\mathbf{x},\xi) = \Phi\{\alpha(t) + \mathbf{x}'\boldsymbol{\beta} + \xi\},$$

  - α is an unknown increasing function.
  - β is the coefficient of predictor $\mathbf{x}$.
  - $\xi \sim N(0, \sigma_\xi^2)$ is the frailty term.

- The marginal distribution of T is a Probit model of Lin and Wang (2010):

$$\tilde{F}(t|\mathbf{x}) = \Phi\{\alpha^*(t) + \mathbf{x}'\boldsymbol{\beta}^*\},$$

$\alpha^*(t) = c\alpha(t)$, $\boldsymbol{\beta}^* = c\boldsymbol{\beta}$, and $c = (1+\sigma_\xi^2)^{-1/2}$.

- The conditional covariate effects given the frailty are proportional to the marginal covaraite effects.

- The intra-cluster association for clustered data under the normal frailty Probit model is characterized by Spearman's correlation coefficient ρ_s and median concordance κ

$$\rho_s = 6\pi^{-1}\sin^{-1}(r/2) \quad \text{and} \quad \kappa = 2\pi^{-1}\sin^{-1}(r),$$

where $r = \sigma_\xi^2/(1+\sigma_\xi^2)$ is the Pearson's coefficient of correlation.

- The same results hold for multivariate survival times under the normal frailty Probit model. In addition, Kendall's τ takes the same from as κ, i.e.,

$$\tau = 2\pi^{-1}\sin^{-1}(r).$$

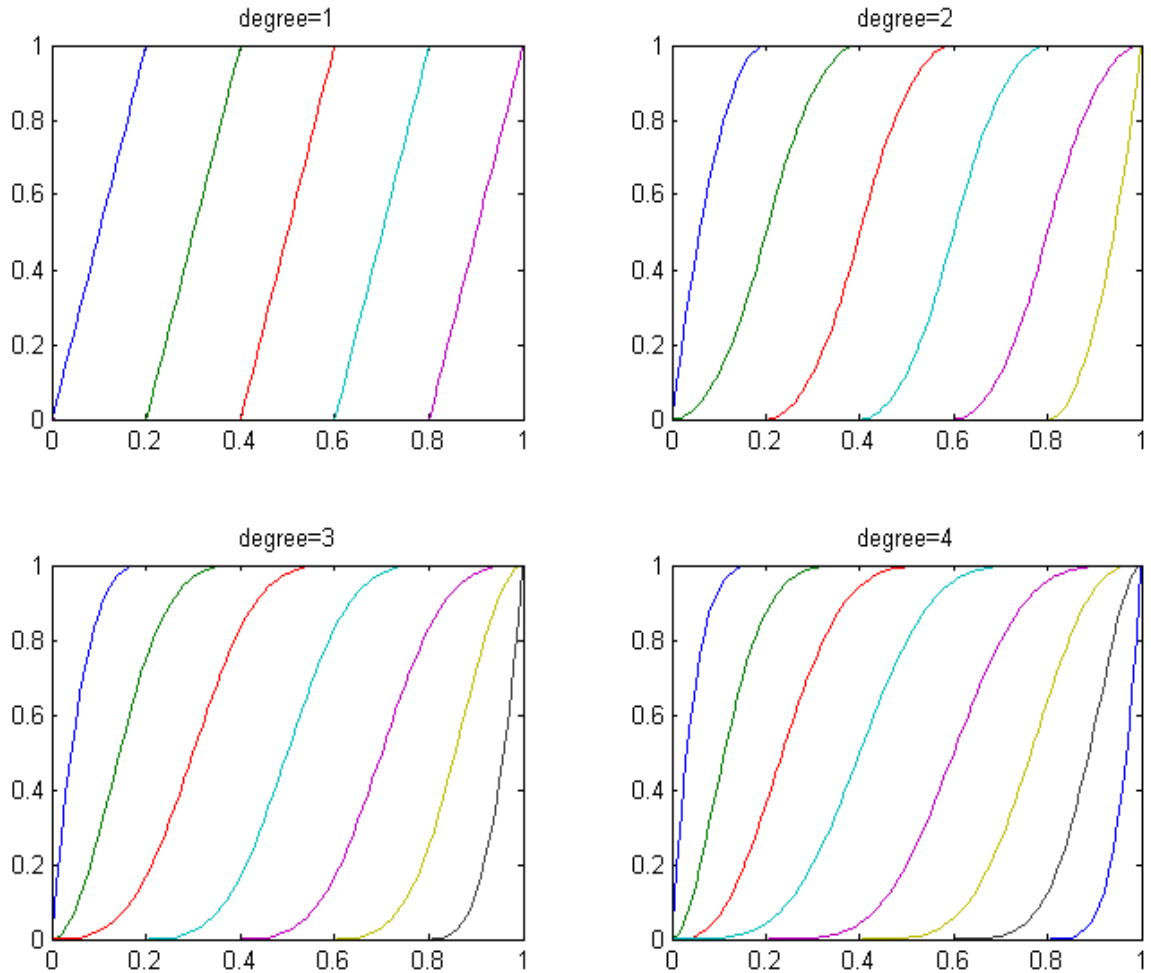
Modeling $\alpha(t)$

- α is an unknown increasing function with $\alpha(0) = -\infty$ and $\alpha(\infty) = \infty$. It is modeled over the observed data range.
- Modeling $\alpha(t)$ with monotone splines (Ramsay, 1988):

$$\alpha(t) = \gamma_0 + \sum_{l=1}^{k} \gamma_l b_l(t),$$

  where $\{b_l, l = 1, \cdots k\}$ are monotone spline bases.
- Spline functions are determine by knots and degree.
- γ_l is restrict to be nonnegative for $l \geq 1$ and γ_0 is unconstrained.

Figure: I-Spline basis functions



Likelihoods

- The observed data $\{\mathbf{x}_{ij}, L_{ij}, R_{ij}\}$, where $\mathbf{x}_{ij}$ is the predictor and $(L_{ij}, R_{ij}]$ is the observed interval for T_{ij} for the jth subject in cluster i, $j = 1, ..., n_i$ and $i = 1, ..., n$.
- The observed likelihood is

$$L_{obs} = \prod_{i=1}^{n} \int \pi(\xi_i) \prod_{j=1}^{n_i} \{F(R_{ij}|\mathbf{x}_{ij}, \xi_i) - F(L_{ij}|\mathbf{x}_{ij}, \xi_i)\} d\xi_i.$$

- The conditional likelihood given the frailties and covariates is

$$L = \prod_{i=1}^{n} \left[\prod_{j=1}^{n_i} F(R_{ij}|\mathbf{x}_{ij}, \xi_i)^{\delta_{ij1}} \{F(R_{ij}|\mathbf{x}_{ij}, \xi_i) - F(L_{ij}|\mathbf{x}_{ij}, \xi_i)\}^{\delta_{ij2}} \right.$$
$$\left. \{1 - F(L_{ij}|\mathbf{x}_{ij}, \xi_i)\}^{\delta_{ij3}} \right] f(\xi_i).$$

.

Data augmentation

- Introduce latent variable z_{ij} for each i and j,

$$z_{ij} \sim N(\alpha(t_{ij}) + \mathbf{x}'_{ij}\boldsymbol{\beta} + \xi_i, 1),$$

where $t_{ij} = R_{ij}1_{(\delta_{ij1}=1)} + L_{ij}1_{(\delta_{ij1}=0)}$.

- The augmented data likelihood function is

$$L_{aug} = \prod_{i=1}^{n}\left[\prod_{j=1}^{m_i}\phi\{z_{ij}-\alpha(t_{ij})-\mathbf{x}'_{ij}\boldsymbol{\beta}-\xi_i\}1_{C_{ij}}(z_{ij})\right]\sigma^{-1}\phi(\sigma^{-1}\xi_i),$$

where C_{ij} is the constrained space of z_{ij} and takes $(0,\infty)$ if $\delta_{ij1}=1$, $(\alpha(L_{ij})-\alpha(R_{ij}),0)$ if $\delta_{ij2}=1$, and $(-\infty,0)$ if $\delta_{ij3}=1$.

Prior specifications

- A multivariate normal prior $N(\boldsymbol{\beta}_0, \Sigma_0)$ for $\boldsymbol{\beta}$.
- Normal prior $N(m_0, \nu_0^{-1})$ for the unconstrained γ_0.
- Independent exponential prior $Exp(\eta)$ for all $\{\gamma_l\}_{l=1}^{k}$.
- Gamma prior $\mathcal{G}a(a_\eta, b_\eta)$ for η.
- Gamma prior $\mathcal{G}a(a, b)$ for σ_ξ^{-2}.

Proposed Gibbs Sampler

- Sample latent vairables z_{ij} from a truncated normal,

$$N(\alpha(t_{ij}) + \mathbf{x}'_{ij}\boldsymbol{\beta} + \xi_i, 1)1_{C_{ij}}(z_{ij}).$$

- Sample $\boldsymbol{\beta}$ from $N(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$, where

$$\hat{\Sigma} = (\Sigma_0^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\mathbf{x}'_{ij}\mathbf{x}_{ij})^{-1}$$

and

$$\hat{\boldsymbol{\beta}} = \hat{\Sigma}\left[\Sigma_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{z_{ij} - \alpha(t_{ij}) - \xi_i\}\mathbf{x}_{ij}\right].$$

Proposed Gibbs sampler (continued)

- Sample γ_0 from $N(E_0, W_0^{-1})$ where $W_0 = \nu_0 + n$ and

$$E_0 = W_0^{-1}\left[\nu_0 m_0 + \sum_{i=1}^{n}\sum_{j=1}^{n_i}[z_{ij} - \sum_{l=1}^{k}\gamma_l b_l(t_{ij}) - \mathbf{x}'_{ij}\boldsymbol{\beta} - \xi_i]\right].$$

- Sample γ_l for each $l \geq 1$, let $W_l = \sum_{i=1}^{n}\sum_{j=1}^{n_i} b_l^2(t_{ij})$.
  - If $W_l = 0$, sample γ_l from the prior $Exp(\eta)$.
  - If $W_l > 0$, sample γ_l from $N(E_l, W_l^{-1})1_{\{\gamma_l > \max(c_l^*, 0)\}}$, where

$$E_l = W_l^{-1}\left[\sum_{i=1}^{n}\sum_{j=1}^{n_i} b_l(t_{ij})[z_{ij} - \gamma_0 - \sum_{l' \neq l}\gamma_{l'} b_{l'}(t_{ij}) - \mathbf{x}'_{ij}\boldsymbol{\beta} - \xi_i] - \eta\right],$$

$$c_l^* = \max_{i:\ \delta_{ij2}=1}\left[\frac{-z_{ij} - \sum_{l' \neq l}\gamma_{l'}(b_{l'}(R_{ij}) - b_{l'}(L_{ij}))}{b_l(R_{ij}) - b_l(L_{ij})}\right].$$

Proposed Gibbs sampler (continued)

- Sample ξ_i for $i = 1, ..., n$ from $N(\mu_i, \sigma_i^2)$ where $\sigma_i^2 = (n_i + \sigma_\xi^{-2})^{-1}$ and

$$\mu_i = \sigma_i^2 \left[\mu_\xi / \sigma_\xi^2 + \sum_{j=1}^{n_i} \{ z_{ij} - \alpha(t_{ij}) - \mathbf{x}'_{ij}\boldsymbol{\beta} \} \right].$$

- Sample σ_ξ^{-2} from $\mathcal{G}a(a + n/2, b + 1/2 \sum_{i=1}^{n} (\xi_i - \mu_\xi)^2)$.
- Sample η from $\mathcal{G}a(a_\eta + k, b_\eta + \sum_{l=1}^{k} \gamma_l)$.

All the unknowns are sampled from their full conditional distributions in closed-form.

Proposed model for multivariate survival times

- The semiparametric normal frailty Probit model:

$$T_j|\xi \sim F_j(t|\mathbf{x},\xi) = \Phi\{\alpha_j(t) + \mathbf{x}'\boldsymbol{\beta} + c_j\xi\},\ j = 1,\cdots,J,$$

- T_j has a marginal semiparametric Probit model.
- c_j's are unknown constants except $c_1 = 1$ for identifiability purpose.
- Having c_js in the model allows different pairs of events to have different correlation.
- Kendall's τ between T_j and T_k is

$$\tau = 2\pi^{-1}\sin^{-1}(r_{jk}),$$

where $r_{jk} = \dfrac{c_j c_k \sigma_\xi^2}{\sqrt{(1+c_j^2\sigma_\xi^2)(1+c_k^2\sigma_\xi^2)}}$ for $j \neq k$.

Simulation setups

- Generate data from $F(t|\mathbf{x}, \xi) = \Phi\{\alpha(t) + \mathbf{x}'\boldsymbol{\beta} + \xi\}$.
- $x_1 \sim N(0, 1)$, $x_2 \sim Bernoulli(0.5)$.
- True $\beta_1 = 1$ or 0, $\beta_2 = -1, 0,$ or 1.
- True $\alpha(t) = 2\log(t) + 1 + t$
- Frailty $\xi_i \sim N(0, 0.4^2)$.
- Generated 100 data sets, each with sample size $N = 200$ or 50 clusters, within each cluster there are 4 observations
- Take 15 equal spaced knots and degree equal to 2 for specifying I spline basis.

Table: simulation result: $\xi \sim N(0, 0.4^2)$

	n=50				n=200			
True	*POINT*	ESE	SSD	CP95	*POINT*	ESE	SSD	CP95
β_1=0	0.0192	0.1077	0.1183	0.94	-0.0005	0.0514	0.0505	0.95
β_2=0	-0.0380	0.2113	0.2278	0.94	0.0121	0.1020	0.0972	0.96
$\sigma = 0.4$	0.6209	0.1263	0.0802	0.37	0.5001	0.0676	0.0494	0.74
β_1=0	0.0053	0.1064	0.1158	0.93	0.0027	0.0507	0.0563	0.92
β_2=-1	-1.1178	0.2282	0.2200	0.94	-0.9906	0.1074	0.1087	0.98
$\sigma = 0.4$	0.6188	0.1230	0.0801	0.47	0.4854	0.0658	0.0491	0.84
β_1=0	-0.0220	0.1100	0.1050	0.96	0.0004	0.0531	0.0505	0.97
β_2=1	1.0828	0.2392	0.25	0.92	1.0414	0.1147	0.1108	0.95
$\sigma = 0.4$	0.6460	0.1327	0.0823	0.34	0.5066	0.0702	0.0540	0.73
β_1=1	1.1008	0.1464	0.1516	0.91	1.0326	0.0690	0.0758	0.91
β_2=0	-0.0002	0.2196	0.2401	0.96	0.0042	0.1048	0.1090	0.91
$\sigma = 0.4$	0.6585	0.1344	0.0895	0.33	0.5075	0.0695	0.0527	0.73
β_1=1	1.1043	0.1450	0.1422	0.91	1.0299	0.0684	0.0802	0.90
β_2=-1	-1.1217	0.2343	0.2661	0.90	-1.0127	0.1112	0.1118	0.95
$\sigma = 0.4$	0.6477	0.1318	0.0765	0.28	0.5038	0.0688	0.0480	0.72
β_1=1	1.1129	0.1567	0.1513	0.91	1.0313	0.0751	0.0719	0.92
β_2=1	1.0539	0.2397	0.2690	0.93	1.0344	0.1191	0.1173	0.95
$\sigma = 0.4$	0.6380	0.1314	0.0911	0.43	0.5528	0.0562	0.0737	0.66

Additional simulation

- Also consider two cases of misspecified frailty distribution,
  - Mixture normal : $0.45 * N(0.5, 0.4^2) + 0.55 * N(-0.5, 0.18^2)$
  - Exponential Gamma: $\exp(\xi_2) \sim \mathcal{G}a(1, 1)$
- All others are the same as in the original simulation setup.

Table: Scenario I: $\xi_1 \sim 0.45 * N(0.5, 0.4) + 0.55 * N(-0.5, 0.18)$ and scenario II:$\exp(\xi_2) \sim \mathcal{G}a(1, 1)$.

	Scenario I				Scenario II			
True	*POINT*	ESE	SSD	CP95	*POINT*	ESE	SSD	CP95
β_1=0	0.0070	0.1067	0.1095	0.93	-0.0148	0.1140	0.1059	0.97
β_2=0	-0.0481	0.2131	0.2157	0.95	-0.0025	0.2258	0.2216	0.94
β_1=0	0.0035	0.1064	0.1129	0.95	0.0098	0.1161	0.1149	0.94
β_2=-1	-1.1037	0.2307	0.2418	0.94	-1.0449	0.2467	0.2338	0.97
β_1=0	0.0056	0.1103	0.1128	0.92	0.0189	0.1154	0.0987	0.96
β_2=1	1.0381	0.2361	0.2535	0.92	1.0185	0.2474	0.2513	0.95
β_1=1	1.0818	0.1444	0.1481	0.93	1.0515	0.1584	0.1785	0.91
β_2=0	0.0321	0.2189	0.2462	0.94	-0.0054	0.2326	0.2365	0.95
β_1=1	1.0877	0.1440	0.1461	0.92	1.0370	0.1585	0.1644	0.96
β_2=-1	-1.0905	0.2360	0.2674	0.91	-1.0283	0.2534	0.2457	0.97
β_1=1	1.0840	0.1488	0.1481	0.94	1.0428	0.1590	0.1729	0.96
β_2=1	1.0269	0.2405	0.2761	0.90	1.0189	0.2537	0.2682	0.94

Mastitis data (Goethals et. al, 2009)

- A total of 100 cows were studied right after giving birth and were examined roughly monthly for udder infection.
- Why to study udder infection: udder infection is known to be associated with reduced milk yield and poor milk quality.
- Each cow has four udder quarters. The infection status of each udder quarter was obtained at each examination.
- The response of interest is the infection time of each udder quarter, which is interval-censored.
- Each cow forms a natural cluster, and the infection times of the four udder quarters from the same cow are correlated.

- Objectives: to study the effects of the number of calvings and the position of udder quarter on the infection time as well as the estimation of the cumulative incidence of the udder infection.
- Covariates to consider:
  - x_1 is the position of the udder quarter (changes within a cow).
  - x_2, x_3 are dummy variables for number of calvings (change between cows).

Table: Result: mastitis data

	Mean	95% CI
β_1	-0.2286	(-0.4624, 0.0036)
β_2	-0.0316	(-0.3621, 0.3002)
β_3	0.1402	(-0.0788, 0.3598)
σ	0.5300	(0.4079, 0.6673)
ρ_s	0.2102	(0.1364, 0.2954)
κ	0.1411	(0.0911, 0.1994)

Lymphatic filariasis data (Williamson et al. 2008)

- A disease caused by Wuchereria bancrofti and transmitted by infectious mosquitoes.
- Bancrofti larvae grow into adult worms living in lymphatic vessels of people.
- Ultrasound was used to visualize the status of the worms.
- A study conducted in Brazil (Dreye et al., 2006).
- 78 worm nests were found in 47 patients that were detected to have lymphatic filariasis.
- Two treatment groups: DEC/ALB combination and DEC alone.
- Ultrasound examinations were taken at 7, 14, 30, 45, 60, 90, 180, 270, and 365 days.
- Objective: to study whether DEC/ALB is more effective to clear the worms than DEC alone.

Consider two covariates

- x_1: 1 for DEC/ALB and 0 for DEC alone.
- x_2: age of patient at the enrollment.

Table: Data analysis of lymphatic filariasis

	Mean	95% CI
β_1	1.1775	(-0.1779, 2.6091)
β_2	0.2160	(-0.5250, 1.0018)
σ	2.1601	(1.2813, 3.3849)
ρ_s	0.7953	(0.6087, 0.9129)
κ	0.6064	(0.4312, 0.7436)

Concluding remarks

- Proposed a normal frailty Probit model for modeling clustered and multivariate survival times.
- The normal frailty Probit model has good properties.
- Proposed fully Bayesian methods for analyzing clustered and multivariate interval-censored data.
- Developed efficient Gibbs samplers that do not involve Metropolis Hastings steps.
- Allow to estimate within-cluster correlation or pairwise correlations in explicit form.

Acknowledgement

- Funded by social science grant program at USC.
- Collaborators
  - Haifeng Wu, PhD student at USC
  - Wanzhu Tu, Professor at Indiana University
  - Iris Lin, Assistant professor at USC
  - Sean Wang, PhD student at USC