1-1-2013

Models and Software Development For Interval-Censored Data

Chun Pan
University of South Carolina

Models and Software Development for Interval-Censored Data

by

Chun Pan

Bachelor of Science
Nanjing University 2001

Master of Science
Nanjing University 2004

Master of Science
University of South Carolina 2009

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Biostatistics

Arnold School of Public Health

University of South Carolina

2013

Accepted by:

Bo Cai, Major Professor

Lianming Wang, Committee Member

Jiajia Zhang, Committee Member

Andrew Ortaglia, Committee Member

Lacy Ford, Vice Provost and Dean of Graduate Studies

Acknowledgments

This dissertation is a summary of the research work I have done under the supervision of my advisor Dr. Bo Cai. I am deeply grateful to him for his insightful, intelligent, and patient guidance and his creative suggestions when I encountered difficuties in the past three years. I sincerely appreciate all of his great support along the way as I would not be able to make it without him.

I am very grateful to Dr. Lianming Wang for instructing me with his expertise on interval-censored data. I have learned a lot about the concept and modeling methodology for this type of data by meeting with him and studying his R programs and publications.

I would like to express my sincere gratitude to Dr. Jiajia Zhang for her enlightening comments on my work. Especially worth to mention, I have been stuck at my second project for months and one day Dr. Zhang's comment solved the riddle for me immediately.

Dr. Andrew Ortaglia has gone through the writing of my dissertation in detail. Also he has been so cordial and supportive both during the completion of the dissertation and during my course study, which has really encouraged me a lot when I came to barriers.

I also would like to thank Dr. Suzanne McDermott and Dr. Joshua Mann for their financial support and advice during my PhD study. Finally, I thank my parents and brother in China for their unconditional love and support.

Abstract

Interval-censored time-to-event data occur naturally in studies of diseases where the symptoms are not directly observable, and periodic lab or clinical examinations are required for detection. Due to the lack of well-established procedures, interval-censored data have been conventionally treated as right-censored data, however, this introduces bias at the first place. This dissertation focuses on methodological research and software development for interval-censored data. Specifically, it consists of three projects. The first project is to create an R package for regression analysis and survival curve estimation of interval-censored data based on several published papers by our research team. In the second project, a Bayesian semiparametric proportional hazards model with spatial random effect is developed for spatially correlated interval-censored data. In the third project, we propose a multivariate frailty model for clustered interval-censored failure times, which is analogous to a mixed model in regression analysis.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Interval-Censoring

Interval-censored data occur naturally in studies of diseases where the symptoms of interest are not directly observable, and laboratrory or clinical examinations are required for detection. The exact time to event of interest T is not directly observed, but is known to fall within a time interval $(L, R]$, such that $0 \leq L < T \leq R \leq \infty$.

Consider in a tumorigenicity study, a lab animal has to be dissected to check whether a tumor has developed. Let C denote the time of dissection, and T denote the true tumor onset time, then the data observed is $(C, 1(T \leq C))$. Then

$$(L, R] = \begin{cases} (0, C], T \leq C \\ (C, \infty), T > C \end{cases} \tag{1.1}$$

This is called case 1 interval-censored or current status data.

Suppose there are two examination times U and V for each subject, then the data observed is $(U, V, \delta_1 = I(T \leq U), \delta_2 = I(U < T \leq V))$. Then

$$(L, R] = \begin{cases} (0, U], & T \leq U \\ (U, V], & U < T \leq V \\ (V, \infty), & T > C \end{cases} \tag{1.2}$$

This is the case 2 interval-censored data. Case k interval-censoring refers to when there are k examination times per subject.

Consider another situation. In an oncology clinical trial for non-small cell lung cancer, the endpoint of interest is progression-free survival (PFS). The patients are

scanned by CT every couple of weeks for evaluation of tumor sizes and new lesions. Then the scan results are read by a diagnostic radiologist to determine if progression has occurred or not. Then PFS is interval-censored as the exact time to progression is not observed but is known to fall within a time interval $(L, R]$. Suppose $O_i = \{O_{i1}, \ldots, O_{i,n_i}\}$ are the examination times for the ith patient, $i = 1, \ldots, n$. Then

$$(L, R] = \begin{cases} (0, O_{i,1}], & T \leq O_{i,1} \\ (O_{i,L}, O_{i,R}], & O_{i,L} < T \leq O_{i,R} \\ (O_{i,n_i}, \infty), & T > O_{i,n_i} \end{cases} \tag{1.3}$$

This is called general interval-censoring, which includes case 1, case 2 and case k interval-censoring as special cases. In the dissertation, we will focus on general interval-censoring.

1.2 Likelihood function

Throughout the dissertation, we assume the following two basic assumptions (Huang and Wellner, 1997). (A1) The failure time is independent of the examinations times given the covariates. (A2) The distribution of the examination times does not involve the parameters of interest.

Under these assumptions we can derive the likelihood function. For case 1, or equivalently current status data, the joint density of a single observation $(C, \delta = I(T \leq C), \boldsymbol{x})$ is:

$$f(\delta, c, \boldsymbol{x}) = f(\delta|c, \boldsymbol{x}) f(c, \boldsymbol{x}) = f(\delta|\boldsymbol{x}) f(c, \boldsymbol{x})$$
$$= F(c|\boldsymbol{x})^{\delta} (1 - F(c|\boldsymbol{x})^{1-\delta} f(c, \boldsymbol{x})),$$

where F is the cdf of T. So for an independent sample of size n from the same distribution, the likelihood function is proportional to:

$$L = \prod_{i=1}^{n} \left\{ F(c_i|\boldsymbol{x}_i)(1 - F(c_i|\boldsymbol{x}_i))^{1-\delta_i} \right\}.$$

For case 2 and general interval-censoring, let δ_1, δ_2, δ_3 denote left-, interval-, and right-censoring, the joint density of a single observation $(\delta_1, \delta_2, \delta_3, L, R, \boldsymbol{x})$ is:

$$f(\delta_1, \delta_2, \delta_3, L, R, \boldsymbol{x}) = f(\delta_1, \delta_2, \delta_3 | L, R, \underline{x}) f(L, R, \boldsymbol{x}) = f(\delta_1, \delta_2, \delta_3 | \boldsymbol{x}) f(L, R, \boldsymbol{x})$$
$$= F(R|\boldsymbol{x})^{\delta_1} (F(R|\boldsymbol{x}) - F(L|\boldsymbol{x}))^{\delta_2} (1 - F(L|\boldsymbol{x}))^{\delta_3} f(L, R, \boldsymbol{x}).$$

The likelihood function for an independent sample of size n from the same distribution is proportional to:

$$L = \prod_{i=1}^{n} \left\{ F(R_i|\boldsymbol{x}_i)^{\delta_{1i}} (F(R_i|\boldsymbol{x}_i) - F(L_i|\boldsymbol{x}_i))^{\delta_{2i}} (1 - F(L_i|\boldsymbol{x}_i))^{\delta_{3i}} \right\}.$$

1.3 Motivations

For interval-censored data, the conventional approach in pharmaceutical industry treats the right-point of the time interval as the observed time, and then apply the standard right-censored methods. However, this approach can lead to biased estimation and invalid inferences (Rücker and Messerer, 1988; Odell et al., 1992; Lindsey and Ryan, 1998; Sun and Chen, 2012). Better estimations can be obtained if the information of interval censorship is taken into account in modeling.

Quite a few new methods for analysis of interval-censored time-to-event data have been proposed in the last two decades. However, most of these methods either rely on parametric assumptions that are hard to verify in practice or are computationally challenging. As a result, none of them has been accepted by the pharmaceutical industry as a standard procedure. We propose to model the survival function semi-parametically through I-splines (Ramsay, 1988) and estimate survival function and regression coefficients through Markov chain Monte Carlo (MCMC) algorithm. Our models are relatively straightforward to implement in practice and give more accurate estimates. Specifically, the dissertation research consists of three projects; each has been initiated to address certain problems as specified below.

Currently, a few packages and SAS macros, mainly the 'interval' and 'glrt' packages in R and the %EMICM and the %ICSTEST macros in SAS, have been developed to provide estimation and comparison for survival functions for interval-censored data. However, there are currently few options available for fitting regression models, except the 'survBayes' package and the 'intcox' package. As the first project of this dissertation, we have built an R package, ICBayes, which fits proportional hazards (PH), proportional odds (PO), and probit regression models from a Bayesian perspective.

It is possible that failure times of interest are both interval-censored and spatially correlated. For instance, in a lung cancer clinical trial, patients are recruited from a number of regions where air quality in a region is similar to that in its neighbors but might differ substantially from that in the other regions. If air quality exerts an effect on treatment outcome, then survival times of patients may be spatially correlated. Only one spatial frailty model has been developed for such type of data under parametric Weibull PH cure rate model (Banerjee and Carlin, 2004). We propose a semiparametric spatial frailty PH model, which provides greater flexibility for modeling failure time.

Clustered interval-censored survival data can easily occur in multicenter clinical trials for cancer, HIV, or other infectious diseases. Tumor progression, HIV progression to AIDS, and presence/absence of an infection normally all need periodic lab examinations for detection. Characteristics that vary by center may affect survival times, which implies either an overall frailty to account for baseline hazard heterogeneity across centers or a frailty corresponding to a certain predictor to account for center-wise variation of this predictor's effect. We propose to model the variance of a potential frailty with a mixture of point mass and gamma distribution, because rather than arbitrarily assuming the heterogeneity structure, we want to actually estimate the probability that a frailty exists.

Chapter 2

Background Knowledge

2.1 The Poisson Distribution and the Poisson Process

2.1.1 Additive Property of Poisson Distribution Suppose $X_1, \ldots, X_n$ are independent with $X_i \sim \text{Poisson}(\lambda_i)$, $i = 1, \ldots, n$, then

$$Y_i = \sum_{i=1}^{n} X_i \sim \text{Poisson}(\sum_{i=1}^{n} \lambda_i).$$

This can be proved through the moment generating function method. This property will be used in introducing Poisson latent variables in our data augmentation.

2.1.2 Poisson Process A Poisson process is a counting process in which events occur continuously and independently of one another. Let $N(t)$ be the number of events in time interval $(0, t]$. The process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate (or intensity) λ, $\lambda > 0$, if

- $N(0) = 0$.
- (independent increment) the number of events in non-overlapping intervals are independent.
- (stationary increment) the number of events in a time interval depends only on the length of the interval.
- $\lim_{h \to 0} \frac{P(N(h)=1)}{h} = \lambda$.
- $\lim_{h \to 0} \frac{P(N(h) \geq 2)}{h} = 0$.

Some useful properties of Poisson process are:

- The number of events in an interval of length t is a Poisson random variable with mean λt.
- The interarrival times are independent $\exp(\lambda)$ random variables.
- The time of the nth event is Gamma(n, λ) random variable, where λ is the rate parameter.

2.1.3 Nonhomogeneous Poisson Process A nonhomogeneous Poisson process relaxes the stationarity assumption of a Poisson process. A process $\{N(t), t >= 0\}$ is a nonhomogeneous Poisson process with rate (intensity) $\lambda(t)$, $t \geq 0$, if

- $N(0) = 0$.
- (independent increment) the number of events in non-overlapping intervals are independent.
- $\lim_{h\to 0} \frac{P(N(t+h)-N(t)=1)}{h} = \lambda(t)$.
- $\lim_{h\to 0} \frac{P(N(t+h)-N(t)\geq 2)}{h} = 0$.

The mean-value function (cumulative intensity function) is defined as:

$$m(t) = \int_0^t \lambda(s)ds, \quad t \geq 0.$$

Some useful properties of a nonhomogeneous Poisson process are:

- For $t > 0$, $N(t) \sim$ Poisson$(m(t))$.
- $N(t + h) - N(t) \sim$ Poisson$(m(t + h) - m(t))$.
- For each $0 \leq t_1 < t_2 < \ldots < t_m$, $N(t_1), N(t_2) - N(t_1), \ldots, N(t_m) - N(t_{m-1})$ are independent Poisson random variables.

- Let S_n be the time of the nth occurrence, then $P\{S_n > t\} = P\{N(t) < n\} = \sum_{j=0}^{n-1} \frac{e^{-m(t)}(m(t))^j}{j!}$. This property will be particularly used to derive data augmentation in our models.

The relationship between Poisson process and nonhomogeneous Poisson process can be illustrated as follows. Suppose that events are occurring according to a Poisson process with rate λ, and the probability that an event at time t is counted is $p(t)$. Then the process of counted events constitutes a nonhomogeneous Poissson process with rate $\lambda(t) = \lambda \cdot p(t)$.

2.2 Monte Carlo Markov Chain Methods

2.2.1 Markov Chains A Markov chain is a stochastic process in which the next state depends only on the current state. This memoryless characteristic is called the Markov property. In mathematical form, consider a draw $\theta^{(t)}$ at iteration t, then the next draw $\theta^{(t+1)}$ at iteration $t+1$ only depends on the current draw $\theta^{(t)}$, and not on any past draws. So the draws in a Markov chain are slightly dependent.

Suppose we have a target distribution $f(\theta)$ and we want to estimate $E(g(\theta))$, where $g()$ is some function. Instead of solving it analytically as $I = \int_S g(\theta)f(\theta)d\theta$, where S is the parameter space (state space). We can approximate the integral via Monte Carlo integration by simulating M values from $f(\theta)$ and calculating $\widehat{I} = \frac{1}{M}\sum_{t=1}^{M} g(\theta^{(t)})$. If the M values are independent, then by Strong Law of Large Numbers (SLLN), $\widehat{I}$ is a consistent estimator of I. The Markov analog to the SLLN is the Ergodic Theorem. It is the reason why Markov chain Monte Carlo methods works: it allows us to use a sample path from a Markov chain to estimate various quantities of interest.

For a discrete Markov chain, it is defined to be irreducible if all its states communicate. A state i is recurrent if $E(V_i|\theta^{(0)} = i) = \infty$, where V_i is the number of visits in state i. Most MCMC methods are based on general state space Markov chains. Here we give the definition of irreducibility and recurrence for them.

Definition 1 (Irreducibility). Given a distribution μ on the state space S, a Markov chain is said to be f-irreducible if for all sets A with $f(A) > 0$ and for all $x \in S$, there exists an $m \in \mathbb{N}_0$ such that

$$P(\theta^{(t+m)} \in A | \theta^{(t)} = x) = \int_A K^{(m)}(x, y)dy > 0,$$

where $K^{(m)}(x, y)$ is the m-step transition kernel.

Definition 2 (Recurrence). A set $A \subset S$ is said to be recurrent for a Markov chain θ if for all $x \in A$

$$E(V_A | \theta^{(0)} = x) = \infty.$$

And a Markov chain is said to be recurrent if

i. The chain is irreducible.

ii. Every measurable set with $f(A) > 0$ is recurrent.

Theorem 1 (Ergodic Theorem). Let θ be a f-irreducible, recurrent $\mathbb{R}^d$-valued Markov chain with stationary distribution f. Then for any integrable function $g : \mathbb{R}^d \to \mathbb{R}$, we have

$$\lim_{M\to\infty} \frac{1}{M} \sum_{t=1}^{M} g(\theta^{(t)}) \to \int_S g(\theta) f(\theta) d\theta,$$

with probability 1 for almost every starting value $\theta^{(0)}$. If θ is Harris-recurrent, then this holds for every starting value $\theta^{(0)}$.

A MCMC method for the simulation of a distribution f is any method producing an ergodic Markov chain $(\theta^{(t)})$ whose stationary distribution is f. The two most commonly used MCMC algorithms are the Metropolis-Hastings algorithm and the Gibbs sampler.

2.2.2 Metropolis-Hastings Algorithm The Metropolis-Hastings algorithm enables generating random numbers from a probability distribution f without directly sampling from it. It constructs an ergodic Markov chain that satisfies the detailed balance property with respect to f, as shown in the following.

The transition kernel K is based on a proposal distribution $q(\theta|\theta^{(t)})$. Specifically, given the current state $\theta^{(t)}$, we generate a candidate value $\theta' \sim q(\theta|\theta^{(t)})$, and set the next state $\theta^{(t+1)}$ as:

$$\theta^{(t+1)} = \begin{cases} \theta', & w.p. \quad a(\theta^{(t)}, \theta') = \min(1, \frac{f(\theta')q(\theta^{(t)}|\theta')}{f(\theta^{(t)})q(\theta'|\theta^{(t)})}) \\ \theta^{(t)}, & w.p. \quad 1 - a(\theta^{(t)}, \theta') \end{cases} \tag{2.1}$$

This leads to the following transition kernel

$$K(\theta^{(t)}, \theta^{(t+1)}) = q(\theta^{(t+1)}|\theta^{(t)})a(\theta^{(t)}, \theta^{(t+1)}) + (1 - \sum_{\theta'} a(\theta^{(t)}, \theta')q(\theta'|\theta^{(t)}))\delta_{\theta^{(t)}}(\theta^{(t+1)}).$$

It is straightforward to show that the two additive items in the equation above satisfy the detailed balance condition, so K is in detailed balance with respect to f:

$$K(\theta^{(t-1)}, \theta^{(t)})f(\theta^{(t-1)}) = K(\theta^{(t)}, \theta^{(t-1)})f(\theta^{(t)}).$$

Thus f is the stationary distribution of the Markov chain $(\underline{\theta}^{(0)}, \underline{\theta}^{(1)}, \ldots)$. Furthermore, if $q(\theta|\theta^{(t)}) > 0$ for all $\theta, \quad \theta^{(t)} \in S$, then the Makov chain is irreducible. And since the chain is Harris recurrent if it is irreducible (Tieney, 1994), then it is Ergodic.

2.2.3 Gibbs Sampler Suppose we have a joint distribution $f(\theta_1, \theta_2, \ldots, \theta_k)$ that we want to sample from. We can use the Gibbs sampler (Geman and Geman, 1984) to sample from it if we know the full conditional distributions for each parameter. The full conditional distribution, $f(\theta_j|\theta_{-j}, y)$, is the distribution of the parameter conditional on data and all the other parameters.

The Gibbs sampler iterates through the following steps:

i Start with initial values $\underline{\theta}^{(0)}$.

ii Draw a value $\theta_1^{(1)}$ from its full conditional distribution $f(\theta_1|\theta_2^{(0)}, \ldots, \theta_k^{(0)}, y)$.

iii Draw a value $\theta_2^{(1)}$ from its full conditional distribution $f(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}, y)$.

iv Continue for $\theta_3, \ldots, \theta_k$. Then we get the first draw $\boldsymbol{\theta}^{(1)}$.

v Repeat steps ii-iv for M iterations.

Then we get the Markov chain $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)})$.

The Hammersley-Clifford Theorem (Robert and Casella, 1999, p298) proved that if $(\theta_1, \theta_2, \ldots, \theta_k)$ satisfy the positivity condition and have joint density, then the full conditionals fully specify the joint distribution. It can be shown that $f(\theta_1, \theta_2, \ldots, \theta_k)$ is the invariant distribution of the Markov chain $(\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots)$ generated by the Gibbs sampler. Also, if the joint density $f(\theta_1, \theta_2, \ldots, \theta_k)$ satisfies the positivity condition, then the Gibbs sampler yields an irreducible, recurrent Markov chain. Hence the chain is ergodic.

Definition 3 (Positivity condition). Random variables $(X_1, \ldots, X_n)$ with joint density $f(x_1, \ldots, x_n)$ and marginal densities $f_{X_i}(x_i)$ is said to satisfy the positivity condition if $f_{X_i}(x_i) > 0$ for all $i = 1, \ldots, n$ implies $f(x_1, \ldots, x_n) > 0$.

2.3 I-Splines

2.3.1 Splines A spline function f is a piecewise polynomial function defined on an interval $[x_{min}, x_{max}]$ with specified continuity constraint. The interval is subdivided by a mesh of points $x_{min} = \xi_1 < \ldots < \xi_q = x_{max}$, and within each subinterval $[\xi_j, \xi_{j+1})$ is a polynomial P_j of order k.

Let n denote the number of free parameters that specify the spline function. A knot sequence $t = \{t_1, \ldots, t_{n+k}\}$ is derived by placing knots at boundary values ξ_i according to the order of continuity at that boundary. The simplest case is to put a single knot at each boundary value, which implies that the order of continuity is $(k-1)$ and that $(k-2)$ derivatives match at boundary points ξ_i. So the knots are $t_1 = \ldots = t_k = x_{min}$, $t_{k+1} = \xi_2$, ..., $t_{n+1} = \ldots = t_{n+k} = x_{max}$. Under this simplest case, the number of free parameters n equals the number of interior knots plus the order k.

2.3.2 M-Splines With a suitable set of basis splines, for instance, the M-spline family $M_i(x|k,t)$, we can construct a spline f as a linear combination:

$$f = \sum_{i=1}^{n} a_i M_i.$$

A set of M-splines can be defined by the following recursions:

For $k = 1$, $M_i(x|k=1,t) = \frac{1}{t_{i+1}-t_i}$; 0 otherwise.

For $k > 1$, $M_i(x|k,t) = \frac{k[(x-t_i)M_i(x|k-1,t)+(t_{i+k}-x)M_{i+1}(x|k-1,t)]}{(k-1)(t_{i+k}-t_i)}$.

As we can see, $M_i(x|k,t) > 0$ only if $t_i \leq x \leq t_{i+k}$. It is an important property, as it implies that the coefficient a_i will only affect f within this subinterval. Other useful properties include: 1) $M_i \geq 0$; 2) M_i integrates to 1; 3) M_i has $k-2$ continuous derivatives at interior knots.

Figure 2.1 shows an M-spline of order 3 on the interval [0,1], with three interior knots at 0.3, 0.5, 0.6, and a spline based on the M-splines.

2.3.3 I-Splines Because M-splines are nonnegative, monotone splines can be derived based on them using integration:

$$I_i(x|k,t) = \int_{x_{min}}^{x} M_i(u|k,t)du.$$

This is called I-splines. Each I_i is a piecewise polynomial of degree k (order k), since each M_i is a piecewise polynomial of degree $k-1$ (order k).

For a simple knot sequence, an I-spline I_i of order k can be expressed in terms of the M-splines M_i of order $k+1$:

$$I_i(x|k,t) = \begin{cases} 0, & i > j \\ \sum_{m=i}^{j}(t_{m+k+1} - t_m)M_m(x|k+1,t)/(k+1), & j-k+1 \leq i \leq j \\ 1, & i < j-k+1. \end{cases} \tag{2.2}$$

Figure 2.2 shows the I-splines that were derived from the M-splines above, and a monotone spline based on the I-splines.

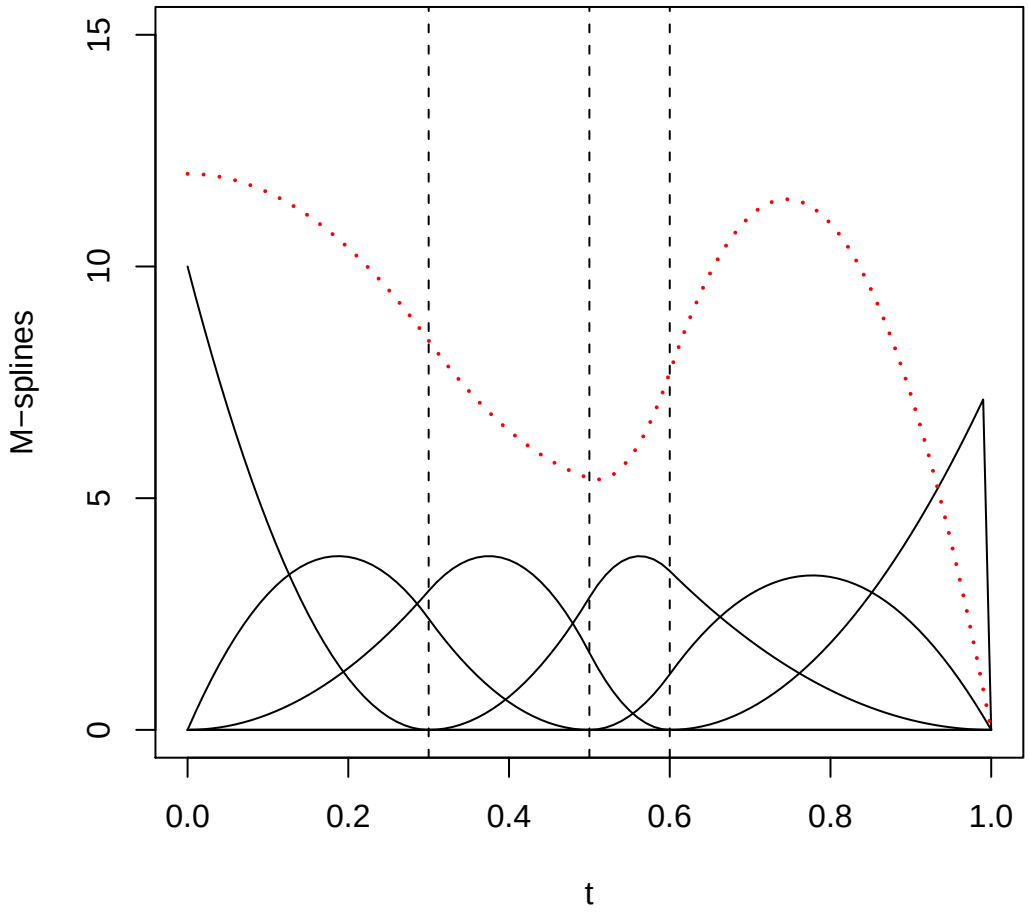

Figure 2.1 A set of M-splines with order = 3, $t = \{0, 0.3, 0.5, 0.6, 1\}$, and $f = 1.2M_1 + 2.0M_2 + 1.2M_3 + 1.2M_4 + 3.0M_5 + 0.0M_6$

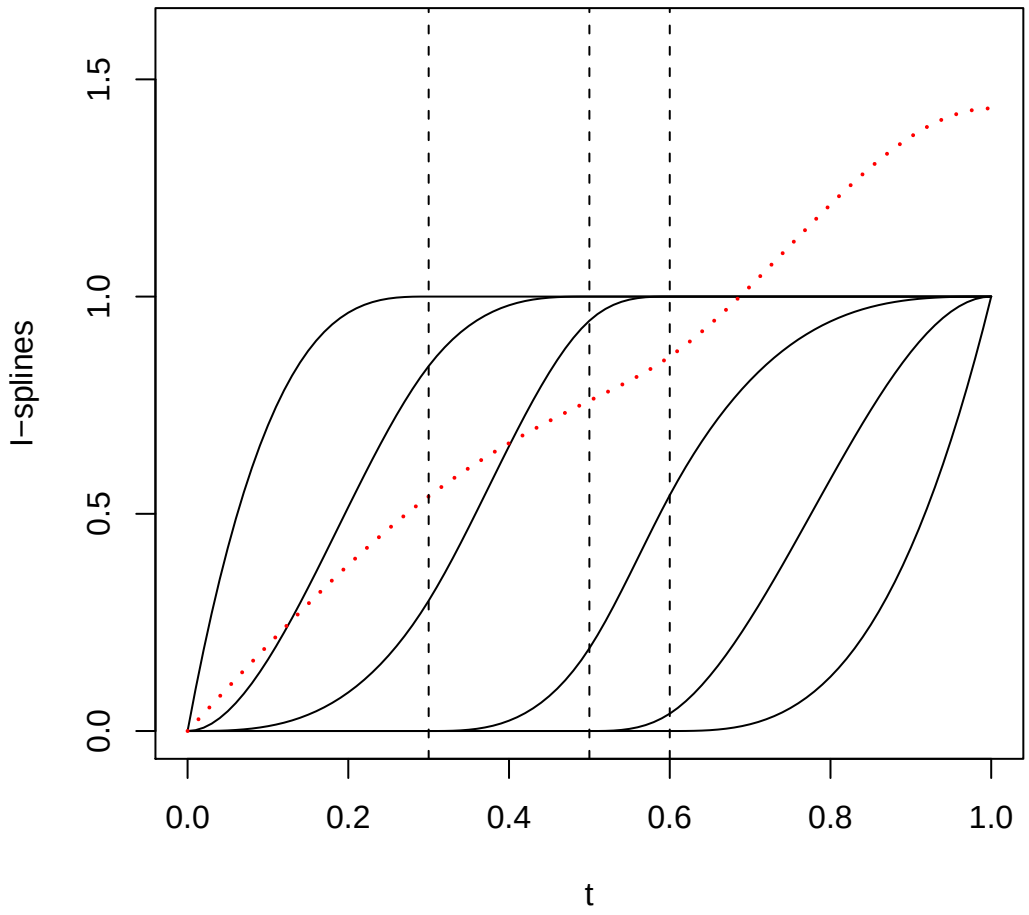

Figure 2.2 A set of I-splines with order = 3, $t = \{0, 0.3, 0.5, 0.6, 1\}$, and $f = (1.2I_1 + 2.0I_2 + 1.2I_3 + 1.2I_4 + 3.0I_5 + 0.0I_6)/6$

Chapter 3

ICBayes: An R Package for Modeling Interval-Censored Data

3.1 Introduction

In clinical trials and medical studies, there are often times where the event of interest is not directly observable and patients are examined periodically for disease occurrence or progression. In such situations, each individual is checked at a sequence of times and the exact survival time is only known to fall within a time interval (L, R]. This time to event is known to be interval-censored (Peto, 1973; Odell et al., 1992; Gentleman and Geyer, 1994). A special case of general interval-censored data is case 1 interval-censored data or equivalently referred to as current status data (Schick and Yu, 2000), where the event of interest is not directly observable and each subject in the study has only one observation time, and hence the failure times are either left-censored or right-censored. Case 1 interval-censored data often occur in tumorigenicity studies where lab animals are sacrificed to see if a certain tumor has developed or not. Three major statistical problems in survival analysis are estimation of survival function, k-sample survival comparison, and estimation of regression coefficients from censored data.

In the past two decades, many new methods have been developed for the analysis of interval-censored data. The survival function for interval-censored data is normally estimated through a classic nonparametric maximum likelihood procedure based on Peto (1973). In his paper, Peto developed a nonparametric method to find the empir-

ical survival curve that maximized the overall likelihood for interval-censored data. Based on observed intervals, a set of distinct intervals was defined. He proved that the likelihood is a function of the magnitude of the decrease of the survival function in these intervals only, and the empirical survival curve should be flat everywhere else. He described a constrained Newton-Raphson algorithm to search for the non-parametric maximum likelihood estimator (NPMLE). Turnbull (1976) proved that the NPMLE could be found by solving a self-consistency equation. This equation can be solved using the EM algorithm (Dempster, Laird, and Rubin, 1977) with the method of Gentleman and Geyer (1994) to ensure global maximum. Later, more efficient algorithms were proposed, for examples, iterative convex minorant (ICM) by Groeneboom and Wellner (1992) and EM iterative convex minorant (EM-ICM) by Wellner and Zhan (1997).

Most of the research on interval-censored data has been focused on subgroup survival comparison. Finkelstein (1986) developed a method for fitting the proportional hazards (PH) model (Cox, 1972) for interval-censored data. The likelihood is written as a function of regression parameters and nuisance parameters that are transformations of baseline survival at distinct time points. Regression coefficients are estimated by the maximum likelihood method and covariance matrix of estimators is estimated by Fisher information. One drawback of this classic method is that the number of nuisance parameters increases as sample size increases and nuisance parameters often approach the boundary of the constrained parameter space, which violates the regularity conditions of maximum likelihood. To avoid the boundary problem associated with likelihood-based tests (e.g., score test, Wald test, and likelihood ratio test) for comparing survival functions, Fay (1996) developed a permutational variance for the score statistic, which requires that censoring is independent of covariates. Also he proposed grouping data to reduce the number of nuisance parameters, such that none of the estimates would approach the parameter space boundary. The methods in Fay

(1996) only consider treatments as covariates and are only able to provide hypothesis testing but not point estimation. Sun (1996) developed a nonparametric test for comparing subgroup survival distributions by assuming that failure time is discrete. Zhao and Sun (2004) presented a generalized log-rank test of subgroup survival function comparison for interval-censored data. The test statistic is constructed nonparametrically and its covariance matrix is estimated based on Rubin's multiple imputation. Sun et al. (2005) proposed a new class of nonparametric generalized log-rank tests for k-sample comparison problem of interval-censored data, and also established the asymptotic distribution of the test statistic. Zhao et al. (2008) presented a class of generalized log-rank tests for partly interval-censored data that include both exact and interval-censored observations, and established their asymptotic properties.

To provide point estimation of treatment effects while controlling for relevant covariates, a regression model is needed. Most regression methods for interval-censored data have been developed under the proportional hazards model, with the first one being Finkelstein (1986). Pan (1999) extended the ICM algorithm, originally developed to compute the NPMLE of survival function, to the PH model for interval-censored data. Compared to the Newton-Raphson iteration suggested in Finkelstein (1986), Pan (1999) used ICM to obtain the MLEs of regression coefficients and baseline survival function. A bootstrap method was applied to estimate standard errors of MLEs of regression coefficients. Recently splines have been used to model baseline hazard function. For instance, Cai and Betensky (2003) modeled the logarithm of baseline hazard function with a linear spline mixed model. Their method produces estimation for regression coefficients and baseline hazard. The covariance matrix of the estimators is obtained through the sandwich method.

Currently, a few R packages and SAS macros have been developed for analyzing interval-censored data. The 'interval' package in R provides the NPMLE of survival function for different subgroups using the EM algorithm and the Gentleman and

Geyer (1994) algorithm, and performs several tests for k-sample comparison based on Sun (1996), Finkelstein (1986), and Fay (1996). The 'glrt' package in R includes functions to perform three generalized logrank tests based on Zhao and Sun (2004), Sun, Zhao, and Zhao (2005), Zhao, Zhao, Sun, and Kim (2008), and Finkelstein's score test. The SAS macro %EMICM uses the EM-ICM algorithm by Wellner and Zhan (1997) to compute the NPMLE of the survival functions for different groups and %ICTEST applies the nonparametric tests developed in Zhao and Sun (2004) and Sun, Zhao, and Zhao (2005). The only software packages available for regression analysis are the two R packages 'intcox' and 'survBayes'. The 'intcox' package gives point estimate for a regression coefficient and estimation for baseline survival curve. However, a bootstrap method is needed to estimate the standard errors of point estimates. The 'survBayes' package gives regression coefficient estimation and baseline survival curve estimaiton, based on a Bayesian approach. We have recently proposed approaches to fit proportional hazards model, proportional odds model, and probit model for interval-censored data (Lin and Wang, 2009; Cai et al., 2011; Wang and Lin, 2011; Lin et al., 2013). The R package 'ICBayes' is created based on these models for estimating survival functions and regression coefficients.

3.2 Models Included in the Package

We have included four models in our package for either case 1 or general interval-censored failure time data. The common parts of our methods are: (1) we used a monotone spline based on a set of I-splines to provide a smooth estimation for survival function; (2) we developed our posteriors through different data augmentations. (3) The MCMC algorithms we proposed are relatively straightforward to implement since most posterior distributions are of standard forms. In the following, we briefly describe the data and model form, spline approximation, data augmentation, and MCMC sampling for each of our models.

3.2.1 Proportional Hazards Model for Case 1 Interval-Censored Data

For the ith subject in study, let T_i denote the failure time of interest, C_i denote the observation time, and δ_i indicate left-censoring, $i = 1, \ldots, n$. Also let $\boldsymbol{x}$ be the vector of covariates, and $\boldsymbol{\beta}$ be the vector of regression coefficients. Under the PH model $\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x})$, Cai et al. (2012) proposed to model the baseline cumulative hazard function with a linear combination of monotone splines (Ramsay, 1988):

$$\Lambda_0(t) = \sum_{l=1}^{K} \gamma_l b_l(t), \tag{3.1}$$

where $\{b_l\}$ is a set of I-splines, each of which is nondecreasing from 0 to 1, and $\{\gamma_l\}$ is a set of nonnegative coefficients. As we have noted in Chapter 2, the set of I-splines are determined by specifying the placement of knots and the order of the I-splines. The number of I-splines K equals the number of interior knots plus the order (Ramsay, 1988).

To facilitate posterior computation, data augmentation with Poisson latent variables is employed. Redefine left-censoring indicator as $\delta_i = I(Z_i > 0)$, where latent variable $Z_i \sim \text{Poisson}(\lambda_0(c_i)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i))$. Furthermore, based on the additive property of Poisson distribution and the form of (3.1), import mutually independent latent variables Z_{il} such that $Z_i = \sum_{l=1}^{k} Z_{il}$ and $Z_{il} \sim \text{Poisson}(\gamma_l b_l(c_i)\exp(\boldsymbol{\beta}'\boldsymbol{x}_i))$. Then the augmented data likelihood in terms of the latent variables is as follows:

$$L = \prod_{i=1}^{n} \left[\prod_{l=1}^{K} \text{Poi}(Z_{il}|\gamma_l b_l(c_i) e^{\boldsymbol{x}_i'\beta}) \right] \left[I(Z_i > 0)\right]^{\delta_i} \left[I(Z_i = 0)\right]^{1-\delta_i}.$$

This is a product of Poisson functions, which allows relatively straightforward derivation of posteriors.

In the Gibbs sampler, the full conditional distributions of the latent variables $\{Z_i\}$ and $\{Z_{il}\}$, and the coefficients $\{\gamma_l\}$ are all of standard forms. The full conditional distributions of regression coefficients β_r, $r = 1, \ldots, p$, are not standard and are sampled using adaptive rejection Metropolis sampling (ARMS) algorithm (Gilks et al., 1995). For more details, please refer to Cai et al. (2012).

3.2.2 Proportional Hazards Model for General Interval-Censored Data

For subject i in a study, let $(L_i, R_i]$ denote the observed interval, δ_{i1}, δ_{i2}, δ_{i3} be left-, interval-, and right-censoring indicators, $i = 1, \ldots, n$. Lin et al. (2013) further extended the method for case 1 interval-censored data to case 2 interval-censored data, under the PH model. The same as (3.1), a linear combination of I-splines is used to provide an approximation of the baseline cumulative hazard function $\Lambda_0(t)$.

In order to derive posteriors that are relatively easy to sample from, a two-step data augmentation is employed. It was shown that the failure time of interest T_i is equivalent to the time of the first occurrence of a recurrent event E for which the number of occurrences $N(t)$ within time interval $(0, t]$ is a nonhomogeneous Poisson process with mean value function $\Lambda_0(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta})$. Let two time points $t_{i1} < t_{i2}$ such that for left-censored observation, $t_{i1} = R_i$, for interval-censored observation, $t_{i1} = L_i$ and $t_{i2} = R_i$, and for right-censored observation, $t_{i2} = L_i$. Then two latent variables $Z_i = N(t_{i1})$ and $W_i = N(t_{2i}) - N(t_{1i})$ are independent Poisson random variables. Furthermore, mutually independent latent variables $\{Z_{il}\}$ and $\{W_{il}\}$ are derived similar to $\{Z_{il}\}$ in the previous section. Then the augmented data likelihood is:

$$L_{aug} = \prod_{i=1}^{n} \left[\prod_{l=1}^{K} \text{Poi}(Z_{il})\text{Poi}(W_{il})^{\delta_{i2}+\delta_{i3}} \right]$$
$$[I(Z_i > 0)]^{\delta_{i1}} [I(Z_i = 0)]^{\delta_{i2}} [I(W_i > 0)]^{\delta_{i2}} [I(Z_i = 0)]^{\delta_{i3}} [I(W_i = 0)]^{\delta_{i3}}.$$

The full conditional distributions of the latent variables $\{Z_i\}$, $\{Z_{il}\}$, $\{W_i\}$, $\{W_{il}\}$, and the coefficients $\{\gamma_l\}$ are all of standard form. The full conditional distributions of regression coefficients β_r, $r = 1, \ldots, p$, are not standard and are sampled using ARMS. For more details, please refer to Lin et al. (2013).

3.2.3 Proportional Odds Model for Case 1 Interval-Censored Data Lin and Wang (2011) developed an approach to analyze case 1 interval-censored data

under the PO model:

$$\frac{F(t|\boldsymbol{x})}{1-F(t|\boldsymbol{x})} = \frac{F_0(t)}{1-F_0(t)}\exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{3.2}$$

where $\omega(t) = F_0(t)/(1-F_0(t))$ is the baseline odds function. A linear combination of I-splines is used to approximate $\omega(t)$ as in (3.1).

A data augmentation with Poisson latent variables was used to obtain posteriors that are easy to sample from. Redefine left-censoring indicator as $\delta_i = I(Z_i > 0)$; latent variable Z_i has conditional probability of $Z_i|\xi_i \sim \text{Poisson}(\omega(c_i)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}\xi_i)$; and $\xi_i \sim \exp(1)$. Furthermore, based on the additive property of Poisson distribution and the form of (3.1), import mutually independent latent variables Z_{il} such that $Z_i = \sum_{l=1}^{k} Z_{il}$ and $Z_{il} \sim \text{Poisson}(\gamma_l b_l(c_i)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}\xi_i)$. Then the data likelihood can be expressed in terms of the latent variables as follows:

$$L_{aug} = \prod_{i=1}^{n}\left[\prod_{l=1}^{k}\text{Poi}(Z_{il}|\gamma_l b_l(c_i)e^{\boldsymbol{x}_i'\boldsymbol{\beta}}\xi_i)\right][I(Z_i>0)]^{\delta_i}[I(Z_i=0)]^{1-\delta_i}.$$

This is a product of Poisson probabilities, which makes most posterior distributions to be of standard forms. In the Gibbs sampler, latent variables $\{Z_i\}$, $\{Z_{il}\}$, $\{\xi_i\}$, and spine coefficients $\{\gamma_l\}$ are sampled from standard distributions, while regression coefficients $\{\beta_r\}$ are sampled by ARMS algorithm.

3.2.4 Probit Model for General Interval-Censored Data The probit model (Lin and Wang, 2009) specifies the cumulative distribution function (CDF) of failure time T is modeled as: $F(t|\boldsymbol{x}) = \Phi(\alpha(t) + \boldsymbol{x}'\boldsymbol{\beta})$, where Φ is the CDF of the standard normal distribution. A linear combination of I-splines is used to model $\alpha(t)$:

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

where γ_0 is an unconstrained constant, $\{\gamma_l\}$ and $\{b_l\}$ are defined as before.

Let $t_i = R_i I(\delta_{i1}=1) + L_i I(\delta_{i1}=0)$, a data augmentation with truncated normal latent variable Z_i is employed, where

$$z_i \sim N(\alpha(t_i) + \boldsymbol{x}_i'\boldsymbol{\beta}, 1), \quad \text{with constraint} z_i \in A_i,$$

where A_i is interval $(0, \infty)$ if $\delta_{i1} = 1$, $(\alpha(L_i) - \alpha(R_i), 0)$ if $\delta_{i2} = 1$, and $(-\infty, 0)$ if $\delta_{i3} = 1$. The augmented likelihood now consists of constrained normal densities:

$$L_{aug} = \prod_{i=1}^{n} N(z_i; \alpha(t_i) + \boldsymbol{x}_i'\boldsymbol{\beta}, 1)$$
$$\{I(z_i > 0)\}^{\delta_{i1}} \{\alpha(L_i) - \alpha(R_i) < z_i < 0\}^{\delta_{i2}} \{I(z_i < 0)\}^{\delta_{i3}} \pi(\lambda_i).$$

The full conditional distributions for $\{Z_i\}$ are truncated normal. Coefficients γ_0, $\{\gamma_l\}$ are sampled from standard distributions. The regression coefficients vector $\boldsymbol{\beta}$ is sampled from a multivariate normal distribution.

3.3 ICBayes Package Design and Use

To use our package, infinity in right-censored observations should be written as NA. There are four modeling methods included in the package: *case2ph*, *case2probit*, *case1po*, and *case1ph*. *case2ph* and *case2probit* fit PH and probit model for general interval-censored data respectively. *case1po* and *case1ph* fit PO and PH model for case 1 interval-censored data only. A user can call these models through the *model* argument using the main function *ICBayes*. The function *ICBayes* allows for a generic input format where a user specifies the left endpoints L, right endpoints R, censoring status, and covariate matrix, etc. Also it allows for a formula input format, where a user inputs a survival object returned by the *Surv* function of the 'survival' package.

Most arguments in the *ICBayes* function are set to reasonable default values to make it more convenient for users. For the I-splines, the *order* is set to be 2. The default *knots* is set to be from minimum observed time point to maximum observed time point, with length = 10. The Normal prior standard deviation for each β in the PH model and case 1 PO model is set to be 10. The support of the target density for sampling β using *arms* in library 'HI' is set to be (-5, 5) (*coef_range = 5*). The Normal prior precision for the coefficient γ_0 in the general interval-censored probit model is set to be 0.1 (*v0 = 0.1*). The confidence level for CI of β is set to be 95% (*conf.int*

= *0.95*). The argument *grids* specifies a sequence of time points to estimate survival probabilities for. Its default value is from minimum time point to maximum time point, with length = 100. It is default that survival probabilities at *grids* are saved in the output for plot (*plot_S* = *TRUE*). One may want to save the MCMC chains for later convergence diagnosis. This can be done by setting *chain.save* = *TRUE* and specifying a local .txt file to store the chains using argument *dd1*. Finally, the default number of iterations is 5000, burn-in is 1000. One can let *thin* be a value greater than 1 to thin the MCMC chains. It is recommended that a user can adjust *knots* and *grids* based on his/her data. It is also recommended that a user adjusts *coef_range* based on his/her rough guess of the possible range for the regression coefficients β.

The available methods for the *ICBayes* object are *print*, $, and *plot*. The *print* method prints as a list the estimated regression coefficients, sample standard deviation of MCMC draws for each regression coefficient, and credible intervals. The $ method allows picking out each component from the *ICBayes* object. The *plot* method plots the estimated baseline survival function at the time points specified by *grids*. To use the *plot* method, one must let the argument *plot_S* = TRUE. Then the estimated baseline survival probabilities at *grids* are stored in the *ICBayes* function output, though not printed by the *print* method.

3.4 Breast Cosmesis Example

To demonstrate the use of our package, we first analyze the general interval-censored breast cancer cosmesis data from Finkelstein and Wolfe (1985). Early breast cancer patients treated with either radiotheraph ($x_1 = 0$) alone or radiotherapy with adjuvant chemotherapy ($x_1 = 1$) were examined periodically for breast retraction. The time unit is month. The following are the first few observations of the data.

```
> library(ICBayes)
> data(bcdata)
```

```
> bcdata<-data.frame(bcdata)
> attach(bcdata)
> head(bcdata)
      L  R status  x1
[1,] 45 NA      2  0
[2,]  6 10      1  0
[3,]  0  7      0  0
[4,] 46 NA      2  0
[5,] 46 NA      2  0
[6,]  7 16      1  0
```

3.4.1 Fit the PH Model for General Interval-Censored Data The following function uses the formula input format and fits a PH model. The input data should be data frame. The range of the observed time points (i.e., L and R) is from 0 to 60, so we set 10 equally spaced knots for I-splines using *knots* = seq(0.1, 60.1, length=10), and we ask for survival estimation at *grids* = seq(0.1, 60.1, by=1). The MCMC chain for β is stored in the file specified by *dd1*. The argument *x_user* = c(0, 1) tells *ICBayes* to estimate survival at *grids* for $x_1 = 0$ and for $x_1 = 1$. For this data, we only have one covaritae x_1. If we have more than one covariate, say, gender and age, and we want to estimate the survival curve at gender = 0 and age = 50 and another survival curve at gender = 1 and age = 50, then we should specify *x_user* = c(0, 50, 1, 50).

```
> try1<-ICBayes(formula = Surv(L, R, type = "interval2") ~ x1,
+ data = bcdata, model = "case2ph", status = bcdata[, 3],
+ order=4, coef_range = 2, x_user=c(0,1), niter=11000,
+ knots=seq(0.1,60.1,length=10),grids=seq(0.1,60.1,by=1),
+ chain.save=TRUE, dd1 = 'C:/MCMC/bcdatapar.txt')
```

The result is as follows. Assume PH model is reasonable, we estimate the regression coefficient β to be 0.71, with a sample standard deviation of 0.27, and 95% credible interval = (0.17, 1.25).

```
coef #Estimted regression coefficients
[1] 0.712153
coef_ssd #Sample standard deviation
[1] 0.2741571
coef_ci #Credible interval of coefficients
       2.5 %CI 97.5 %CI
[1,] 0.1749649 1.251057
```

As shown in the code above, we have saved the MCMC chain for β in the local file 'C:/MCMC/bcdatapar.txt'. Now we can perform convergence diagnosis. In Figure 3.1, a traceplot of the MCMC chain for β is created to check if it is mixed well and stable. Our chain seems to converge.

```
> wd1<-paste('C:/MCMC/')
> setwd(wd1)
> f.name<-paste('bcdatapar','.txt', sep='')
> parest<-data.matrix(read.table(f.name))
> plot.ts(parest)
```

3.4.2 Fit the Probit Model for General Interval-Censored Data Since for general interval-censored data, we have two types of models availabe in our package: the PH model and the probit model, now, the probit model is fit to compare with the PH model. In order to do that, one only needs to change *model = "case2ph"* to *model = "case2probit"*.

```
> try6<-ICBayes(formula = Surv(L, R, type = "interval2") ~ x1,
```

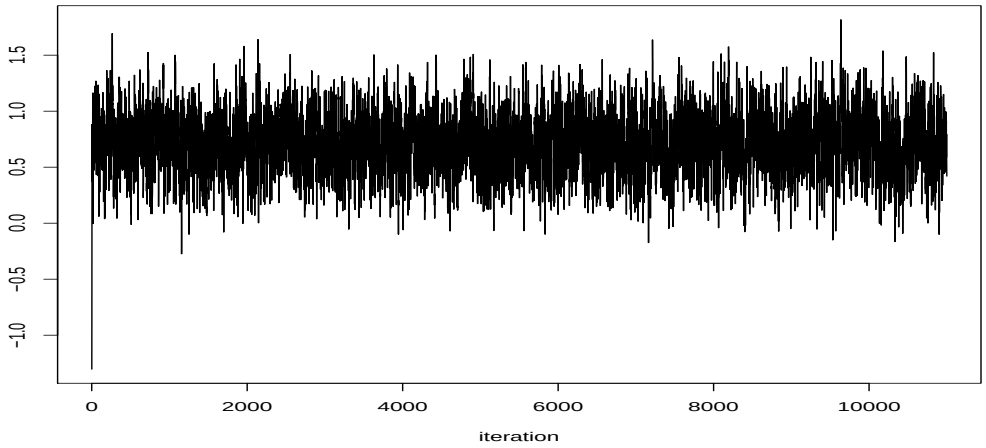

Figure 3.1 Traceplot of MCMC chain for β of case 2 PH model in breast cancer study

```
+ data = bcdata, model = "case2probit", status = bcdata[, 3],
+ order=4, x_user=c(0,1), niter=11000,
+ knots=seq(0.1,60.1,length=10),grids=seq(0.1,60.1,by=1),
+ chain.save=TRUE, dd1 = 'C:/MCMC/bcdatapar2.txt')
```

The result is as follows. Under the probit model assumption, the regression coefficient is estimated to be 0.48, with a sample standard deviation of 0.23, and a 95% credible interval = (0.04, 0.93).

```
coef #Estimted regression coefficients
[1] 0.4818546
coef_ssd #Sample standard deviation
[1] 0.2254858
coef_ci #Credible interval of coefficients
        2.5 %CI  97.5 %CI
[1,] 0.04456837 0.9277489
```

The traceplot of the MCMC chain of β in the probit model is presented in Figure 3.2. It shows that our chain mixes well and has converged to its stationary distribution.

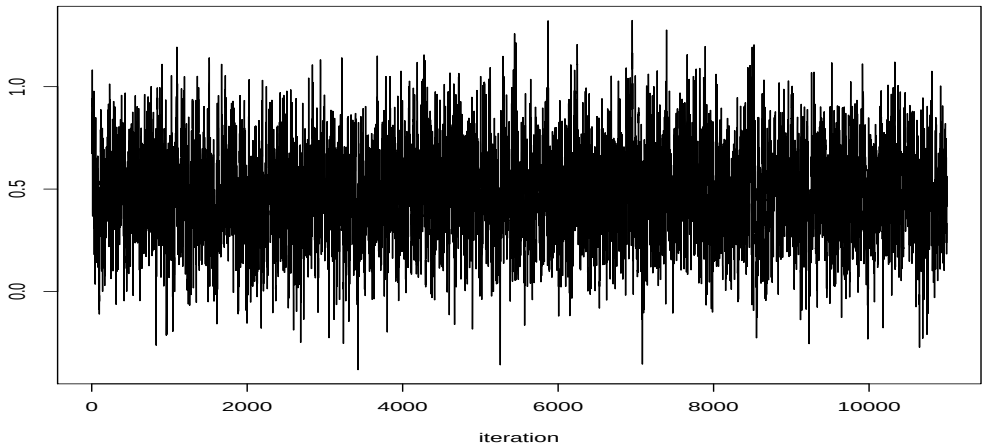

Figure 3.2 Traceplot of MCMC chain for β of case 2 probit model in breast cancer study

3.4.3 Plot of Survival Functions for the Breast Cancer Data In the input for both PH and probit models, we used c(0, 1) as the user-specified covariate vector, which means estimating survival probabilities at *grids* for $x_1 = 0$ and $x_1 = 1$. Of course, here $S(t|x_1 = 0) = S_0(t)$. Based on the survival probabilities stored in the output, we can plot the estimated survival functions for the two treatment groups. Using the code shown below, we plotted the estimated survival curves based on the NPMLE method using the *icfit* function in the 'interval' package, and the case 2 PH model and case 2 probit model in our package. In the *ICBayes* object, the vector that stores the estimated survival probabilities at *grids* for x_*user* = c(0, 1) is S_m. Its length is G∗2, where G = length(*grids*). The first G elements are the $\hat{S}(t|x_1 = 0)$ and the rest G elements are the $\hat{S}(t|x_1 = 1)$ at *grids*. From Figure 3.3, we can see that our two models produce similar results for the data. Compared to the step function based on NPMLE (overlayed), our estimation is smooth since we connect point estimates

with polynomials.

```
# compared to NPMLE plot
library(interval)
R2<-ifelse(is.na(R),Inf,R)
fit1<-icfit(Surv(L,R2,type = 'interval2')~x1, data = bcdata)
plot(fit1)
S_m1<-matrix(try1$S_m,ncol=length(try1$grids),byrow=TRUE)
lines(try1$grids,S_m1[1,],col=2)
lines(try1$grids,S_m1[2,],col=2,lty=2)
S_m2<-matrix(try6$S_m,ncol=length(try6$grids),byrow=TRUE)
lines(try6$grids,S_m2[1,],col=3)
lines(try6$grids,S_m2[2,],col=3,lty=2)
rect(3,-0.035,20,0.2,col='white',border=NA)
legend(1,0.3,c('x1=0 (NPMLE)','x1=1 (NPMLE)','x1=0 (case2PH)',
'x1=1 (case2PH)','x1=0 (case2probit)','x1=1 (case2probit)'),
lty=c(1,2,1,2,1,2),col=c(1,1,2,2,3,3),cex=0.8)
```

3.5 Lung Cancer Example

Hoel and Walberg (1972) studied time to onset of lung cancer for two groups of mice: one group living in conventional environment (96 mice, treatment = 1) and the other group in germfree environment (48 mice, treatment = 0). Each mouse was sacrificed at a random time to check for lung tumors. Hence it is case 1 interval-censored data. The following are the first few observations from the data set. The time unit is day.

```
> library(ICBayes)
> data(lungdata)
> lungdata<-data.frame(lungdata)
```

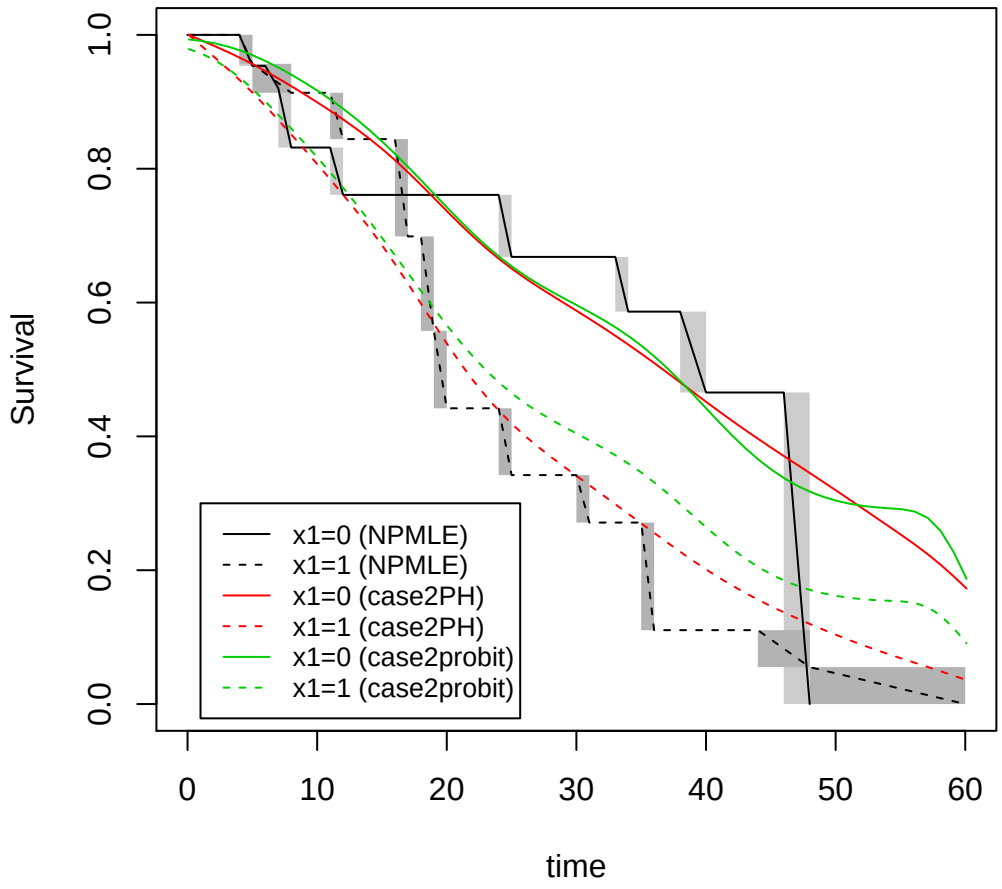

Figure 3.3 Estimated survival curves for two groups of patients in breast cancer study

```
> head(lungdata)

  L   R status treatment

1 0 381      1         1

2 0 477      1         1

3 0 485      1         1

4 0 515      1         1

5 0 539      1         1
```

```
6 0 563      1          1
```

3.5.1 Fit the PH Model for Case 1 Interval-Censored Data Suppose we want to fit a PH model, and for illustration purpose, this time we use the generic input format. Now instead of putting L, R, and covariate matrix in a formula, we specify them separately, and we do not have *data* argument. We observe that the observed time points (i.e., L and R) range from 0 to 1008, ruthermroe, the observed time points are not evenly distributed. So we set *knots* and *grids* as unequally spaced to try to match with the data structure. Other arguments are the same.

```
> # generic form
> try2<-ICBayes(model="case1ph", L=lungdata[, 1],
+ R=lungdata[, 2], status=lungdata[, 3], xcov=lungdata[, 4],
+ niter=11000, x_user=c(0,1), order=2,
+ knots=c(0,100,200,300,400,seq(450,1008,length=50)),
+ grids=c(0,100,200,300,400,seq(402,1008,by=2)),
+ chain.save=TRUE,dd1 = "C:/MCMC/lungpar.txt")
```

The output is as follows. Assume the PH model is reasonable, we estimate the regression coefficient β to be -1.11, with a sample standard deviation of 0.27, and 95% credible interval = (-1.66, -0.57). This is consistent with Huang (1996) which concludes that mice in conventional environment seems to have lower risk of tumor (p-value = 0.054).

```
coef #Estimted regression coefficients
[1] -1.113273
coef_ssd #Sample standard deviation
[1] 0.2749167
coef_ci #Credible interval of coefficients
       2.5 %CI   97.5 %CI
```

```
[1,] -1.659517 -0.5730999
```

We can look at the traceplot of the MCMC chain of β to check for convergence. As shown in Figure 3.4, the chain seems to converge.

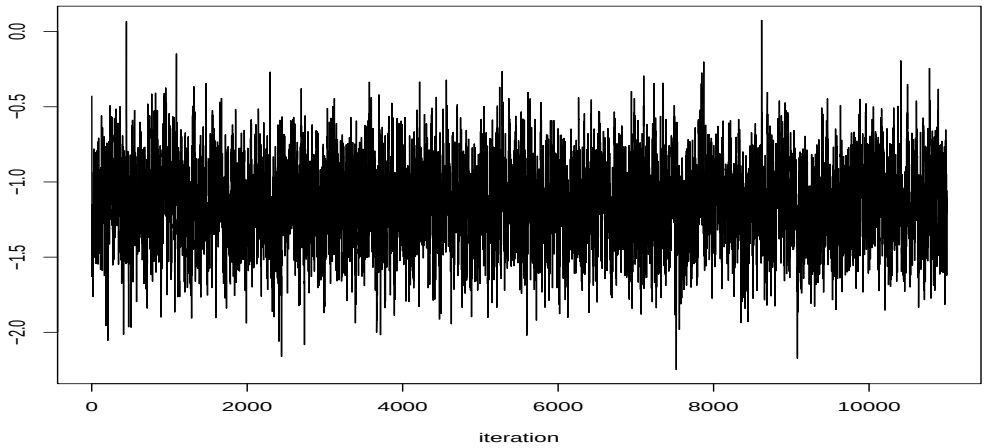

Figure 3.4 Traceplot of MCMC chain for β in case 1 PH model for lung cancer data

3.5.2 Fit the PH Model for General Interval-Censored Data Since case 1 data is a special case of general interval-censored data, then we expect that our case 2 PH model should produce very similar results as the case 1 PH model. The results will not be exactly the same, because our model construction processes are different. For comparison, the case 2 PH model is fit using the code below:

```
# case2ph
> status_new=2*(lungdata[,3]==0)+0*(lungdata[,3]==1) # important!
> try7<-ICBayes(formula=Surv(L,R,type='interval2')~treatment,
+ data=lungdata,model='case2ph',status=status_new,
+ order=2,x_user=c(0,1),niter=11000,
+ knots=c(0,100,200,300,400,seq(450,1008,length=50)),
+ grids=c(0,100,200,300,400,seq(402,1008,by=2)),
+ chain.save=TRUE,dd1 = "C:/MCMC/lungpar2.txt")
```

To be consistent with the methodology papers this package is based on, we have used 1 to indicate left-censoring and 0 to indicate right-censoring for case 1 data. We have used 0, 1, and 2 to indicate left-, interval-, and right-censoring for general interval-censored data. So it is important to note that when we fit case 2 models for case 1 data, our censorship variable need to be redefined using the first line of code shown above.

We obtained the following result:

```
coef #Estimted regression coefficients
[1] -1.139954
coef_ssd #Sample standard deviation
[1] 0.2668109
coef_ci #Credible interval of coefficients
       2.5 %CI   97.5 %CI
[1,] -1.675733 -0.6267516
```

The traceplot for our β suggests convergence, as seen in Figure 3.5. By comparing with the results from fitting case 1 PH model, we can see that the point estimate, the SSD, and the credible interval are all very close to their counterparts from for the two models.

3.5.3 Plot of Survival Functions for the Lung Cancer Data In the input for both case 1 and case 2 PH models, we used c(0, 1) as the user-specified covariate vector, which means calculating survival probabilities at *grids* for treatment = 0 and treatment = 1. Of course, $S(t|\text{treatment} = 0) = S_0(t)$. Based on the survival probabilities stored in the *ICBayes* output element S_m, we can plot the estimated survival functions for the two treatment groups, using the code shown below. From Figure 3.6, we can see that the curves from case1PH and case2PH models are not very close to that from the NPMLE. Actually, the NPMLE is based on Turnbull intervals. The first Turnbull interval = (524, 546] for the treatment = 1 group. So

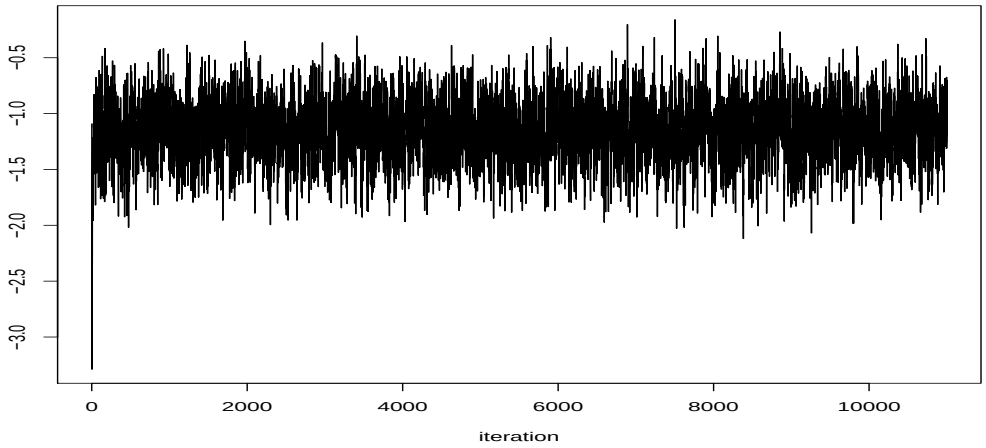

Figure 3.5 Traceplot of the MCMC chain of β in case 2 PH model for lung cancer data

the estimated survival probability before time = 524 days is 1. Similarly, the first Turnbull interval = (371, 381] for the treatment = 0 group. So the corresponding estimated survival probability befoer time = 371 is 1. However, our estimates are smooth rather than step functions, so the estimated survival probabilities during the two early periods are not going to be constantly 1.

```
# compared to NPMLE plot
library(interval)
L<-lungdata[,1]; R<-lungdata[,2]; treatment<-lungdata[,4]
R2<-ifelse(is.na(R),Inf,R)
fit1<-icfit(Surv(L,R2,type = 'interval2')~treatment, data = lungdata)
plot(fit1)
S_m1<-matrix(try2$S_m,ncol=length(try2$grids),byrow=TRUE)
lines(try2$grids,S_m1[1,],col=2)
lines(try2$grids,S_m1[2,],col=2,lty=2)
S_m2<-matrix(try7$S_m,ncol=length(try7$grids),byrow=TRUE)
lines(try7$grids,S_m2[1,],col=3)
```

```
lines(try7$grids,S_m2[2,],col=3,lty=2)
rect(400,-0.035,800,0.1,col='white',border=NA)
legend(0,0.4,c('trt=0 (NPMLE)','trt=1 (NPMLE)',
'trt=0 (case1PH)','trt=1 (case1PH)',
'trt=0 (case2ph)','trt=1 (case2ph)'),
lty=c(1,2,1,2,1,2),col=c(1,1,2,2,3,3),cex=0.8)
```

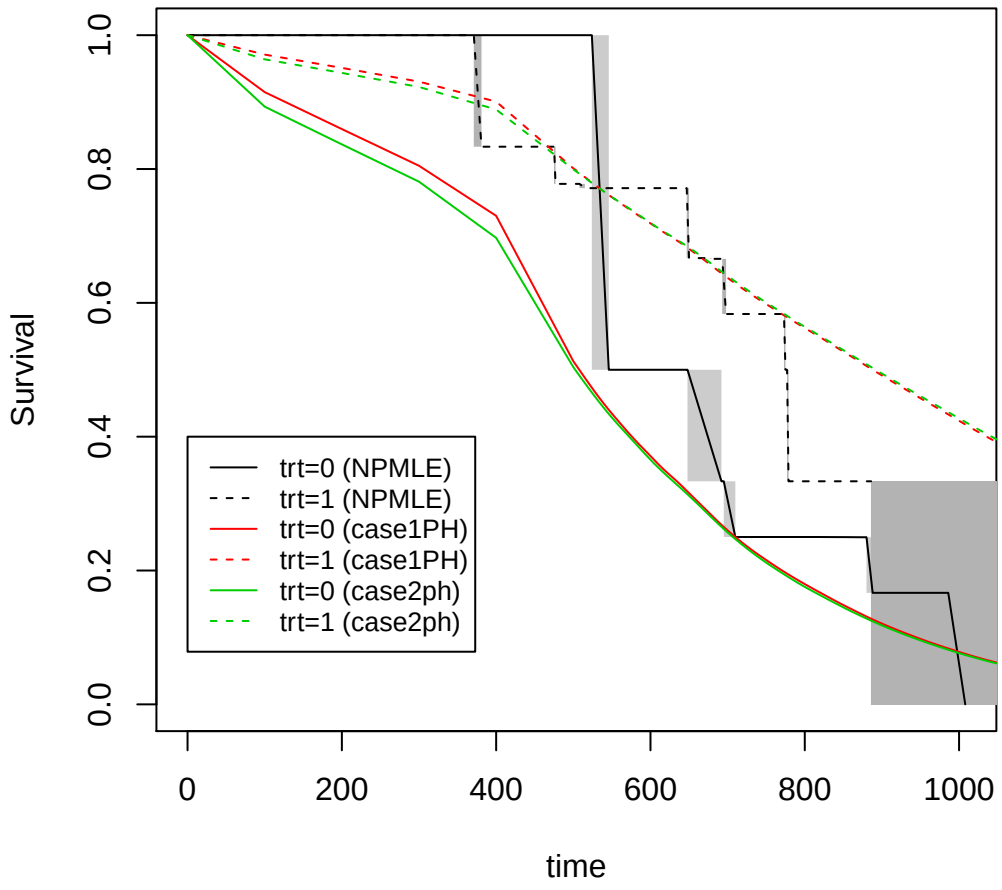

Figure 3.6 Estimated survival curves for two groups of patients in lung cancer study

Chapter 4

Modeling Interval-Censored Survival Data with Spatial Correlation

In randomized clinical trials, subjects are recuited from different geographical regions. Uncontrolled factors that vary spatially and exert an effect on study outcome may cause dependency among subjects. Suppose a clinical trial for lung cancer is carried out in China with patients being recruited from all the provinces. Then from this mapping of air pollution particles in China (Figure 4.1) , we can see that particle level varies greatly across the country, and two provinces closer to each other probably have similar particle level. Since air quality may well affect a person's lung function, it suggests that patients may be spatially correlated. Patients from the same province may be correlated because they share the same air quality, and a province's air quality may be affected by its neighboring provinces. In this project, we will model the described spatial dependency through the conditional autoregressive (CAR) distribution. Furthermore, suppose an endpoint of the trial is progression for lung cancer patients, then the time-to-event is interval-censored, since progression can only be examined by CT scan very few weeks.

The conditional autoregressive (CAR) distribution is the most commonly used model for region-specific random effects. Different from commonly used random effects that are assumed to be independently and identially distributed, the spatial random effects are dependent and they jointly follow the CAR distribution. Hodges et al. (2003) derived the correct power for the precision parameter in the CAR model.

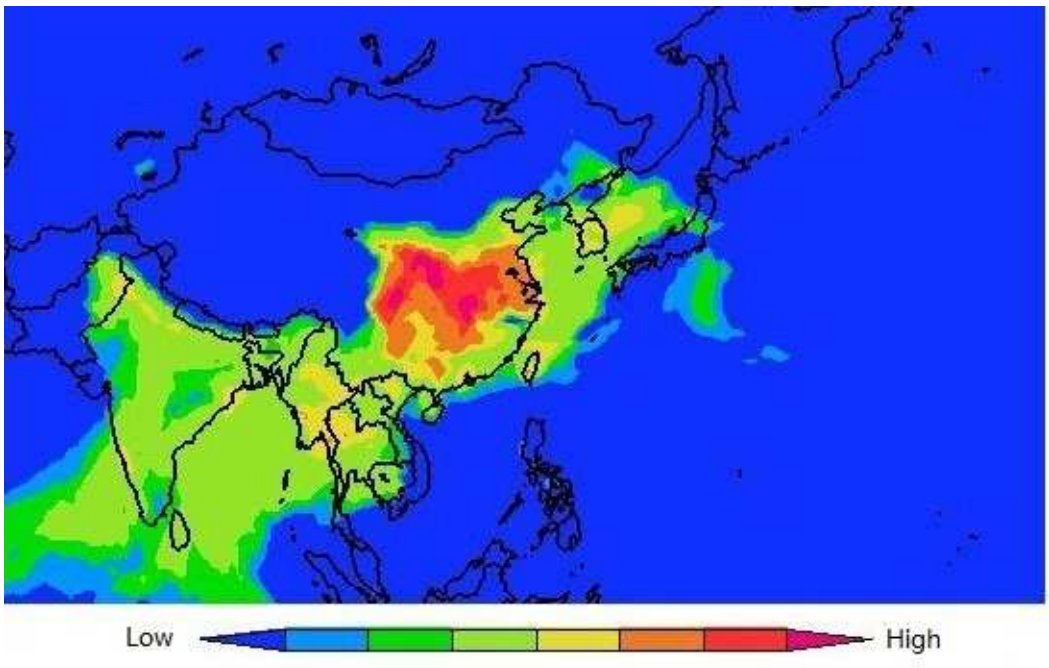

Figure 4.1 A recent mapping of air pollution particle levels in China

Banerjee et al. (2003) modeled spatially correlated frailties using both geostatistical approach and lattice approach under Weibull PH model. Banerjee and Carlin (2004) modeled spatially correlated time-to-relapse in a Minnesota smoking clinical trial through parametric cure rate models with multivariate conditional autoregressive (MCAR) prior for spatial random effects. Jin et al. (2005) further extended MCAR model for lattice data to generalized MCAR where they demonstrated its use under Gaussian model and Poisson model. Not so much work has been done for spatially correlated data, and most of them has been conducted under parametric settings, such as normal, Poisson, or Weibull PH models. Because real-life survival data are usually complicated, parametric models are typically not flexible enough to provide a good fit for the true survivorship. In this project, we introduce an efficient and easy-to-implement Bayesian approach for analyzing general interval-censored data with spatial correlation under semiparametric PH model.

4.1.1 Data and Likelihood Function Suppose there are I areas and the ith area contains n_i subjects. Let T_{ij} denote the failure time for the jth person in the ith area and $(L_{ij}, R_{ij}]$ the observed interval for T_{ij}, where $i = 1, \ldots, I$ and $j = 1, \ldots, n_i$. Let $\boldsymbol{x}_{ij}$ denote the $p \times 1$ covariate vector for person j in area i. Let ϕ_i denote the spatial random effect for area i. We consider a mixed PH model for T_{ij}, for which the survival function of T_{ij} given x_{ij} and ϕ_i is

$$S(t|\boldsymbol{x}_{ij}, \phi_i) = \exp\left\{-\Lambda_0(t)\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i)\right\},$$

where $\Lambda_0(t)$ is the baseline cumulative hazards function. Furthermore, let δ_{1ij}, δ_{2ij}, δ_{3ij} be the left-, interval-, and right-censoring indicators for person j in area i. Let $(\boldsymbol{\beta}, \boldsymbol{\Lambda}_0)$ denote the unknown parameters in the model, then the observed data likelihood is:

$$Lk_{obs}(\boldsymbol{\Lambda}_0, \boldsymbol{\beta}) = \prod_{i=1}^{I}\left\{\int Lk_i(\boldsymbol{\Lambda}_0, \boldsymbol{\beta}|\phi_i)\pi(\phi_i)d\phi_i\right\}, \tag{4.1}$$

where $Lk_i(\boldsymbol{\Lambda}_0, \boldsymbol{\beta}|\phi_i) = \prod_{j=1}^{n_i} Lk_{ij}(\boldsymbol{\Lambda}_0, \boldsymbol{\beta}|\phi_i)$ is the conditional likelihood contributed by the subjects in the ith area given the random effect ϕ_i, and:

$$L_{ij}(\boldsymbol{\Lambda}_0, \boldsymbol{\beta}|\phi_i) = \tag{4.2}$$
$$\{F(R_{ij}|\boldsymbol{x}_{ij}, \phi_i)\}^{\delta_{ij1}}\{F(R_{ij}|\boldsymbol{x}_{ij}, \phi_i) - F(L_{ij}|\boldsymbol{x}_{ij}, \phi_i)\}^{\delta_{ij2}}\{1 - F(L_{ij}|\boldsymbol{x}_{ij}, \phi_i)\}^{\delta_{ij3}}.$$

Using the observed likelihood for Bayesian computation is difficult because the integral in the observed likelihood (4.1) does not have an explicit form. To overcome this, we consider the following form of likelihood function treating all ϕ_i's as unknow parameters:

$$Lk(\boldsymbol{\Lambda}_0, \boldsymbol{\beta}, \boldsymbol{\phi}) = \prod_{i=1}^{I}\{Lk_i(\boldsymbol{\beta}, \boldsymbol{\Lambda}_0)\pi(\phi_i)\}.$$

4.1.2 Estimating $\Lambda_0(t)$ The above model is semiparametric as $\Lambda_0(t)$ is a totally unspecified non-decreasing function. Estimating $\Lambda_0(t)$ is challenging since it is infinite dimensional. For right-censored data, there exists a partial likelihood that allows us

to estimate the regression parameters $\boldsymbol{\beta}$ directly without the need of estimating $\Lambda_0(t)$ under the PH model (Cox, 1972). However, the partial likelihood does not exist for interval-censored data (Sun, 2006), and we need to estimate both $\boldsymbol{\beta}$ and $\Lambda_0(t)$ simultaneously. Following Joly et al. (1998) and Cai et al. (2011), we model $\Lambda_0(t)$ with a linear combination of monotone splines (Ramsay, 1988):

$$\Lambda_0(t) = \sum_{l=1}^{K} \gamma_l b_l(t), \tag{4.3}$$

where $\{b_l\}$ is a set of basis splines (I-splines), each of which is nondecreasing from 0 to 1, and $\{\gamma_l\}$ is a set of non-negative coefficients. To construct the set of basis splines, we need to specify the order for them and an increasing sequence of knots within the data range. The basis splines are fully determined once knots and order are specified. The number of basis splines equals the number of interior knots plus the order. The order taking values of 1, 2, and 3 produces linear, quadratic, and cubic basis splines respectively. We recommend to take 2 or 3 as order value for adequate smoothness and to take 10-30 equally spaced knots for adequate modeling flexibility from our intensive experiences (Lin and Wang, 2010; Wang and Dunson, 2011; Cai et al., 2011).

4.1.3 Data augmentation Although one may use Metropolis-Hastings algorithms to sample all the parameters, it is difficult to find good proposal distributions to obtain reasonable acceptance rates and well mixed MCMC chains. In the following, we propose a two-step data augmentation to facilitate posterior computation by taking advantage of the PH model structure and the spline modeling form of $\Lambda_0(t)$. Assume that there is an underlying recurrent event E, for which the number of occurrences $N(t)$ within time interval $(0, t]$ is a nonhomogeneous Poisson process with mean function $\Lambda_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta} + \phi)$. Define T as the the time of the first occurrence of E. This latent process totally determines T, since $P(T > t) = P(N(t) = 0) = \exp\{-\Lambda_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta} + \phi)\}$ is our survival function.

Let time points $t_1 < t_2$. Specifically, for left-censoring, $t_1 = R$; for interval-censoring, $t_1 = L$ and $t_2 = R$; and for right-censoring, $t_2 = L$, where (L, R] is the observed interval for T. By the properties of nonhomogeneous Poisson process, random variable $Z = N(t_1) \sim \text{Poi}(\Lambda_0(t_1)\exp(\boldsymbol{x}'\boldsymbol{\beta} + \phi))$, random variable $W = \{N(t_2) - N(t_1)\} \sim \text{Poi}(\{\Lambda_0(t_2) - \Lambda_0(t_1)\}\exp(\boldsymbol{x}'\boldsymbol{\beta} + \phi))$, and Z and W are independent conditional on ϕ. For left-censored data, W is not defined. The reason is that t_2 is some point greater than $t_1 = R$, so W can take any integer value and will not contribute any information about the failure time T. For interval-censored data, $Z = 0$ and $W > 0$. For right-censored data, t_1 is some point that is less than $t_2 = L$, so $Z = W = 0$. According to our data structure, for the jth subject in the ith area, we define t_{ij1} and t_{ij2} analogous to t_1 and t_2, and Z_{ij} and W_{ij} analogous to Z and W. Then $Z_{ij} \sim \text{Poi}(\Lambda_0(t_{ij1})\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i))$ and $W_{ij} \sim \text{Poi}(\{\Lambda_0(t_{ij2}) - \Lambda_0(t_{ij1})\}\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i))$. The augmented data likelihood for subject j in area i is:

$$Lk_{aug1_{ij}}(\boldsymbol{\theta}|Z_{ij}, W_{ij}, \phi_i) =$$

$$\text{Poi}(Z_{ij})\text{Poi}(W_{ij})^{\delta_{ij2}+\delta_{ij3}}[Z_{ij} > 0]^{\delta_{ij1}}[Z_{ij} = 0]^{\delta_{ij2}}[W_{ij} > 0]^{\delta_{ij2}}[Z_{ij} = 0]^{\delta_{ij3}}[W_{ij} = 0]^{\delta_{ij3}}.$$

Integrating out Z_{ij} and W_{ij} will lead to the conditional likelihood $Lk_{ij}(\boldsymbol{\theta}|\phi_i)$ in (4.2).

Furthermore, based on the additive property of Poisson distribution and the spline modeling form of $\Lambda_0(t)$ in (4.3), we decompose Z_{ij} and W_{ij} as follows: $Z_{ij} = \sum_{l=1}^{K} Z_{ijl}$ and $W_{ij} = \sum_{l=1}^{K} W_{ijl}$, where Z_{ijl}'s and W_{ijl}'s are mutually independent, $Z_{ijl} \sim \text{Poi}(\gamma_l b_l(t_{ij1})\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i))$, $W_{ijl} \sim \text{Poi}(\{\gamma_l b_l(t_{ij2}) - \gamma_l b_l(t_{ij1})\}\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i))$, with constraints $\sum_{l=1}^{K} Z_{ijl} > 0$ if $\delta_{ij1} = 1$, $\sum_{l=1}^{K} Z_{ijl} = 0$ and $\sum_{l=1}^{K} W_{ijl} > 0$ if $\delta_{ij2} = 1$, and $\sum_{l=1}^{K} Z_{ijl} = 0$ and $\sum_{l=1}^{K} W_{ijl} = 0$ if $\delta_{ij3} = 1$. Then for subject j in area i, the augmented data likelihood based on Z_{ijl}'s and W_{ijl}'s is:

$$Lk_{aug2_{ij}}(\boldsymbol{\theta}|Z'_{ijl}s, W'_{ijl}s, \phi_i) = \left\{\prod_{l=1}^{K} \text{Poi}(Z_{ijl})\text{Poi}(W_{ijl})^{\delta_{ij2}+\delta_{ij3}}\right\}$$

$$[Z_{ij} > 0]^{\delta_{ij1}}[Z_{ij} = 0]^{\delta_{ij2}}[W_{ij} > 0]^{\delta_{ij2}}[Z_{ij} = 0]^{\delta_{ij3}}[W_{ij} = 0]^{\delta_{ij3}}.$$

This likelihood function is simply a product of Poisson probability mass functions, which will lead to a straightforward posterior computation to be presented later.

4.1.4 Spatial Random Effects A common model for lattice data is conditional autoregressive (CAR) distribution, originally developed by Besag (1974). Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_I)$ be the spatial random effect vector, then the general form of CAR model is (Besag and Kooperberg, 1995),

$$p(\boldsymbol{\phi}) \propto \tau_\phi^{(I-B)/2} \exp\left\{-\frac{\tau_\phi}{2}\boldsymbol{\phi}' Q \boldsymbol{\phi}\right\}, \tag{4.4}$$

where Q is an $I \times I$ positive definite symmetric matrix. B is the number of disconnected groups of areas (Hodges et al., 2003). For the South Carolina map and Minnesota map used in our study, there is only one group of areas respectively, so $B = 1$. While if there is one or more islands that are disconnected from other areas, then B is greater than 1. For example, in the Scottish lip cancer study (Clayton and Kaldor, 1987), there are three islands (Orkneys, Shetland, and Outer Hebrides islands) in addition to the mainland of Scotland, so the total number of disconnected groups of areas is $B = 4$.

By Brook expansion (Brook, 1964), (4.4) is also equivalent to:

$$\phi_i | \phi_{-i} \sim \mathrm{N}(\sum_j \zeta_{ij}\phi_j, \kappa_i), \quad \text{for } i = 1, \ldots, I, \tag{4.5}$$

with $\zeta_{ii} = 0$, $\zeta_{ij} = -(Q_{ij}/Q_{ii})$ for $i \neq j$, and $\kappa_i = 1/Q_{ii}$. $i \sim j$ denotes that areas i and j are neighbors, and ϕ_{-i} is the set of all spatial random effects except for the one for area i.

If we specify Q in a way such that $Q\mathbf{1} = \mathbf{0}$, then we get the so called intrinsic conditional autoregressive (ICAR) model (Besag and Kooperberg, 1995). Now equations (4.4) and (4.5) still hold, however, Q is positive semi-definite and the variance matrix Q^{-1} no longer exists. An algebraic decomposition of the power term in (4.4) leads us to

$$p(\boldsymbol{\phi}) \propto \tau_\phi^{(I-B)/2} \exp\left\{\frac{1}{2}\sum_{i<j} Q_{ij}(\phi_i - \phi_j)^2\right\}. \tag{4.6}$$

Note that now ϕ_i's are actually non-identifiable, as we can add any same value to all ϕ_i's, and (4.6) still remains the same. In Bayesian implementation, we impose an identifying sum-to-zero constraint by centering the ϕ_i's around zero after each iteration (Carlin and Louis, 2000, p.263).

Spatial statistics practice usually furthermore specifies $\zeta_{ij} = \frac{1}{m_i} 1_{(i\sim j)}$, and $\kappa_i = 1/(m_i\tau_\phi)$, where m_i is the number of neighbors for area i, and τ_ϕ is a precision parameter. Then $Q = \tau_\phi W$, where $W_{ii} = m_i$ and $W_{ij} = -1_{(i\sim j)}$. Now (4.6) becomes

$$p(\boldsymbol{\phi}) \propto \tau_\phi^{(I-B)/2} \exp\left\{ -\frac{\tau_\phi}{2} \sum_{i<j} (\phi_i - \phi_j)^2 1_{(i\sim j)} \right\}, \tag{4.7}$$

Again by Brook expansion, (4.7) is equivalent to

$$\phi_i | \phi_{-i} \sim \mathrm{N}(\overline{\phi}_{\partial_i}, 1/(m_i\tau_\phi)), \quad \text{for } i = 1, \ldots, I, \tag{4.8}$$

where $\overline{\phi}_{\partial_i}$ is area i's neighbor mean of random effects. This conditional distribution (4.8) is to be used as prior for ϕ_i in our MCMC sampling algorithm.

4.2 Prior specifications and posterior computation

For coefficients γ_l, $l = 1, \ldots, K$, we choose an exponential prior, $\gamma_l \sim \exp(\eta)$, and a hyperprior for η, $\eta \sim \mathrm{Ga}(a_\eta, b_\eta)$. This specification leads to conjugate posteriors for both γ_l's and η. For spatial precision parameter τ_ϕ, we choose a gamma prior, $\tau_\phi \sim \mathrm{Ga}(a_\tau, b_\tau)$, which also leads to a conjugate posterior. For regression coefficients β_r, $r = 1, \ldots, p$, we assume a normal prior $\beta_r \sim \mathrm{N}(0, \sigma_0^2)$. The posterior full consitional for β_r is not conjugate, and we use the adaptive rejection Metropolis sampling (ARMS) algorithm (Gilks et al., 1995) for sampling. The posterior for each ϕ_i is not conjugate either. We use the Metropolis-Hastings algorithm for sampling.

After initializing values for the parameters, the proposed MCMC algorithm iterates through the following steps.

- Set $Z_{ij} = 0$ and $W_{ij} = 0$ for all i and j, $Z_{ijl} = 0$ and $W_{ijl} = 0$ for all i, j, and l. If $\delta_{ij1} = 1$, then sample

$$Z_{ij} \sim \text{Poi}(\Lambda_0(R_{ij})\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i))1_{(Z_{ij}>0)},$$

$$(Z_{ij1}, \ldots, Z_{ijK}) \sim \text{Multinomial}(Z_{ij}, \boldsymbol{p}_{ij}), \quad \boldsymbol{p}_{ij} \propto (\gamma_1 b_1(R_{ij}), \ldots, \gamma_K b_K(R_{ij})).$$

If $\delta_{ij2} = 1$, then sample

$$W_{ij} \sim \text{Poi}(\{\Lambda_0(R_{ij}) - \Lambda_0(L_{ij})\}\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i))1_{(W_{ij}>0)},$$

$$(W_{ij1}, \ldots, W_{ijK}) \sim \text{Multinomial}(W_{ij}, \boldsymbol{q}_{ij}),$$

$$\boldsymbol{q}_{ij} \propto (\gamma_1 \{b_1(R_{ij}) - b_1(L_{ij})\}, \ldots, \gamma_K \{b_K(R_{ij}) - b_K(L_{ij})\}).$$

- Sample β_r, $r = 1, \ldots, p$, using ARMS method from its full conditional distribution

$$p(\beta_r | Z'_{ij}s, W'_{ij}s, \boldsymbol{\Lambda}_0, \beta_{-r}, \boldsymbol{\phi}) \propto$$
$$\exp(\sum_{i=1}^{I}\sum_{j=1}^{n_i}(\boldsymbol{x}'_{ij}\boldsymbol{\beta})(Z_{ij}\delta_{ij1} + W_{ij}\delta_{ij2})$$
$$- \sum_{i=1}^{I}\sum_{j=1}^{n_i} e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\phi_i} \{\Lambda_0(R_{ij})(\delta_{ij1} + \delta_{ij2}) + \Lambda_0(L_{ij})\delta_{ij3}\})p(\beta_r),$$

where $p(\beta_r) = N(0, \sigma_0^2)$ is the prior used for β_r, and β_{-r} denotes all the β's except for β_r.

- Sample γ_l, $l = 1, \ldots, K$, from $\text{Ga}(a_{\gamma l}, b_{\gamma l})$, where

$$a_{\gamma l} = 1 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Z_{ijl}\delta_{ij1} + W_{ijl}\delta_{ij2}),$$

$$b_{\gamma l} = \eta + \sum_{i=1}^{I}\sum_{j=1}^{n_i} e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\phi_i} \{b_l(R_{ij})(\delta_{ij1} + \delta_{ij2}) + b_l(L_{ij})\delta_{ij3}\}.$$

- Sample η from $\text{Ga}(a_\eta + K, b_\eta + \sum_{l=1}^{K}\gamma_l)$.

- Sample ϕ_i, $i = 1, \ldots, I$, using MH from its full conditional distribution

  $p(\phi_i | Z'_{ij}s, W'_{ij}s, \boldsymbol{\Lambda}_0, \boldsymbol{\beta}, \phi_{-i}) \propto$
  $\exp(\sum_{j=1}^{n_i} \phi_i(Z_{ij}\delta_{ij1} + W_{ij}\delta_{ij2}) - \sum_{j=1}^{n_i} e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}+\phi_i} \{\Lambda_0(R_{ij})(\delta_{ij1} + \delta_{ij2}) + \Lambda_0(L_{ij})\delta_{ij3}\})$
  $p(\phi_i | \phi_{-i})$,

  where $p(\phi_i|\phi_{-i})$ denotes the prior in (4.8).

- Based on prior (4.7), sample τ_ϕ from $\text{Ga}(\frac{I-1}{2} + a_\tau, b_\tau + \frac{1}{2}\sum_{i<j}(\phi_i - \phi_j)^2 1_{(i\sim j)})$.

4.3 Model Comparison

To evaluate the performance of our proposed model, we compared it to a Weibull proportional hazards model with spatial random effects. The later can be easily fitted in WinBUGS, which also allows the use of the CAR distribution as prior for the spatial random effect vector. We consider the following two Bayesian model selection criteria: the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and pseudo-marginal likelihood (PsML) (Geisser and Eddy, 1979; Gelfand et al., 1992).

DIC is derived for the purpose of comparing complex hierarchical models. It is a Bayesian analogue of AIC, and can be readily computed in Markov chain Monte Carlo analysis. Let $\boldsymbol{\omega}$ be the set of all parameters in the model of interest, the Bayesian deviance is defined as

$$\text{D}(\boldsymbol{\omega}) = -2\text{log}f(\boldsymbol{y}|\boldsymbol{\omega}) + 2\text{log}h(\boldsymbol{y}), \tag{4.9}$$

where $f(\boldsymbol{y}|\boldsymbol{\omega})$ is the likelihood for the observed data $\boldsymbol{y}$ and $h(\boldsymbol{y})$ is some standardizing function of data alone. Since the second term in (4.9) is free of $\boldsymbol{\omega}$, then for our comparison, it can be ignored in deviance calculation given that we are comparing models based on the same data set. Based on Bayesian deviance, a measure for the effective number of parameters in a model is defined as

$$p_D = \overline{D} - D(\overline{\boldsymbol{\omega}}),$$

where $\overline{D} = E(D|\boldsymbol{y})$ is the posterior mean of deviance, and $D(\overline{\boldsymbol{\omega}}) = D(E(\boldsymbol{\omega}|\boldsymbol{y}))$ is the deviance at posterior means of parameters. DIC is defined as the sum of $\overline{D}$ and p_D in parallel of the definition of AIC and BIC, where $\overline{D}$ measures the model fit. Similar to AIC and BIC, a smaller value of DIC indicates a better model.

We also compared the two models with the PsML based on conditional predictive ordinate (CPO). For each observation i, CPO is defined as the cross-validation posterior predictive density conditional on the rest of the observations (Dey, et al., 1997):

$$\text{CPO}_i = f(y_i|y_{-i}) = \int f(y_i|\boldsymbol{\omega}, y_{-i}) f(\boldsymbol{\omega}|y_{-i}))d\boldsymbol{\omega},$$

where y_{-i} denotes all the observations but observation i. A Monte Carlo estimate of CPO_i is given by

$$\text{CPO}_i = \left[\frac{1}{T}\sum_{t=1}^{T}\left(1/f(y_i|\boldsymbol{\omega}^{(t)})\right)\right]^{-1}$$

where $f(y_i|\boldsymbol{\omega}^{(t)})$ is the ith individual likelihood evaluated at iteration t. The product of CPOs, $\prod_{i=1}^{n} f(y_i|y_{-i})$, called psdudo-marginal likelihood (PsML), has been proposed to be a surrogate for $f(\boldsymbol{y})$ by Geisser and Eddy (1979). The negative natural log of PsML is called negative cross-validatory log likelihood (NLLK) and is used for model comparison (Cooper et al., 2003). A smaller value of NLLK implies a better model. Note that for comparing two models $M1$ and $M2$,

$$\text{NLLK}_1 - \text{NLLK}_2 = -\log(\frac{\prod_{i=1}^{n} f(y_i|y_{-i}, M_1)}{\prod_{i=1}^{n} f(y_i|y_{-i}, M_2)}),$$

where the term

$$\frac{\prod_{i=1}^{n} f(y_i|y_{-i}, M_1)}{\prod_{i=1}^{n} f(y_i|y_{-i}, M_2)}$$

is the pseudo-Bayes factor (Geisser and Eddy, 1979; Gelfand and Dey, 1994). An advantage of the pseudo-Bayes factor and the NLLK is that they can be used with improper priors.

4.4 A Simulation Study

We evaluated the performance of our model through a simulation study. A total of 100 data sets were generated. For each data set, the spatial layout is based on the 46 counties in South Carolina, with 10 subjects in each county. For each data set, failure times were generated from the mixed PH model:

$$S(t|x_{ij1}, x_{ij2}) = \exp\left\{-\Lambda_0(t)\exp(\beta_1 x_{ij1} + \beta_2 x_{ij2} + \phi_i)\right\}, \tag{4.10}$$

where $\Lambda_0(t) = \log(1+t)$, $\beta_1 = 1$, $\beta_2 = 1$, $\tau_\phi = 4$, x_{ij1}'s $\sim$ Bernoulli(0.5), and x_{ij2}'s $\sim \mathrm{N}(0, 0.5^2)$. We assume that the random number of medical examinations performed for each person is 1 plus a Poisson random number with mean 2. The gap times between adjacent medical examinations follow an exponential distribution with mean 1. The observed interval is formed by the consecutive examination times (including 0 and ∞) that contains the true failure time. To generate spatial random effect ϕ_i's, we first generated ϕ_i*'s from multivariate normal $N_I(\mathbf{0}, (\tau_\phi W^*)^{-1})$, where $W^* = W + \mathrm{diag}(0.0001, I)$, and then centered ϕ_i*'s around zero to get ϕ_i's. We imported W^* to make the precision matrix invertible, and put the sum to zero constraint to make the ϕ_i's identifiable.

To construct the monotone splines, we set the order of I-splines as 2 and chose 15 equally spaced knots within the range of observed times. For hyper-parameters, we set $\sigma_0^2 = 100$, $a_\eta = 1$, and $b_\eta = 1$. For the spatial precision parameter τ_ϕ, we select $a_\tau = 8$ and $b_\tau = 2$ to give a priror with over 99% densities between 0 and 10. We also fitted the same 100 data sets using a spatial Weibull PH model in WinBUGS. For each MCMC sampling of both our model and the Weibull model, we drew 11,000 iterations and discard the first 1,000 as a burn-in. We compared the two models based on frequentist operating characteristics of estimated coefficients, model goodness of fit, and estimation of the baseline survival function $S_0(t)$.

From the theory of Markov chains, we would expect our chains to converge to

their stationary distributions. Although we can never be sure of convergence, there are several tests available in the 'coda' package in R to see if the chains appear to be converged. Here we show the traceplots (Figure 4.2, Figure 4.3, and Figure 4.4) and Gewke's diagnostic (Geweke, 1992) for regression coefficients β_1, β_2, and spatial precision parameter τ_ϕ, from the first simulated dataset. Basically, the Geweke's diagnostic produces a Z-score for a test of equality of means between the first and last parts of the chain. Based on the Z-score, a corresponding p-value can be calculated. Table 4.1 presents the z-scores and p-values for the Geweke's tests. All the p-values are greater than 0.05. Based on both graphic and non-graphic diagnostics, it seems that the MCMC chains mix well and converge.

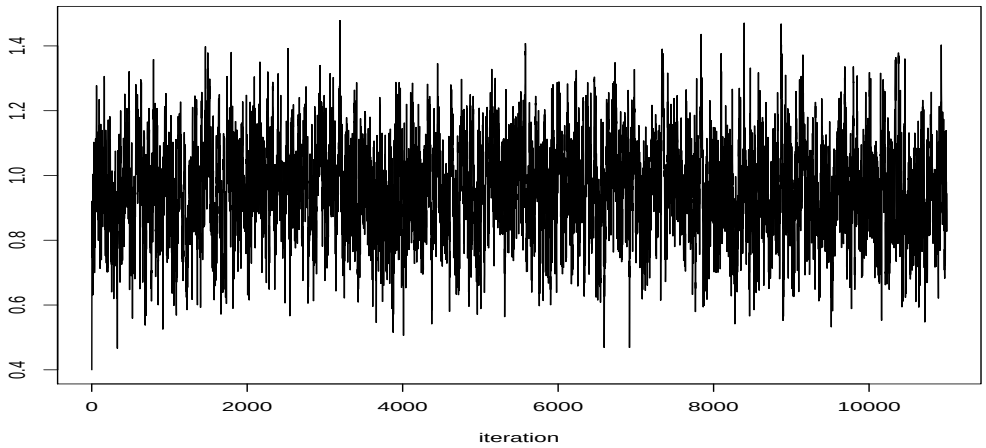

Figure 4.2 Traceplot of fixed effect β_1 in the 1st simulated dataset.

Table 4.2 summarizes the estimation results of our model versus the Weibull model. For each parameter, the point estimate is the average of the 100 posterior means, the empirical standard error (ESE) is the average of the 100 estimated standard errors, the sample standard deviation (SSD) is the sample standard deviation of the 100 posterior means, and the 95% coverage probability (95CP) is the percent of the 100 credible intervals for each β_r that contains the true parameter value. We can see that the proposed model works very well. The percent bias is only 4.7% for

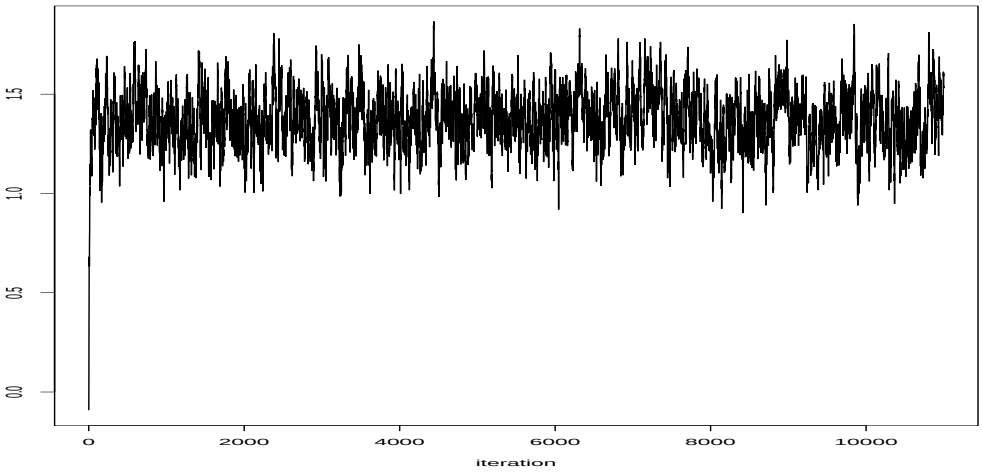

Figure 4.3 Traceplot of fixed effect β_2 in the 1st simulated dataset.

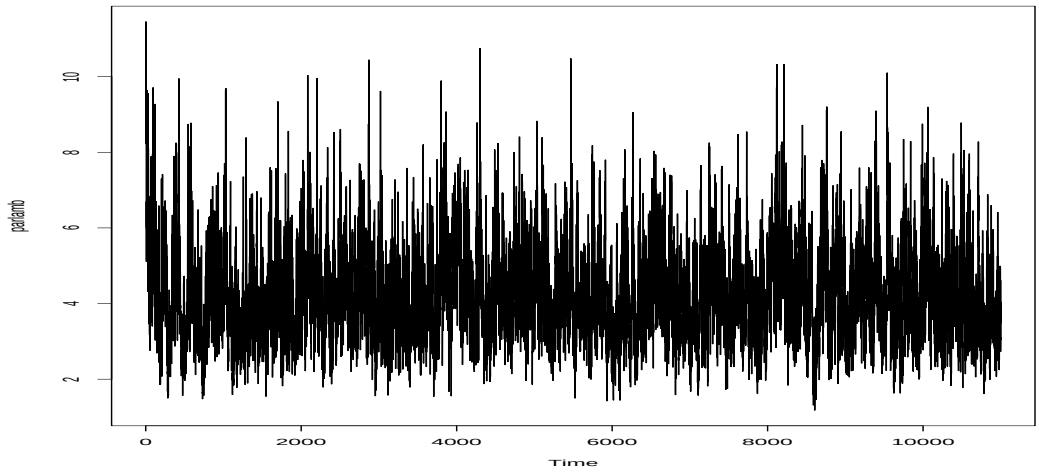

Figure 4.4 Traceplot of spatial precision τ_ϕ in the 1st simulated dataset.

β_1 and 1.4% for β_2. The coverage probabilities are close to the nominal level of 0.95. Both models provide similar precision in estimation. In contrast, the point estimates from the Weibull model are very biased, as the percent bias is 32.6% for β_1 and 32.0% for β_2. The magnitudes of DIC and NLLK under the Weibull model are much larger than those under the proposed model, indicating a decisive evidence of favoring the proposed model (Kass and Raftery, 1995).

We estimated $S_0(t)$ at 100 equally spaced points $\{g_1, \ldots, g_{100}\}$ within the data

Table 4.1 Geweke's convergence diagnostic for MCMC chains of the model parameters in the 1st simulated dataset

Paramter	Z-score	p-value
β_1	-1.5372	0.1242
β_2	-0.5737	0.5662
τ_ϕ	0.7077	0.4791

Table 4.2 Estimation results for simulation 1

		Proposed Model				Weibull Model			
	True	Estimate	SSD	ESE	95CP	Estimate	SSD	ESE	95CP
β_1	1	1.047	0.156	0.141	0.93	0.674	0.133	0.112	0.21
β_2	1	1.014	0.121	0.133	0.97	0.680	0.117	0.113	0.19
τ_ϕ	4	3.954	0.547	1.247	1.00	4.431	0.507	1.323	1.00
DIC			726					1149	
NLLK			364					520	

range using the proposed model and the Weibull model. The estimated baseline survival functions and the true baseline survival function are plotted in Figure 4.5. For the proposed model, the 95% credible interval for each $S_0(g_d)$ averaged over the 100 data sets are also plotted (dotted lines), where $d = 1, \ldots, 100$. The estimated $S_0(t)$ curve from our model provides a very good approximation to the true $S_0(t)$ curve. Also our 95% pointwise credible intervals totally cover the true baseline survival

curve. On the other hand, the estimated $S_0(t)$ curve from the Weibull model only provides a reasonably good estimation between time unit 0 and 1, then it diverges from the true curve very fast. This graph demonstrates that parametric assumption of baseline survival time can be too rigid. On the contrary, the proposed semi-parametric formulation is fairly data adaptive.

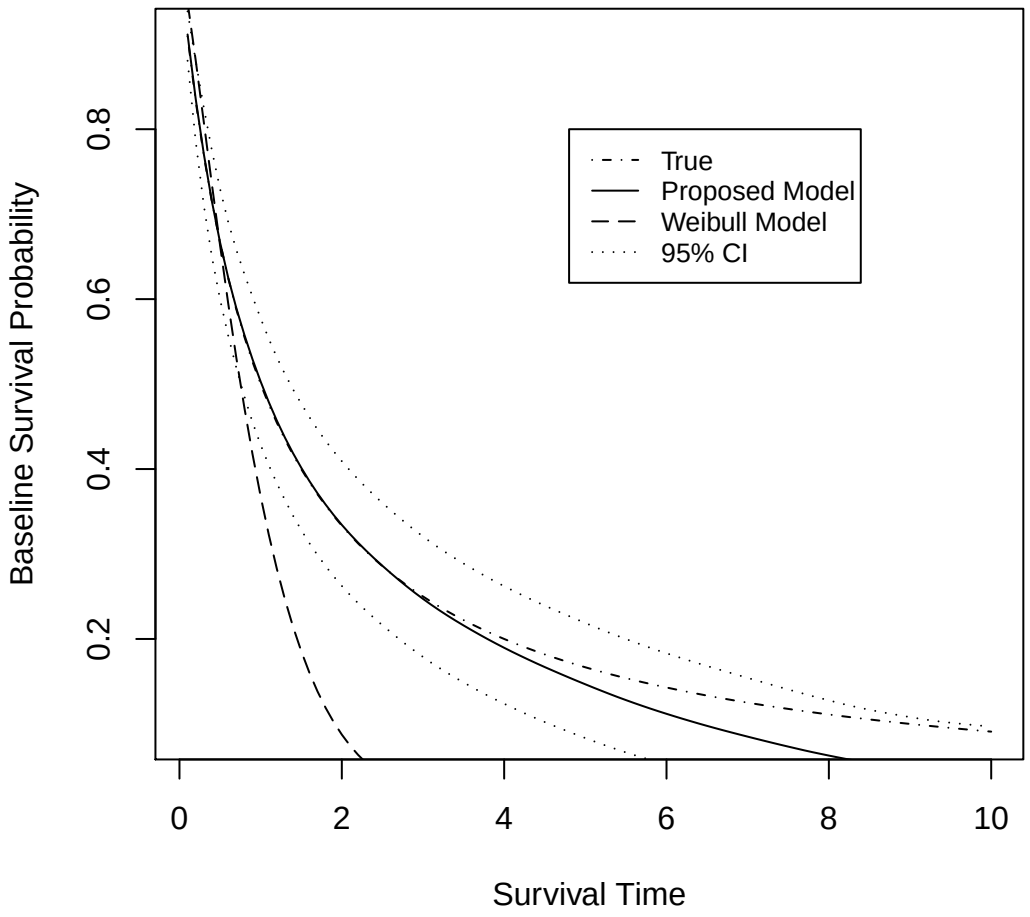

Figure 4.5 Plot of estimated baseline survival curve based on 100 simulated data sets using proposed model (with 95% pointwise credible intervals) and Weibull model, compared to true baseline survival curve (simulation 1).

To investigate the spatial pattern of survivorship, we plotted the posterior means

of the spatial random effects (ϕ_i's) based on both the proposed model and the Weibull model for a randomly selected simulated data set (Figure 4.6). Note that each frailty corresponds to a county in South Carolina in the map. The two maps in Figure 4.6 use the same grayscale. Both maps show higher values of ϕ_i in the middle part of the state.

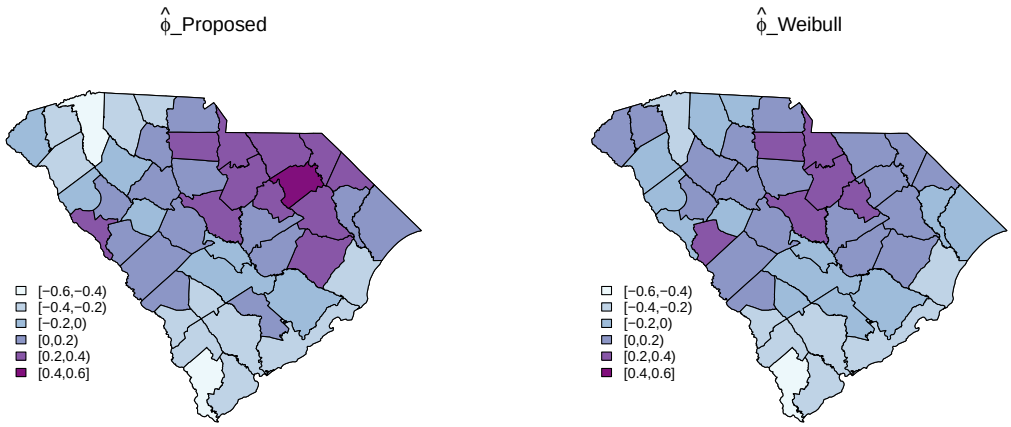

Figure 4.6 Maps of posterior means for the spatial random effects ϕ_i over 46 counties of SC based on proposed model and Weibull model (simulation 1).

To give the Weibull PH model with spatial random effect in WinBUGS an advantage so as to further test our proposed model, we performed a second simulation by generating data based on a Weibull PH model with spatal random effect. We let $\lambda_0(t) = rt^{r-1}$, $r = 1.5$ in model (4.10). Other parameters are the same as in the first data model. We fit both our model and the spatial Weibull PH model in WinBUGS, with $r \sim \text{Ga}(1, 1)$ as prior for r, and other priors and hyperparameters are the same as in the first simulation setting. The convergence of the MCMC chains are

checked with traceplots and Geweke's diagnostic (not shown). The estimation results are summarized in Table (4.3). As we expected, the estimates from the Weibull PH model with spatial random effect are very close to the true values, since the simulated data are generated from it. For this different failure time distribution, our model still gives accurate estimation for the regression coefficients and the spatial precision. The coverage probabilities for β_1 and β_2 are close to the nominal 95% level. This implies the robustness of our method to different underlying distributions for T. The DIC and NLLK values from the two models are very close, which furthermore shows our model is as good as the Weibull model for this parametric setting.

Table 4.3 Estimation results for simulation 2

		Proposed Model				Weibull Model			
	True	Estimate	SSD	ESE	95CP	Estimate	SSD	ESE	95CP
β_1	1	0.993	0.187	0.158	0.92	1.001	0.210	0.155	0.90
β_2	1	1.006	0.155	0.154	0.95	1.015	0.150	0.154	0.94
τ_ϕ	4	4.057	0.514	1.297	1.00	4.085	0.506	1.299	1.00
r	1.5					1.505	0.123	0.106	0.96
DIC			525					527	
NLLK			263					264	

The baseline survvial $S_0(t)$ at 100 equally spaced points $\{g_1, \ldots, g_{100}\}$ within the data range are also estimated using our model and the Weibull model. Figure 4.7 is the plot of the estimated survival functions and the true function, with 95% pointwise credible intervals from our model for each $S_0(g_d)$, $d = 1, \ldots, 100$. The estimated $S_0(t)$ curves from our model and the Weibull model almost match exactly with the

true $S_0(t)$ curve. And the pointwise credible intervals totally cover the true baseline survival curve.

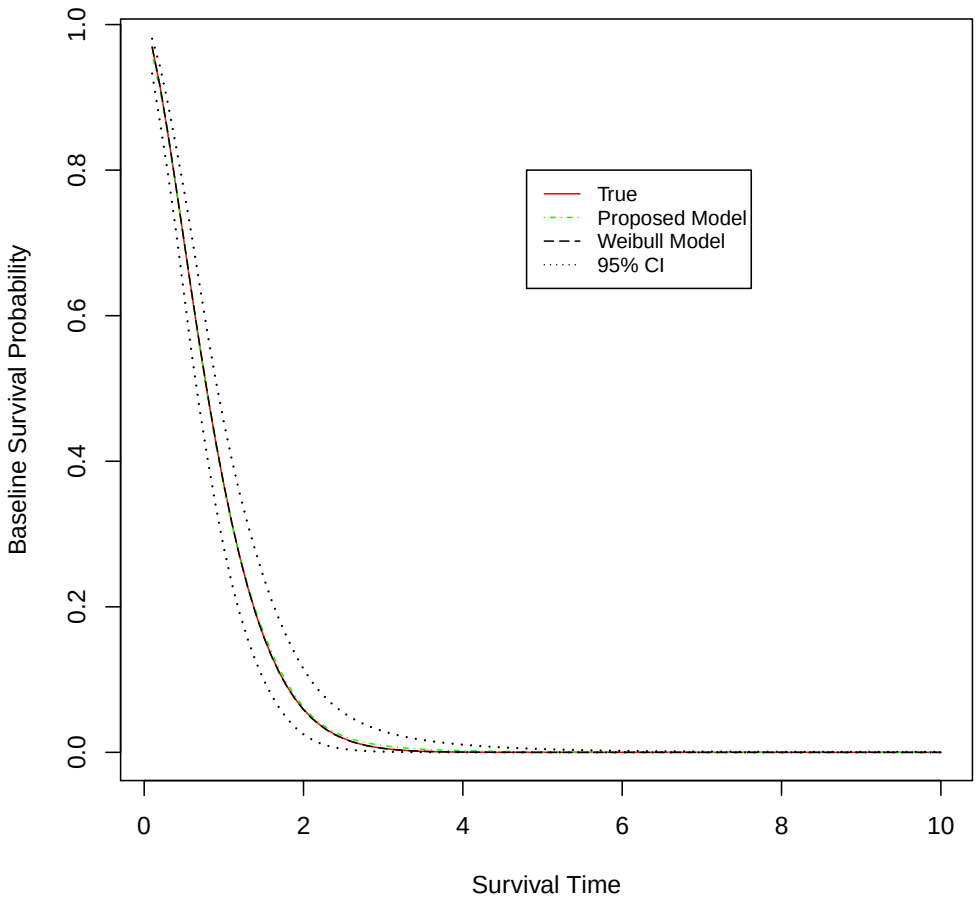

Figure 4.7 Plot of estimated baseline survival curve based on 100 simulated data sets using proposed model (with 95% pointwise credible intervals) and Weibull model, compared to true baseline survival curve (simulation 2).

4.5 Smoking-Relapse Data Application

We now apply our proposed model to a geographically referenced smoking-relapse data set consisting of 223 subjects living in 51 zip code areas in the southeastern corner of Minnesota. This data is a subset from a lung health study carried out by Murray

et al. (1998) on the effect of intermittent smoking on pulmonary function. The 223 smokers selected all have quit smoking at least once during the study period and have a Minnesota zip code of residence. The event of interest is relapse to smoking. Each zip code area forms a spatial cluster. Subjects were monitored at annual visits for 5 years. The time origin for each subject is the study entry time. There are four prognostic factors of interest: gender, duration as a smoker (years), treatment (smoking intervention (SI), usual care (UC)), and average number of cigarettes per day over the last 10 years. Each subject has two time points referring to the observed interval where true time to relapse falls in, so it is an interval-censored data set. One thing to point out is that there is actually no left-censoring in this particular data set. However, our method is certainly applicable for data sets that contain all three types of censorship.

For the analysis, our goal is to estimate the effect of the four prognostic factors on time to relapse to smoking, after adjusting for the spatial dependency within and among clusters. We analyzed the data using both our proposed model and the spatial Weibull PH model in WinBUGs. For the MCMC chains of regression parameters β_1 through β_4, we looked at traceplots (Figure 4.8 to Figure 4.11) and the Geweke's diagnostic (Table 4.4) to examine convergence. The traceplots show that all of the MCMC chains mix very well and are very stable. Also, all of the p-values for Geweke's diagnostic are greater than 0.05, indicating convergence for the MCMC chains.

The regression estimation results are presented in Table 4.5. The estimated regression coefficients from both models have same direction and similar magnitude. However, smoking intervention is significantly effective in reducing relapse risk based on our model, while the Weibull model fails to detect this significance. Duration of smoking is also significant based on our model, although the magnitude is small. Duration as a smoker may be confounded with age, and younger people may tend to relapse more often. The DIC and NLLK values are lower for the proposed model

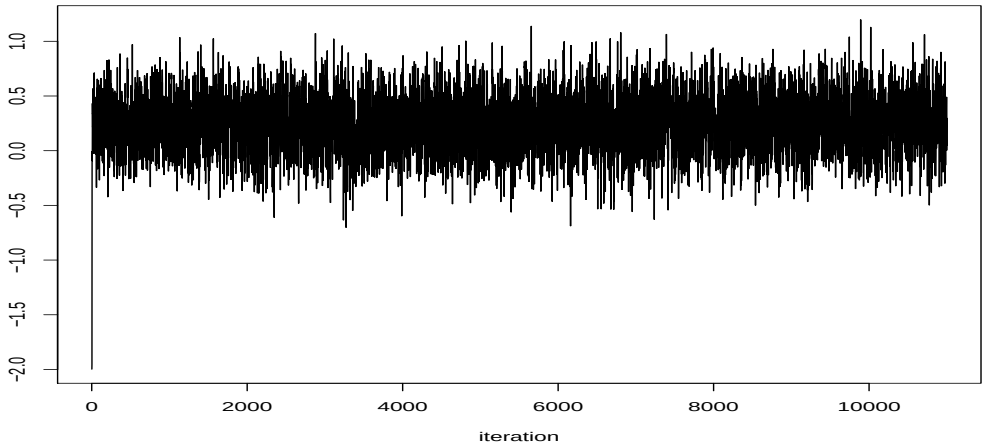

Figure 4.8 Traceplot of fixed effect β_1 in the smoking-relapse data.

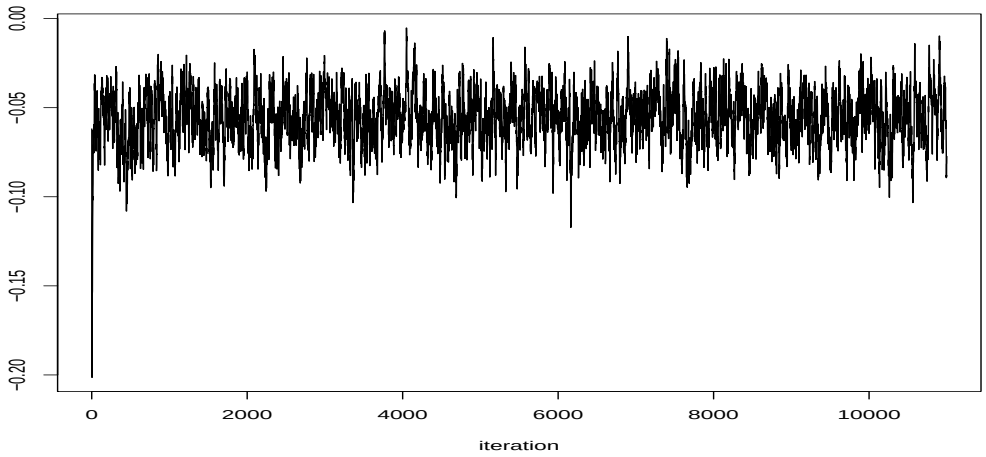

Figure 4.9 Traceplot of fixed effect β_2 in the smoking-relapse data.

than for the Weibull model, which implies that our model has a better fit for the data.

We also checked the performance of the proposed model from the perspective of survivorship estimation. In Figure 4.12, we plotted the nonparametric maximum likelihood estimates using the Turnbull method (Turnbull, 1976; Giolo, 2004), the proposed estimates, and the Weibull estimates of the survival functions for the four gender and treatment interaction groups. Turnbull (1976) proposed a nonparametric

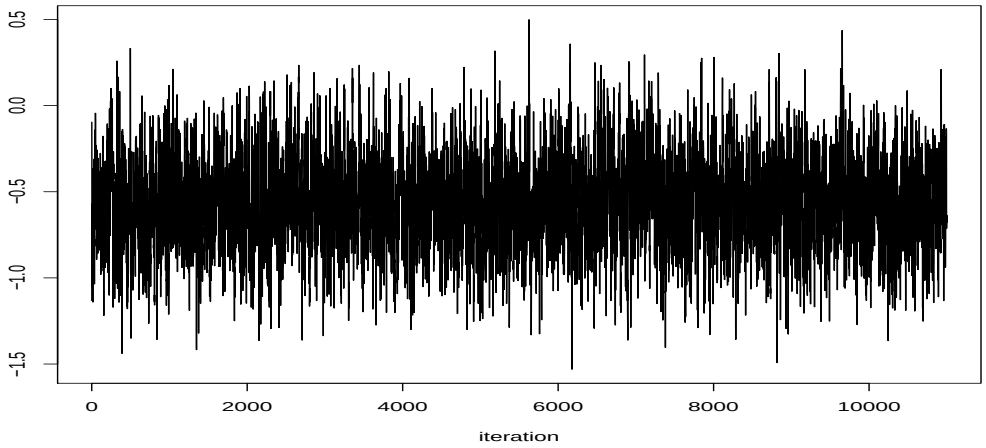

Figure 4.10 Traceplot of spatial precision τ_ϕ in the smoking-relapse data.

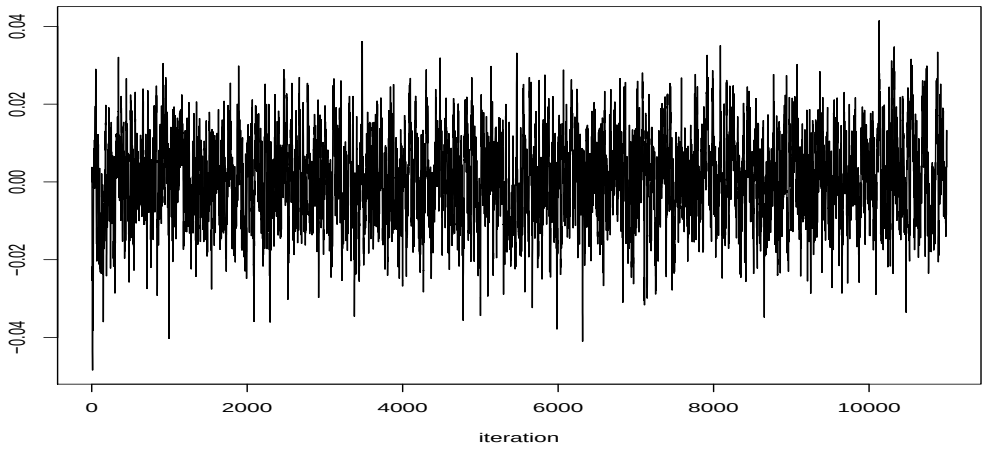

Figure 4.11 Traceplot of spatial precision τ_ϕ in the smoking-relapse data.

maximum likelihood estimator of survival function for interval-censored data. The estimated curves from the proposed model are close to the Turnbull curves, indicating that our model fit the data well. In contrast, the Weibull model seems to constantly underestimate the true survival probabilities, and the magnitude of divergence from the other two methods increases quickly after around 1.5 years.

From the plot of Turnbull estimates, males with smoking intervention has the highest survivorship, and the rest three groups are roughly similar to each other.

Table 4.4 Geweke's convergence diagnostic for MCMC chains of the model parameters in the smoke-relapse data

Paramter	Z-score	p-value
β_1	-1.4203	0.1556
β_2	-1.4716	0.1411
β_3	-1.2607	0.2074
β_4	-0.8535	0.3934

From the curves based on the proposed model and the Weibull model, the survival probability is the highest for males with smoking intervention, followed by females with smoking intervention, males with usual care, and finally females with usual care. Figure 4.12 suggests that smoking intervention is effective in reducing the risk of relapse, and women may be more likely to relapse than men. This agrees with the regression estimates in Table 4.5.

To explore the spatial pattern, in Figure 4.13, we mapped the posterior means of spatial frailties (ϕ_i's) for the smoking data, using the proposed model and the Weibull model. Note that a same grayscale is used for both maps. There are 83 zip code areas in total, with only 51 of them containing data. The 32 zip code areas without data are plotted in white color. The map based on the proposed model shows higher values for ϕ_i's in the northwest region, which indicates higher risks of relapse in this region. This pattern also agrees with the finding from Banerjee and Carlin (2004), where they fitted a spatial MCAR Weibull cure rate model. The map based on the Weibull model also show somewhat higher values for ϕ_i in the same region, but in a less accentuated way.

Table 4.5 Estimation results for the smoking cessation study

	Proposed Model		Weibull Model	
	Estimate	95%CI	Estimate	95%CI
Gender (male=0)	0.255	(-0.246,0.747)	0.374	(-0.122, 0.870)
Duration as a smoker	-0.056	(-0.084, -0.029)	-0.030	(-0.065, 0.004)
Treatment(UC=0)	-0.555	(-1.065, -0.025)	-0.405	(-0.929, 0.153)
Cigarettes per day	0.001	(-0.020, 0.022)	0.009	(-0.013, 0.030)
r	—	—	2.218	(1.652, 2.782)
DIC		399		456
NLLK		200		224

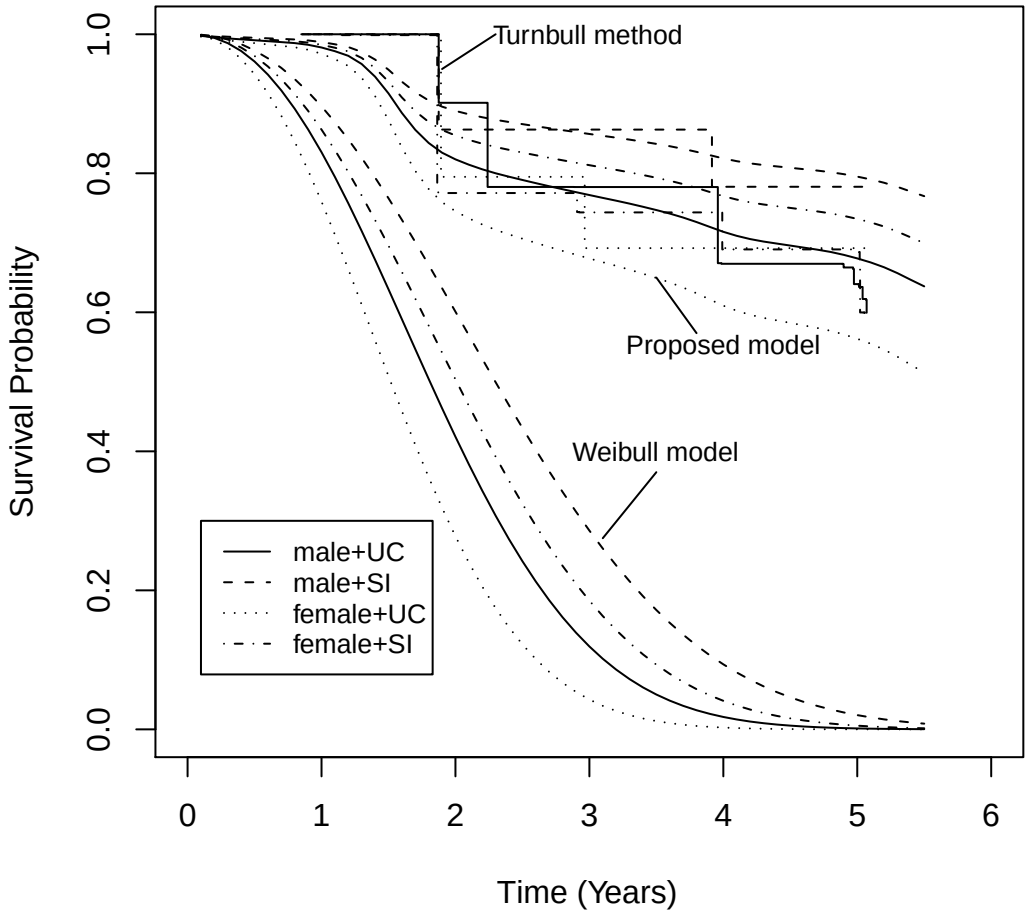

Figure 4.12 Estimated survival curves for the smoking cessation study, using Turnbull method, proposed model, and Weibull model; event of interest is time to relapse to smoking. Four curves are plotted for each method based on the four subgroups formed by gender and treatment.

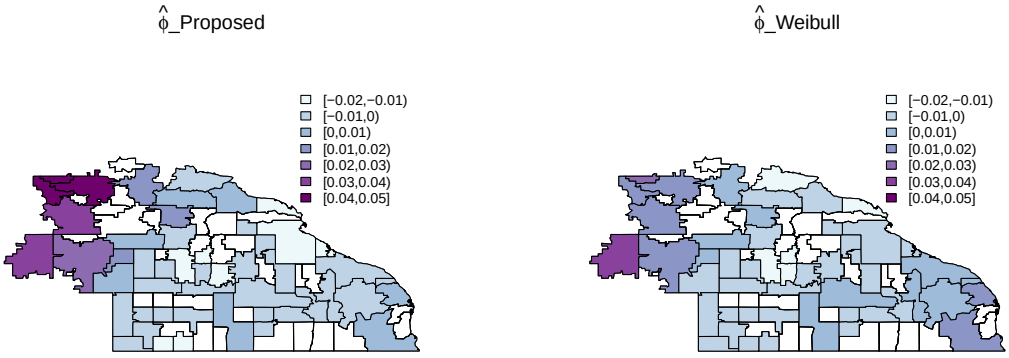

Figure 4.13 Maps of posterior means for the spatial random effects ϕ_i over 51 zip codes areas in southeast Minnesota based on proposed model and Weibull model. The other 32 zip code areas without data are plotted in white color.

Chapter 5

Multivariate Frailty Model for Clustered Interval-Censored Data

5.1 Introduction

In medical studies, time-to-event of interest may not be directly observed but is known to fall within a certain interval. Such data are referred to as interval-censored (Finkelstein and Wolfe, 1986) and they occur naturally in infectious disease studies or oncology clinical trials where patients are periodically checked by lab examinations for infection or cancer progression. Furthermore, event times in a cluster may be correlated. For example, in multi-center clinical trials, patients and clinical practice characteristics that vary by center may exert an influence on treatment outcome. If these effects are sufficiently powerful, then inferences that ignore the clustering can be misleading.

The concept of frailty, first introduced by Vaupel et al. (1979), is widely used in survival analysis to characterize the heterogeneity among clusters. Goethals et al. (2009) modeled interval-censored clustered cow udder quarter infection times by proportional hazards (PH) model with a gamma frailty. Zhang and Sun (2010) proposed a multiplier for survival function based on cluster size or a within-cluster resampling procedure to avoid the correlation issue with clusters. Kim (2010) modeled failure times and cluster size jointly using PH model with normal random effect and a mixed ordinal regression model. Those methods have focused on one shared frailty or equivalently one random intercept term to account for variation in baseline risk across

clusters. A natural extension is to also allow for variation of treatment effect across clusters. This has been done for right-censored failure times (Ripatti and Palmgren, 2000; Vaida and Xu, 2000). Furthermore, instead of directly presuming a random effect term, one may first want to test its existence.

To address the problem of identifying random effects in survival analysis, Dunson and Chen (2004) developed a multivariate frailty model for right-censored event times. Defining a PH model with both random intercept and random effect for binary predictors, they assigned priors for the frailty variances that are mixtures of point mass at zero and inverse-Gamma densities. Cai (2010) extended Dunson and Chen (2004) method to allow for nonparametric modeling for log of the multivariate frailties. Here we modified the Dunson and Chen (2004) approach to the problem of selecting random effects in interval-censoring setting when there is a subset of predictors with possibly heterogeneous coefficients. In addition, we model the cumulative hazard function nonparametrically. A Markov chain Monte Carlo (MCMC) algorithm is developed to estimate the probability that a frailty exists and to obtain the Bayes Factor for testing. The inferences on population parameters are model-averaged in the sense that we have accounted for the uncertainty in the frailty selection. Although we follow the frailty selection procesure in Dunson and Chen (2004), the specific interval-censoring data structure necessitates a fundamentally different model construction process.

5.2 Prior and Augumentation for Frailties

Consider the following multivariate frailty proportional hazards model:

$$\lambda_{ij}(t|\boldsymbol{x}_{ij}) = \lambda_0(t)\left\{\xi_{i1}\prod_{h=1}^{H}\xi_{i,h+1}^{x_{ijh}}\right\}exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}) \tag{5.1}$$

where there are $i = 1, \ldots, I$ clusters, and $j = 1, \ldots, n_i$ subjects within each cluster i. The x_{ijh}, $h = 1, \ldots, H$, indicate binary predictors, ξ_{i1} is the baseline frailty, ξ_{i2} to $\xi_{i,H+1}$ are the frailties related to x_{ij1} to x_{ijH}.

For each frailty ξ_{ih}, $h = 1, \ldots, H+1$, we specify its prior as

$$\pi(\xi_{ih}; \kappa_h) = \begin{cases} 1(\xi_{ih} = 1), & if \kappa_h = 0 \\ \text{Ga}(\xi_{ih}; \kappa_h^{-1}, \kappa_h^{-1}), & if \kappa_h > 0 \end{cases}$$

where the Gamma density has mean 1 and variance κ_h. The mean of 1 makes the average hazard identifiable. And the variance κ_h measures the degree of between-cluster variability and thus the level of within-cluster dependence (Glidden and Vittinghoff, 2004). As we know, if $\kappa_h = 0$, then $\xi_{ih} \equiv 1$ for all i. This corresponds to the null hypothesis of homogeneity, or equivalently, zero within-cluster correlation. We choose the following mixed prior for κ_h:

$$\pi(\kappa_h) = 1(\kappa_h = 0)\pi_{0h} + 1(\kappa_h > 0)(1 - \pi_{0h})\text{IG}(\kappa_h; a_h, b_h), \tag{5.2}$$

where $\pi_{0h} = \Pr(H_0 : \kappa_h = 0)$ is the prior probability of the null hypothesis of homogeneity for the hth frailty, and $\text{IG}(\cdot; a_h, b_h)$ is the inverse Gamma prior of κ_h under the alternative hypothesis $H_1 : \kappa_h > 0$. We will refer (5.2) as a zero-inflated inverse Gamma density, ZI-IG$(\cdot; a_h, b_h)$.

We employ the data augmentation introduced in Dunson and Chen (2004) to prevent MCMC samples of κ_h from "stuck" at zero for a long period of iterations. Let $\delta_h = 1(\kappa_h = 0)$ be an indicator that equals 1 if $H_0 : \kappa_h = 0$ holds and 0 otherwise. Then for computational purpose, introduce latent variables $\tilde{\kappa}_h$ and $\tilde{\xi}_{ih}$ as follows:

$$\kappa_h = (1 - \delta_h)\tilde{\kappa}_h \quad \xi_{ih} = \tilde{\xi}_{ih}^{1-\delta_h}, \quad \text{for} \quad i = 1, \ldots, I, \quad h = 1, \ldots, H+1.$$

The following prior density results in the original model formulation:

$$\pi(\tilde{\kappa}_h, \tilde{\boldsymbol{\xi}}_h, \delta_h) = \text{IG}(\tilde{\kappa}_h; a_h, b_h) \left\{ \prod_{i=1}^{I} \text{Ga}(\tilde{\xi}_{ih}; \tilde{\kappa}_h^{-1}, \tilde{\kappa}_h^{-1}) \right\} \pi_{0h}^{\delta_h} (1 - \pi_{0h})^{1-\delta_h}.$$

5.3 Modeling Interval-Censored Data

For interval-censored data, we need to estimate baseline cumulative hazard function . We approximate it using a linear combination of I-splines (Cai et al, 2011):

$$\Lambda_0(t) = \sum_{l=1}^{K} \gamma_l b_l(t), \tag{5.3}$$

where $\{b_l(t)\}$ is a set of I-splines (Ramsay, 1988), each of which is nondecreasing from 0 to 1, and $\{\gamma_l\}$ is a set of nonnegative coefficients. So $\Lambda_0(t)$ is nondecreasing.

Let $(L_{ij}, R_{ij}]$ denote the observed time interval for the jth subject in the ith cluster, then conditional on frailties, the likelihood function is: $L(\boldsymbol{\Lambda}_0, \boldsymbol{\beta}, \boldsymbol{\xi}) =$

$$\prod_{i=1}^{I} \prod_{j=1}^{n_i} \left\{F(R_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\xi}_i)\right\}^{\delta_{ij1}} \left\{F(R_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\xi}_i) - F(L_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\xi}_i)\right\}^{\delta_{ij2}} \left\{1 - F(L_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{\xi}_i)\right\}^{\delta_{ij3}}.$$

We formulate the likelihood using a two-step data augmentation to facilitate posterior computation (Lin et al., 2013). It can be shown that under our model in (5.1), the failure time of interest T_{ij} is equivalent to the time of the 1st occurrence of a nonhomogeneous Poisson counting process $\{N(t) : t > 0\}$ with mean function $\Lambda_0(t)\left\{\xi_{i1}\prod_{h=1}^{H}\xi_{i,h+1}^{x_{ijh}}\right\}\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta})$. Let $t_{ij1} < t_{ij2}$ be two time points such that $t_{ij1} = R_{ij}$ if left-censoring, $t_{ij1} = L_{ij}$ and $t_{ij2} = R_{ij}$ for interval censoring, and $t_{ij2} = L_{ij}$ for right-censoring, we introduce two independent latent variables: $Z_{ij} = N(t_{ij1}) \sim \text{Poi}(\Lambda_0(t_ij1)\left\{\xi_{i1}\prod_{h=1}^{H}\xi_{i,h+1}^{x_{ijh}}\right\}\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}))$ and $W_{ij} = N(t_{ij2}) - N(t_{ij1}) \sim \text{Poi}((\Lambda_0(t_{ij2}) - \Lambda_0(t_{ij1}))\left\{\xi_{i1}\prod_{h=1}^{H}\xi_{i,h+1}^{x_{ijh}}\right\}\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}))$.

Based on the additive property of Poisson distribution and the form of $\Lambda_0(t)$ in (5.3), we further decompose as follows: $Z_{ij} = \sum_{l=1}^{K} Z_{ijl}$, $W_{ij} = \sum_{l=1}^{K} W_{ijl}$, Z_{ijl}'s, and W_{ijl}'s are mutually independent, $Z_{ijl} \sim \text{Poi}(\gamma_l b_l(t_{ij1})\xi_{i1}\left\{\prod_{h=1}^{H}\xi_{i,h+1}^{x_{ijh}}\right\}\exp(\boldsymbol{x}_{ij}'\boldsymbol{\beta}))$, $W_{ijl} \sim \text{Poi}((\gamma_l b_l(t_{ij2}) - \gamma_l b_l(t_{ij1}))\xi_{i1}\left\{\prod_{h=1}^{H}\xi_{i,h+1}^{x_{ijh}}\right\}\exp(\boldsymbol{x}'\boldsymbol{\beta}))$, with $\sum_{l=1}^{K} Z_{ijl} > 0$ if $\delta_{ij1} = 1$, $\sum_{l=1}^{K} Z_{ijl} = 0$ and $\sum_{l=1}^{K} W_{ijl} > 0$ if $\delta_{ij2} = 1$, and $\sum_{l=1}^{K} Z_{ijl} = 0$ and $\sum_{l=1}^{K} W_{ijl} = 0$ if $\delta_{ij3} = 1$.

Then the augmented data likelihood is:

$$L_{aug} = \prod_{i=1}^{I}\prod_{j=1}^{n_i}\left\{\prod_{l=1}^{K}\text{Poi}(Z_{ijl})\text{Poi}(W_{ijl})^{\delta_{ij2}+\delta_{ij3}}\right\} \tag{5.4}$$

$$\{1(Z_{ij}>0)\}^{\delta_{ij1}}\{1(Z_{ij}=0)\}^{\delta_{ij2}}\{1(W_{ij}>0)\}^{\delta_{ij2}}\{1(Z_{ij}=0)\}^{\delta_{ij3}}\{1(W_{ij}=0)\}^{\delta_{ij3}}$$

5.4 Gibbs Sampling Algorithm

Based on the augmented data likelihood (5.4), we can derive the full conditional distributions for our parameters. The Gibbs sampler for posterior computation can be outlined as the following three parts: i) updating Poisson latent variables Z_{ij}'s, Z_{ijl}'s, W_{ij}'s, and W_{ijl}'s; ii) updating the parameters related to baseline cumulative hazard $\Lambda_0(t)$; iii) updating the regression coefficient parameters; iv) updating $\boldsymbol{\delta}$, $\boldsymbol{\kappa}$, and $\boldsymbol{\xi}$ using the data augmentation algorithm in section 5.2.

For Poisson latent variables, at each iteration, first set Z_{ij}'s, Z_{ijl}'s, W_{ij}'s, and W_{ijl}'s all equal to zero, then:

- If $\delta_{ij1}=1$, then sample

$$Z_{ij}\sim\text{Poi}(\Lambda_0(R_{ij})\xi_{i1}\left\{\prod_{h=1}^{H}\xi_{i,h+1}^{x_{ijh}}\right\}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij}))1(Z_{ij}>0),$$

$$(Z_{ij1},\ldots,Z_{ijK})\sim\text{Multinomial}(Z_{ij},\boldsymbol{p}_{ij}),\quad \boldsymbol{p}_{ij}\propto(\gamma_1 b_1(R_{ij}),\ldots,\gamma_K b_K(R_{ij})).$$

- If $\delta_{ij2}=1$, then sample

$$Wij\sim\text{Poi}((\Lambda_0(R_{ij})-\Lambda_0(L_{ij}))\xi_{i1}\left\{\prod_{h=1}^{H}\xi_{i,h+1}^{x_{ijh}}\right\}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij}))1(W_{ij}>0),$$

$$(W_{ij1},\ldots,W_{ijK})\sim\text{Multinomial}(W_{ij},\boldsymbol{q}_{ij}),$$

$$\boldsymbol{q}_{ij}\propto(\gamma_1(b_1(R_{ij})-b_1(L_{ij})),\ldots,\gamma_K(b_K(R_{ij})-b_K(L_{ij}))).$$

For the coefficient γ_l involved in approximating baseline cumulative hazard $\Lambda_0(t)$, we assign a prior $\gamma_l\sim\exp(\eta)$ and a hyperprior $\eta\sim\text{Ga}(a_\eta,b_\eta)$. Then we got the

conjugate full conditional density:

$$\gamma_l|\cdot \sim \text{Ga}(1+\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Z_{ijl}\delta_{ij1}+W_{ijl}\delta_{ij2}),$$
$$\eta+\sum_{i=1}^{I}\sum_{j=1}^{n_i}\xi_{i1}\left\{\prod_{h=1}^{H}\xi_{i,h+1}^{x_{ijh}}\right\}\exp(\boldsymbol{\beta}'\boldsymbol{x}_{ij})(b_l(R_{ij})(\delta_{ij1}+\delta_{ij2})+b_l(L_{ij})\delta_{ij3}))$$

For sampling of regression parameters β_r, $r=1,\ldots,p$, since all our covariates are binary in this project, we define parameter $\omega_r = e^{\beta_r}$, $r=1,\ldots,p$, which makes the regression term: $\exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta}) = \prod_{r=1}^{p}\omega_r^{x_{ijr}}$. By specifying a Gamma prior $\omega_r \sim \text{Ga}(c,d)$, we get conjugate gamma full conditional distributions for the parameters ω_r:

$$\omega_r|\cdot \sim \text{Ga}(c+\sum_{i=1}^{I}\sum_{j=1}^{n_i}x_{ijr}(Z_{ij}\delta_{ij1}+W_{ij2}\delta_{ij2}),$$
$$d+\sum_{i=1}^{I}\sum_{j=1}^{n_i}x_{ijr}(\prod_{l\neq r}\omega_l^{x_{ijl}})(\xi_{i1}\prod_{h=1}^{H}\xi_{i,h+1}^{x_{ijh}})(\Lambda_0(R_{ij})\delta_{ij1}+\Lambda_0(R_{ij})\delta_{ij2}+\Lambda_0(L_{ij})\delta_{ij3}))$$

Step iv) of our Gibbs sampling algorithm is a little complicated and iterates through the following substeps:

(1) Sample δ_h, $h=1,\ldots,H$, from Bernoulli($\tilde{\pi}_h$), where

$$\tilde{\pi}_h = \frac{\pi_{0h}L(\Lambda_0,\boldsymbol{\beta},\boldsymbol{\xi}_h\equiv 1,\boldsymbol{\xi}_{(-h)})}{\pi_{0h}L(\Lambda_0,\boldsymbol{\beta},\boldsymbol{\xi}_h\equiv 1,\boldsymbol{\xi}_{(-h)})+(1-\pi_{0h})L(\Lambda_0,\boldsymbol{\beta},\boldsymbol{\xi}_h=\tilde{\boldsymbol{\xi}}_h,\boldsymbol{\xi}_{(-h)})},$$

with $\boldsymbol{\xi}_{(-h)}$ denotes all frailties except for the hth.

(2) Sample $\tilde{\kappa}_h$, $h=1,\ldots,H+1$, from its full conditional distribution using adaptive rejection Metropolis sampling (ARMS) algorithm (Gilks et al., 1995):

$$\tilde{\kappa}_h|\cdot \propto \left\{\frac{\tilde{\kappa}_h^{-1}**\tilde{\kappa}_h^{-1}}{\Gamma(\tilde{\kappa}_h^{-1})}\right\}^{I}\exp\left[-\tilde{\kappa}_h^{-1}\left\{\sum_{i=1}^{I}\tilde{\xi}_{ih}-\sum_{i=1}^{I}\log(\tilde{\xi}_{ih})+b_h\right\}\right](\tilde{\kappa}_h^{-1})^{a_h-1}.$$

(3) Sample $\tilde{\xi}_{i1}$, $i=1,\ldots,I$ from

$$\text{Ga}(\tilde{\kappa}_1+(1-\delta_1)\sum_{j=1}^{n_i}(Z_{ij}\delta_{ij1}+W_{ij}\delta_{ij2}),$$
$$\tilde{\kappa}_1^{-1}+(1-\delta_1)\sum_{j=1}^{n_i}e^{\boldsymbol{x}'_{ij}\beta}\xi_{i2}^{x_{ij1}}\ldots\xi_{i,H+1}^{x_{ijH}})\left\{\Lambda_0(R_{ij})(\delta_{ij1}+\delta_{ij2})+\Lambda_0(L_{ij})\delta_{ij3}\right\}.$$

(4) For each $h = 2, \ldots, H+1$, sample $\tilde{\xi}_{ih}$, $i = 1, \ldots, I$, from

$$\mathrm{Ga}(\tilde{\kappa}_h^{-1} + (1-\delta)\sum_{j=1}^{n_i} x_{ij1}(Z_{ij}\delta_{ij1} + W_{ij}\delta_{ij2}),$$
$$\tilde{\kappa}_h^{-1} + \sum_{j=1}^{n_i}\left[\xi_{i1}e^{\boldsymbol{x}'_{ij}\boldsymbol{\beta}}(\prod_{l\neq h}\xi_{i,l+1}^{x_{ijl}})\left\{{}_0(R_{ij})(\delta_{ij1}+\delta_{ij2}) + \Lambda_0(L_{ij})\delta_{ij3}\right\}\right](1-\delta_h)x_{ij,h-1}).$$

For each frailty, we record the estimated probability of homogeneity for at each iteration as $\tilde{\pi}_h^{(s)}$. Then we can estimate the posterior probability of homogeneity using the average over all iterations:

$$\hat{\pi}_h = \frac{1}{S}\sum_{s=1}^{S}\tilde{\pi}_h^{(s)}, \quad h = 1, \ldots, H+1.$$

Then we can calculate the corresponding Bayes factor of homogeneity for each frailty on the basis of posterior probabilities of indicators ($\delta_h, \quad h = 1, \ldots, 3$):

$$BF_h = \frac{Pr(\delta_h = 1|\boldsymbol{L}, \boldsymbol{R}, \boldsymbol{x})/Pr(\delta_h = 1)}{Pr(\delta_h = 0|\boldsymbol{L}, \boldsymbol{R}, \boldsymbol{x})/Pr(\delta_h = 0)}, \quad h = 1, \ldots, H+1,$$

where $Pr(\delta_h = 1)$ denotes the prior probability of homogeneity of the hth frailty and $Pr(\delta_h = 1|\boldsymbol{L}, \boldsymbol{R}, \boldsymbol{x})$ denotes the correponding posterior probability of homogeneity. Because we assume that the prior probabilities of homogeneity and heterogeneity are equal (i.e., 0.5), the Bayes factor reduces to:

$$\mathrm{BF}_h = \frac{Pr(\delta_h = 1|\boldsymbol{L}, \boldsymbol{R}, \boldsymbol{x})}{\mathrm{Pr}(\delta_h = 0|\boldsymbol{L}, \boldsymbol{R}, \boldsymbol{x})}$$

.

5.5 A Simulation Study

Our simulation study is conceptualized under a typical multi-center randomized clinical trial setting. We are interested in examine institutional variation in baseline hazard and covariate effects. Suppose there are I = 25 institutions, each has recruited $n_i = 20$ patients. There are two binary predictors x_1 (treatment effect) and

x_2. Suppose there is institutional variation of x_1 effect only, then we generated data from the following model:

$$\lambda_{ij}(t|x_{ij}) = \lambda_0(t)\xi_{i1}\xi_{i2}^{x_{ij1}}\xi_{i3}^{x_{ij2}}\exp(\beta_1 x_{ij1} + \beta_2 x_{ij2}), \tag{5.5}$$

where $\xi_{i1} \equiv 1$ (no heterogeneity in baseline risk), $\xi_{i2} \sim \text{Ga}(2,2)$ (treatment-institution interaction), and $\xi_{i3} \equiv 1$ (no heterogeneity among the x_2 effect). We set $\Lambda_0(t) = \log(1+t)$, $\beta_1 = \beta_2 = 1$, and generated $x_{ij1} \sim$ Bernoulli(0.5), $x_{ij2} \sim$ Bernoulli(0.5).

Since we want to test whether a frailty term exists or not, we would pretend that we do not know that $\xi_{i1} \equiv 1$ and $\xi_{i3} \equiv 1$. Ideally, we would expect our model to produce a high probability that the variances κ_1 and κ_3 for the first and third frailties are from the zero point mass, and a low probability that the variance κ_2 for the second frailty is from the zero point mass.

For the heterogeneity structure, we chose prior of $\kappa_h \sim \text{ZI-IG}(\pi_{0h}, a_h, b_h)$, with $\pi_{0h} = 0.5$, $a_h = 0.001$, $b_h = 0.001$, for $h = 1, 2, 3$. For exponentiated regression coefficient ω_r, we chose prior $\omega_r \sim \text{Ga}(3,1)$. For coefficients γ_l, we chose $\gamma_l \sim \exp(\eta)$, with mean $\frac{1}{\eta}$, and hyperprior $\eta \sim \text{Ga}(1,1)$. We generated a total of 100 datasets. For each dataset, we run a total of 13,000 iterations, with the first 3000 discarded as a burn-in. The simulation results are summarized over the 100 datasets.

We evaluated the convergence of the MCMC chains for our model parameters with traceplot and the Geweke's convergence diagnostic. Here we present the diagnostics for our first simulated dataset. Figures 5.1 to Figure 5.5 are the traceplots for β_1, β_2, π_1, π_2, and π_3. The chains seem to mix well. Table 5.1 contains the z-scores and the correspoinding p-values from the Geweke's test. Since all the p-values are greater than 0.05, it is reasonable to say that the chains have converged to their stationary distributions.

Table 5.2 lists the posterior probability of homogeneity for each frailty and the related Bayes factors. Kass and Raftery (1995) suggests the cutoff points for positive, strong, and very strong evidence for a Bayes factor as: 3, 20, and 150. The Bayes

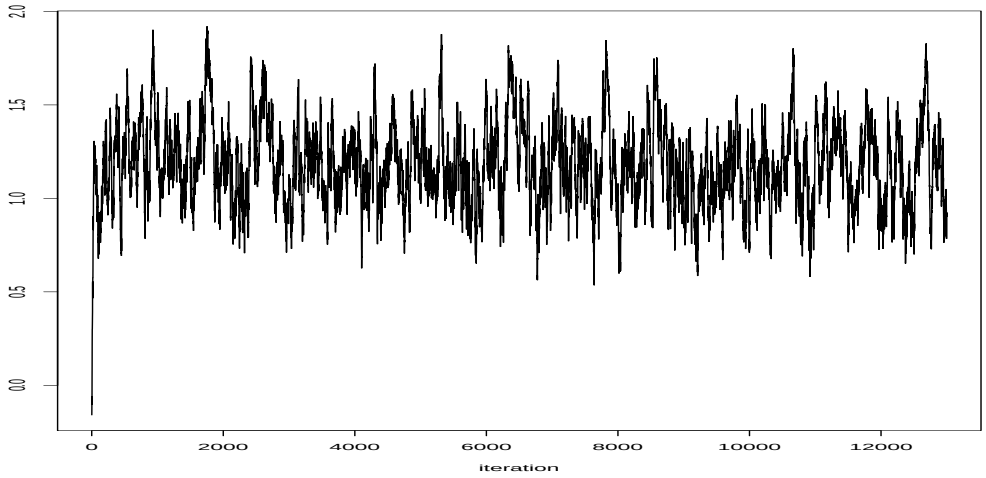

Figure 5.1 Traceplot of fixed effect β_1 in simulation study.

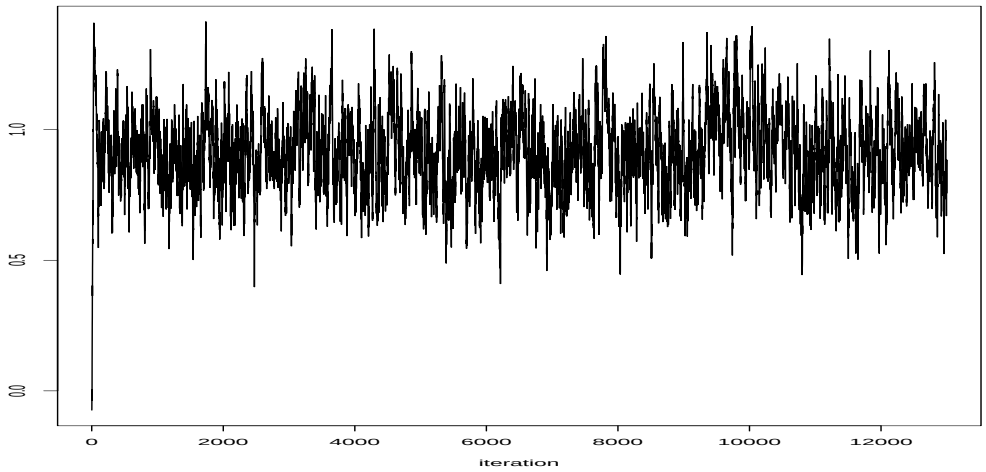

Figure 5.2 Traceplot of fixed effect β_2 in simulation study.

factor in support of homogeneity for the first frailty is 30, which is a strong evidence that there is no institutional variation in baseline risk. For the second frailty, the Bayes factor for homogeneity is less than 1, so we calculated the Bayes factor for heterogeneity as $(1 - 0.017)/0.017 = 59 > 20$, so we have strong evidence that the effect of covariate x_1 varies across institutions. Finally, for the third frailty, the Bayes factor for homogeneity is $14 > 3$, which provides a positive evidence that the effect of covariate x_2 does not have institutional variation.

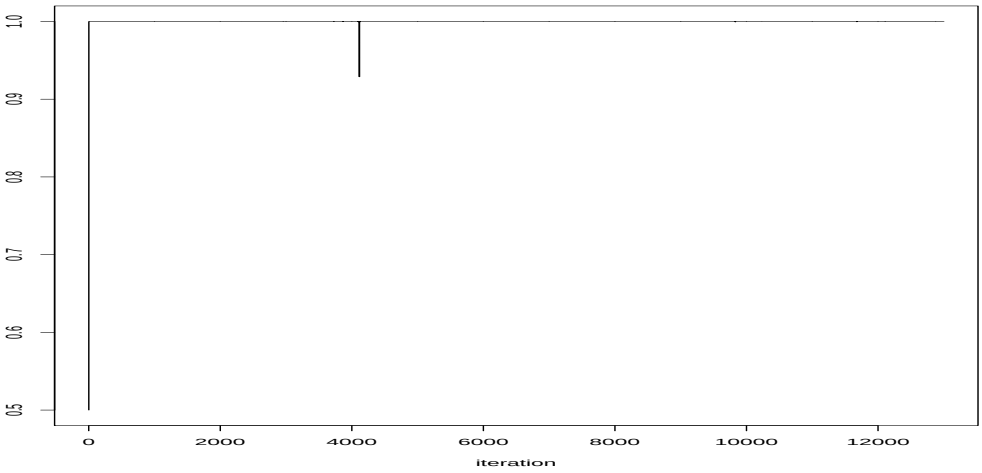

Figure 5.3 Traceplot of the probability of homogeneity for ξ_{i1} in simulation study.

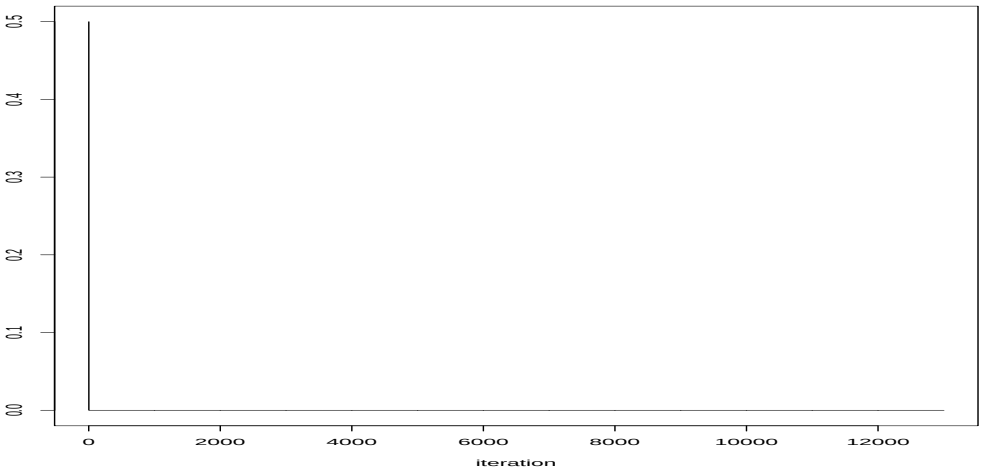

Figure 5.4 Traceplot of the probability of homogeneity for ξ_{i2} in simulation study.

Since there are three potential frailties in the model, then based on the absence/presence of each frailty, there can be 8 possible models. Each model corresponds to one combination of the three homogeneity indicators δ_1, δ_2, and δ_3. **Table 5.3** lists the posterior probability of each model with corresponding combination of δ's. The models are listed in the order of estimated posterior probabilities from high to low. Our true model is the one with only the second frailty exists, i.e., $(\delta_1, \delta_2, \delta_3) = (1, 0, 1)$. Our analysis correctly places a high probability (0.8729) on

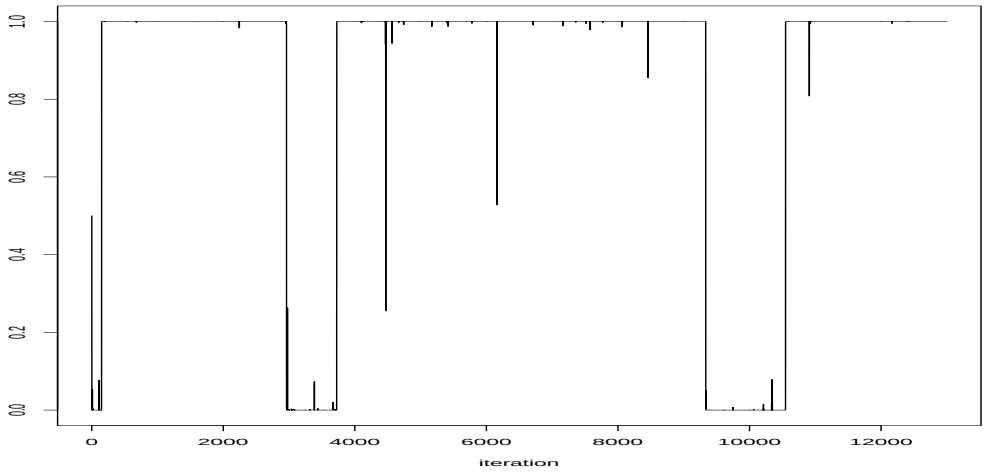

Figure 5.5 Traceplot of the probability of homogeneity for ξ_{i3} in simulation study.

Table 5.1 Geweke's convergence diagnostic for MCMC chains of the model parameters in the 1st simulated dataset

Paramter	Z-score	p-value
β_1	-1.5372	0.8724
β_2	-0.5737	0.4862
π_1	-0.9982	0.3182
π_2	1	0.3173
π_3	-0.7091	0.4783

this model.The other models all have very low estimated posterior probabilities.

In addition to the selection of frailties, our approach produces model-averaged inference on the fixed effects of covariates after accounting for uncertainty in the frailty selection process (**Table 5.4**). The point estimates for β_1 and β_2 are very close to the true values of 1. And the 95% coverage probabilities are close to the nominal level.

Table 5.2 Posterior homogeneity probabilities and Bayes factors of frailties in the simulation study

Paramter	Estimate	Bayes factor
$\pi_1 = \Pr(H_{01}\|\text{data})$	0.968	30
$\pi_2 = \Pr(H_{02}\|\text{data})$	0.017	< 1
$\pi_3 = \Pr(H_{03}\|\text{data})$	0.933	14

Table 5.3 Posterior probabilities of the 8 possible models in the simulation study

Model $(\delta_1, \delta_2, \delta_3)$	1 (1,0,1)	2 (1,0,0)	3 (0,0,1)	4 (1,1,1)	5 (0,1,1)	6 (0,0,0)	7 (1,1,0)	8 (0,1,0)
Probability	0.8729	0.0648	0.0370	0.0093	0.0088	0.0035	0.0020	0.0019

Table 5.4 Estimated fixed effects and coverage probabilities in the simulation study

Paramter	Estimate	Bias	95%$C.P.$
$\beta_1(1)$	1.02	0.02	0.95
$\beta_2(1)$	1.03	0.03	0.98

5.6 Lymphatic Filariasis Example

Lymphatic filariasis is a parasitic disease caused my microscopic, thread-like worms. It is spread from person to person by mosquitoes. The adult worms live in human lymph system in the form of nests. Their characteristic movement called "filarial dance sign" can be detected by ultrasound. When all the worms in a nest are killed

by antifilarial drugs, the filarial dance sign ceases. A randomized clinical trial was conducted in Recife, Bazil (Dreyer et al., 2006) to compare the effect of co-administration of diethylcarbamazine and albendazole (DEC/ALB) (new treatment) against DEC alone (standard treatment). Among the 47 patients, 22 were randomized to the DEC/ALB treatment arm ($x_1 = 0$) and 25 to the DEC ($x_1 = 1$) treatment arm. During the one-year follow-up period, ultrasound examinations were performed at 7, 14, 30, 45, 60, 90, 180, 270, and 360 days after drug administration. The outcome of interest is the nest-specific time of clearance of all its worms. So the data is interval-censored and each patient forms a cluster with cluster size being the number of nests of adult filial worms residing in this patient.

This data has been analyzed by Williamson et al. (2008), Zhang and Sun (2010), and Kim (2010) using different modeling methods. Their analysis results suggest that the time-to-clearance of a nest depends on the number of nests in the corresponding patient. So in addition to the main variable of interest: the treatment effect (x_1), we also defined another covariate $x_2 = 1$ if cluster size > 1, and 0 otherwise. There are 23 patients with 1 nest each, 20 patients with 2 nests, 2 patients with 3 nests, 1 patient with 4 nests, and another 1 patient with 5 nests. We want to investigate whether there is any patient-wise variation of baseline hazard, treatment effect, or cluster size effect. Also, we are interested in evaluating the fixed effect (within-cluster effect) of treatment and cluster size. Consider the multivariate frailty model of the form (5.5). We let $\pi_{0h} = 0.5$, $a_h = 1$, $b_h = 1$, for $h = 1, 2, 3$; $\omega_r \sim \text{Ga}(1, 1)$; and $\eta \sim \text{Ga}(1, 1)$. The MCMC algorithm was carried out through 11000 iterations with the first 1000 as a burn-in.

We assessed the MCMC chains for our model parameters with traceplot and Geweke's convergence diagnostic. Figure 5.6 to Figure 5.10 are the traceplots for model parameters, we can see that the chains are well mixed and are stable. **Table 5.5** shows the Geweke's diagnostic for the chains of our model parameters: fixed ef-

fects β_1, β_2, and frailty homogeneity indicator π_1, π_2, π_3. As we can see, the p-values are all greater than 0.05. By looking at the traceplots and the p-values, we have confidence that the MCMC chains have converged.

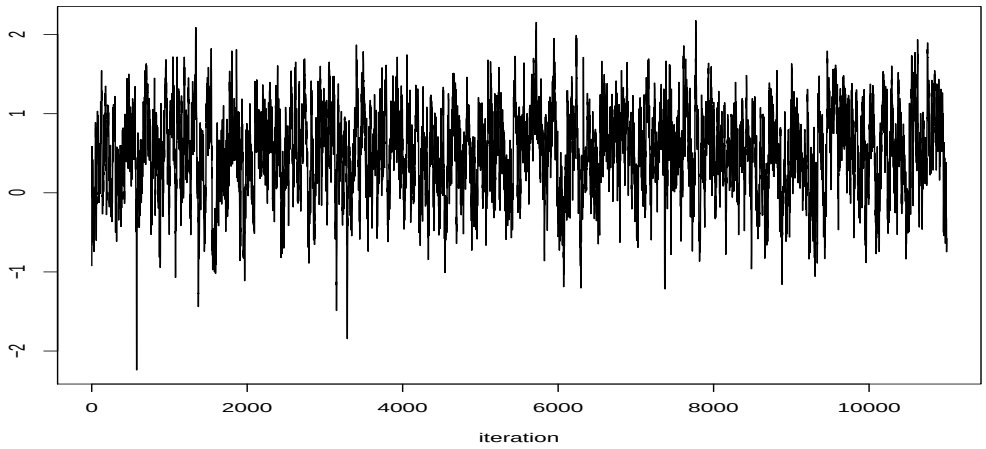

Figure 5.6 Traceplot of fixed effect β_1 in lymphatic filariasis study.

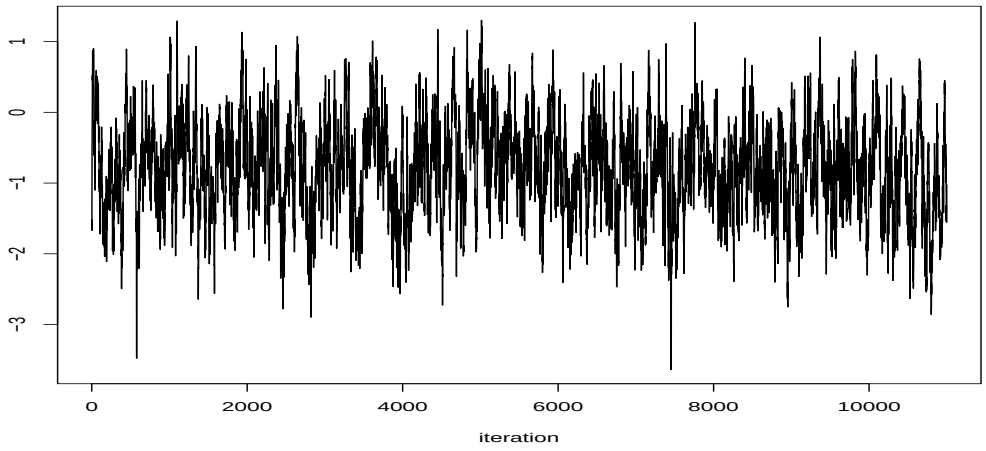

Figure 5.7 Traceplot of fixed effect β_2 in lymphatic filariasis study.

Table 5.6 shows the posterior probability of homogeneity for each of our three frailties and the corresponding Bayes factor. According to the Bayes factor scale suggested by Kass and Raftery (1995), we have very strong evidence (BF $>$ 150) of

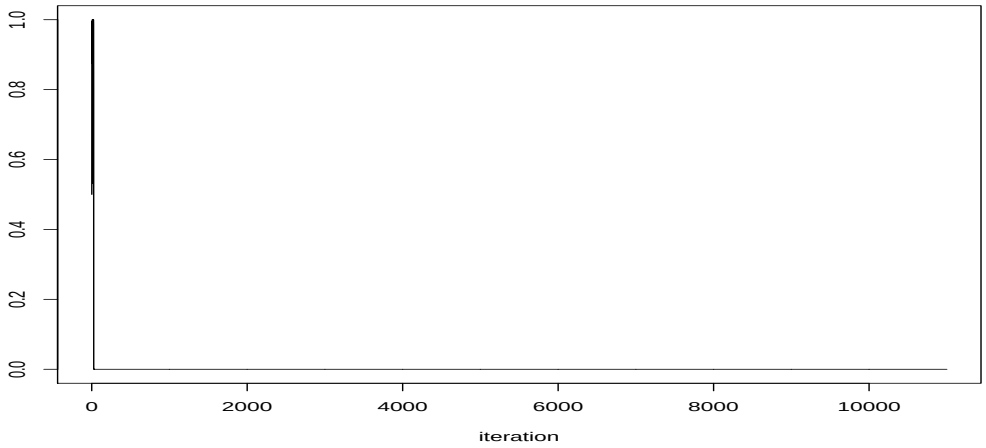

Figure 5.8 Traceplot of the probability of homogeneity for ξ_{i1} in lymphatic filariasis study.

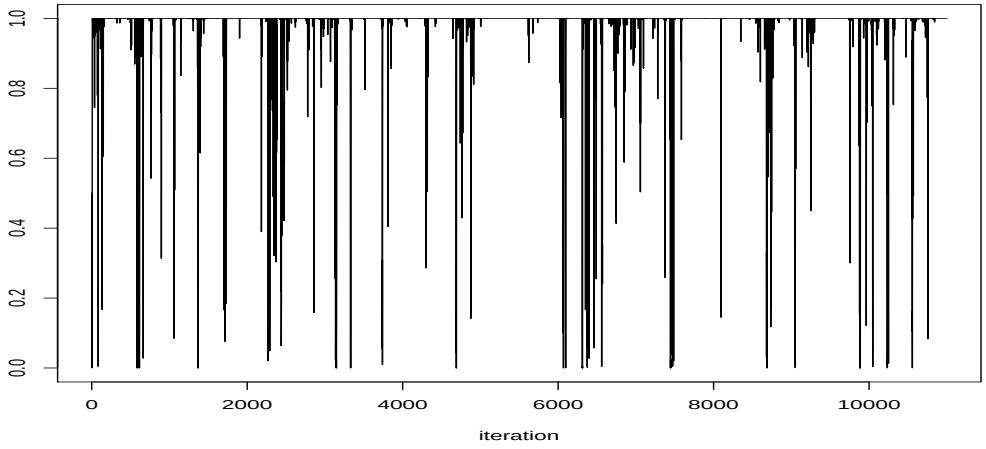

Figure 5.9 Traceplot of the probability of homogeneity for ξ_{i2} in lymphatic filariasis study.

patient-wise variation in baseline hazard function, and strong evidence (BF $>$ 20) of homogeneity in treatment effect and cluster size effect across patients.

Table 5.7 shows the posterior probability for each of the 8 possible models based on the absence/presence of the 3 potential frailties. The models are listed in the order of high posterior probability to low posterior probability. There is a very high probability (0.9595) for model 1, which specifies that there exists a frailty for

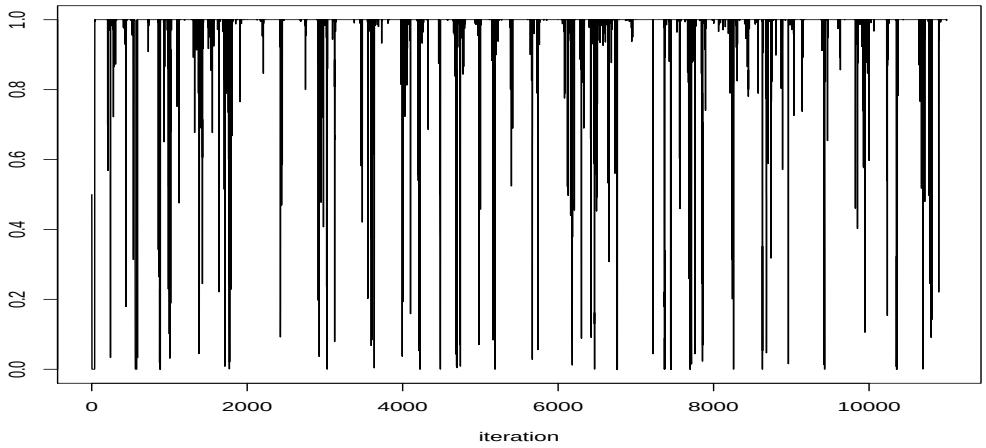

Figure 5.10 Traceplot of the probability of homogeneity for ξ_{i3} in lymphatic filariasis study.

Table 5.5 Geweke's convergence diagnostic for MCMC chains of model parameters in the lymphatic filariasis study

Paramter	Z-score	p-value
β_1	-0.9598	0.3372
β_2	0.7836	0.4333
π_1	0.876	0.3810
π_2	-0.7706	0.4409
π_3	-1.248	0.2120

baseline hazard but no frailty for either treatment effect or cluster size. The estimated posterior probabilities for all the other models are very low. As expected, this model selection result agrees closely with the frailty selection result.

Table 5.8 presents the model-averaged estimation for fixed effects of treatment and cluster size. The estimate for treatment effect ($\hat{\beta}_1 = 0.536$) agrees with the one

($\hat{\beta}_1 = 0.585$, p-value = 0.089) from Williamson et al. (2008). Although the influence of cluster size is insignificant too, the negative sign implies possibly slower elimination in a patient with multiple nests. This agrees with the previous studies.

Table 5.6 Posterior homogeneity probabilities and Bayes factors of frailties in the lymphatic filariasis study

Paramter	Estimate	Bayes factor
$\pi_1 = \Pr(H_{01}\|\text{data})$	6e-18	< 1
$\pi_2 = \Pr(H_{02}\|\text{data})$	0.981	49
$\pi_3 = \Pr(H_{03}\|\text{data})$	0.977	49

Table 5.7 Posterior probabilities of the 8 possible models in the lymphatic filariasis study

Model $(\delta_1, \delta_2, \delta_3)$	1 (0,1,1)	2 (0,1,0)	3 (0,0,1)	4 (0,0,0)	5 (1,1,1)	6 (1,1,0)	7 (1,0,1)	8 (1,0,0)
Probability	0.9595	0.0211	0.0186	0.0008	0.0000	0.0000	0.0000	0.0000

Table 5.8 Estimated fixed effects and 95% CIs in the lymphatic filariasis study

Parameter	Estimate	95%$C.I.$
Treatment	0.516	(-0.58, 1.48)
Cluster size	-0.785	(-2.05, 0.52)

Chapter 6

Conclusions

In this dissertation, we have developed an R package for modeling case 1 or general interval-cenosred data under the PH, PO, or probit models, a PH model with spatial random effect for spatially correlated interval-censored data, and a multivariate frailty PH model for clustered interval-censored data. Methodology and software development for interval-censored data are important for many medical studies in practice (e.g., infection, AIDS, and oncology clinical trials). They would allow for more accurate estimation of treatment effect while accounting for interval-censorship. For all of the three projects, our nonparametric formulation of a monotonic spline provides more flexible approximation for baseline cumulative hazard function compared to parametric assumptions. Data augmentation makes the MCMC sampling straightforward to implement compared to other methods in the literature. In the following, we conclude the three projects respectively.

In the first project, we first review the methods and software available for analyzing interval-censored data. Considering the lack of very useful software for fitting regression models for such data, we aim to provide a comprehensive package which implement several popular survival regression models under Bayesian framework. We provide inference for regression coefficients and smooth estimation for both baseline survival curves and particular survival curves based on user-specified covariate(s). The package is straightforward to use, as a user only needs to provide interval endpoints, covariates, censoring indicator, knots for I-splines, and grids for survival estimation. On the other hand, it is also flexible, as a user can adjust model fitting

through multiple arguments based on her/his prior knowledge about the data. Furthermore, MCMC chains for critical parameters can be saved for later convergence diagnosis. Our package is still under development. For example, one future plan is to include extra functions to analyze clustered interval-censored data and spatial interval-censored data.

For the second project, we developed an efficient semiparametric method under the PH model to deal with spatially correlated general interval-censored data. To account for our assumption that subjects within an area are correlated and the random effect for an area depends on the random effects of its neighbors, we employ the conditional autoregressive model that is commonly used in spatial statistics for lattice data. The capture of spatial heterogeneity and clustering by structured random effects reduces estimation biases for regression parameters that are often of main interest. An area of future work is to evaluate the robustness of spatial random effects to violation of normality assumption. One may need to consider nonparametric priors to relax the normality assumption if spatial random effects are sensitive to model misspecification. Also, one may incorporate the more general spatial random effects to allow for potential variation of the effects of predictors across areas. In addition, the proposed model can be extended to accommodate spatial and temporal data.

In the third project, we propose a Bayesian approach for inferences on clustered interval-censored failure time data. We generalize the commonly used shared frailty model for interval-censored data to include multiple frailties to account for cluster-wise variation of covariate effects. Furthermore, our method allows for frailty selection by estimating the posterior probability of a frailty and calculating the related Bayes factor. After accounting for uncertainty in frailty selection and random effects, the inference for proportional hazards fixed regression coefficients are less prone to bias. The proposed Gibbs sampler algorithm has been proved to be efficient and have good performance through our simulation study and real data analysis. This approach is

especially useful in multi-center clinical trials for cancer or infectious diseases, since the detection of cancer progression or infection is normally made through periodic lab examinations. Since current literature for clustered interval-censored data are mainly involving shared-frailty models, our developed model is a meaningful extension to the methodology for this type of data. The future research may focus on incorporating nonparametric modeling for frailties with selection, while allowing for nonparametric modeling for the hazard function.

Bibliography

[1] Banerjee, S. and Carlin, B. P. (2004). Parametric spatial cure rate models for interval-censored time-to-relapse data. *Biometrics* **60**, 268-275.

[2] Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, Florida: Chapman & Hall/CRC.

[3] Banerjee, S., Wall, M., and Carlin, B. P. (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics* **4**, 123-142.

[4] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society. Series B. (Methodologies)* **36**, 192-236.

[5] Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82**, 733-746.

[6] Brook, D. 1964. On the distinction between the conditional probability and the joint prob- ability approaches in the specification of nearest-neighbour systems. *Biometrika* **51**, 481-483.

[7] Cai, B. (2010). Bayesian semiparametric frailty selection in multivariate event time data. *Biometrical Journal*. **52**, 171-185.

[8] Cai, B., Lin, X., and Wang, L. (2011). Bayesian proportional hazards model for current status data with monotone splines. *Computational Statistics and Data Analysis* **55**, 2644-2651.

[9] Cai, T. and Betensky, R. A. (2003). Hazard regression for interval-censored data with penalized spline. *Biometrics* **59**, 570-579.

[10] Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edition. Boca Raton, Florida: Chapman & Hall/CRC Press.

[11] Clayton, D. G. and Kaldor, J. M. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43**, 671-681.

[12] Cooper, N. J., Sutton, A. J., Mugford, M., and Abrams, K. R. (2003). Use of Bayesian Markov Chain Monte Carlo methods to model cost-of-illness data. *Medical Decision Making* **23**, 38-53.

[13] Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* **34**, 187-220.

[14] Dey, D. K., Chen, M., and Chang, H. (1997). Bayesian approach for nonlinear random effects models. *Biometrics* **53**, 1239-1252.

[15] Dreyer, G, Addiss, D, Williamson, J. M., and Noroes, J. (2006). Efficacy of co-administered diethylcarbamazine and albendazole against adult Wuchereria bancrofti. *Transactions of the Royal Society for Tropical Medicine and Hygiene* **100**, 1118-1125.

Dunson04 Dunson, D. B. and Chen, Z. (2004). Selecting factors predictive of heterogeneity in multivariate event time data. *Biometrics* **60**, 352-358.

[16] Fay, M. P. (1996). Rank invariant tests for interval censored data under the grouped continuous model. *Biometrics* **52**, 811-822.

[17] Fay, M. P. and Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: the **interval** R package. *Journal of Statistical Software* **36**, 1-34.

[18] Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845-854.

[19] Finkelstein, D. M. and Wolfe, R. A. (1985). A Semiparametric Model for Regression Analysis of Interval-Censored Failure Time Data. *Biometrics* **41**, 933-945.

[20] Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association* **74**, 153-160.

[21] Gelfand, A. E. and Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 501-514.

[22] Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Technical Report*. Department of Statistics, Standford University.

[23] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transaction on Pattern Analysis and Machine Intelligence* **6**, 721-741.

[24] Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: consistency and computation. *Biometricka* **81**, 618-623.

[25] Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. *Bayesian Statistics* **4**, 169-193.

[26] Gilks, W. R., Best, N., and Tan, K. (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Applied Statistics* **44**, 455-472.

[27] Gilks, W. R. and Wild, P. (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statististics* **41**, 337-348.

[28] Giolo, S. R. (2004). Turnbull's nonparametric estimator for interval-censored Data. *Technical Report*. Department of Statistics, Federal University of Paraná.

[29] Goethals, K., Ampe, B., Berkvens, D., Laevens, H., Janssen, P., and Duchateau, L. (2009). Modeling interval-censored, clustered cow udder quarter infection times through the shared gamma frailty model. *Journal of Agricultural, Biological, and Environmental Statistics* **14**, 1-14.

[30] Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*. DMV Seminar Band 19, Birkhäuser, Basel.

[31] Hodges, J. S., Carlin, B. P., and Fan, Q. (2003). On the precision of the conditionally autoregressive prior in spatial models. *Biometrics* **59**, 317-322.

[32] Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress.

[33] Jin, X., Carlin, B. P., and Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics* **61**, 950-961.

[34] Joly, P., Commernges, D., and Letenneur, L. (1998). A penalized likelihood approach for arbitrily censored data and truncated data: application to age-specific incidence of mementia. *Biometrics* **54**, 185-194.

[35] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773-795.

[36] Kim, Y. (2010). Regression analysis of clustered interval-censored data with informative cluster size. *Statistics in Medicine* **29**, 2956-2962.

[37] Lin, X., Cai, B., Wang, L., and Zhang, Z. (2013). A novel Bayesian approach for analyzing general interval-censored failure time data under the proportional hazards model.

[38] Lin, X. and Wang, L. (2010). A semiparametric probit model for case 2 interval-censored failure time data. *Statistics in Medicine* **29**, 972-981.

[39] Lin, X. and Wang, L. (2011). Bayesian proportional odds models for analyzing current status data: univariate, clustered, and multivariate. *Communications in Statistics - Simulation and Computation* **40**, 1171-1181.

[40] Lindsey J. C. and Ryan L. M. (1998). Tutorial in biostatistics methods for interval-censored data. *Statistics in Medicine* **17**, 219-238.

[41] Murray, R., Anthonisen, N. R., Connett, J. R., Wise, R. A., Lindgren, P. G., Greene, P. G., and Nides, M. A. (for the Lung Health Study Research Group) (1998). Effects of multiple attempts to quit smoking and relapses to smoking on pulmonary function. *Journal of Clinical Epidemiology* **51**, 1317-1326.

[42] Odell, P. M., Anderson, K. M., and D'Agostino, R. B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics* **48**, 951-959.

[43] Pan, W. (1999). Extending the iterative convex minorant algorithm to the Cox model for interval-censored data. *Journal of Computational and Graphical Statistics* **8**, 109-120.

[44] Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics* **56**, 199-203.

[45] Peto, R. (1973). Experimental survival curves for interval-censored data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **22**, 86-91.

[46] Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science* **3**, 425-441.

[47] Ripatti, S. and Palmgren, J. (2000). Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics* **56**, 1016-1022.

[48] Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer.

[49] Rücker, G. and Messerer, D. (1988). Remission duration: an example of interval-censored observations. *Statistics in Medicine* **7**, 1139-1145.

[50] Schick, A. and Yu, Q. (2000). Consistency of the GMLE with Mixed Case Interval-censored Data. *Scandinavian Journal of Statistics* **27**, 45-55.

[51] So, Y., Johnston, G., and Kim, S. H. (2010). Analyzing interval-censored survival data with SAS software. *Proceedings of the SAS Global Forum 2010 Conference*, Cary, NC: SAS Institute Inc.

[52] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* **64**, 583-679.

[53] Sun, J. (1996). A non-parametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine* **15**, 1387-1395.

[54] Sun, J., Zhao, Q., and Zhao, X. (2005). Generalized log-rank test for interval-censored data. *Scandinavian Journal of Statistics* **32**, 45-57.

[55] Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data.* Springer.

[56] Sun, X. and Chen, C. (2012). Comparison of Finkelstein's method with the conventional approach for interval-censored data analysis. *Statistics in Biopharmaceutical Research* **2**, 97-108.

[57] Turnbull, B. W. (1976). The Empirical distribution function with arbitrarily grouped, censored and truncated Data. *Journal of the Royal Statistical Society. Series B (Methodological)* **38**, 290-295.

[58] Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine* **19**, 3309-3324.

[59] Vaupel, J. M., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* **16**, 439-454.

[60] Wang, L. and Dunson, D. B. (2011). Semiparametric Bayes proportional odds models for current status data with under-reporting. *Biometrics* **67**, 1111-1118.

[61] Wang, L. and Lin, X. (2011). A Bayesian approach for analyzing case 2 interval-censored failure time data under the semiparametric proportional odds model. *Statistics and Probability Letters* **81**, 876-883.

[62] Wellner, J. A. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association* **92**, 945-959.

[63] William, J. M., Kim, H., Manatunga, A., and Addiss, D. G. (2008). Modeling survival data with informative cluster size. *Statistics in Medicine* **27**, 543-555.

[64] Yavuz, A. C. and Lambert, P. (2011). Smooth estimation of survival functions and hazard ratios from interval-censored data using Bayesian penalized B-splines. *Statistics in Medicine* **30**, 75-90.

[65] Yu, Q., Schick, A., Li, L., Wong, G. (1998). Asymptotic properties of the GMLE with case 2 interval-censored data. *Statistics Probability Letters* **37**, 223-228.

[66] Zhang, X. and Sun, J. (2010). Regression analysis of clustered interval-censored failure time data with informative cluster size. *Computational Statistics and Data Analysis* **54**, 1817-1823.

[67] Zhao, Q and Sun, J (2004). Generalized log-rank test for mixed-censored failure time data. *Statistics in Medicine* **23**, 1621-1629.

[68] Zhao, X., Zhao, Q., Sun, J., and Kim, J. S. (2008). Generalized log-rank tests for partly interval-censored failure time data. *Biometrical Journal* **50**, 375-385.