12-15-2014

Semiparametric Regression Analysis of Bivariate Interval-Censored Data

Naichen Wang
University of South Carolina - Columbia

Semiparametric Regression Analysis of Bivariate interval-censored Data

by

Naichen Wang

Bachelor of Science
Dalian University of Technology, China, 2007
Master of Arts
University of South Carolina, 2009

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2014

Accepted by:

Lianming Wang, Major Professor

Timothy E. Hanson, Committee Member

David B. Hitchcock, Committee Member

Jiajia Zhang, Committee Member

Lacy Ford, Vice Provost and Dean of Graduate Studies

Acknowledgments

I gratefully acknowledge the help and encouragement of Professor Lianming Wang, my thesis supervisor. Without his help, I cannot finish the doctoral work at this time. I extend my gratitude to the members of my advisory committee: Dr. Timothy E. Hanson, Dr. David B. Hitchcock and Dr. Jiajia Zhang for their insightful comments and suggestions on my work.

I am indebted to all faculty members, staff, colleagues in the Department of Statistics at university of South Carolina for providing me such an amazing research environment. I really enjoy studying here.

I want to thank my parents and friends who have helped me during the writing process.

Finally I would like to present my deep gratitude to my wife Shanshan Wu and my son Suian Wang for their love and support in my life.

Abstract

Survival analysis is a long-lasting and popular research area and has numerous applications in all fields such as social science, engineering, economics, industry, and public health. Interval-censored data are a special type of survival data, in which the survival time of interest is never exactly observed but is known to fall within some observed interval. Interval-censored data arise commonly in real-life studies, in which subjects are examined at periodical or irregular follow-up visits. In this dissertation, we develop efficient statistical approaches for regression analysis of bivariate interval-censored data, in which the two survival times of interest are correlated and both have an interval-censored data structure.

Chapter 1 first describes the structure of interval-censored data in detail, and four real-life data sets are presented for illustrations. A literature review is provided regarding the existing semiparametric regression models and methods on interval-censored data. The last section of this chapter provides some important background knowledge to be used in later chapter of this dissertation, such as Kendall's τ and Dirichlet process mixture model.

Chapter 2 proposes a novel and fast EM algorithm for regression analysis of bivariate current status data based on the Gamma-frailty proportional hazards (PH) model. Monotone splines are adopted to approximate the unknown conditional baseline cumulative functions. A three-stage data augmentation is proposed and leads to a complete data likelihood in a simple form. An EM algorithm is further derived utilizing this complete likelihood. The resulting algorithm is easy to implement, robust to initialization, and enjoys quick convergence. The proposed method has excellent

performance in estimating the regression parameters, the baseline survival functions, and the statistical association between the both failure times through simulation studies. The method is also robust to the misspecifications of frailty distribution. Moreover, the method is much faster than existing approaches in the literature. Our method is illustrated by a real-life application about the prevalence of antibodies to hepatitis B and HIV among Irish prisoners.

In Chapter 3, I revisit the topic on bivariate current status data but from a Bayesian perspective. Two Bayesian methods are proposed: one for Gamma-frailty PH model and one for frailty PH model with unknown frailty distribution. A Dirichlet process Gamma mixture model is proposed for modeling the unknown frailty distribution. Efficient Gibbs samplers are proposed for these two models. Simulation results suggest that both of the two proposed methods work well in the cases of correctly specified and misspecified frailty distributions. The method based on the Gamma-frailty PH model is preferred because of its simpler model structure and robust performance in addition to providing Kendall's τ in closed form.

Chapter 4 investigates Bayesian regression analysis of bivariate interval-censored data. First, an efficient method is proposed based on the Gamma-frailty PH model, and simulation studies show that the proposed method works well when the model is correctly specified. It is also observed that the method leads to biased estimates when the two failure times are independent or weakly correlated. To handle both dependent and independent cases, a mixture of gamma and point mass at one is proposed for the frailty distribution. An efficient Gibbs sampler is proposed and is shown to have good performance in both cases through simulation studies. A read-life data set from an AIDs clinical trial is analyzed for illustration.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Survival data involve the time to some event of interest such as the time to onset (or relapse) of a disease. A special feature of survival data is the possibility of censoring (Cox and Oakes 1984) because the exact failure time of the event is not available to observe directly. For example, the time to the instant status change of some type of tumors is not obtainable due to the loss to follow-up or non-occurrence of the tumor at the end of the study. Alternatively, some incomplete observation of failure time is available such as the examination time, the dropout time, or the ending time of the study. Such the observation time is referred to as censoring time. In some studies, there may be more than one censoring times for each subject due to the design of the study.

Various of censoring types can be found in the censored data: left-censoring, right-censoring and interval-censoring, which are defined by the relationship between the failure time and the censoring time. A left- or right-censored observation arises when the failure time is smaller or greater than the censoring time. An interval-censored observation appears when the failure time event is known to have occurred between two examination times with changed status. General interval-censored data contain all these three types of censored observations, and such data naturally result from studies with periodical examinations. A special type of interval-censored data is called current status data, in which all the failure times are either left censored or right censored. Current status data appear commonly in cross-sectional studies, in which there is only one examination time for all subjects. Current status data are

also referred to as case 1 interval-censored data, and general interval-censored data are also referred to as case 2 interval-censored data in the literature.

The remainder of this chapter is organized as follows: Sections 1.1 and 1.2 demonstrate different types of interval-censored data with four real examples. Section 1.3 reviews existing approaches for the regression analysis of the multivariate interval-censored data. Section 1.4 discusses the Gamma-frailty proportional hazards (PH) model, which is the fundamental model throughout the dissertation. Section 1.5 reviews the most common method to estimate the association among correlated survival data. Section 1.6 provides background knowledge about Dirichlet process and Dirichlet process mixture models. Such knowledge will help readers understand the proposed nonparametric approach allowing a random distribution for frailty in Chapter 3.

1.1 Current status data

Current status data commonly arise in many epidemiological, social, and medical studies. For example, in tumorigenicity studies conducted by the National Toxicology Program, rats are exposed to different test agents in an effort to assess their toxicity. Researchers then examine the animals' organs for tumors at the time of their death. Consequently, the tumor onset time for a particular animal is never known exactly, but rather is known relative to the animal's time of death; i.e., the tumor onset time is either before or after the time of death. Hoel and Walberg (1972) described a tumorigenicity experiment which was designed to investigate lung tumors on 144 mice. The experiment consisted of two treatments: conventional environment(CE) and germ-free environment(GE). The researchers were interested in discovering whether there is any significant difference of tumor occurrence time under two environments. In the experiment, these mice were examined for lung tumors only at the time when they died during the study or were sacrificed at the end of the study. Consequently, the

Table 1.1 Death times and lung tumor status for 144 male RFM mice

Group	Tumor status	Death times
CE	1	381, 477, 485, 515, 539, 563, 565, 582, 603, 616
		624, 650, 651, 656, 659, 672, 679, 698, 702, 709
		723, 731, 775, 779, 795, 811, 839
	0	45, 198, 215, 217, 257, 262, 266, 371, 431, 447
		454, 459, 475, 479, 484, 500, 502, 503, 505, 508
		516, 531, 541, 553, 556, 570, 572, 575, 577, 585
		588, 594, 600, 601, 608, 614, 616, 632, 632, 638
		642, 642, 642, 644, 644, 647, 647, 653, 659, 660
		662, 663, 667, 667, 673, 673, 677, 689, 693, 718
		720, 721, 728, 760, 762, 773, 777, 815, 886
GE	1	546, 609, 692, 692, 710, 752, 773, 781, 782, 789
		808, 810, 814, 842, 846, 851, 871, 873, 876, 888
		888, 890, 894, 896, 911, 913, 914, 914, 916, 921
		921, 926, 936, 945, 1008
	0	412, 524, 647, 648, 695, 785, 814, 817, 851, 880
		913, 942, 986

tumor onset times were never known exactly but were known to be smaller or larger than their death or sacrifice times according to their tumor status. From Table 1.1, the tumor status 1 means the lung tumor was found at the death and the onset time is left-censored by the death time and 0 indicates the onset time is right-censored. Therefore, this tumor data can be regarded as case 1 interval-censored data or current status data.

Bivariate current status data occur in the situation when two correlated failure times are recorded but each subject is observed only once. The failure times of interest are either left or right-censored. Allright et al. (2000) explored a real life data from a survey carried out among prisoners in the Republic of Ireland in 1999. The survey was conducted to determine the prevalence of antibodies to hepatitis B and HIV. The experimental designer examined many possible risk factors such as: current sentence status, time spent in prison in the past 10 years, history of smoking heroin, the history of injecting the drug, the age when first injecting the drug, history of sexual

transmitted disease etc. The responses were collected in the form of questionnaires. The infection status of each of the participants, for both diseases, were determined according to an oral fluid test. From the data description, the occurrence time of the infection for each prisoner was not observed directly but is either smaller or greater than the age of the prisoner at the time of testing. Hence, the data set can be considered as bivariate current status data.

1.2 Case 2 interval-censored data

Case 2 interval-censored data often arise when periodic scheduled follow-ups for patients are implemented in some clinical trials or studies such as monitoring the progress of chronic diseases like AIDS or cancer. A typical example of such data was reported in Finkelstein and Wolfe (1985) reproduced from a breast cancer study. It contained 94 patients subject to early breast cancer which were assigned to one of two treatment groups, radiotherapy alone or radiation therapy together with adjuvant chemotherapy. In the study, the patients were monitored to detect the cosmetic appearance like breast retraction during the clinic visits every 4 to 6 months. The actual examination times differ from patient to patient since some of them missed their visits. The goal of the study is to investigate the difference between the two treatments for their effects on the rate of change of deterioration of the cosmetic state. From the description of the study, we notice that the exact event time for the instant change of the cosmetic appearance was not observable but was within some interval between visits. Therefore, The failure times for the appearance of breast retraction of each patient were interval-censored, which are presented in Table 1.2.

In Chapter 4 we focus on bivariate interval-censored data. The data set we use to evaluate our approaches is the AIDS data from the ACTG 181 study (Goggins and Finkelstein , 2000; Finkelstein, et al., 2002). In this study, patients were required to provide their urine samples every 4 weeks and blood samples every 12 weeks. The

Table 1.2 Interval times (in months) of cosmetic deterioration (retraction) for early breast cancer patients

Radiotherapy alone				Radiation with adjuvant chemotherapy			
(45,]	(25,37]	(37,]	(4,11]	(8,12]	(0,5]	(30,34]	(16,20]
(17,25]	(6,10]	(46,]	(0,5]	(13,]	(0,22]	(5,8]	(13,]
(33,]	(15,]	(0,7]	(26,40]	(30,36]	(18,25]	(24,31]	(12,20]
(18,]	(46,]	(19,26]	(46,]	(10,17]	(17,24]	(18,24]	(17,27]
(46,]	(24,]	(11,15]	(11,18]	(17,27]	(11,]	(8,21]	(17,26]
(46,]	(27,34]	(36,]	(37,]	(35,]	(17,23]	(33,40]	(4,9]
(22,]	(7,16]	(36,44]	(5,12]	(16,60]	(33,]	(24,30]	(31,]
(38,]	(34,]	(17,]	(46,]	(11,]	(15,22]	(35,39]	(16,24]
(19,35]	(46,]	(5,12]	(9,14]	(13,39]	(15,19]	(23,]	(11,17]
(36,48]	(17,25]	(36,]	(46,]	(13,]	(19,32]	(4,8]	(22,]
(37,44]	(37,]	(24,]	(0,8]	(44,48]	(11,13]	(34,]	(34,]
(40,]	(33,]			(22,32]	(11,20]	(14,17]	(10,35]

samples were tested for the presence of the cytomeglovirus (CMV) virus referred to as shedding of virus. Since the existence of the CMV shedding can not be observed directly from any symptoms, the actual onset time is then not available but is related to the testing time according to the test result. For example, some patients returned with changed CMV shedding status which produced the interval-censored failure time. Some patients did not have the changed CMV shedding until the last visit, which resulted in right-censored failure time. Some already started the CMV shedding from the entrance to the study, which indicated the failure time was left-censored. This type of data is called general interval-censored data or case 2 interval-censored data.

1.3 Existing approaches

The primary goals of analyzing multivariate data are usually centered around the estimation of the survival functions, assessing the significance of covariate effects, and estimating the correlation between failure times. Typically, the analysis of correlated survival data is approached from either one of two perspective; i.e., from either the marginal likelihood or frailty model approach. Existing work which does

not fall within these two categories includes Wang et al. (2008)) based on copula models and Kim (2014) based on multistate models, among others. The marginal likelihood approach (Wei, et al., 1989; Guo and Lin, 1994; Cai and Prentice, 1995), which ignores the correlation between each of the failure times, has been a popular method of analyzing multivariate survival data. Various regression models have been proposed and investigated for studying current status and interval-censored data using the marginal likelihood approach. For example, Goggins and Finkelstein (2000) and Kim et al. (2002) studied correlated interval-censored data using the marginal proportional hazards (PH) model. Chen, et al. (2007) and Tong et al. (2008) investigated the marginal proportional odds model and the marginal additive hazards model, respectively. Even though the likelihood approach can provide robust inference in general, it does not account for the correlation that naturally exists between the multiple failure times.

In order to acknowledge the underlying correlation structure, the frailty model approach is commonly used to jointly model multivariate survival data; i.e., one or more frailty terms are introduced in order to describe the dependence structure between the multiple responses. For example, Vaida and Xu (2000) used the the PH model with normal frailty to analyze the clustered survival data. Huang and Wolf (2002) treated censoring as dependent in the clustered data for the frailty model. Dunson and Dinse (2002) proposed a probit model with normal frailty for bivariate current status data with informative censoring. Komarek and Lesaffre (2007) proposed a frailty accelerated failure time model for correlated interval-censored data. Zuma (2007) explored the Gamma-frailty Weibull model for multivariate interval-censored data. Chen et al. (2009) studied multivariate current status data under the PH model with a normal frailty. Lin and Wang (2011) proposed a Bayesian proportional odds models with a shared gamma frailty for bivariate or clustered current status data. Callegaro and Lacobeli (2012) proposed a Cox model with log-skewed-normal frailties for clustered

right-censored data. For a more in depth review of frailty modeling techniques in survival analysis please see Hougaard (2000), Ibrahim et al. (2008), and Wienke (2012).

1.4 The Gamma-frailty proportional hazards (PH) model

The PH model was first introduced by Cox in 1972 (Cox,1972). It defines the hazard function $\lambda(t;\boldsymbol{x})$ as

$$\lambda(t;\boldsymbol{x}) = \lambda_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}), \tag{1.1}$$

where the $\lambda_0(t)$ is an arbitrary unknown baseline hazard function and $\boldsymbol{\beta}$ is the vector of regression parameters corresponding with covariate $\boldsymbol{x}$. with the specification of hazard function (1.1), the survival function can be expressed as

$$S(t;\boldsymbol{x}) = \exp(-\Lambda_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta})), \tag{1.2}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is the cumulative baseline function.

Clayton (1978) first proposed to use gamma frailty in modeling correlated failure times, and since then the Gamma-frailty model has been used extensively. The Gamma-frailty PH model, as a special Gamma-frailty model, has been widely used for studying correlated survival data. It incorporates the random effect with a multiplicative gamma distributed latent variable on the baseline hazard:

$$\lambda(t;\boldsymbol{x},\eta) = \lambda_0(t)\exp(\boldsymbol{x}'\boldsymbol{\beta})\eta. \tag{1.3}$$

The frailty term η describes the heterogeneity among the subjects in the study. A Gamma distribution $\mathcal{G}(\nu,\nu)$ is commonly used with mean 1 and variance ν^{-1}. The common shape and rate parameters in the gamma distribution guarantees mean 1 and avoids the non-identifiability issue between η and unspecified λ_0.

In literature, we can find a substantially large number of papers which contribute to developing methods for multivariate survival data under Gamma-frailty PH model. Guo and Rodriguez (1992) provided an EM algorithm for clustered data under

Gamma-frailty PH model. For the purposes of analyzing multivariate right-censored data the Gamma-frailty PH model has become also quite popular; e.g., see Klein (1992), Andersen et al. (1997), Rondeau et al. (2003), Cui and Sun (2004), and Yin and Ibrahim (2005) among many others. In contrast, the Gamma-frailty PH is less frequently adopted for studying multivariate current status or interval-censored data, due to the complex structure of this data. Chang et al. (2007) and Wen and Chen (2011) studied the Gamma-frailty PH model from a theoretical and computational perspectives but for clustered current status data. Hens et al. (2009) proposed a correlated Gamma-frailty PH model for bivariate current status data. Considering the popularity of the frailty PH model, further efforts to develop flexible techniques of analyzing multivariate current status and interval-censored data under this model are obviously needed. In this dissertation, we propose Bayesian and frequentist approaches to analyze the bivariate current status data and general interval-censored data under Gamma-frailty PH model.

1.5 Estimation of the association of bivariate interval-censored data

As mentioned above, one of primary goals for the bivariate interval-censored data analysis is to quantify the association of the two failure times. The copula model is one of the commonly used models for bivariate data to estimate the association. Shih and Louis (1995) and Hsu and Prentice (1996) applied a copula model to derive the dependence for bivariate right-censored data. Sun et al. (2006) proposed an estimation of the association between two correlated interval-censored failure times followed by a copula model.

Copula model defines the joint survival function as

$$S(s,t) = C_{\alpha_c}\{S_1(s), S_2(t)\},$$

where s and t are two survival times with respect to the two events, C_{α_c} is a genuine

survival function mapping the unit square $[0,1]^2$ to $[0,1]$, and α_c is a global association of parameter. S_1 and S_2 are the marginal survival functions of two interested failure times, respectively. The flexibility of choosing different genuine functions is one of the attractive features to when applying the copula model in survival analysis. For example, Archimedean copula is one of the many bivsariate survival model families which represents the joint distribution as $C_{\alpha_c}\{S_1(s), S_2(t)\} = \phi_{\alpha_c}\Big[\phi^{-1}_{\alpha_c}\{S_1(s)\}, \phi^{-1}_{\alpha_c}\{S_2(t)\}\Big]$. If specifying ϕ as the Laplace transformation of Gamma distribution, we obtain

$$C_{\alpha_c}\{S_1(s), S_2(t)\} = \Big[\{S_1(s)\}^{1-\alpha_c} + \{S_2(t)\}^{1-\alpha_c} - 1\Big]^{1/(1-\alpha_c)},\ \alpha_c > 1, \tag{1.4}$$

which is commonly referred to as Clayton copula (Clayton, 1978). In the following chapters, one may notice the joint survival function derived from Gamma-frailty PH model is a special case of Clayton copula.

A worth noting from the different copula models is that the association parameter α_c is difficult to directly interpret the correlation due to different genuine function. Commonly, Kendall's τ, the rank correlation coefficient, is derived to measure the correlation. It is defined as

$$\tau = E[\text{sign}\{(T_{i1} - T_{j1})(T_{i2} - T_{j2})\}],$$

where (T_{i1}, T_{i2}) and (T_{j1}, T_{j2}) are two independent and identically distributed copies of (T_1, T_2) and sign$(\cdot)$ is the usual sign function; i.e., this function takes values 1, 0, and -1 when the argument is positive, zero, and negative, respectively. Consequently, one can re-express τ as

$$\tau = Pr\{(T_{i1} - T_{j1})(T_{i2} - T_{j2}) > 0\} - Pr\{(T_{i1} - T_{j1})(T_{i2} - T_{j2}) < 0\}.$$

In the copula model, α_c is linked to Kendall's τ through $\tau = 4\int_0^1\int_0^1 C_{\alpha_c}(u,v)dudv - 1$. Under Clayton copula (1.4), Kendall's τ is explicitly represented via $\tau = \dfrac{\alpha_c - 1}{\alpha_c + 1}$. Under Gamma-frailty PH model, it can be represented as $\tau = \dfrac{1}{1 + 2\nu}$.

1.6 Dirichlet process mixture model

In Chapter 3, we propose a Dirichlet process Gamma mixture distribution to allow random distribution of the frailty in the frailty PH model. Dirichlet process (D-P) is currently one of the most popular Bayesian nonparametric models which was first introduced in (Ferguson, 1973) for general Bayesian statistical modeling in non-parametric problems. DP, denoted by $\mathcal{DP}(\alpha H_0)$, is defined by two parameters a positive concentration parameter α and a probability base measure H_0. Suppose a random distribution H is distributed according to $\mathcal{DP}(\alpha H_0)$ over a parameter space Θ: $H \sim \mathcal{DP}(\alpha H_0)$. Then for any finite measurable partition $\{A_1, A_2, ..., A_q\}$ of Θ,

$$\{H(A_1), H(A_2), ..., H(A_q)\} \sim Dirichlet\ \{\alpha H_0(A_1), \alpha H_0(A_2), ..., \alpha H_0(A_q)\}.$$

For any set A, $H(A)$ has a Beta distribution with mean $H_0(A)$ and variance $H_0(A)\{1-H_0(A)\}/\alpha+1$. As a result, if α goes large, the DP will concentrate on its mean distribution H_0.

To be more precise about the what a random draw from a DP looks like, Sethuraman (1981, 1994) developed a constructive definition of the DP called the stick-breaking construction. It represents the Dirichlet process $H \sim \mathcal{DP}(\alpha H_0)$ as a sum of infinity components:

$$H(\cdot) = \sum_{q=1}^{\infty} \pi_q \delta_{v_q}(\cdot),$$

where $v_q \sim H_0$ and $\delta_{v_q}(\cdot)$ is a dirac function with mass on v_q. An infinite sequence of weights $\{\pi_q\}_{q=1}^{\infty}$ is distributed according to a GEM (Griffiths-Engen-McCloskey) process with concentration parameter α (Pitman and Yor, 1997): $\boldsymbol{\pi}|\alpha \sim \text{GEM}(\alpha)$ that is

$$\begin{aligned} \pi_1 &= v_1, \pi_q = (1-v_1)(1-v_2)...(1-v_{q-1})v_q,\ q \geq 2, \\ v_q &\sim \text{Beta}(1, \alpha). \end{aligned}$$

Dirichlet process mixture model (DPM) was first formalized by Antoniak (1974) and Ferguson (1983) in which the DP was used as a prior over the distribution of the parameters. Suppose η_i is distributed according to some distribution $F(\nu_i)$ parameterized by ν_i: $\eta_i|\nu_i \sim F(\nu_i)$ and each ν_i is independently and identically drawn from the Dirichlet process $\mathcal{DP}(\alpha H_0)$: $\nu_i|H \sim H,\ H|(\alpha, H_0) \sim \mathcal{DP}(\alpha H_0)$. Hence, we generate a Dirichlet process mixture for the distribution of η_i. Following the perspective of stick-breaking process which constructs the DP with countably infinite sum of atomic measures, DPM can be rebuilt as a mixture model. By introducing a cluster variable d_i which is assigned a probability π_q to be the integer q, the DPM can be rewritten as

$$\begin{aligned} \boldsymbol{\pi}|\alpha &\sim \text{GEM}(\alpha), \qquad \nu_q|H_0 \sim H_0, \\ d_i|\boldsymbol{\pi} &\sim \text{Mult}(\boldsymbol{\pi}),\ \eta_i|\{\nu_q\}, d_i \sim F(\nu_{d_i}). \end{aligned} \tag{1.5}$$

In literature, Dirichlet mixture process has been used extensively in different types of modelling problems when conventional parametric priors would impose unreasonably stiff constraints on the distributional assumptions. In the context of density estimation and clustering data, DPM received a lot of attention (Lo, 1984; Escobar and West, 1995; Rasmussen, 2000; Neal, 2000). In the failure time modelling when the support of the density is on the positive reals, Hanson (2006) pointed out that few researches contributed to that topic and he proposed an AFT model with baseline survival function modelled by Dirichlet process Gamma mixture. In chapter 3, we assign the Dirichlet process Gamma mixture as a prior to the frailty distribution to avoid the parametric assumption.

Chapter 2

Regression analysis of bivariate current status data under the Gamma-frailty proportional hazards model using the EM algorithm

2.1 Introduction

Current status data commonly arise in many epidemiological, social, and medical studies and it is characterized by the fact that the failure time of interest is not directly observed, but is known to occur either before or after an examination time. In other words, the failure times are either left- or right-censored. We illustrate the current status data with a couple of examples in the introduction chapter. We notice that the statistical literature is replete with methods of analyzing current status data pertaining to a single failure time of interest. Huang (1996) applied the proportional hazards model to the univariate current status data to obtain the maximum likelihood estimator and asymptotic variance matrix for regression parameter. Huang and Rossini (1997) fit with the proportional odds model using sieve method. McMahan et al (2013) developed an EM algorithm for the analysis of univariate current status data .Some review of methods can be found in Huang and Wellner (1997); Jewell and van der Laan (2003); Sun (2006); Ding-Geng Chen, et al.(2012).

The primary goal of this chapter is to develop a precise, flexible, and computationally efficient method that can be used to analyze correlated bivariate current status data under the Gamma-frailty PH model. To provide adequate modeling flexibility,

the monotone splines of Ramsay (1988) are adopted to model the unknown conditional cumulative baseline hazard functions. Through a three-stage data augmentation procedure, involving latent Poisson random variables, a novel EM algorithm is derived for the purposes of model fitting. At each iteration of our algorithm the spline coefficients are updated in closed form, with the regression parameters being updated through solving a low-dimensional system of equations. Additionally, all of the expectations involved in the E-step of our algorithm are available in closed form. These features allow our approach to be very computationally efficient, when compared to other competing methods, especially for the analysis of large data sets as is illustrated in our simulation and data analysis sections. Further our proposed technique is easy to implement and robust to initialization.

The remainder is organized as follows. In Section 2.2, we present the Gamma-frailty PH model, the observed data likelihood, and describe how monotone splines are used to approximate the unknown conditional cumulative baseline hazard functions. In Section 2.3, we present the derivation of our EM algorithm. In Section 2.4 we illustrate the performance of our method through a simulation study and in Section 2.5 we present the results from a real data application. Section 2.6 concludes with a summary and discussion.

2.2 Model, data, and likelihood.

Gamma-frailty PH model

Let T_1 and T_2 denote two failure times of interest. Under the Gamma-frailty PH model the conditional cumulative hazard function for T_j, given the frailty η, can be expressed as

$$\Lambda_j(t|\boldsymbol{x},\eta) = \Lambda_{0j}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\eta, \text{ for } j = 1,2 \tag{2.1}$$

where $\eta \sim \mathcal{G}a(\nu, \nu)$, Λ_{0j} is the conditional cumulative baseline hazard function, $\boldsymbol{x}$ is a vector of covariates, and $\boldsymbol{\beta}_j$ is the corresponding vector of regression parameters. This model assumes that T_1 and T_2 are conditionally independent given the frailty term η. Further, under the Gamma-frailty PH model the conditional cumulative distribution function for T_j is given by $F_j(t|\boldsymbol{x}, \eta) = 1 - \exp\{-\Lambda_{0j}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\eta\}$. Theoretically for $F_j(t|\boldsymbol{x}, \eta)$ to be proper, Λ_{0j} must be a nondecreasing function such that $\Lambda_{0j}(0) = 0$ and $\lim_{t\to\infty} \Lambda_{0j}(t) = \infty$. The latter characteristic can be relaxed when one considers a finite time domain; e.g., when conducting data analysis.

The Gamma-frailty PH model has several desirable properties. First, greater modeling flexibility can be obtained since the conditional cumulative baseline hazard functions (i.e., Λ_{0j} for $j = 1, 2$) are not required to have a specific form. A discussion on how to exploit this characteristic is provided in Section 2.3. Second, the marginal and joint survival functions can be expressed in closed form as

$$S_j(t|\boldsymbol{x}) = P(T_j > t|\boldsymbol{x}) = \left\{1 + \nu^{-1}\Lambda_{0j}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\right\}^{-\nu}, \text{ for } j = 1, 2, \tag{2.2}$$

$$\begin{aligned} S(t_1, t_2|\boldsymbol{x}) &= P(T_1 > t_1, T_2 > t_2|\boldsymbol{x}) \\ &= \left\{1 + \nu^{-1}\Lambda_{01}(t_1)\exp(\boldsymbol{x}'\boldsymbol{\beta}_1) + \nu^{-1}\Lambda_{02}(t_2)\exp(\boldsymbol{x}'\boldsymbol{\beta}_2)\right\}^{-\nu}, \end{aligned}$$

respectively. Notice, it can be ascertained from (2.2) that T_j marginally follows a generalized odds-rate hazards model (Scharfstein et al., 1998; Banerjee et al., 2007). Consequently, the regression parameter $\boldsymbol{\beta}_j$ can also be interpreted as the marginal covariate effects on T_j under the generalized odds-rate hazards model. Finally, the Gamma-frailty PH model also provides a closed-form expression for the correlation between the two failure times in terms of Kendall's τ with $\tau = (1+2\nu)^{-1}$ (Hougaard, 2000; Wang and Ding, 2000; Sun et al., 2006). Kendall's τ is defined as

$$\tau = E[\text{sign}\{(T_{i1} - T_{j1})(T_{i2} - T_{j2})\}],$$

where (T_{i1}, T_{i2}) and (T_{j1}, T_{j2}) are two independent and identically distributed copies of (T_1, T_2) and sign$(\cdot)$ is the usual sign function; i.e., this function takes values 1, 0,

and -1 when the argument is positive, zero, and negative, respectively. It is important to notice that under the Gamma-frailty PH model τ is a deterministic function of ν. Therefore, the correlation structure (i.e., τ) between the two failure times can be accurately estimated as long as as a good estimate of ν is available.

Data structure and observed likelihood function

In order to derive the observed data likelihood, we proceed under several common assumptions as in Chen et al. (2009), Hens et al. (2009), and Wen and Chen (2011) among others. In particular, we assume that the two failure times are subject to univariate censoring at the observation time C, and moreover that T_j, given the covariates, is independent of C, for j=1,2. Let $\delta_j = 1_{(T_j \leq C)}$ denote the censoring indicator for T_j, where $1_{(\cdot)}$ denotes the usual indicator function; i.e., $\delta_j = 1$ if the failure time is left-censored and $\delta_j = 0$ if it is right-censored. Consequently, the observed data from a study consisting of n subjects can be succinctly expressed as $\{(c_i, \boldsymbol{x}_i, \delta_{i1}, \delta_{i2}), i = 1, ..., n\}$, where each $(c_i, \boldsymbol{x}_i, \delta_{i1}, \delta_{i2})$ is an independent realization of $(C, \boldsymbol{x}, \delta_1, \delta_2)$. Under the Gamma-frailty PH model and the aforementioned assumptions, the observed data likelihood is

$$\mathcal{L}_{obs} = \prod_{i=1}^{n} \int \Big[\prod_{j=1}^{2} \{1 - S_j(c_i|\boldsymbol{x}_i, \eta_i)\}^{\delta_{ij}} \{S_j(c_i|\boldsymbol{x}_i, \eta_i)\}^{(1-\delta_{ij})} \Big] g(\eta_i|\ \nu, \nu) d\eta_i,$$

where $S_j(t|\boldsymbol{x}_i, \eta_i) = 1 - F_j(t|\boldsymbol{x}_i, \eta_i)$ and $g(\cdot|\ \nu, \nu)$ is the probability density function of a gamma random variable whose shape and rate parameters both equal ν. Using the expressions that were presented in Section 2.1 for the marginal and joint survival functions one can rewrite the observed data likelihood, after integrating over the frailty terms, as

$$\begin{aligned} \mathcal{L}_{obs} \ = \ & \prod_{i \in A_1} S(c_i, c_i|\boldsymbol{x}_i) \prod_{i \in A_2} \{S_1(c_i|\boldsymbol{x}_i) - S(c_i, c_i|\boldsymbol{x}_i)\} \prod_{i \in A_3} \{S_2(c_i|\boldsymbol{x}_i) - S(c_i, c_i|\boldsymbol{x}_i)\} \\ & \times \prod_{i \in A_4} \{1 - S_1(c_i|\boldsymbol{x}_i) - S_2(c_i|\boldsymbol{x}_i) + S(c_i, c_i|\boldsymbol{x}_i)\}, \end{aligned} \tag{2.3}$$

where $A_1 = \{i : \delta_{i1} = 0, \delta_{i2} = 0\}$, $A_2 = \{i : \delta_{i1} = 0, \delta_{i2} = 1\}$, $A_3 = \{i : \delta_{i1} = 1, \delta_{i2} = 0\}$, and $A_4 = \{i : \delta_{i1} = 1, \delta_{i2} = 1\}$.

Monotone splines for $\Lambda_{0j}(t)$

The unknown parameters in (2.3) involve the regression parameters $\boldsymbol{\beta}_j$ and the nondecreasing functions Λ_{0j}, for $j = 1, 2$. It is important to note that Λ_{0j} is an infinite dimensional parameter. Following the work of Cai et al. (2011) and McMahan et al. (2013), we propose to model these unknown functions using the monotone splines of Ramsay (1988). Specifically, we assume that Λ_{0j} can be written as,

$$\Lambda_{0j}(t) = \sum_{l=1}^{k} \gamma_{jl} I_{jl}(t), \tag{2.4}$$

where $I_{jl}(\cdot)$, for $l = 1, ..., k$, is a monotone spline basis function and γ_{jl} is the corresponding spline coefficient. In particular, the spline basis functions are nondecreasing piecewise polynomials ranging from 0 to 1 and the spline coefficients are restricted to be positive (i.e., $\gamma_{jl} \geq 0$). Proceeding in this fashion ensures the monotonicity of Λ_{0j}.

The basis functions of Ramsay (1988) are fully determined once their degree and knot set have been specified. The degree of the basis functions controls the overall smoothness of the splines; e.g., specifying the degree to be 1, 2, or 3 corresponds to the splines being linear, quadratic, or cubic, respectively. The knot set is typically comprised of an increasing sequence of values within the data range, and in conjunction with the degree controls the shape of the splines. Given the degree and knot set the number of corresponding basis functions (k) is equal to the degree plus the number of interior knots.

For our purposes, it is reasonable to use the same set of basis functions to model both Λ_{01} and Λ_{02}. That is to say, since both failure times are subject to the same censoring time we are modeling both of the conditional cumulative baseline hazard functions over the same time domain. Consequently, to simplify our notation we

write $I_{jl}(\cdot)$, for $j = 1, 2$, as $I_l(\cdot)$ from henceforward. It has been our experience that specifying the degree of the basis functions to be either 2 or 3 provides adequate smoothness. Further, we recommend that the knot set consist of a fixed number of equally spaced points between the minimum and maximum of the censoring times. Model selection criteria, such as Akaike information criterion (AIC) or the Bayesian information criterion (BIC), can be used to determine the appropriate number of knots, as was demonstrated in Rosenberg (1995) and McMahan et al. (2013). An alternate approach would be to treat both the number and position of the knots as unknown parameters and optimize over them according to some selection criterion as in Shen (1988) or use Bayesian reversible jump Markov chain Monte Carlo method (Green, 1995). Methods based on this strategy are usually computationally burdensome and time consuming.

2.3 The proposed method

A data augmentation

In lieu of the monotone spline representation of the conditional cumulative baseline hazard functions, the unknown parameters in (2.3) are $\boldsymbol{\theta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2', \nu)'$, where $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jk})'$ for $j = 1, 2$. Consequently, one could obtain an estimator of $\boldsymbol{\theta}$ by directly maximizing the observed data likelihood; i.e., the maximum likelihood estimate (MLE) could be obtained as $\widehat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}} \mathcal{L}_{obs}(\boldsymbol{\theta})$. However, numerically maximizing (2.3) with respect to $\boldsymbol{\theta}$ is challenging because those spline coefficients are constrained to be nonnegative. From our experience, it is very difficult to set up good initial values for this constrained optimization , and common numerical optimization techniques including Newton related algorithms fail to provide converged results. To obviate these difficulties, we have developed a novel EM algorithm for the purposes of obtaining the MLE of $\boldsymbol{\theta}$.

The derivation of our proposed EM algorithm relies on a three-stage data augmentation involving latent Poisson random variables. In the first stage we introduce the individual frailty terms (i.e., η_i for $i = 1, ..., n$) as latent random variables, and obtain the following conditional likelihood

$$\mathcal{L}_1(\boldsymbol{\theta}) = \prod_{i=1}^{n} g(\eta_i|\ \nu, \nu) \prod_{j=1}^{2} \{1 - S_j(c_i|\boldsymbol{x}_i, \eta_i)\}^{\delta_{ij}} \{S_j(c_i|\boldsymbol{x}_i, \eta_i)\}^{(1-\delta_{ij})}. \tag{2.5}$$

The second stage involves relating the censoring indicator δ_{ij} to a latent Poisson random variable z_{ij}, where $z_{ij}|\eta_i \sim \mathcal{P}\{\cdot|\Lambda_{0j}(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\}$ and $\mathcal{P}(\cdot|a)$ denotes the cumulative distribution function of the Poisson distribution with mean a. Consequently, the conditional likelihood of the observed data and the latent variables can be expressed as

$$\mathcal{L}_2 = \prod_{i=1}^{n} g(\eta_i|\ \nu, \nu) \prod_{j=1}^{2} \delta_{ij}^{1_{(z_{ij}>0)}} (1 - \delta_{ij})^{1_{(z_{ij}=0)}} \mathcal{P}\{z_{ij}|\Lambda_{0j}(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\}. \tag{2.6}$$

In the final stage, we decompose each z_{ij} as the sum of k independent latent Poisson random variables; i.e., we let $z_{ij} = \sum_{l=1}^{k} z_{ijl}$, where $z_{ijl}|\eta_i \sim \mathcal{P}\{\cdot|\gamma_{jl}I_l(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\}$ for $l = 1, ..., k$. At this layer, we arrive at the conditional likelihood that we refer to as the complete data likelihood, which is given by

$$\mathcal{L}_c = \prod_{i=1}^{n} g(\eta_i|\ \nu, \nu) \prod_{j=1}^{2} \delta_{ij}^{1_{(z_{ij}>0)}} (1 - \delta_{ij})^{1_{(z_{ij}=0)}} \prod_{l=1}^{k} \mathcal{P}\{z_{ijl}|\gamma_{jl}I_l(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\}. \tag{2.7}$$

It is important to notice that by integrating all of the z_{ijl}'s out of (2.7) one would obtain (2.6), similarly integrating all of the z_{ij}'s out of (2.6) results in (2.5), and finally integrating all of the η_i's out of (2.5) leads back to the observed data likelihood.

The EM algorithm

To develop our EM algorithm we will view (2.7) as our complete data likelihood, in which the latent variables (i.e., the z_{ijl}'s, z_{ijl}'s, and η_i's) are viewed as missing. The E-step in our EM algorithm involves taking the expectation of $\log\mathcal{L}_c(\boldsymbol{\theta})$ with respect

to all of the latent variables conditional on the observed data, which we denote by $\mathcal{D}$, and the current parameter estimate $\boldsymbol{\theta}^{(m)} = (\boldsymbol{\beta}_1^{(m)\prime}, \boldsymbol{\beta}_2^{(m)\prime}, \boldsymbol{\gamma}_1^{(m)\prime}, \boldsymbol{\gamma}_2^{(m)\prime}, \nu^{(m)})'$. This yields $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) = H_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) + H_3(\boldsymbol{\theta}^{(m)})$, where

$$\begin{aligned} H_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) &= \sum_{i=1}^{n}\sum_{j=1}^{2}\sum_{l=1}^{k}\left[\{\log(\gamma_{jl}) + \boldsymbol{x}_i'\boldsymbol{\beta}_j\}E(z_{ijl}) - \gamma_{jl}I_l(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)E(\eta_i)\right], \\ H_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)}) &= n\nu\log(\nu) - n\log\{\Gamma(\nu)\} + \nu\sum_{i=1}^{n}[E\{\log(\eta_i)\} - E(\eta_i)], \end{aligned}$$

and $H_3(\boldsymbol{\theta}^{(m)})$ is a function of $\boldsymbol{\theta}^{(m)}$ but is free of $\boldsymbol{\theta}$. Notice, for notational brevity we have suppressed the conditioning arguments in the above expectations; i.e., it is understood that $E(\cdot) = E(\cdot|\mathcal{D}, \boldsymbol{\theta}^{(m)})$. All of the conditional expectations in $H_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ and $H_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ have closed form expressions and are provided in the appendix, along with a brief sketch of their derivation. The M-step in our algorithm then finds $\boldsymbol{\theta}^{(m+1)}$ as the maximizer of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$; i.e., $\boldsymbol{\theta}^{(m+1)} = \text{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$. First we note that $H_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ is free of ν, consequently one can obtain $\nu^{(m+1)}$ by directly maximizing $H_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$, and it is easy to show that this maximizer is unique. Similarly, $H_2(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ is free of the regression parameters and spline coefficients, so we need only maximize $H_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})$ with respect to the $\boldsymbol{\beta}_j$'s and $\boldsymbol{\gamma}_j$'s. To this end, we consider the following partial derivatives,

$$\begin{aligned} \frac{\partial H_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{n}\{E(z_{ij}) - E(\eta_i)\Lambda_{0j}(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\}\boldsymbol{x}_i', \text{ for } j = 1, 2, \\ \frac{\partial H_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})}{\partial \gamma_{jl}} &= \sum_{i=1}^{n}{\gamma_{jl}}^{-1}E(z_{ijl}) - I_l(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)E(\eta_i), \text{ for } l = 1, ..., k \text{ and } . \end{aligned}$$

Setting $\partial H_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})/\partial\gamma_{jl}$ equal to 0 and solving for γ_{jl} we obtain the maximizer as a function of $\boldsymbol{\beta}_j$, which can be expressed as

$$\gamma_{jl}^{*(m)}(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n} E(z_{ijl})}{\sum_{i=1}^{n} E(\eta_i)I_l(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)} \tag{2.8}$$

for $l = 1, ..., k$ and $j = 1, 2$. To find $\boldsymbol{\beta}_j^{(m+1)}$, one would then replace γ_{jl} by (2.8) in the system of equations given by $\partial H_1(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m)})/\partial\boldsymbol{\beta}_j = 0$ and solve for $\boldsymbol{\beta}_j^{(m+1)}$. In doing so, $\gamma_{jl}^{(m+1)}$ is simultaneously determined as $\gamma_{jl}^{(m+1)} = \gamma_{jl}^{*(m)}(\boldsymbol{\beta}^{(m+1)})$, for $l = 1, ..., k$

and $j = 1, 2$. Notice the expression of $\gamma_{jl}^{(m+1)}$ automatically satisfies the nonnegative constraint for each j and l.

We now succinctly state our EM algorithm. First, initialize $\boldsymbol{\theta}^{(0)}$ and set $m = 0$, then repeat the following steps until convergence:

1. Calculate $\boldsymbol{\beta}_j^{(m+1)}$ by solving the following system of equations for each j,

$$\sum_{i=1}^{n} \left\{ E(z_{ij}|\mathcal{D}, \boldsymbol{\theta}^{(m)}) - E(\eta_i|\mathcal{D}, \boldsymbol{\theta}^{(m)}) \sum_{l=1}^{k} \gamma_{jl}^{*(m)}(\boldsymbol{\beta}) I_l(c_i) \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_j) \right\} \boldsymbol{x}_i = 0.$$

2. Calculate $\gamma_{jl}^{(m+1)} = \gamma_{jl}^{*(m)}(\boldsymbol{\beta}^{(m+1)})$, for $l = 1, \ldots, k$ and $j = 1, 2$.

3. Obtain $\nu^{(m+1)} = \text{argmax}_{\nu} H_2(\boldsymbol{\theta}; \boldsymbol{\theta}^{(m)})$ and update $m = m + 1$.

Denote the final value $\boldsymbol{\theta}^{(m+1)}$ at convergence as $\widehat{\boldsymbol{\theta}}$. It is easy to show that $\widehat{\boldsymbol{\theta}}$ satisfies the score equations associated with the observed data likelihood, and it is therefore the MLE of $\boldsymbol{\theta}$.

Variance Estimate

In order to draw inference, an estimate of the variance-covariance matrix, $\boldsymbol{\Sigma}$, of $\widehat{\boldsymbol{\theta}}$ has to be obtained. Since the observed data likelihood exists in closed form, a natural estimator of $\boldsymbol{\Sigma}$ would be $\{I(\widehat{\boldsymbol{\theta}})\}^{-1}$, where $I(\boldsymbol{\theta})$ is the observed information matrix; i.e., $I(\boldsymbol{\theta}) = \partial^2 \log \mathcal{L}_{obs} / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'$. As one might expect, analytically expressing the mixed partial derivatives involved in $I(\boldsymbol{\theta})$ can be quite tedious, and likewise evaluating them can be computationally burdensome. A plausible alternative would involve using Louis's method to evaluate $I(\widehat{\boldsymbol{\theta}})$, but this approach is also fraught with the same complexities. To circumvent these issues, we suggest implementing the approach outlined in Zeng et al. (2007) and Lin and Wang (2010). In particular, we propose to approximate the (s, l)th element of $I(\widehat{\boldsymbol{\theta}})$ using

$$I_{s,l}(\widehat{\boldsymbol{\theta}}) = h_n^{-2} \Big[\log\{\mathcal{L}(\widehat{\boldsymbol{\theta}} + h_n \vec{e}_s - h_n \vec{e}_l)\} - \log\{\mathcal{L}(\widehat{\boldsymbol{\theta}} + h_n \vec{e}_s)\} - \log\{\mathcal{L}(\widehat{\boldsymbol{\theta}} - h_n \vec{e}_l)\} + \log\{\mathcal{L}(\widehat{\boldsymbol{\theta}})\} \Big]$$

where h_n is a tuning constant of order $n^{-1/2}$ and $\vec{e}_s$ is a binary vector whose sth element is 1 with all others being 0. Proceeding in this fashion provides a straightforward, reliable, and efficient method of estimating $\boldsymbol{\Sigma}$.

2.4 Simulation Study

In order to evaluate the finite sample performance of the proposed methodology, several extensive simulation studies were conducted. In particular, three scenarios were considered: Scenario (1) considers the situation in which the assumed model (i.e., the Gamma-frailty PH model) is correctly specified; Scenario (2) investigates the situation in which the frailty distribution is misspecified; and Scenario (3) examines the situation in which the two failure times share both a common cumulative baseline hazard function and set of regression parameters. The two former scenarios investigate the performance of the proposed methodology across a wide variety of settings, while the latter allows for a direct comparison between the proposed approach and the methodology presented in Wen and Chen (2011).

Under Scenario (1), the following models for T_1 and T_2 were considered,

$$F_j(t|x_1, x_2, \eta) = 1 - \exp\{-\Lambda_{0j}(t)\exp(\beta_{j1}x_1 + \beta_{j2}x_2)\eta\}, \text{ for } j = 1, 2,$$

where $x_1 \sim$ Bernoulli(0.5), $x_2 \sim N(0, 0.5^2)$, $\eta \sim \mathcal{G}(1, 1)$, $\Lambda_{0j}(t) = \log(1 + t) + t^{3/2}$, and each of the regression parameters took on values 0.5 or -0.5, with $\beta_{11} = \beta_{21}$ and $\beta_{12} = \beta_{22}$. The censoring time C was generated from a truncated exponential distribution $\mathcal{E}(\lambda)$ with support $(0, 10)$. In order to examine different censoring rates, several values of the rate parameter were considered; i.e., $\lambda \in \{1, 2, 5\}$. The censoring indicator, δ_j, was sampled according to a Bernoulli distribution with success probability $F(C|x_1, x_2, \eta)$, for each j. Proceeding in this fashion circumvents sampling the failure times directly. For each of the regression parameter configurations we generated 500 data sets, each containing $n = 200$ observations.

To fit the proposed model, the degree of the monotone splines was specified to be 3 and a knot set consisting of 5 equally spaced knots within the minimum and maximum of the censoring times for each data set was considered. The EM algorithm outlined in Section 2.2 was then used to estimate both the regression and spline coefficients as well as the frailty variance parameter for each of the data sets. Under all simulation settings, the regression parameters were initialized to be 0, the initial value of the frailty variance ν was taken to be 2, and the initial values of the spline coefficients were randomly generated according to an $\mathcal{E}(1)$ distribution. Convergence of the algorithm was declared when all of the absolute differences between consecutive updates for the regression parameters and the frailty variance were less than 10^{-4}. To approximate the observed information matrix, the tuning parameter h_n was taken to be $0.01n^{-1/2}$. Table 2.1 and 2.2 summarizes the parameter estimates resulting from the proposed approach, across a variety of the considered simulation settings. In particular, the settings chosen for these two tables provide a summary of the performance of the proposed methodology across a variety of censoring rates. The summarized results in Table 2.1 and 2.2 include the average of the 500 MLEs minus their true value (`BIAS`), the average of the estimated standard errors (`ESE`), the sample standard deviation of the 500 MLEs (`SSD`), and empirical coverage probabilities associated with 95% Wald confidence intervals (`CP95`). As seen in Table 2.1 and 2.2, the parameter estimates exhibit little, if any, evidence of bias, the averaged standard errors are in agreement with the sample standard deviations of the MLEs, and the coverage probabilities are close to their nominal levels, for all the regression parameters and the frailty variance parameter under all considered configurations.

Scenario (2) was aimed at investigating how sensitive the proposed techniques is to the gamma frailty assumption. This simulation considered exactly the same model specifications as in Scenario (1), with the exception that a mixture log-normal

distribution was specified for the frailty term; i.e.,

$$f(\eta) = 0.25\mathcal{LN}(-1, 2) + 0.50\mathcal{LN}(-1, 0.61) + 0.25\mathcal{LN}(0.5, 0.39),$$

where $\mathcal{LN}(\mu, \sigma^2)$ denotes the log-normal distribution with location parameter μ and scale parameter σ. The proposed model was again fit using the same settings as were described in Scenario (1). Table 2.3 and 2.4 presents a summary of the regression parameter estimates obtained by the proposed methodology for the same simulation configurations as were considered in Table 2.3 and 2.4 . These results again exhibit little, if any, bias in the point estimates, the average estimated standard errors remain in agreement with the sample standard deviation of the MLEs, and the coverage probabilities are again close to 0.95 for all regression parameters, under all considered settings. This suggests that our proposed methodology can provide accurate point and variance estimates for all of the regression parameters even when the frailty distribution is egregiously misspecified; i.e., the proposed methodology is robust to the misspecification of the frailty distribution.

Scenario (3) was designed to compare the proposed approach with the methodology presented in Wen and Chen (2011). In particular, the technique proposed by Wen and Chen (2011) was designed to analyze clustered current status data under the Gamma-frailty PH model. This method can also be used to analyze bivariate current status data, under the assumption that the two correlated failure times share both a common cumulative baseline hazard function and set of regression coefficients. In order to facilitate this comparison, slight modifications of the methods presented in Section 2.4 were made so that these two techniques could be compared. In particular, the modifications allow for different censoring times for the two events and require $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}$ and $\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \boldsymbol{\gamma}$. Simulation settings were chosen to emulate the studies conducted in Wen and Chen (2011). Specifically, the common cumulative baseline hazard function was taken to be $\Lambda_0(t) = t$, the frailty distribution was specified to be $\mathcal{G}(1, 1)$, and the two censoring times were independently generated

according to $\mathcal{U}_{(0,1)}$. The settings for the regression coefficients and covariate distributions were specified to be the same as in Scenario (1). Convergence for both methods was declared when the difference between consecutive updates of the parameters of interest were less than 10^{-5}. This stricter convergence criterion was chosen because it was the default value used in Wen and Chen's Matlab package for implementing their method. Table 2.5 summarizes the parameter estimates obtained by the two competing techniques. In summary, both methods seem to perform well; however, the proposed method seems to perform slightly better; i.e., the proposed technique results in parameter estimates that exhibit both a smaller bias and variance.

In summary, through simulation the proposed methodology has been shown to be very reliable for the purposes of analyzing bivariate current status data. Further, through the derivation of a novel EM algorithm, the proposed model can be fit at a minimal computational expense; e.g., the average time required for the EM algorithm to both converge and estimate the variance-covariance matrix under Scenarios (1) and (2) was approximately 3 to 4 seconds per data set. Time trials were also conducted between the proposed approach and the methodology presented in Wen and Chen (2011). The results from these additional studies, which are provided in Table 2.6 , indicate that the proposed method is far more computationally efficient when compared to this competing technique, especially for larger sample sizes. For example, when $N = 2000$ the proposed approach is approximately 470 times faster than the methodology proposed in Wen and Chen (2011). This finding suggests that our methodology, unlike many competing approaches, can be used to analyze relatively large data sets in a timely fashion.

2.5 An application

Now we apply the proposed EM algorithm to a set of bivariate current status data we introduced in the introduction part. The initial analysis of this data set, which

was reported in Allright et al. (2000), identified risk factors associated with HIV and hepatitis B using logistic regression; i.e., these authors treated the HBV and HIV infection statuses of the individuals as binary responses and modeled each response individually. As both of these infection can be asymptomatic, many afflicted subjects were likely infected long before the time of the survey, a feature that was not accounted for in the original analysis. Further, the original analysis of this data modeled the incidence of HBV and HIV separately, and therefore cannot estimate the correlation that natural exists between these two diseases. In what follows, we will focus on jointly modeling the infection times of these diseases, both of which are not directly observed but are rather known to be either left- or right-censored relative to the age of the participant at the time of testing. Proceeding in this fashion will allow us to estimate the cumulative incidence of these two diseases, assess different risk factors, and quantify the statistical correlation between the two diseases. The censoring time is taken to be the participants age at the time of the survey. In total, there were 865 prisoners who provided completed questionnaires and were tested for both diseases. Table 2.7 summarizes the detailed information of the data sets. Figure 2.1 displays the histogram for the age of the prisoners.

In the analysis, we considered the following covariates: whether the participants had ever been treated for a sexually transmitted disease (TreatSTD), participated in intravenous drug use (EVRINJ), smoked heroin (SMHeroin), had a sexual relationship with another man before/after committal (ASM/ASMIP), and whether they used condoms during heterosexual intercourse (UseCondoms) as well as the amount of time the participants spent in prison in the past 10 years (TSL10Y). With the exception of TSL10Y, all of the aforementioned covariates are binary. To implement our proposed methodology, we took the degree of the monotone splines to be 3 and tried several knot sets each consisting of m equally spaced interior knots. It was observed that the estimation results are relatively robust to the number of knots. Table

2.9 presents AIC and BIC values corresponding to different values of m. Since both AIC and BIC are smallest when $m = 5$, we chose our final model to be the one that makes use of 5 interior knots to approximate the the conditional cumulative baseline hazards function. For the purpose of comparison, we also performed a complimentary univariate analysis of this data. More specifically, using the R function `glm` we fit a binary regression model for each of the diseases, separately, using a cloglog link function. This regression analysis uses the age at the time of testing as a continuous covariate and results in a special parametric PH model with $\log(\Lambda_0(t)) = ct$ for some unknown constant c. The cloglog link is chosen because the regression parameters can be interpreted as log hazard ratios, which are comparable to those provided by our joint analysis.

Table 2.8 summarizes the estimates of the regression parameters obtained from our joint analysis under our final model (i.e., when $m = 5$) and compares these results to the estimates obtained from the univariate binary regression analysis. Both of these techniques identify the same set of significant risk factors; i.e., participating in intravenous drug use increases a subjects risk of contracting both of these diseases, the risk of contracting HIV increases with the amount of time spent in prison, and condom use decreases the risk of contracting HIV. However, it is worthwhile to point out that our joint modeling approach is able to identify these covariates at a higher level of statistical significance when compared to the estimates obtained by the univariate approach.

Further, the joint analysis yields an estimate of ν to be $\hat{\nu} = 0.61$ with standard error 0.09. This leads to an estimate of Kendall's concordance τ between the two failure times to be $\hat{\tau} = 0.45$ with a standard error of 0.037. The standard error of $\hat{\tau}$ was obtained through an application of the delta method. This suggests that there is a moderate correlation between the two failure times. In comparison, it is not possible to estimate the correlation between the onset time of HBV and HIV using

the univariate modeling approach. Our proposed model also allows us to estimate the marginal cumulative incidence $1 - S_j(t|\boldsymbol{x})$ of HBV and HIV for subgroups of prisoners using equation (2). For example, Figure 2.2 presents the estimated marginal cumulative incidences of HBV for the baseline group (i.e., controlling all covariates equal to 0) and for those subjects who had participated in intravenous drug use (still controlling all other covariates equal to 0) from our joint analysis as well as the corresponding curves from the binary regression. It is reasonable that the estimated baseline cumulative incidences are similar from the two methods. However, there is a substantial difference between the estimated curves for the intravenous drug users from the two methods. The curve from our joint analysis is believed to be more accurate due to borrowing information between the correlated failure times in the joint modeling and more flexible semiparametric modeling in the marginal distributions.

2.6 Concluding remark

In this chapter, we develop a computationally efficient method of analyzing bivariate current status data under the Gamma-frailty PH model, by generalizing the work of McMahan et al. (2013). Our formulation approximates the unknown conditional cumulative baseline hazard functions with monotone splines, which significantly reduces the number of unknown parameters while maintaining adequate modeling flexibility. A three-stage data augmentation procedure is used to facilitate the derivation of our EM algorithm. The resulting algorithm involves solving a low-dimensional system of equations for updating the regression parameters and the frailty variance parameter, with the spline coefficients being updated in closed form. All of the expectations involved in the EM algorithm can be expressed in closed form. The EM algorithm is easy to implement and enjoys fast convergence. Through simulation, we have shown that the proposed method accurately and efficiently estimates all of the unknown parameters (and thus the correlation between the two failure times) when the model

is correctly specified and that the estimation of the regression parameters is robust to the misspecification of the frailty distribution.

The approach of Chen et al. (2009) can also be used to analyze bivariate current status data. Specifically their approach is based on an EM algorithm under a frailty PH model with correlated normal frailties. The EM algorithm proposed by these authors is definitively more complicated than ours since it uses a series of numerically intensive approximations to evaluate the conditional expectations involved in the E-step of their algorithm. Such numerical approximations cause not only computational burdens but convergence problems when a strict convergence criteria is used. Consequently, this method, because of its computational nature, may not be appropriate for analyzing large data sets, due to the time required to complete model fitting. On the other hand, the approach of Chen et al. (2009) allows for correlated normal frailties for multivariate current status data, while our work focuses only on bivariate current status data. Our method can naturally be extended to multivariate data with a shared frailty, but extensions allowing for correlated frailties do not appear to be straightforward.

Topics for future work include the development of hypothesis testing procedures that can be used to evaluate our modeling assumptions. For example, in our data application it may or may not be reasonable to assume that the infection times are conditionally independent of the censoring times, given the covariates. Consequently, the development of a formal method of testing this assumption would be discernibly beneficial. Further, for situations in which the aforementioned modeling assumption does not hold, we plan to extend our proposed methodology to allow for informative censoring, a modeling attribute that would extend the utilitarian nature of our work to many other epidemiological and medical research areas.

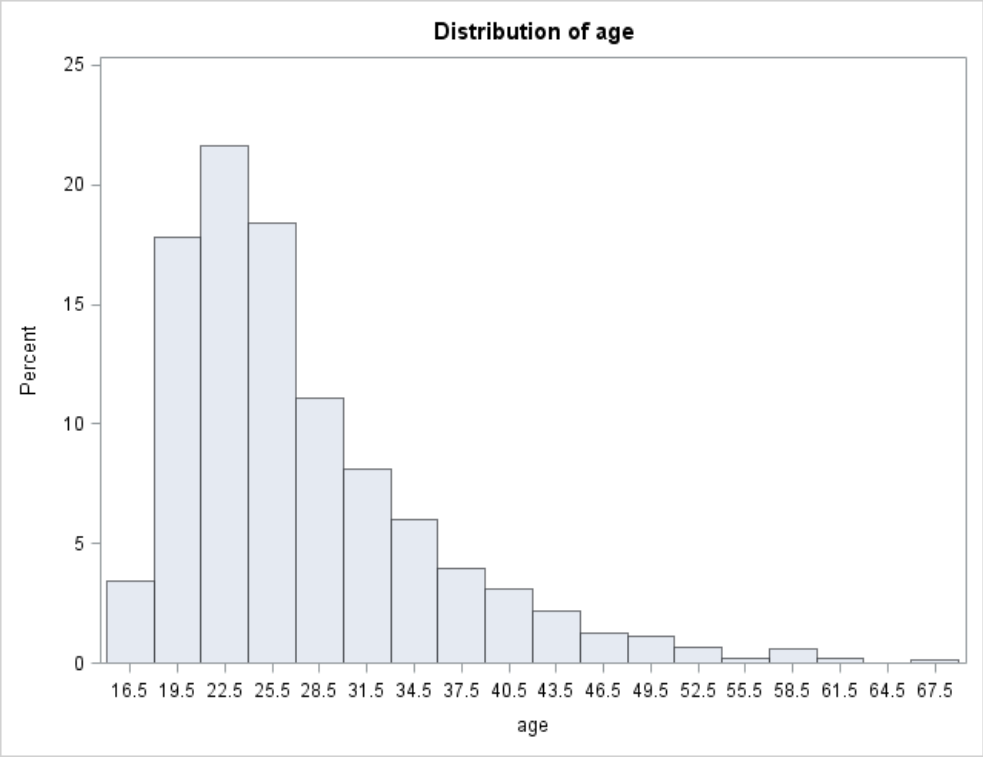

Figure 2.1 The histogram of the age.

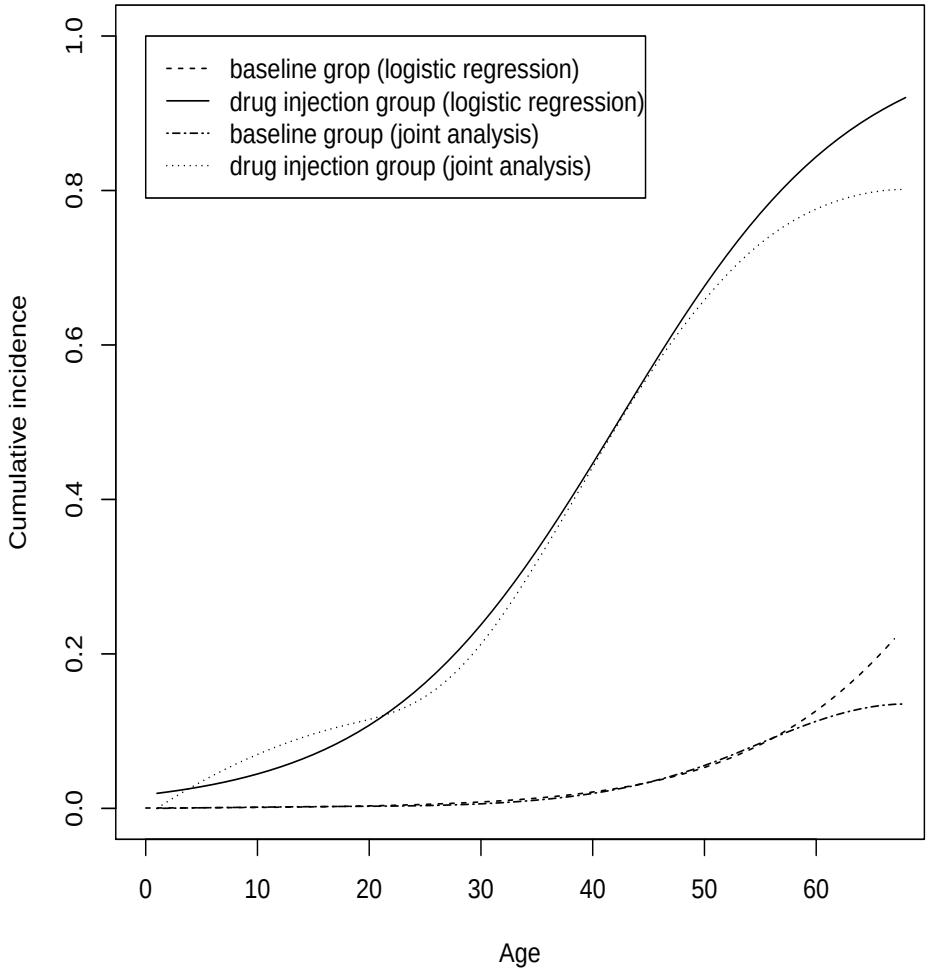

Figure 2.2 The estimated marginal cumulative incidence functions of hepatitis B for the baseline group (controlling all covariates equal to 0) and drug injection subgroup (controlling all other covariates equal to 0).

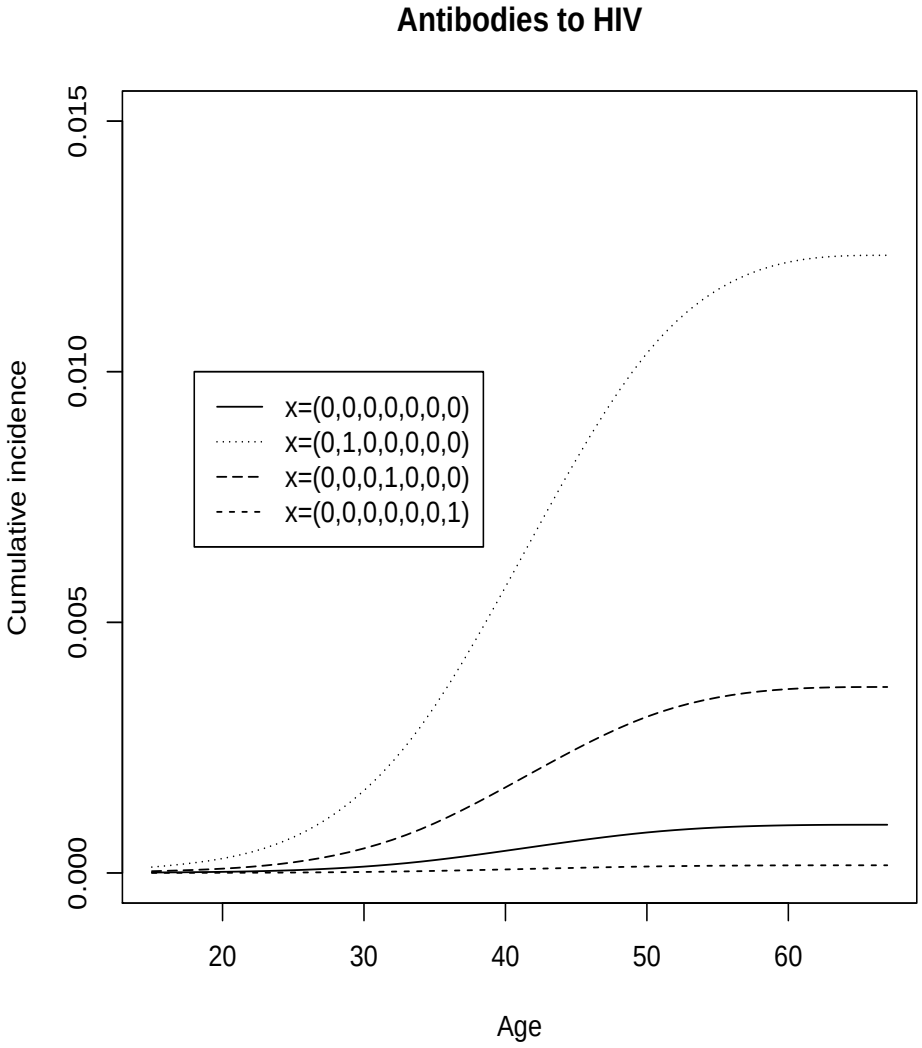

Figure 2.3 The estimated marginal cumulative incidence functions of hiv for the baseline group (controlling all covariates equal to 0) and other significant risk factor subgroups (controlling all other covariates equal to 0).

Table 2.1 Simulation results under Scenario (1) with different left-censoring rates (LR). Summarized results include the average of the 500 MLEs minus their true value (BIAS), the sample standard deviation of the 500 MLEs (SSD), the average of the estimated standard errors (ESE), and empirical coverage probabilities associated with 95% Wald confidence intervals (CP95), for each parameter.

$\beta_{j1} = -0.5$ and $\beta_{j2} = 0.5$						$\beta_{j1} = 0.5$ and $\beta_{j2} = 0.5$					
LR		BIAS	SSD	ESE	CP95	LR		BIAS	SSD	ESE	CP95
45.2%	$\hat{\beta}_{11}$	0.011	0.347	0.348	0.952	54.3%	$\hat{\beta}_{11}$	-0.007	0.354	0.346	0.936
	$\hat{\beta}_{12}$	0.012	0.340	0.352	0.956		$\hat{\beta}_{12}$	-0.005	0.370	0.351	0.942
45.2%	$\hat{\beta}_{21}$	-0.001	0.335	0.350	0.952	54.1%	$\hat{\beta}_{21}$	0.013	0.330	0.344	0.968
	$\hat{\beta}_{22}$	0.030	0.374	0.352	0.946		$\hat{\beta}_{22}$	-0.014	0.355	0.350	0.954
	$\hat{\nu}^{-1}$	0.034	0.336	0.367	0.982		$\hat{\nu}^{-1}$	-0.018	0.302	0.346	0.982
30.9%	$\hat{\beta}_{11}$	-0.040	0.366	0.364	0.952	39.9%	$\hat{\beta}_{11}$	0.011	0.361	0.350	0.956
	$\hat{\beta}_{12}$	0.021	0.395	0.372	0.946		$\hat{\beta}_{12}$	0.026	0.355	0.355	0.952
30.8%	$\hat{\beta}_{21}$	-0.029	0.368	0.363	0.946	39.7%	$\hat{\beta}_{21}$	0.010	0.363	0.352	0.946
	$\hat{\beta}_{22}$	0.038	0.384	0.372	0.956		$\hat{\beta}_{22}$	0.044	0.354	0.355	0.966
	$\hat{\nu}^{-1}$	0.077	0.416	0.414	0.978		$\hat{\nu}^{-1}$	0.039	0.355	0.385	0.992
16.2%	$\hat{\beta}_{11}$	-0.037	0.482	0.459	0.934	22.7%	$\hat{\beta}_{11}$	0.046	0.410	0.403	0.954
	$\hat{\beta}_{12}$	0.037	0.532	0.460	0.928		$\hat{\beta}_{12}$	0.048	0.410	0.405	0.956
16.0%	$\hat{\beta}_{21}$	-0.060	0.507	0.457	0.906	22.7%	$\hat{\beta}_{21}$	0.010	0.400	0.403	0.942
	$\hat{\beta}_{22}$	0.017	0.487	0.458	0.926		$\hat{\beta}_{22}$	0.014	0.406	0.402	0.952
	$\hat{\nu}^{-1}$	0.187	0.716	0.635	0.974		$\hat{\nu}^{-1}$	0.088	0.539	0.509	0.976

Table 2.2 Simulation results under Scenario (1) with different left-censoring rates (LR). Summarized results include the average of the 500 MLEs minus their true value (BIAS), the sample standard deviation of the 500 MLEs (SSD), the average of the estimated standard errors (ESE), and empirical coverage probabilities associated with 95% Wald confidence intervals (CP95), for each parameter.

$\beta_{j1} = -0.5 \quad \beta_{j2} = -0.5$						$\beta_{j1} = 0.5 \quad \beta_{j2} = -0.5$					
LR		BIAS	SSD	ESE	CP95	LR		BIAS	SSD	ESE	CP95
45.1%	$\hat{\beta}_{11}$	-0.006	0.324	0.345	0.965	54.1%	$\hat{\beta}_{11}$	-0.006	0.335	0.347	0.956
	$\hat{\beta}_{12}$	0.005	0.361	0.351	0.942		$\hat{\beta}_{12}$	-0.028	0.356	0.348	0.943
44.8%	$\hat{\beta}_{21}$	0.002	0.341	0.345	0.945	54.1%	$\hat{\beta}_{21}$	-0.012	0.333	0.347	0.965
	$\hat{\beta}_{22}$	-0.029	0.364	0.350	0.948		$\hat{\beta}_{22}$	-0.003	0.346	0.348	0.945
	$\hat{\nu}^{-1}$	0.022	0.332	0.361	0.979		$\hat{\nu}^{-1}$	-0.003	0.288	0.334	0.993
31.1%	$\hat{\beta}_{11}$	-0.001	0.366	0.366	0.953	39.6%	$\hat{\beta}_{11}$	0.033	0.334	0.354	0.963
	$\hat{\beta}_{12}$	-0.011	0.374	0.368	0.96		$\hat{\beta}_{12}$	-0.035	0.365	0.356	0.944
31.0%	$\hat{\beta}_{21}$	-0.030	0.377	0.366	0.949	39.7%	$\hat{\beta}_{21}$	0.030	0.341	0.352	0.957
	$\hat{\beta}_{22}$	-0.020	0.380	0.367	0.940		$\hat{\beta}_{22}$	-0.009	0.334	0.350	0.957
	$\hat{\nu}^{-1}$	0.050	0.423	0.422	0.983		$\hat{\nu}^{-1}$	0.043	0.369	0.377	0.980
16.20%	$\hat{\beta}_{11}$	-0.053	0.480	0.452	0.945	22.4%	$\hat{\beta}_{11}$	0.048	0.414	0.405	0.937
	$\hat{\beta}_{12}$	-0.046	0.488	0.455	0.946		$\hat{\beta}_{12}$	-0.043	0.421	0.405	0.938
16.11%	$\hat{\beta}_{12}$	-0.031	0.477	0.454	0.935	22.7%	$\hat{\beta}_{12}$	0.015	0.412	0.407	0.948
	$\hat{\beta}_{22}$	-0.045	0.467	0.455	0.944		$\hat{\beta}_{22}$	-0.038	0.412	0.403	0.942
	$\hat{\nu}^{-1}$	0.181	0.650	0.637	0.980		$\hat{\nu}^{-1}$	0.134	0.553	0.511	0.981

Table 2.3 Simulation results under Scenario (2) with different left-censoring rates (LR) when the gamma frailty distribution is misspecified. Summarized results include the average of the 500 MLEs minus their true value (BIAS), the sample standard deviation of the 500 MLEs (SSD), the average of the estimated standard errors (ESE), and empirical coverage probabilities associated with 95% Wald confidence intervals (CP95), for each parameter.

$\beta_{j1} = -0.5$ and $\beta_{j2} = 0.5$						$\beta_{j1} = 0.5$ and $\beta_{j2} = 0.5$					
LR		BIAS	SSD	ESE	CP95	LR		BIAS	SSD	ESE	CP95
43.3%	$\hat{\beta}_{11}$	0.034	0.340	0.340	0.956	52.1%	$\hat{\beta}_{11}$	-0.020	0.327	0.332	0.948
	$\hat{\beta}_{12}$	0.005	0.358	0.348	0.948		$\hat{\beta}_{12}$	0.013	0.339	0.335	0.956
43.4%	$\hat{\beta}_{21}$	0.025	0.342	0.337	0.948	52.2%	$\hat{\beta}_{21}$	-0.030	0.344	0.334	0.950
	$\hat{\beta}_{22}$	0.013	0.348	0.346	0.954		$\hat{\beta}_{22}$	0.002	0.332	0.336	0.950
29.0%	$\hat{\beta}_{11}$	-0.015	0.391	0.370	0.952	29.2%	$\hat{\beta}_{11}$	-0.006	0.335	0.357	0.968
	$\hat{\beta}_{12}$	-0.004	0.398	0.377	0.938		$\hat{\beta}_{12}$	-0.009	0.368	0.362	0.938
37.7%	$\hat{\beta}_{21}$	-0.042	0.399	0.372	0.940	37.6%	$\hat{\beta}_{21}$	0.013	0.375	0.360	0.930
	$\hat{\beta}_{22}$	0.011	0.394	0.377	0.956		$\hat{\beta}_{22}$	0.006	0.361	0.355	0.958
15.0%	$\hat{\beta}_{11}$	0.005	0.485	0.471	0.944	21.1%	$\hat{\beta}_{11}$	0.017	0.426	0.414	0.946
	$\hat{\beta}_{12}$	0.060	0.504	0.480	0.944		$\hat{\beta}_{12}$	0.022	0.428	0.417	0.958
15.0%	$\hat{\beta}_{21}$	0.009	0.486	0.464	0.938	21.4%	$\hat{\beta}_{21}$	0.011	0.444	0.420	0.948
	$\hat{\beta}_{22}$	0.029	0.505	0.478	0.944		$\hat{\beta}_{22}$	0.034	0.426	0.415	0.950

Table 2.4 Simulation results under Scenario (2) with different left-censoring rates (LR) when the gamma frailty distribution is misspecified. Summarized results include the average of the 500 MLEs minus their true value (BIAS), the sample standard deviation of the 500 MLEs (SSD), the average of the estimated standard errors (ESE), and empirical coverage probabilities associated with 95% Wald confidence intervals (CP95), for each parameter.

$\beta_{j1} = -0.5$ and $\beta_{j2} = -0.5$						$\beta_{j1} = 0.5$ and $\beta_{j2} = -0.5$					
LR		BIAS	SSD	ESE	CP95	LR		BIAS	SSD	ESE	CP95
43.1%	$\hat{\beta}_{11}$	-0.009	0.333	0.335	0.943	52.3%	$\hat{\beta}_{11}$	-0.005	0.346	0.332	0.947
	$\hat{\beta}_{12}$	-0.001	0.360	0.342	0.937		$\hat{\beta}_{12}$	-0.005	0.345	0.334	0.953
43.1%	$\hat{\beta}_{21}$	0.020	0.333	0.333	0.947	52.5%	$\hat{\beta}_{21}$	-0.010	0.317	0.332	0.949
	$\hat{\beta}_{22}$	-0.011	0.370	0.342	0.953		$\hat{\beta}_{22}$	0.003	0.345	0.336	0.953
29.0%	$\hat{\beta}_{11}$	-0.017	0.372	0.381	0.947	29.2%	$\hat{\beta}_{11}$	-0.025	0.365	0.352	0.941
	$\hat{\beta}_{12}$	-0.007	0.387	0.380	0.943		$\hat{\beta}_{12}$	-0.0163	0.373	0.352	0.944
37.7%	$\hat{\beta}_{21}$	0.009	0.380	0.373	0.947	37.5%	$\hat{\beta}_{21}$	0.007	0.355	0.352	0.946
	$\hat{\beta}_{22}$	-0.009	0.383	0.380	0.959		$\hat{\beta}_{22}$	-0.0164	0.352	0.351	0.949
15.1%	$\hat{\beta}_{11}$	-0.056	0.476	0.462	0.932	21.3%	$\hat{\beta}_{11}$	-0.009	0.445	0.417	0.936
	$\hat{\beta}_{12}$	-0.031	0.482	0.471	0.945		$\hat{\beta}_{12}$	-0.0536	0.463	0.416	0.924
15.0%	$\hat{\beta}_{21}$	-0.071	0.531	0.467	0.9050	21.2%	$\hat{\beta}_{21}$	0.000	0.418	0.421	0.943
	$\hat{\beta}_{22}$	-0.036	0.495	0.475	0.939		$\hat{\beta}_{22}$	-0.014	0.425	0.416	0.966

Table 2.5 Simulation results from the proposed method and the approach of Wen and Chen (2011) under Scenario (3). Summarized results include the average of the 500 MLEs minus the true value (BIAS), the sample standard deviation of the 500 MLEs (SSD), the average of the estimated standard errors (ESE), and empirical coverage probabilities associated with 95% Wald confidence intervals (CP95), for each parameter.

			Proposed method				Wen and Chen (2011)			
β_1	β_2		BIAS	SSD	ESE	CP95	BIAS	SSD	ESE	CP95
0.5	-0.5	$\hat{\beta}_1$	0.018	0.232	0.230	0.954	0.043	0.247	0.240	0.954
		$\hat{\beta}_2$	-0.010	0.246	0.230	0.936	-0.035	0.261	0.238	0.934
		$\hat{\nu}^{-1}$	0.019	0.404	0.427	0.990	0.097	0.445	0.461	0.966
0.5	0.5	$\hat{\beta}_1$	0.024	0.239	0.231	0.950	0.045	0.255	0.241	0.952
		$\hat{\beta}_2$	0.018	0.234	0.232	0.942	0.038	0.245	0.245	0.944
		$\hat{\nu}^{-1}$	0.048	0.438	0.428	0.990	0.120	0.473	0.473	0.956
-0.5	0.5	$\hat{\beta}_1$	-0.023	0.244	0.257	0.950	-0.048	0.261	0.258	0.942
		$\hat{\beta}_2$	-0.020	0.265	0.250	0.950	0.060	0.275	0.267	0.954
		$\hat{\nu}^{-1}$	0.119	0.522	0.542	0.976	0.214	0.572	0.594	0.994
-0.5	-0.5	$\hat{\beta}_1$	-0.025	0.253	0.257	0.954	-0.044	0.260	0.255	0.952
		$\hat{\beta}_2$	0.035	0.261	0.253	0.948	-0.040	0.280	0.257	0.944
		$\hat{\nu}^{-1}$	0.058	0.525	0.534	0.990	0.124	0.570	0.569	0.980

Table 2.6 Average convergence time (in seconds) associated with the proposed methodology and the approach of Wen and Chen (2011) per data set.

Approach	n = 100	n = 200	n = 500	n = 1000	n = 2000
Wen and Chen	18.5	26.8	163.8	1416.1	11937
Proposed	12.1	11.2	27.1	35.4	24.7

Table 2.7 The summarized demographic and covariate information including the frequency of all binary covariate variables and the responses. 0 indicates the negative status and 1 is positive.

		0	1
Covariate	TreatSTD	756(87.40%)	109(12.60%)
	EVRINJ	489(56.53%)	376(43.47%)
	SMHeroin	452(52.25%)	413(47.75%)
	ASM	847(97.92%)	18(2.08%)
	ASMIP	853(98.61%)	12(1.39%)
	UseCondom	278(32.14%)	587(67.86%)
Response	HBV	792(91.56%)	73(8.44%)
	HIV	848(98.03%)	17(1.97%)

Table 2.8 Estimated covariates effects, their standard errors, the corresponding 95% confidence intervals and the p-value from the significance test. The covariates included in the both models are: ever treated for sexually transmitted infection; injecting drug use; Smoking heroin; time spent in prison in the past 10 years; ever had sex with a man; ever had sex with a man inside prison; use condoms during heterosexual intercourse; age at the survey.

		Logistic regression				Gamma frailty PH model			
	Covariate	Point	S.E.	95%CI	p-value	Point	S.E.	95%CI	p-value
hepatitis B	TreatSTD	0.53	0.34	(-0.13,1.19)	0.12	0.56	0.31	(-0.06,1.17)	0.07
	EVRINJ	3.71	0.59	(2.56,4.85)	0	3.90	0.43	(3.05,4.75)	0
	SMHeroin	0.07	0.35	(-0.62,0.76)	0.84	0.07	0.32	(-0.56,0.70)	0.83
	TSL10Y	0.00	0.15	(-0.29,0.29)	1	0.01	0.11	(-0.20,0.22)	0.93
	ASM	0.67	0.86	(-1.01,2.36)	0.44	0.80	0.90	(-0.97,2.57)	0.37
	ASMIP	-1.06	1.33	(-3.67,1.55)	0.43	-1.30	1.35	(-3.95,1.35)	0.34
	UseCondom	-0.32	0.29	(-0.88,0.25)	0.27	-0.27	0.25	(-0.77,0.23)	0.28
s	Age	0.10	0.02	(0.06,0.13)	0	-	-	-	-
HIV	TreatSTD	0.55	0.61	(-0.64,1.74)	0.36	0.57	0.46	(-0.32,1.47)	0.22
	EVRINJ	2.54	0.86	(0.86,4.23)	0.0031	2.57	0.54	(1.48,3.65)	1.94×10^{-4}
	SMHeroin	-0.66	0.66	(-1.95,0.63)	0.32	-0.66	0.47	(-1.57,0.26)	0.16
	TSL10Y	1.51	0.66	(0.22,2.80)	0.02	1.35	0.12	(1.12,1.59)	0
	ASM	-0.19	1.74	(-3.61,3.22)	0.91	-0.39	1.41	(-3.16,2.38)	0.78
	ASMIP	1.18	1.75	(-2.24,4.60)	0.50	1.03	1.39	(-1.69,3.75)	0.46
	UseCondom	-1.86	0.55	(-2.94,-0.79)	0.0007	-1.84	0.38	(-2.58,-1.10)	1.28×10^{-6}
	Age	0.10	0.04	(0.03,0.17)	0.01	-	-	-	-

Table 2.9 Estimation of the significant risk factors effect as well as their standard deviations with different number of knots; the last two column are AIC and BIC.

	Hepatitis B	HIV					
knots	EVRINJ	EVRINJ	TSL10Y	UseCondom	ν	AIC	BIC
3	4.03(0.42)	2.34(0.52)	1.42(0.11)	-1.77(0.36)	0.56(0.08)	7217.59	7327.13
4	4.09(0.42)	2.22(0.50)	1.42(0.11)	-1.72(0.36)	0.48(0.06)	7194.40	7313.46
5	3.90(0.43)	2.57(0.54)	1.35(0.12)	-1.84(0.38)	0.61(0.09)	7101.23	7229.83
6	4.19(0.42)	2.38(0.52)	1.29(0.11)	-1.76(0.38)	0.49(0.06)	7281.98	7420.10
7	4.10(0.41)	2.19(0.49)	1.31(0.11)	-1.71(0.35)	0.47(0.06)	7208.64	7356.24
10	4.03(0.51)	2.27(0.79)	1.28(0.17)	-1.73(0.56)	0.54(0.26)	7209.78	7386.01

Chapter 3

Bayesian regression analysis of bivariate current status data

3.1 Introduction

In the this chapter, we develop an easy-to-implement Bayesian estimation method to analyze bivariate current status data under the Gamma-frailty PH model. We use the same modeling techniques as in chapter 2 such as the monotone spline approximation and the data augmentation through Poisson latent variables. In section 3.2, we propose an efficient Gibbs sampler under Gamma-frailty PH model involving only easy-to-facilitate posterior computations. In section 3.3 we provide a Dirichlet process gamma mixture prior for the frailty distribution so that we can allow for an unknown distribution for frailty. We develop a Gibbs sampler based on exact block approach. Section 3.4 shows the simulation performance and section 3.5 applies the methods to the same real-life data as in chapter 2.

3.2 Bivariate Gamma-frailty PH model

A Gibbs sampler is a MCMC (Markov chain Monte Carlo) algorithm to approximate the joint distribution of all variables by sampling the variables sequentially from their univariate conditional distributions. It requires to have a posterior distribution for each unknown parameter. The three-staged augmentation scheme developed in the previous chapter does a great favor in obtaining the common parametric distributions in the posterior computation. It substantially reduces the computation effort and

leads the MCMC easy to facilitate. The following MCMC algorithm is developed based on the augmented likelihood (2.7).

Prior and posterior computation

We assign conditionally independent exponential priors $\mathcal{E}xp(\lambda_j)$ for the spline coefficients γ_{jl}s and take a gamma distribution $\mathcal{G}(a_\lambda, b_\lambda)$for λ_j. We use the shrinkage priors for the spline coefficients to avoid the over-fitting problem when applying too many knots. We adopt independent priors for each component of $\boldsymbol{\beta}_j$ which $\pi(\cdot) = N(\mu_j, \sigma_j^2)$ and we use the $\mathcal{E}xp(1)$ as the prior for the variance parameter ν of gamma frailty distribution. The adaptive rejection Metropolis sampling (ARMS) (Gilks and Wild, 1992) is appropriate and convenient to sample $\boldsymbol{\beta}_j$ and v since their posterior densities are log-concave. The MCMC algorithm iterates through the following steps:

1. Sample each $z_{ij} \sim \mathcal{P}\Big\{\Lambda_{0j}(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\Big\}$ for $z_{ij} > 0$ if $\delta_{ij} = 1$
   and $\{z_{ijl}\}_{l=1}^k \sim \mathcal{M}\Big\{z_{ij}, (p_{ij1}, p_{ij2}, ..., p_{ijk})\Big\}$
   where $\mathcal{M}$ denotes a multinomial distribution and $p_{ijl} = \gamma_{jl}I_l(c_i)/\sum_{l=1}^k \gamma_{jl}I_l(c_i)$ for $l = 1, ...k$, $i = 1, 2, ...n$, $j = 1, 2$.

2. Sample $\gamma_{jl} \sim \mathcal{G}\Big\{1 + \sum_{i=1}^n z_{ijl}, \lambda_j + \sum_{i=1}^n I_l(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\Big\}$.
   Sample the $\lambda_j \sim \mathcal{G}(a_\lambda + k, b_\lambda + \sum_{l=1}^k \gamma_{jl})$.

3. Sample each η_i from $\mathcal{G}\Big[\nu + \sum_{q=1}^h (Z_{i1l} + Z_{i2l}), \nu + \sum_{l=1}^k \Big\{\gamma_{1l}I_l(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_1) + \gamma_{2l}I_l(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_2)\Big\}\Big]$.

4. Sample ν and $\boldsymbol{\beta}_j$ by ARMS. The full conditional distribution of ν and $\boldsymbol{\beta}_j$ is

$$\begin{aligned} \nu|\cdot &\propto \prod_{i=1}^n g(\eta_i; \nu, \nu)e^{-\nu} \\ \boldsymbol{\beta}_j|\cdot &\propto \pi(\boldsymbol{\beta}_j)\exp\Big\{\sum_{i=1}^n z_i\boldsymbol{x}_i'\boldsymbol{\beta}_j - \sum_{i=1}^n \Lambda_j(C_i)e^{\boldsymbol{x}_i'\boldsymbol{\beta}_j}\eta_i\Big\}. \end{aligned}$$

We can observe from the above algorithm that the procedure is efficient and all full conditionals have simple closed forms except for $\boldsymbol{\beta}_j$ and ν, which can be updated by ARMS. In the simulation section we observe the good performance of this Bayesian method as well as the good rate of convergence and mixing.

3.3 Dirichlet Process Gamma Mixture

Gamma frailty model gives us several promising properties but we concern about the validity of the assumption for the frailty distribution. As introduced in chapter 1, Dirichlet process mixture is widely used in modeling the unknown distributions. In this section, to relax the gamma frailty assumption, we assign the Dirichlet process Gamma mixture as a prior distribution for the frailty η:

$$f(\eta_i) = \int \mathcal{G}(\eta_i; \boldsymbol{\nu}) dH(\boldsymbol{\nu}) \tag{3.1}$$

where H is a random distribution drawn from the Dirichlet Process $\mathcal{DP}(\alpha H_0)$ with base measure H_0 and spread α. Here, We adopt the stick-breaking representation to express the Dirichlet process and rewrite (3.1) as:

$$\begin{aligned}
\eta_i|(\boldsymbol{\nu}, d_i) &\sim \mathcal{G}(\eta_i|\nu_{d_i}, \nu_{d_i}) \; i = 1, 2, ..., n \\
d|\pi_{1:\infty} &\sim \sum_{q=1}^{\infty} \pi_q \delta_q(\cdot) \\
\nu_q &\sim H_0(\boldsymbol{\nu}), \; q = 1, 2, ..., \\
\pi_1 &= v_1, \quad \pi_q = (1 - v_1)(1 - v_2)....(1 - v_{q-1})v_q, \; q \geq 2 \\
v_q &\sim Beta(1, \alpha)
\end{aligned} \tag{3.2}$$

where $\{d_i\}$s are allocation parameters used to identify the ν_{d_i} associated with ν_q and locate the measurements η_i to the parameter value ν_{d_i}.

Gibbs sampler

Gibbs sampling methods for Dirichlet process of mixture based on the hierarchical structure were being developed recently. Ishwaran and James (2001) proposed a blocked Gibbs sampler with resorting to a finite dimensional Dirichlete priors. Walker (2007) presented a slice sampling method and Papaspiliopoulos and Roberts (2008) developed a retrospective sampling method. Both of them provided an approach to the posterior simulation without approximation. Papaspiliopoulos (2008) composed those two methods and proposed an efficient exact block Gibbs sampler which we would like to utilize in the following MCMC algorithm. The latent variable u_i is constructed to make the conditional distribution of the allocation parameter d convenient to achieve. We only describe the sampling procedure without derivation.

In out our algorithm, $H_0(\cdot)$ is specified to be an exponential distribution $\mathcal{E}xp(1)$. The gamma distribution $\mathcal{G}(a_\alpha, b_\alpha)$ is provided as the prior for the concentration parameter α. The dimension of Dirichlete priors is set at N at the beginning and update through the iterates of the algorithm. The algorithm of posterior computation is carried out via the following steps after specifying the initial values for the hyper parameters.

1. Sample each $z_{ij} \sim \mathcal{P}\Big\{\Lambda_{0j}(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\Big\}$ for $z_{ij} > 0$ if $\delta_{ij} = 1$
   and $\{z_{ijl}\}_{l=1}^k \sim \mathcal{M}\Big\{z_{ij}, (p_{ij1}, p_{ij2}, ..., p_{ijl})\Big\}$,
   where $\mathcal{M}$ denotes a multinomial distribution and $p_{ijl} = \gamma_{jl}I_l(c_i)/\sum_{l=1}^k \gamma_{jl}I_l(c_i)$
   for $l = 1, ...k$, $i = 1, 2, ...n$, $j = 1, 2$.

2. Sample $\gamma_{jl} \sim \mathcal{G}\Big[1 + \sum_{i=1}^n z_{ijl}, \lambda_j + \sum_{i=1}^n I_l(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\Big]$,
   Sample the $\lambda_j \sim \mathcal{G}(a_\lambda + k, b_\lambda + \sum_{l=1}^k \gamma_{jl})$.

3. Sample $\nu_q \sim H(\cdot)\prod_{d_i=q}\mathcal{G}(\eta_i|\nu_{d_i}, \nu_{d_i})$ for $q = 1, 2, ...N$.

4. Sample v_q from $Beta(1 + m_q, n + \alpha - \sum_{i=1}^q m_i)$, where $m_q = \sum_{i=1}^n 1(d_i = q)$
   Calculate $\pi_1 = v_1$, $\pi_q = (1 - v_1)(1 - v_2)...v_q$, for $q = 2, ...N$.

5. Sample u_i from uniform$(0, w_{d_i})$. Let $u^* = \min(u_1, ...u_n)$.

   If $\sum_{q=1}^{N} \pi_q > 1 - u^*$, $P(d_i = q) \propto \mathcal{G}(\eta_i|\nu_q, \nu_q)1(u_i < w_{d_i=q})$ $i = 1, 2, ...n$, otherwise, $N = N + 1$, $v_N \sim Beta(1, \alpha), \nu_N \sim G(\cdot), \pi_N = v_N \Pi_{q<N}(1 - v_q)$.

6. Sample α from $\mathcal{G}\Big\{a_\alpha + N, b_\alpha - \sum_{q=1}^{N} \log(1 - v_q)\Big\}$.

7. Sample η_i from

$$\mathcal{G}\Big[\nu_{d_i} + \sum_{l=1}^{k}\sum_{j=1}^{2}\Big\{z_{ijl}1(I_l(c_i) > 0)\Big\}, \nu_{d_i} + \sum_{l=1}^{k}\sum_{j=1}^{2}\Big\{\gamma_{jl}I_l(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\Big\}\Big]$$

   .

8. Sample each component of $\boldsymbol{\beta}_j$ by ARMS.

3.4 Simulation

Based on the proposed approaches, simulation studies are conducted to evaluate the performance. The exact failure time T_j is generated from the true model

$$F_j(t|x_1, x_2, \eta) = 1 - \exp\{-\Lambda_{0j}(t)\exp(\beta_1^{(j)}x_1 + \beta_2^{(j)}x_2)\eta\}, \quad j = 1, 2,$$

where x_1 is a Bernoulli(0.5) random variable and x_2 is a normal variable with mean 0 variance 0.5^2, and the regression parameters $\beta_1^{(j)}$s take 1 or 0 and $\beta_2^{(j)}$s take 0 or -1. The true cumulative baseline function is $\Lambda_{0j}(t) = \log(1 + t + t^2) + t^{3/2}$. The censored time C is generated from truncated exponential distribution $\mathcal{E}(1)$ with support on $[0, 10]$, since the true survival probability is too small when failure time is greater than 10. We generate the censoring indicator δ_j by sampling from a Bernoulli distribution with probability of success $F(C|x_1, x_2, \eta)$ for each j. We consider two scenarios for the true frailty distribution: one is $\mathcal{G}(1, 1)$; the other is mixture log-normal distribution:

$$f(\eta) = 0.25\mathcal{LN}(-1, 2) + 0.50\mathcal{LN}(-1, 0.61) + 0.25\mathcal{LN}(0.5, 0.39),$$

where $\mathcal{LN}(\mu, \sigma^2)$ denotes the log-normal distribution with location parameter μ and scale parameter σ. For each parameter configuration, we generated 100 data sets with each containing 300 observations.

In the specifying of monotone splines, we chose degree 3 and took 15 equally spaced knots spanning in the range of observation times. To implement the computation, We specified $a_\lambda = b_\lambda = 1$ give a diffuse gamma prior for the hyper-parameter λ. In the setting of the exact gibbs sampler of DPGM, besides the above specification, the start value for N is 10 and $a_\alpha = b_\alpha = 1$. For each data set, we executed 6000 iterations and burn in the first 1000 ones.

Table 3.1 summarizes the results based on the first scenario in which the true frailty distribution is $\mathcal{G}(1, 1)$. As we can observe, we apply two approaches and both of them provide good estimation. The biases are small, ESDs are close to SSDs, and the CP95s are close to the nominal value 0.95. Table 3.2 presents the results under the second scenario. The small biases, the agreement between ESDs and SSDs, CP95s close to 0.95 prove that two approaches have good performance. It suggests that the Gamma frailty PH model is robust to misspecification of the frailty distribution. Moreover, because Gamma-frailty PH model carries the capability to quantify the association of the events, we prefer the gamma frailty model to Dirichlet process mixture gamma model.

3.5 Prevalence of antibodies to hepatitis B and HIV

In this analysis, we apply the proposed methods with the same real data as chapter 2 but consider two risk factors, whether a participant had drug injection before (x_1 taking 1 for yes and 0 for no), whether a participant had been treated for STIs before (x_2 taking 1 for yes and 0 for no). The censoring time is taken to be the age at the survey. There are 1097 prisoners who had complete information for this survey for this analysis. For the monotone spline specification, we take 3 for the degree and try different numbers of equally spaced knots within the data range. Table 3.3 presents the estimation results when using different numbers of knots together with the LPML (Log pseudo marginal likelihood) which is obtained from the summation

of log conditional predictive ordinates (Gelsser and Eddy, 1979; Gelfand and Dey, 1994). The gamma frailty model gives relatively robust estimation results when using different number knots. We choose the results using 4 interior knots since it has the largest PML. Having drug injection history and being treated for STIs both increase the risk of hepatitis B and HIV. These conclusions are more clear from Figure 3.1 and Figure 3.2, which plot the estimated marginal cumulative incidences $(1 - S(t|\boldsymbol{x}))$ of hepatitis B and HIV respectively for different subgroups with $\boldsymbol{x} = (0, 0)$, $(1, 0)$, $(0, 1)$, and $(0, 0)$. We obtain $\hat{\tau} = 0.52$ with 90% confidence interval $(0.28, 0.68)$. This suggests that there is medium to strong correlation between hepatitis B and HIV. Ignoring this correlation typically will lose efficiency.

Further more, we apply the DPGM to the data with the optimal number of knots in Gamma-frailty PH model. Table 3.4 provides a close result to Table 3.3. Figure 3.3 and 3.4 show us the good mixing for the regression parameter estimation.

3.6 Discussion

In this chapter, we have developed an easy-to-implement Bayesian approach to analyze the bivariate current status data. We provide a flexible and convenient framework based on Gamma frailty model to estimate the regression parameters as well as the correlation component. The monotone spline approximation of the baseline and the data augmentation greatly improve the computation efficiency. We adopt a nonparametric approach (DPGM)to relax the frailty distribution assumption. Our MCMC algorithms work well and have good mixing property. Based on the results in simulation and real data, the gamma frailty model is robust to the missspecification of true frailty distribution. The ability of summarizing the correlation encourages us to fit the proposed model in the application study.

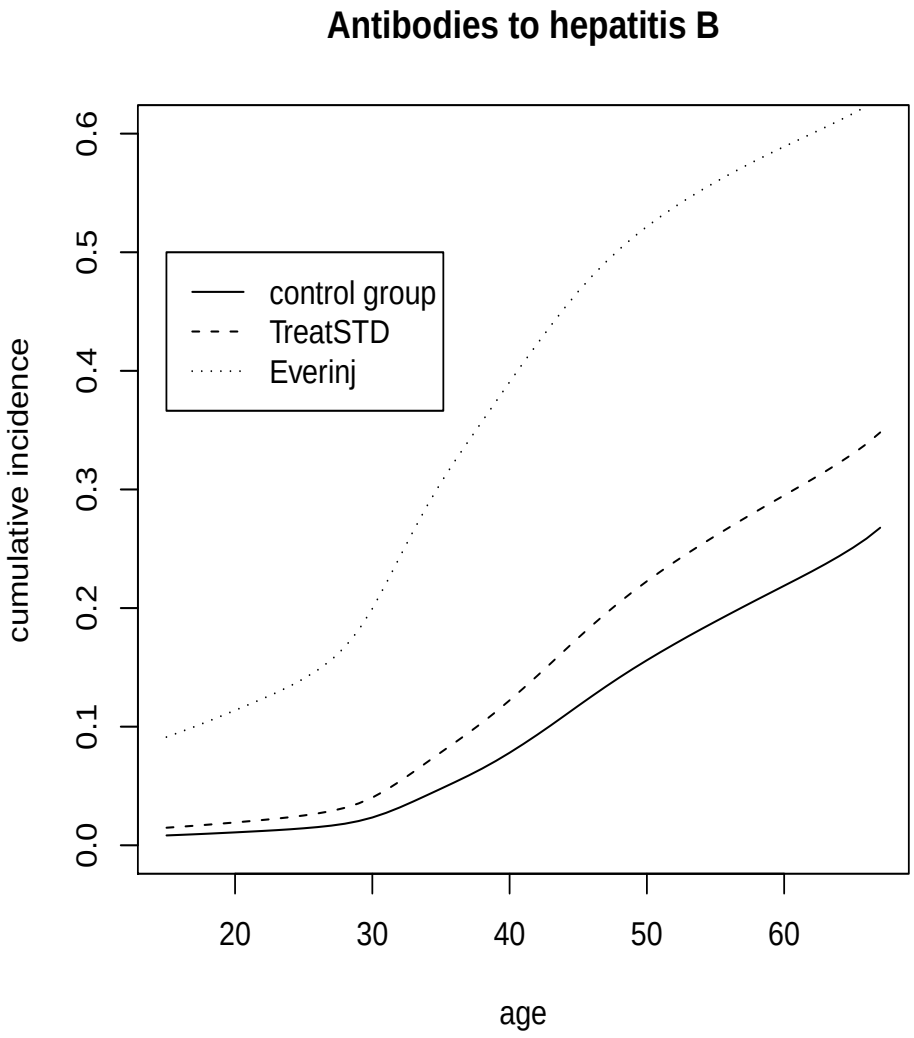

Figure 3.1 The estimated marginal cumulative incidence functions of hepatitis B for different subgroups. Here $\boldsymbol{x} = (x_1, x_2)$, where x_1 denotes whether a person had drug infection history, x_2 denotes whether a person had been treated for STIs

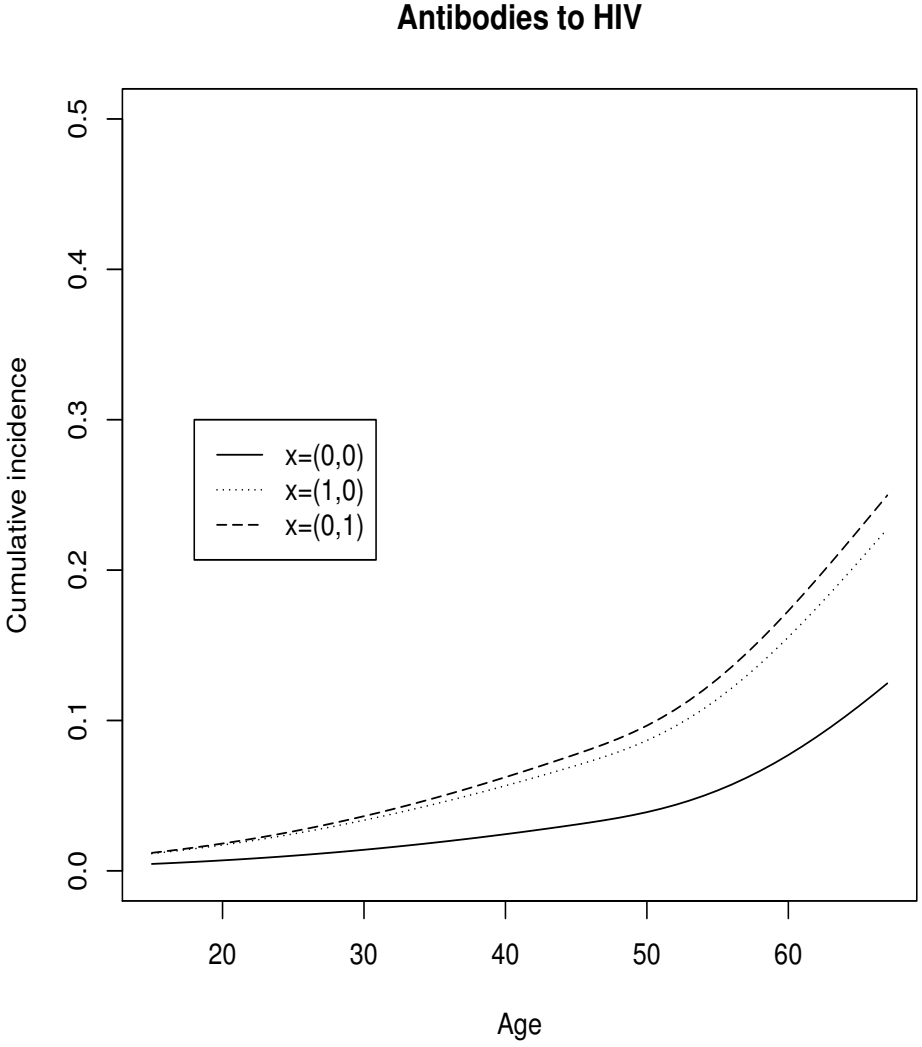

Figure 3.2 The estimated marginal cumulative incidence functions of HIV for different subgroups. Here $\boldsymbol{x} = (x_1, x_2)$, where x_1 denotes whether a person had drug infection history, x_2 denotes whether a person had been treated for STIs.

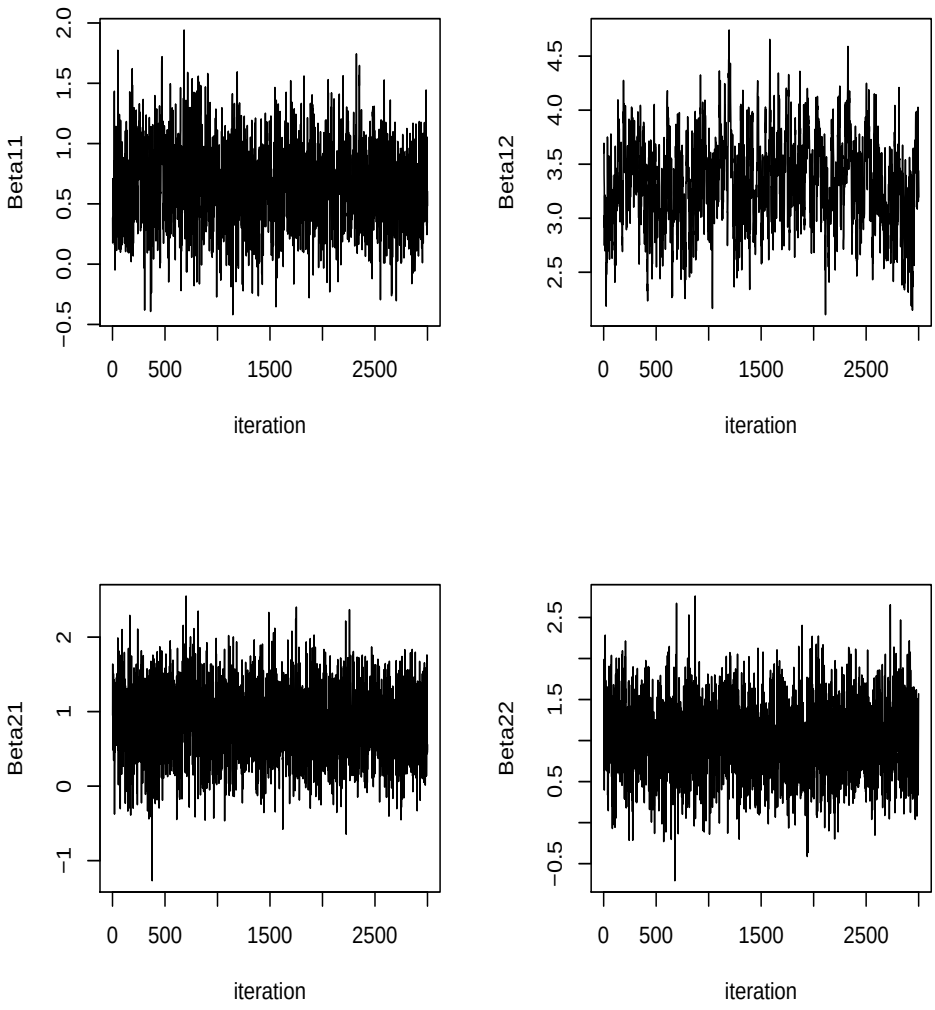

Figure 3.3 The traceplots of the iteration number against the values of the draw of the regression parameters at each iteration based on Gamma-frailty PH model with 4 equal spaced knots.

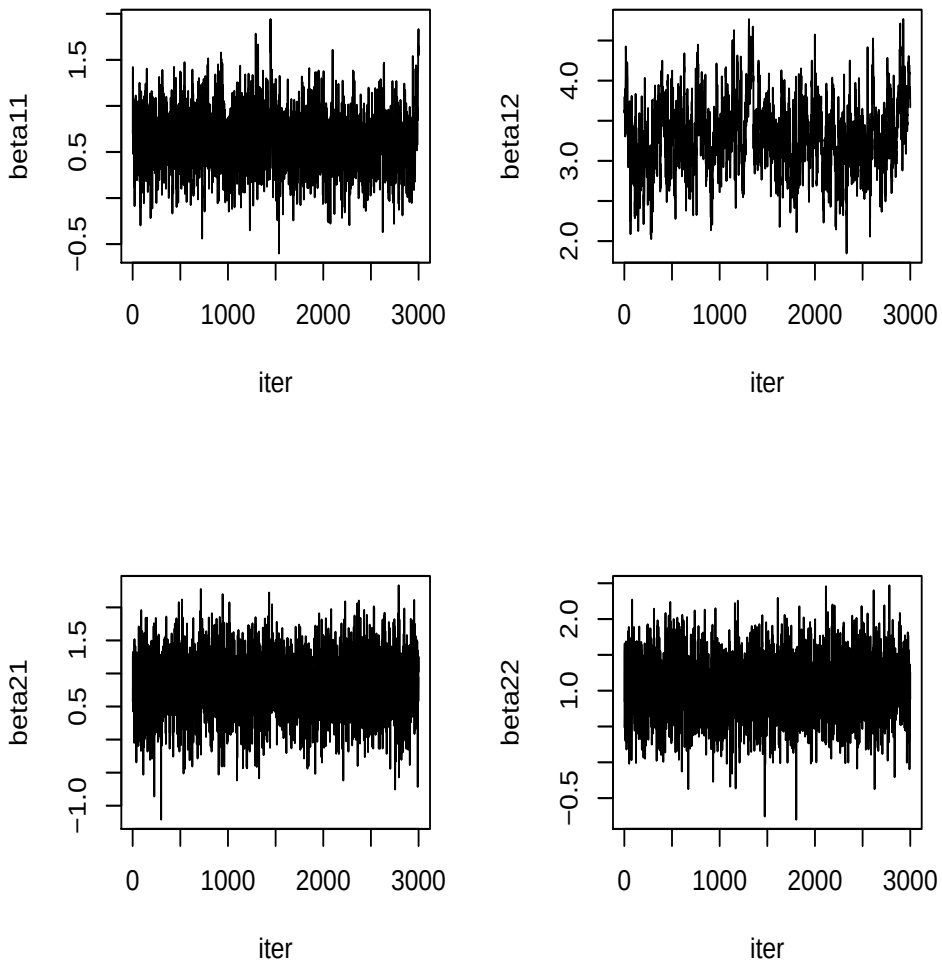

Figure 3.4 The traceplots of the iteration number against the values of the draw of the regression parameters at each iteration based on DPGM with 4 equal spaced knots.

Table 3.1 Simulation results for the regression parameters and the variance parameter of the gamma frailty when the true frailty distribution is $\mathcal{G}(1,1)$. BIAS denotes the average of the point estimates minus the true value, ESD is the average of estimated standard deviations, SSD is the sample standard deviation of the point estimates, and CP95 is the 95% coverage probability.

		Gamma frailty model					DPGM				
$\beta_1^{(j)}$	$\beta_2^{(j)}$	Est	Bias	SSD	ESD	CP95	Est	Bias	SSD	ESD	CP95
1	0	$\hat{\beta}_1^{(1)}$	-0.0389	0.4265	0.3631	0.92	$\hat{\beta}_1^{(1)}$	-0.0036	0.3724	0.3707	0.97
		$\hat{\beta}_2^{(1)}$	-0.0530	0.3748	0.3572	0.92	$\hat{\beta}_2^{(1)}$	-0.0279	0.4027	0.3636	0.96
		$\hat{\beta}_1^{(2)}$	-0.0374	0.4146	0.3647	0.92	$\hat{\beta}_1^{(2)}$	0.0341	0.4192	0.3712	0.89
		$\hat{\beta}_2^{(2)}$	-0.0273	0.3866	0.3568	0.94	$\hat{\beta}_2^{(2)}$	0.0144	0.3635	0.3634	0.95
		$\hat{\nu}$	0.0986	0.3516	0.3314	0.9	-	-	-	-	-
1	-1	$\hat{\beta}_1^{(1)}$	-0.0493	0.3524	0.3606	0.93	$\hat{\beta}_1^{(1)}$	-0.0351	0.3745	0.3819	0.94
		$\hat{\beta}_2^{(1)}$	-0.014	0.3928	0.3765	0.93	$\hat{\beta}_2^{(1)}$	-0.0899	0.4299	0.4006	0.94
		$\hat{\beta}_1^{(2)}$	-0.0312	0.3633	0.3691	0.93	$\hat{\beta}_1^{(2)}$	-0.0026	0.3294	0.3855	0.99
		$\hat{\beta}_2^{(2)}$	-0.0383	0.4150	0.3786	0.91	$\hat{\beta}_2^{(2)}$	-0.0778	0.4013	0.3974	0.94
		$\hat{\nu}$	0.0461	0.3135	0.3055	0.95	-	-	-	-	-
0	0	$\hat{\beta}_1^{(1)}$	-0.0301	0.3333	0.3325	0.94	$\hat{\beta}_1^{(1)}$	0.0012	0.3189	0.3374	0.98
		$\hat{\beta}_2^{(1)}$	-0.0123	0.3398	0.3364	0.95	$\hat{\beta}_2^{(1)}$	0.0279	0.2752	0.3491	1
		$\hat{\beta}_1^{(2)}$	-0.0522	0.2877	0.3332	0.99	$\hat{\beta}_1^{(2)}$	-0.0385	0.3382	0.3420	0.94
		$\hat{\beta}_2^{(2)}$	-0.0470	0.3032	0.3355	0.96	$\hat{\beta}_2^{(2)}$	0.0034	0.3218	0.3516	0.96
		$\hat{\nu}$	0.0536	0.3422	0.3458	0.96	-	-	-	-	-
0	-1	$\hat{\beta}_1^{(1)}$	-0.0087	0.3107	0.3378	0.95	$\hat{\beta}_1^{(1)}$	-0.0525	0.3583	0.3448	0.93
		$\hat{\beta}_2^{(1)}$	0.0276	0.3802	0.3635	0.97	$\hat{\beta}_2^{(1)}$	-0.0488	0.4060	0.3781	0.91
		$\hat{\beta}_1^{(2)}$	-0.0091	0.3136	0.3366	0.95	$\hat{\beta}_1^{(2)}$	-0.0871	0.3519	0.3425	0.94
		$\hat{\beta}_2^{(2)}$	0.0338	0.3325	0.3603	0.97	$\hat{\beta}_2^{(2)}$	-0.0397	0.3880	0.3768	0.93
		$\hat{\nu}$	0.0630	0.3919	0.3406	0.91	-	-	-	-	-

Table 3.2 Simulation results for the regression parameters and the variance parameter of the gamma frailty when the true frailty distribution mixture of log-Normal distributions with three components. BIAS denotes the average of the point estimates minus the true value, ESD is the average of estimated standard errors, SSD is the sample standard deviation of the point estimates, and CP95 is the 95% empirical coverage probability.

		Gamma frailty model					DPGM				
β_1	β_2	Est	Bias	SSD	ESD	CP95	Est	Bias	SSD	ESD	CP95
1	0	$\hat{\beta}_1^{(1)}$	-0.0054	0.2787	0.2771	0.96	$\hat{\beta}_1^{(1)}$	-0.0146	0.2655	0.2799	0.96
		$\hat{\beta}_2^{(1)}$	0.0095	0.2530	0.2705	0.96	$\hat{\beta}_2^{(1)}$	0.0086	0.2564	0.2700	0.95
		$\hat{\beta}_1^{(2)}$	-0.0244	0.2893	0.2779	0.95	$\hat{\beta}_1^{(2)}$	-0.0347	0.2796	0.2787	0.93
		$\hat{\beta}_2^{(2)}$	0.0161	0.2609	0.2723	0.97	$\hat{\beta}_2^{(2)}$	0.0104	0.2595	0.2715	0.97
1	-1	$\hat{\beta}_1^{(1)}$	-0.0337	0.2722	0.2813	0.96	$\hat{\beta}_1^{(1)}$	0.0101	0.2875	0.2865	0.94
		$\hat{\beta}_2^{(1)}$	-0.0277	0.2581	0.2904	0.97	$\hat{\beta}_2^{(1)}$	-0.0140	0.2987	0.2938	0.95
		$\hat{\beta}_1^{(2)}$	-0.0599	0.2878	0.2836	0.94	$\hat{\beta}_1^{(2)}$	-0.0289	0.2702	0.2858	0.97
		$\hat{\beta}_2^{(2)}$	-0.0347	0.2739	0.2900	0.94	$\hat{\beta}_2^{(2)}$	0.0059	0.2911	0.2936	0.96
0	0	$\hat{\beta}_1^{(1)}$	-0.0022	0.2829	0.2674	0.95	$\hat{\beta}_1^{(1)}$	-0.0273	0.2686	0.2637	0.95
		$\hat{\beta}_2^{(1)}$	-0.0075	0.2839	0.2667	0.91	$\hat{\beta}_2^{(1)}$	-0.0466	0.2177	0.2233	0.94
		$\hat{\beta}_1^{(2)}$	0.0353	0.2583	0.2629	0.95	$\hat{\beta}_1^{(2)}$	-0.0327	0.2418	0.2678	0.99
		$\hat{\beta}_2^{(2)}$	0.0317	0.2831	0.2692	0.9	$\hat{\beta}_2^{(2)}$	-0.0125	0.2680	0.2663	0.95
0	-1	$\hat{\beta}_1^{(1)}$	-0.0182	0.2670	0.2606	0.92	$\hat{\beta}_1^{(1)}$	0.0055	0.2496	0.2615	0.97
		$\hat{\beta}_2^{(1)}$	-0.0013	0.2899	0.2857	0.94	$\hat{\beta}_2^{(1)}$	0.0160	0.2856	0.2902	0.96
		$\hat{\beta}_1^{(2)}$	-0.0419	0.2795	0.2659	0.93	$\hat{\beta}_1^{(2)}$	-0.0107	0.2613	0.2652	0.96
		$\hat{\beta}_1^{(2)}$	-0.0302	0.3422	0.2875	0.94	$\hat{\beta}_2^{(2)}$	-0.0028	0.3292	0.2892	0.93

Table 3.3 Point estimates and 90% confidence interval of the covariate effects on hepatitis and HIV based on Gamma frailty model when using different numbers of knots. The two covariates in order are whether having drug injection, whether being treated for sexually transmitted infections. Column 6 shows the estimation of gamma variance parameter ν, column 7 is the estimation of Kendall's τ and its 90% confidence interval, and columns 8 provides the LPML (Log pseudo marginal likelihood) for different knots.

	Hepatitis B		HIV				
knots	$\hat{\beta}_1^{(1)}$	$\hat{\beta}_2^{(1)}$	$\hat{\beta}_1^{(2)}$	$\hat{\beta}_2^{(2)}$	$\hat{\nu}$	$\hat{\tau}$	LPML
3	0.62	3.04	0.87	1.08	0.72	0.45	-357.86
	(0.02,1.27)	(2.30,3.99)	(-0.08,1.76)	(0.23,1.96)	(0.24,1.51)	(0.22,0.71)	
4	0.66	3.33	0.86	1.02	0.53	0.52	**−355.63**
	(0.03,1.33)	(2.55,4.10)	(-0.08,1.75)	(0.17,1.90)	(0.24,1.29)	(0.24,0.70)	
5	0.60	3.03	0.80	0.98	0.77	0.45	-357.42
	(-0.03,1.24)	(2.31,3.85)	(-0.13,1.70)	(0.13,1.85)	(0.28,1.98)	(0.15,0.67)	
7	0.62	3.09	0.75	0.86	0.55	0.51	-356.08
	(0.01,1.28)	(2.32,3.87)	(-0.19,1.66)	(0.03,1.68)	(0.22,1.17)	(0.25,0.71)	
10	0.60	2.97	0.69	0.69	0.52	0.53	-358.43
	(0.08,1.13)	(2.33,3.68)	(-0.12,1.42)	(0.02,1.37)	(0.22,1.10)	(0.31,0.69)	
12	0.58	2.86	0.65	0.61	0.48	0.55	-362.35
	(0.04,1.16)	(2.22,3.57)	(-0.12,1.34)	(-0.05,1.29)	(0.20,1.07)	(0.32,0.71)	
15	0.57	2.74	0.57	0.45	0.40	0.58	-362.97
	(-0.10,1.27)	(2.08,3.47)	(-0.36,1.45)	(-0.37,1.22)	(0.16,0.81)	(0.34,0.77)	

Table 3.4 Point estimates and 90% confidence interval of the covariate effects on hepatitis and HIV based on DPGM when using different numbers of knots. The two covariates in order are whether having drug injection, whether being treated for sexually transmitted infections.

	Hepatitis B		HIV	
knots	$\hat{\beta}_1^{(1)}$	$\hat{\beta}_2^{(1)}$	$\hat{\beta}_1^{(2)}$	$\hat{\beta}_2^{(2)}$
3	0.60	2.98	0.88	1.08
	(0.12,1.10)	(2.36,3.68)	(0.10,1.62)	(0.38,1.80)
4	0.63	3.32	0.85	1.02
	(0.13,1.17)	(2.6,4.08)	(0.06,1.58)	(0.33,1.75)
5	0.60	3.10	0.82	1.01
	(0.08,1.13)	(2.5,3.77)	(0.05,1.55)	(0.26,1.78)
7	0.61	3.10	0.75	0.88
	(0.07,1.15)	(2.44,3.77)	(-0.05,1.49)	(0.21,1.58)
10	0.59	2.90	0.67	0.71
	(0.06,1.11)	(2.30,3.50)	(-0.12,1.44)	(0.02,1.42)
12	0.57	2.85	0.63	0.64
	(0.03,1.12)	(2.25,3.54)	(-0.13,1.38)	(-0.04,1.3)
15	0.56	2.72	0.55	0.49
	(0,1.12)	(2.12,3.35)	(-0.23,1.31)	(-0.15,1.14)

Chapter 4

Frailty PH models for bivariate general interval-censored data allowing weak dependence and independence

4.1 Introduction

When investigating chronic diseases, researchers often adopt periodic clinical examinations or laboratory tests on each patient to monitor the time to onset (or relapse) of the disease. Interval-censored data are generated if the exact failure time is not available to observe but known within some interval due to an observation of the changed status in the study. Repeated in the introduction chapter, interval censored data involve three censoring types: right-censoring, left-censoring and interval-censoring. Current status data or Case I interval-censored data is an important special case of interval-censored data which contains either left- or right-censored observations (Groeneboom and Wellner 1992). Case II interval-censored or general interval censored-data contain observations with all these censoring types and hence are more complicated.

An example of interval-censored data was described in Finkelstein and Wolfe (1985) on breast cancer patients. Multivariate interval censored data often arise when subjects experience multiple correlated events. For example, Goggins and Finkelstein (2000) gave details of the ACTG 181 study, which was designed to investigate the opportunistic infection cytomegalovirus referred to as shedding of HIV in the blood

and urine samples of patients.

In literature, types of methods were contributed to analyze the interval-censored data. Without considering covariates, Turnbull (1976) proposed an NPMLE estimator (Turnbull's estimator) of the survival function for interval-censored data. Gentleman and Geyer (1994) proved a necessary and sufficient condition (Kuhn-Tucker conditions) for a self-consistent estimator to be the NPMLE. Some other papers related to nonparametric analysis of interval-censored data include Andersen and Ronn (1995); Yu et al. (2000). In other regression analysis of interval-censored data, the primary interests are to estimate covariate effects and the baseline survival or hazard function. Available approaches for the regression analysis of interval-censored data include Finkelstein (1986), Huang and Rossini (1997), Betensky et al. (2002), Lin and Wang (2009)and McMahan et al. (2013) among many others.

For multivariate interval-censored data, it is of interest to quantify the correlation among those failure events in addition to estimating the survival functions and covariate effects. Typically, the analysis of correlated survival data is approached from either one of two perspective; i.e., from either the marginal likelihood or frailty model approach. In the introduction part, we present a brief review of the current methods for multivariate interval-censored data.

In this chapter, we first propose a Bayesian approach to analyze the multivariate interval-censored data under Gamma-frailty PH model. Monotone spline approximation to cumulative baseline hazard functions is adopted. An efficient Bayesian MCMC with data augmentation is developed to implement the model. Notice from the simulation study that if the two interested events are independent, the proposed method provides a large bias in the estimation of regression coefficients. We extend the Gamma-fraily PH model to a PH model with a mixture distributed frailty allowing to handle both independent and dependent cases. The remainder of this chapter organizes as follows: in section 4.2, we introduce Gamma frailty PH model

along with a Bayesian framework to implement it; In section 4.3, Gamma mixture frailty is present with an adjusted MCMC algorithm. An extensive simulation study is conducted to evaluate and compare the performance with two PH frailty models in section 4.4. In section 4.5, an application study is used to illustrate two methods.

4.2 Gamma frailty PH model

Data Structure, Likelihood

Suppose two interested failure times on n subjects awaiting for investigation. Let T_{ij} be the exact failure time for the jth event on the ith subject which is denoted by a $(p \times 1)$ vector $\boldsymbol{x}_i$. Assuming we have a sequence of J_i independent observation times $(c_1, c_2, ...c_{J_i})$ on each subject for both events where J_i is set to be any random number to avoid making the unrealistic assumption of requiring the same number of observational times for all subjects in the study. Owing to the interval-censored data structure, T_{ij} is not available to observe exactly but rather lies in some interval generated by some observation times: $T_{ij} \in (L_{ij}, R_{ij})$, where L_{ij} and R_{ij} are chosen from $\{0, c_1, c_2, ...c_{J_i}, \infty\}$. Three censoring types can appear in interval-censored data: right censoring, left censoring and interval censoring. We use three indicators δ_{ij1} δ_{ij2} and δ_{ij3} to identify the corresponding censoring type for T_{ij} and they are subject to $\delta_{ij1} + \delta_{ij2} + \delta_{ij3} = 1$.

By multiplying the shared frailty $\boldsymbol{\eta}$, the univariate Proportional Hazards model is extended to bivariate PH frailty model defining the hazard function given $\boldsymbol{x}_i$ and η_i as

$$\lambda_{ij}(t, \boldsymbol{x}, \eta) = \lambda_{0j}(t) \exp(\boldsymbol{x}_i' \boldsymbol{\beta}_j) \eta_i,$$

where $\lambda_{0j}(t)$ is the conditional cumulative baseline hazard for the jth event and $\boldsymbol{\beta}_j$ is the corresponding vector of regression parameters. This model suggests that the two interested failure times are conditional independent given $\boldsymbol{\eta}$. Let $F_{ij}(\cdot|\boldsymbol{x}_i, \eta_i)$ be the

conditional cumulative distribution function for the jth event, thus, under the PH frailty model,

$$F_{ij}(t|\boldsymbol{x},\eta) = 1 - \exp(-\Lambda_{0j}(t)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i). \tag{4.1}$$

To derive the observed data likelihood, we use several common assumptions as in Chen et.al.(2009), Hens et al.(2009) that is the two failure time are subject to the univariate censoring among the observation times and non-informative censoring scheme is adopted for which the failure time is independent of censoring time given the covariates. Consequently, the observations consisting of n subjects can be represented succinctly as $\{(L_{ij}, R_{ij}, \boldsymbol{x}_i, \delta_{ij1}, \delta_{ij2}, \delta_{ij3}), i = 1, ..., n, j = 1, 2\}$. Under the specify assumptions above and PH frailty model, the observed likelihood function is

$$L_{obs} = \prod_{i=1}^{n} \int \Big[\prod_{j=1}^{2} F(R_{ij}|\boldsymbol{x}_i,\eta_i)^{\delta_{ij1}} \Big\{ F(R_{ij}|\boldsymbol{x}_i,\eta_i) - F(L_{ij}|\boldsymbol{x}_i,\eta_i) \Big\}^{\delta_{ij2}} \Big\{ 1 - F(L_{ij}|\boldsymbol{x}_i,\eta_i) \Big\}^{\delta_{ij3}} \Big] \pi(\eta_i|\nu) d\eta_i, \tag{4.2}$$

where L_{ij} and R_{ij} are the left- and right- ending point of the censoring interval and $\pi(\eta_i|\nu)$ is the density distribution of η_i associated with parameter ν. If η follows Gamma distribution $\mathcal{G}(\nu,\nu)$, (4.1) is specified as Gamma frailty PH model.

Monotone splines to approximate $\Lambda_{0j}(t)$

To conquer the challenge to approximate the proper cumulative baseline hazard function $\Lambda_{0j}(t)$ without losing the flexibility or computational burden, we still apply the monotone splines approximation (2.4) encouraged by the succeed in the previous chapters. Generally, we intend to use the monotone splines with degree 3 and equally placed knots in the range of the censoring times.

Data Augmentation

To implement a Bayesian sampling method, we are required to derive the posterior distribution for the unknown parameters, particularly, $\boldsymbol{\theta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \nu)$ when applying the spline representation. However, the complicated form of observed likelihood (4.2) makes it impossible to obtain the standard posterior distribution when plus a prior. In such a situation, Metropolis-Hastings steps are proposed to sample any nonstandard probability distribution with a chance of poor mixing of Markov chains happening and a great effort in computation.

Followed the novel augmenting scheme developed in Lin et al.(2013), we connect the Gamma PH frailty model and the nonhomogeneous Poisson process $\{N_j(t) : t \geq 0\}$ with intensity function $\lambda_{0j}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\eta$, where η follows the Gamma distribution. Assume T_j as the first arrival epoch in the process,

$$P(T_j > t) = P\{N_j(t) = 0\} = \exp\{-\Lambda_{0j}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\eta\},$$

from which, this is the survival function for T_j under the Gamma frailty PH model. Further more, for any two time points $t_1 < t_2$, if $t_1 > T_j$, then $N_j(t) > 0$; if $t_1 < T_j < t_2$, then $N_j(t_1) = 0$ and $N_j(t_2) > 0$ and if $t_2 < T_j$, then $N_j(t_2) = 0$. Therefore, by the property of the Poisson process we have

$$\begin{aligned}
P(T_j < t_1|\eta) &= 1 - \exp\{-\Lambda_{0j}(t_1)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\eta\};\\
P(t_1 < T_j < t_2|\eta) &= P\{N_j(t_1) = 0 \& N_j(t_2) > 0\}\\
&= P\{N_j(t_1) = 0\} - P\{N_j(t_2) = 0\}\\
&= \exp\{-\Lambda_{0j}(t_1)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\eta\} - \exp\{-\Lambda_{0j}(t_2)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\eta\}.\\
P(T_j > t_2|\eta) &= \exp\{-\Lambda_{0j}(t_2)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\eta\}
\end{aligned}$$

Moreover, we can produce z_j and w_j whereas $z_j \sim \mathcal{P}\{\Lambda_{0j}(t_1)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\eta\}$ and $w_j \sim \mathcal{P}[\{\Lambda_{0j}(t_2) - \Lambda_{0j}(t_1)\}\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\eta]$ to present the relationship between the PH model

and Poisson process. That is, we have

$$\begin{aligned} P(T_j < t_1) &= P(z_j > 0); \\ P(t_1 < T_j < t_2) &= P(z_j = 0 \ \& \ w_j > 0) \\ P(T_j > t_2) &= P(z_j = 0 \ \& \ w_j = 0). \end{aligned}$$

To start the first augmentation step, we let $t_{ij1} = R_{ij}1_{(\delta_{ij1}=1)} + L_{ij}1_{(\delta_{ij1}=0)}$ and $t_{ij2} = R_{ij}1_{(\delta_{ij2}=1)} + L_{ij}1_{(\delta_{ij3}=1)}$. We introduce latent Poisson variables as $z_{ij} \sim \mathcal{P}\{\Lambda_0(t_{ij1})\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\}$ and $w_{ij} \sim \mathcal{P}[\{\Lambda_0(t_{ij2}) - \Lambda_0(t_{ij1})\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i]$. Apparently, We have only one constraint $z_{ij} > 0$ in the case of left-censoring and in the case of interval-censoring, there are constraints $z_i = 0$ and $w_i > 0$. For right-censoring, there are constraints $z_i = 0$ and $w_i = 0$. As a result, the augmented likelihood function can be written as

$$L_{aug1} = \prod_{i=1}^{n}\left[\prod_{j=1}^{2}\mathcal{P}(z_{ij})\mathcal{P}(w_{ij})^{\delta_{ij2}+\delta_{ij3}}\{\delta_{ij1}1_{(z_{ij}>0)} + \delta_{ij2}1_{(z_{ij}=0,w_{ij}>0)} + \delta_{ij3}1_{(z_{ij}=0,w_{ij}=0)}\}\right]\pi(\eta_i).$$

It is apparent that integrating out z_{ij}s, w_{ij}s, and η_is will result in (4.2).

Further more, applying the spline formulation, we could decompose z_{ij}, w_{ij} into independent Poisson variables z_{ijl}s and w_{ijl}s by defining $z_{ij} = \sum_{l=1}^{k} z_{ijl}$ and $w_{ij} = \sum_{l=1}^{k} w_{ijl}$, where

$$z_{ijl} \sim \mathcal{P}\{\gamma_{jl}I_l(t_{ij1})\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\}\eta_i; \ w_{ijl} \sim \mathcal{P}[\gamma_{jl}\{I_l(t_{ij2}) - I_l(t_{ij1})\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i].$$

Hence, under the specification above, we arrive at the conditional likelihood function that we refer to as the complete data likelihood as following:

$$L_{aug2} = \prod_{i=1}^{n}\left\{\prod_{j=1}^{2}\prod_{l=1}^{k}\mathcal{P}(z_{ijl})\mathcal{P}(w_{ijl})^{\delta_{ij2}+\delta_{ij3}}\right\}\pi(\eta_i), \tag{4.3}$$

which is subject to the constraints $\sum_l z_{ijl} > 0$ if $\delta_{ij1} = 1$, $\sum_l z_{ijl} = 0$ and $\sum_l w_{ijl} > 0$ if $\delta_{ij2} = 1$, and $\sum_l z_{ijl} = 0$ and $\sum_l w_{ijl} = 0$ if $\delta_{ij3} = 1$. MCMC assisted posterior inferences will be provided in the next section.

MCMC algorithm

We assign independent exponential priors $Exp(\lambda_j)$ for the coefficients γ_{jl}s of monotone spline, and adopt a gamma distribution $\mathcal{G}(a_\lambda, b_\lambda)$ as a hyper prior for each λ_j. The use of shrinkage priors for the spline coefficients is to address the over-fitting problems when a large quantity of knots are placed. For regression parameters β_js we apply an independent normal distribution $N(\mu_j, \delta_j^2)$ by which is log-concave. The adaptive rejection Metropolis sampling (ARMS) (Gilks and Wild, 1992) is used when direct sampling is not feasible. After assigning the initial values for hyper parameters, our MCMC algorithm iterates with the following steps:

1. Sample z_{ij}s, z_{ijl}s, w_{ij}s, and w_{ijl}s. Let $z_{ij} = 0$ and $w_{ij} = 0$ for all i and j, and $z_{ijl} = 0$ and $w_{ijl} = 0$ for all l. If $\delta_{ij1} = 1$, then sample

$$\begin{aligned} z_{ij} &\sim \mathcal{P}\{\Lambda_{0j}(R_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\}1_{(z_{ij}>0)}, \\ (z_{ij1}, \cdots, z_{ijk}) &\sim \mathcal{M}(z_{ij}, \boldsymbol{p}_{ij}), \quad \boldsymbol{p}_{ij} \propto \{\gamma_{j1}I_1(R_i), \ldots, \gamma_{jk}I_k(R_i)\}, \end{aligned}$$

   where $\mathcal{M}$ denotes a multinomial distribution. If $\delta_{ji2} = 1$, then sample

$$\begin{aligned} w_{ij} &\sim \mathcal{P}[\{\Lambda_0(R_i) - \Lambda_0(L_i)\}\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i]1_{(w_{ij}>0)}, \\ (w_{ij1}, \cdots, w_{ijk}) &\sim \mathcal{M}(w_{ij}, \boldsymbol{q}_{ij}), \end{aligned}$$

   where

$$\boldsymbol{q}_{ij} \propto [\gamma_{j1}\{I_1(R_i) - I_1(L_i)\}, \ldots, \gamma_{jk}\{I_k(R_i) - I_k(L_i)\}]$$

   .

2. Sample γ_{jl} from Gamma distribution $\mathcal{G}(a_{\gamma_{jl}}, b_{\gamma_{jl}})$, where

$$\begin{aligned} a_{\gamma_{jl}} &= 1 + \sum_i \{z_{ijl}\delta_{ij1} + w_{ijl}\delta_{ij2}\}, \\ b_{\gamma_{jl}} &= \lambda_j + \sum_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\{(\delta_{ij1} + \delta_{ij2})I_l(R_i) + \delta_{ij3}I_l(L_i)\}. \end{aligned}$$

3. Sample λ_j from Gamma distribution $\mathcal{G}(a_{\lambda_j} + k, b_{\lambda_j} + \sum_l \gamma_{jl})$.

4. Sample η_i from a Gamma distribution $\mathcal{G}(a_{\eta_i}, b_{\eta_i})$, where

$$\begin{aligned} a_{\eta_i} &= \nu + \sum_j (z_{ij}\delta_{ij1} + w_{ij}\delta_{ij2}) \\ b_{\eta_i} &= \nu + \sum_j \{\Lambda_{0j}(R_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)(\delta_{ij1}+\delta_{ij2}) + \delta_{ij3}\Lambda_{0j}(L_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\}. \end{aligned}$$

5. Sample ν and $\boldsymbol{\beta}_j$ by ARMS. The full conditional distribution of ν and $\boldsymbol{\beta}_j$ are

$$\begin{aligned} \nu|\cdot &\propto \prod_{i=1}^{n} g(\eta_i;\nu,\nu)e^{-\nu}, \text{and} \\ \beta_{jl}|\cdot &\propto \exp\sum_i \Big[\boldsymbol{x}_i'\boldsymbol{\beta}_j(z_{ij}\delta_{ij1} + w_{ij}\delta_{ij2}) \\ &\quad - \exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\eta_i\{\Lambda_{0j}(R_i)(\delta_{ij1}+\delta_{ij2}) + \Lambda_{0j}(L_i)\delta_{ij3}\}\Big] f(\beta_{jl}) \end{aligned}$$

   respectively, where $f(\beta_{jl})$ is a normal prior distribution.

One may notice that our algorithm mostly contains direct sampling steps from standard posterior distributions except for regression parameters and Gamma variance parameter which can be updated quickly by ARMS. The MCMC algorithm is implement using R and the code is available upon request.

4.3 Mixture frailty PH model

To be more general, in the phenomenon that the two failure times are barely correlated or almost completely independently, the Gamma distribution is not appropriate to be chosen as the frailty distribution. Motivated by the simulation performance in the following section, we extended the Gamma-frailty PH model to the Gamma mixture frailty model a compromising to address the problem arisen by that the random effects have zero variance. We assume the following frailty density:

$$\eta \sim \pi\delta_1(\cdot) + (1-\pi)\mathcal{G}(\cdot;\nu,\nu).$$

The case $\eta = 1$ means there is no heterogeneity among the two events. The mixture-frailty PH model retains the good properties. The marginal and joint survival function

can be obtained in a closed form as

$$S_j(t|\boldsymbol{x}) = \pi \exp\Big\{-\Lambda_{0j}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\Big\} + (1-\pi)\Big\{1+\nu^{-1}\Lambda_{0j}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\Big\}^{-\nu};$$

$$S(t_1,t_2|\boldsymbol{x}) = \pi \exp\Big\{-\textstyle\sum_{j=1}^{2}\Lambda_{0j}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\Big\} + (1-\pi)\Big\{1+\sum_{j=1}^{2}\nu^{-1}\Lambda_{0j}(t)\exp(\boldsymbol{x}'\boldsymbol{\beta}_j)\Big\}^{-\nu}.$$

Even with the slight change of the frailty distribution, the Kendall's τ is available to quantify the correlation structure via $\tau = \dfrac{1-\pi}{1+2\nu}$ after the estimation of ν and π.

The MCMC algorithm is developed based on the one under Gamma frailty PH model. The sampling steps are all exact same except for sampling η. We adjust the step and add one more step to sample π.

1. sample η_i from a mixture distribution

$$\mathcal{C}\delta_1(\cdot) + (1-\mathcal{C})\mathcal{G}(\cdot;\tilde{a}_{\eta_i},\tilde{b}_{\eta_i}),$$

where

$$\mathcal{C} = \left\{1 + \frac{(1-\pi)\nu^{\nu}\,\Gamma(\tilde{a}_{\eta_i})}{\pi\exp(\nu-\tilde{b}_{\eta_i})\;\tilde{b}_{\eta_i}^{\tilde{a}_{\eta_i}}\;\Gamma(\nu)}\right\}^{-1}$$

and

$$\begin{aligned}\tilde{a}_{\eta_i} &= \nu + \sum_j (z_{ij}+w_{ij})\\ \tilde{b}_{\eta_i} &= \nu + \sum_j \{\Lambda_{0j}(R_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)(\delta_{ij1}+\delta_{ij2}) + \delta_{ij3}\Lambda_{0j}(L_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)\}.\end{aligned}$$

2. Sample an additional parameter π from

$$\pi \sim Beta(1+\sum_i 1(\eta_i=1), n+1-\sum_i 1(\eta_i\neq 1)). \tag{4.4}$$

4.4 Simulation

We conduct a simulation study to evaluate the proposed approaches for two PH frailty models. We generate the paired failure times from $F(t_j|x_1,x_2,\eta) = 1 - \exp\{-\Lambda_{0j}(t)\exp(\beta_{j1}x_1+\beta_{j2}x_2)\eta\}$, where x_1 is a Bernoulli(0.5) random variable and x_2 is a $N(0,0.5^2)$ random variable. The true cumulative hazard is defined as $\Lambda_{0j}(t) =$

$\log(1+t+t^2)+t^{3/2}$ and the true value for β_{j1} is 1 or 0 and β_{j2} is -1, 0, or 1. Based on the assumption of univariate censoring that is the two events share the same random observation, we define the number of randomness as equal to 1 plus a Poisson random variable with mean 2. We generate the gaps between the adjacent observations through independent exponential distributions with mean 1 and the observation times are produced by cumulating the gaps. Accordingly the censoring indicator can be inferred by comparing the observations with the exact failure time.

To examine the performance of the MCMC algorithm on the two proposed PH frailty models, we consider two scenarios for frailty generation based on two distributions. One is Gamma distribution $\mathcal{G}(1,1)$ and the other is η_is are all 1. Hence, we obtain the dependent bivariate failure times and independent ones respectively.

For the monotone spline specifications, we choose degree 3 for adequate smoothness and select 15 equally spaced knots. We specify the normal priors $N(0,10^2)$ for $\boldsymbol{\beta}$s and take $a_\lambda = b_\lambda = 1$ for hyper parameters γ_{jl}s. We generate 100 data sets with each containing 300 observations for each scenario. For each data set, we execute 6000 iterations and burn in the first 1000 ones. Fast convergence and good mixing property are observed in all the simulations.

Table 4.1 shows the behaviour under dependence in which the frailty is generated from gamma distribution. Table 4.2 exhibits the results of estimation of $\boldsymbol{\beta}_j$s and the correlation measurement Kendall's τ under the independence case. Bias denotes the average of the 100 point estimates minus the true value, ESE is presenting the average of the estimated standard errors, SSD is the sample standard deviation of the 100 point estimates, and CP95 expresses 95% coverage probability.

From Table 4.1, one can observe the two models performed equivalently well in the dependent situation. From Table 4.2 We notice that the mixture frailty model outperform the gamma frailty model with smaller bias and smaller variance.

To formally test the dependence, Bayes factor is calculated during the simulation

for each set up. Consider the hypothesis test with H_0: failure times are dependent and H_1: failure times are independent, the Bayes factor B_{10} is obtained by calculating the ratio of posterior odds to the prior odds:

$$B_{10} = \frac{Pr(H_1|\mathbf{D})}{Pr(H_0|\mathbf{D})} \left\{ \frac{Pr(H_1)}{Pr(H_0)} \right\}^{-1}.$$

In the independence simulation scenario, the average Bayes factor value of 100 data sets is approximately 7.86 which sufficiently supports the independence.

Summarily, from two tables, it is concluded that the Bayesian approach for mixture frailty PH model works very well in all three conditions since the Biases are all close to zeros, ESEs are close to the corresponding SSDs and the CP95s are close to 0.95. The Gamma frailty PH model has a good performance for the correlated failure times.

4.5 AIDS example

We apply the proposed methods to the bivariate interval-censored data ACTG 181 which arose from the AIDS Clinical Trials Group protocol. The study was carried by following the patients with observing the instant status of test for CMV shedding in their blood and urine samples. As discussed in the introduction chapter before, the exact occurrence time can not be obtained and only known to fall in some observation interval. From the data set, we found that for time to blood shedding 15 patients were left censored, 15 were interval censored and 174 were right censored; for the urine shedding 69 were left censored, 47 were interval censored and 88 were right censored. The covariate is a binary vector to indicate the CD4 counts. The purpose of the study is to investigate the effect of CD4 on CMV shedding in blood and urine. Moreover, we wanted to explore the correlation between the blood shedding and urine shedding.

We implemented the data with the approaches developed above. Table 4.3 presents

the results by fitting with the gamma frailty model and with the mixture frailty model. We found that for both models, the factor effects for blood and urine are all significant and positive. To describe the degree of correlation, the estimation of Kendall's τ were provided and it is 0.5275 for gamma frailty model and 0.4615 for mixture model. We may conclude that the blood shedding and urine shedding are strongly correlated. Bayes factor is around 0.37 which provides no evidence against the dependence.

To make a model selection, we formally compared them from Bayesian perspective by adopting the log scale of pseudo marginal likelihood (LPML) given by the sum of log of CPO(conditional predictive ordinates), where CPO is defined as

$$CPO_i = \left[E\{ \frac{1}{P(T_{i1} \in \Delta_{i1}, T_{i2} \in \Delta_{i2} | \boldsymbol{\beta}, \boldsymbol{x}_i, \delta_{i1}, \delta_{i2})} \} \right]^{-1}, \tag{4.5}$$

where the Δ_{i1} and Δ_{i2} are two intervals containing the failure time t_i, and the expectation is taken with respect to the full posterior distribution of $\boldsymbol{\beta}$ given the observation data. It can be evaluated by taking the inverse of average $\{P(T_{i1} \in \Delta_{i1}, T_{i2} \in \Delta_{i2} | \boldsymbol{\beta}, \boldsymbol{x}_i, \delta_{i1}, \delta_{i2})\}^{-1}$ simultaneously with estimating the parameters in Gibbs sampler. The criterion of selecting model is to choose the larger log scale of PML as $\ln(PML) = \sum_i \ln(CPO_i)$. For gamma frailty model $\ln(PML)$ is -389.59 which is smaller than -388.63 for mixture frailty model. Based on the rule, mixture frailty model is preferable to gamma frailty. Figure 4.1 displayed the discrete values of $\ln \frac{CPO_{mixture}}{CPO_{gamma}}$ for each subject. There are about 56% of ratios are positive which support the mixture frailty model.

4.6 Conclusion

In this paper, we have developed a Bayesian method to analyze bivariate interval censored data under PH frailty model. Our method allows to estimate the regression parameters and baseline survival function coincidently. To explore the correlation between two failure times, we made assumptions about the frailty distribution: Gamma

distributed or Mixture distribution with point mass at 1. Our approach is able to estimate the correlation by evaluating the Kendall's τ under both frailty models. By examining the distribution form and simulation results, one can learn that the mixture frailty PH model performs better than gamma frailty PH model when the two follow-up events are independent. We applied the method to the shedding time data from AIDS study and presented the comparison of models by calculating the PML which supported the preference of mixture frailty model as well.

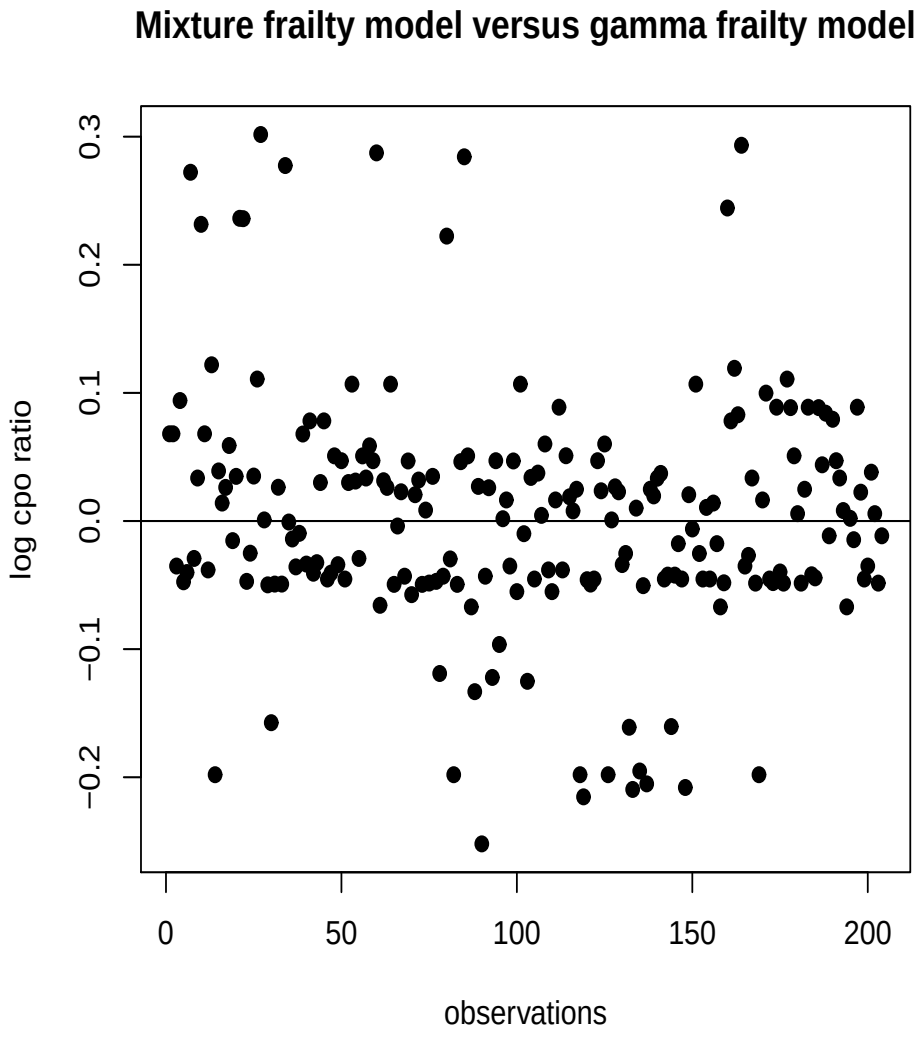

Figure 4.1 The discrete point ratio of CPO in log scale between the two frailty PH models

Table 4.1 Simulation results for the regression parameters and the variance parameter when the true frailty distribution is $\mathcal{G}(1,1)$. BIAS denotes the average of the point estimates minus the true value, ESD is the average of estimated standard errors, SSD is the sample standard deviation of the point estimates, and CP95 is the 95% empirical coverage probability.

		Gamma frailty PH model					Mixture frailty PH model				
β_1	β_2	Est	Bias	SSD	ESE	CP95	Est	Bias	SSD	ESE	CP95
1	1	$\hat{\beta}_1^{(1)}$	-0.0476	0.2643	0.2624	0.90	$\hat{\beta}_1^{(1)}$	-0.0803	0.2575	0.2559	0.94
		$\hat{\beta}_2^{(1)}$	-0.0989	0.2506	0.2597	0.93	$\hat{\beta}_2^{(1)}$	-0.1283	0.2439	0.2550	0.93
		$\hat{\beta}_1^{(2)}$	-0.0324	0.2799	0.2616	0.90	$\hat{\beta}_1^{(2)}$	-0.0642	0.2676	0.2537	0.92
		$\hat{\beta}_2^{(2)}$	-0.0066	0.2706	0.2561	0.94	$\hat{\beta}_2^{(2)}$	-0.0359	0.2704	0.2550	0.95
		$\hat{\nu}$	0.0388	0.1682	0.1833	0.96	$\hat{\nu}$	-0.0474	0.1489	0.1869	0.93
1	0	$\hat{\beta}_1^{(1)}$	0.0116	0.2636	0.2631	0.93	$\hat{\beta}_1^{(1)}$	-0.0237	0.2543	0.2595	0.93
		$\hat{\beta}_2^{(1)}$	-0.0238	0.2629	0.2618	0.99	$\hat{\beta}_2^{(1)}$	-0.0162	0.2567	0.2493	0.96
		$\hat{\beta}_1^{(2)}$	-0.0389	0.2128	0.2606	0.98	$\hat{\beta}_1^{(2)}$	-0.0495	0.2319	0.2554	0.95
		$\hat{\beta}_2^{(2)}$	0.0131	0.2529	0.2602	0.93	$\hat{\beta}_2^{(2)}$	0.0176	0.2430	0.2488	0.94
		$\hat{\nu}$	0.0345	0.1729	0.1841	0.94	$\hat{\nu}$	-0.0466	0.1666	0.1876	0.91
1	-1	$\hat{\beta}_1^{(1)}$	-0.0586	0.2649	0.2639	0.94	$\hat{\beta}_1^{(1)}$	-0.0774	0.2680	0.2574	0.90
		$\hat{\beta}_2^{(1)}$	-0.0138	0.2645	0.2642	0.96	$\hat{\beta}_2^{(1)}$	0.0067	0.2579	0.2598	0.94
		$\hat{\beta}_1^{(2)}$	-0.0399	0.2304	0.2602	0.98	$\hat{\beta}_1^{(2)}$	-0.0664	0.2237	0.2578	0.98
		$\hat{\beta}_2^{(2)}$	0.0159	0.2198	0.2667	0.96	$\hat{\beta}_2^{(2)}$	0.0455	0.2114	0.2610	0.97
		$\hat{\nu}$	0.0693	0.1772	0.1912	0.97	$\hat{\nu}$	-0.0408	0.1691	0.1922	0.96
0	1	$\hat{\beta}_1^{(1)}$	-0.0476	0.2643	0.2624	0.90	$\hat{\beta}_1^{(1)}$	-0.0472	0.2241	0.2310	0.96
		$\hat{\beta}_2^{(1)}$	-0.0989	0.2506	0.2597	0.93	$\hat{\beta}_2^{(1)}$	-0.0575	0.2972	0.2439	0.90
		$\hat{\beta}_1^{(2)}$	-0.0324	0.2799	0.2616	0.90	$\hat{\beta}_1^{(2)}$	-0.0390	0.2298	0.2273	0.93
		$\hat{\beta}_2^{(2)}$	-0.0066	0.2706	0.2561	0.94	$\hat{\beta}_2^{(2)}$	-0.0481	0.2718	0.2439	0.87
		$\hat{\nu}$	0.0388	0.1681	0.1833	0.96	$\hat{\nu}$	-0.0785	0.1455	0.1909	0.95
0	0	$\hat{\beta}_1^{(1)}$	-0.0576	0.2264	0.2405	0.95	$\hat{\beta}_1^{(1)}$	-0.0650	0.2261	0.2344	0.95
		$\hat{\beta}_2^{(1)}$	-0.0130	0.2441	0.2731	0.98	$\hat{\beta}_2^{(1)}$	-0.0137	0.2442	0.2675	0.98
		$\hat{\beta}_1^{(2)}$	-0.0466	0.2488	0.2420	0.95	$\hat{\beta}_1^{(2)}$	-0.0517	0.2349	0.2350	0.96
		$\hat{\beta}_2^{(2)}$	-0.0292	0.2375	0.2742	0.98	$\hat{\beta}_2^{(2)}$	-0.0297	0.2375	0.2665	0.97
		$\hat{\nu}$	0.0157	0.1544	0.1794	0.97	$\hat{\nu}$	-0.0832	0.1339	0.1844	0.96

Table 4.2 Simulation results for the regression parameters and the association measurement when the failure events are independent. BIAS denotes the average of the point estimates minus the true value, ESD is the average of estimated standard deviations, SSD is the sample standard deviation of the point estimates, and CP95 is the 95% empirical coverage probability.

		Gamma frailty PH model					Mixture frailty PH model				
β_1	β_2	Est	Bias	SSD	ESE	CP95	Est	Bias	SSD	ESE	CP95
1	1	$\hat{\beta}_1^{(1)}$	0.1527	0.2354	0.2449	0.91	$\hat{\beta}_1^{(1)}$	0.0632	0.2397	0.2236	0.93
		$\hat{\beta}_2^{(1)}$	0.1378	0.2436	0.2474	0.92	$\hat{\beta}_2^{(1)}$	0.0531	0.2371	0.2282	0.94
		$\hat{\beta}_1^{(2)}$	0.1170	0.2404	0.2312	0.93	$\hat{\beta}_1^{(2)}$	0.0325	0.2285	0.2229	0.94
		$\hat{\beta}_2^{(2)}$	0.1248	0.2721	0.2424	0.88	$\hat{\beta}_2^{(2)}$	0.0401	0.2657	0.2288	0.91
		$\hat{\tau}$	0.0989	0.0111	0.0201	-	$\hat{\tau}$	0.0234	0.0201	0.0143	-
1	0	$\hat{\beta}_1^{(1)}$	0.1812	0.2208	0.2401	0.90	$\hat{\beta}_1^{(1)}$	0.0762	0.2250	0.2223	0.95
		$\hat{\beta}_2^{(1)}$	-0.0109	0.2085	0.2186	0.96	$\hat{\beta}_2^{(1)}$	-0.0093	0.1890	0.1935	0.96
		$\hat{\beta}_1^{(2)}$	0.1411	0.2193	0.2411	0.93	$\hat{\beta}_1^{(2)}$	0.0373	0.2069	0.2240	0.96
		$\hat{\beta}_2^{(2)}$	-0.0334	0.2243	0.2177	0.95	$\hat{\beta}_2^{(2)}$	-0.0250	0.1947	0.1965	0.96
		$\hat{\tau}$	0.1019	0.01169	0.0264	-	$\hat{\tau}$	0.0237	0.0216	0.0130	-
1	-1	$\hat{\beta}_1^{(1)}$	0.1688	0.2222	0.2398	0.96 s	$\hat{\beta}_1^{(1)}$	0.0677	0.2175	0.2255	0.94
		$\hat{\beta}_2^{(1)}$	-0.1548	0.1950	0.2293	0.92s	$\hat{\beta}_2^{(1)}$	-0.0699	0.1846	0.2157	0.99
		$\hat{\beta}_1^{(2)}$	0.1180	0.2313	0.2347	0.91	$\hat{\beta}_1^{(2)}$	0.0302	0.2219	0.2221	0.96
		$\hat{\beta}_2^{(2)}$	-0.1418	0.2137	0.2274	0.94	$\hat{\beta}_2^{(2)}$	-0.0576	0.2153	0.2143	0.96
		$\hat{\tau}$	0.1002	0.0117	0.0256	-	$\hat{\tau}$	0.0253	0.0167	0.0215	-
0	1	$\hat{\beta}_1^{(1)}$	0.0294	0.1633	0.1935	0.97	$\hat{\beta}_1^{(1)}$	0.0222	0.1466	0.1737	0.97
		$\hat{\beta}_2^{(1)}$	0.1210	0.2152	0.2067	0.91	$\hat{\beta}_2^{(1)}$	0.0318	0.2011	0.1943	0.92
		$\hat{\beta}_1^{(2)}$	-0.0120	0.2058	0.1939	0.93	$\hat{\beta}_1^{(2)}$	-0.0146	0.1898	0.1737	0.91
		$\hat{\beta}_2^{(2)}$	0.1074	0.1862	0.2073	0.91	$\hat{\beta}_2^{(2)}$	0.0151	0.1692	0.1917	0.96
		$\hat{\tau}$	0.0954	0.0124	0.0238	-	$\hat{\tau}$	0.0225	0.0160	0.0194	-
0	0	$\hat{\beta}_1^{(1)}$	-0.0120	0.1806	0.1933	0.97	$\hat{\beta}_1^{(1)}$	-0.0159	0.1611	0.1730	0.97
		$\hat{\beta}_2^{(1)}$	0.0121	0.1957	0.1839	0.91	$\hat{\beta}_2^{(1)}$	0.0092	0.1785	0.1656	0.92
		$\hat{\beta}_1^{(2)}$	-0.0344	0.2016	0.1929	0.92	$\hat{\beta}_1^{(2)}$	-0.0332	0.1798	0.1726	0.93
		$\hat{\beta}_2^{(2)}$	0.1074	0.1862	0.2073	0.91	$\hat{\beta}_2^{(2)}$	-0.0113	0.1636	0.1640	0.97
		$\hat{\tau}$	0.0979	0.0096	0.0247	-	$\hat{\tau}$	0.0236	0.0135	0.0236	-

Table 4.3 Point estimates and 95% confidence interval of the covariate(the count of CD4) effect on the occurrences of CMV shedding in blood and urine and the association via Kendall's τ through the two frailty PH models

	Gamma frailty			Mixture frailty		
Parameter	Mean	SD	95%CI	Mean	SD	95%CI
β_b	0.9575	0.4350	(0.1296, 1.8407)	0.8019	0.4321	(0.0691, 1.6657)
β_u	1.4729	0.3814	(0.7808, 2.2664)	1.2250	0.3735	(0.5297, 1.9887)
τ	0.5275	0.0648	(0.3898, 0.6429)	0.4615	0.0709	(0.3228, 0.5935)

Bibliography

[1] Allwright, S., Bradley, F., Long, J., Barry, J., Thornton, L. and Parry, J. V. (2000). Prevalence of antibodies to hepatitis B, hepatitis C, and HIV and risk factors in Irish prisoners: results of a national cross sectional survey. *British Medical Journal* **321**, 78.

[2] Andersen, P. K., Klein, J. P., Knudsen, K. M. and Palacios, R. T. (1997). Estimation of variance in Cox's regression model with shared gamma frailties. *Biometrics* **53**, 1475-1484.

[3] Andersen, P. K. and Ronn, B. B. (1995). A nonparametric test for comparing two samples where all observations are either left- or right-censored. *Biometrics* **51**, 323-329.

[4] Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**, 1152-1174.

[5] Banerjee, T., Chen, M-H., Dey, D. K. and Kim, S. (2007). Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime Data Analysis* **13**, 241-260.

[6] Betensky, R. A., Lindsey, J. C., Ryan, L. M. and Wand, M. P. (2002). A local likelihood proportional hazards model for interval censored data. *Statistics in Medicine* **21**, 263-275.

[7] Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Polya Urn schemes. *Annals of Statistics* **1**, 353-355.

[8] Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**, 151-164.

[9] 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.

[10] Callegaro, A. and Lacobeli, S. (2012). The Cox shared frailty model with log-skew-normal frailties. *Statistical Modeling* **12**, 399-418.

[11] Chang, I-S., Wen, C-C., and Wu, Y-J. (2007). A profile likelihood theory for the correlated Gamma-frailty model with current statust family data. *Statistica Sinica* **17**, 1023-1046.

[12] Chen, C. M., Wei, J. C., Hsu, C. M., Lee, M. Y. (2014). Regression analysis of multivariate current status data with dependent censoring: application to ankylosing spondylitis data. *Statistics in Medicine* **33**, 772-785

[13] Chen, M-H., Tong, X. and Sun, J. (2009). A frailty model approach for regression analysis of multivariate current status data. *Statistics in Medicine* **28**, 3424-3436.

[14] Chen, M-H., Tong X. W. and Sun, J. (2007). The proportional odds model for multivariate interval-censored failure time data. *Statistics in Medicine* **26**, 5147-5161.

[15] Clayton, D. G. (1978). A model of association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141-151.

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

[17] Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman & Hall.

[18] Cui, S. and Sun, Y. (2004). Checking for the Gamma frailty distribution under the marginal proportional hazards frailty model. *Statistica Sinica* **14**, 249-267.

[19] Chen, D-G, Sun, J., Peace, K. E. (2012). *Interval-Censored Time-to-Event Data: Methods and Applications*. Chapman & Hall.

[20] Dunson, D. B. and Dinse, G. (2002). Bayesian models for multivariate current status data with informative censoring. *Biometrics* **58**, 79-88.

[21] Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577-588.

[22] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209-230.

[23] Ferguson, T.S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*. New York: Academic Press.

[24] Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 933-945.

[25] Finkelstein, D. M., Goggins, W. B. and Schoenfeld, D. A. (2002).Analysis of failure time data with dependent interval censoring. *Biometrics* **58**, 298-304.

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

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

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

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

[30] Goggins W. B., Finkelstein, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics* **56**, 940-943.

[31] Green, P. J. (1995). Reversible jump markov chain monte carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

[32] Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhauser: Basel.

[33] Guo, G. and Rodriguez, G. (1992). Estimating a multivariate proportional hazards model for clustered data using the EM algorithm, with an application to child survival in guatemala. *Journal of the American Statistical Association* **87**, 969-976.

[34] Guo, S. W. and Lin, D. Y. (1994). Regression analysis of multivariate grouped survival data. *Biometrics* **50**, 632-639.

[35] Hanson, T. (2006). Modeling censored lifetime data using a mixture of gammas baseline. *Bayesian Analysis* **1**, 575-594.

[36] Hoel, D. G. and Walberg, H. E. (1972). Statistical analysis of survival experiments. *Journal of National Cancer Institute* **49**, 361-372.

[37] Hens, N., Wienke, A., Aerts, M. and Molenberghs, G. (2009). The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. *Statistics in Medicine* **28**, 2785-2800.

[38] Hougaard, P. (2000). *Analysis of Multivariate Survival Data.* Springer: New York.

[39] Hsu, L. and Prentice, R. L. (1996). On assessing the strength of dependency between failure time variates. **1996**, 491-506.

[40] Huang, J. (1996). Effcient estimation for the Cox model with interval censoring. *Annals of Statistics* **24**, 540-568.

[41] Huang, J. and Rossini, A. J. (1997). Sieve estimation for the proportional odds failuretime regression model with interval censoring. *Journal of the American Statistical Association* **92**, 960-967.

[42] Huang, J. and Wellner, J. (1997). *Interval Censored Survival Data: A Review of Recent Progress.* Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis, Springer, New York, 123-170.

[43] Huang, X. and Wolf, R. A. (2002). A frailty model for informative censoring. *Biometrics* **58**, 510-520.

[44] Ibrahim, J. S., Chen, M-H. and Sinha, D. (2008). *Bayesian Survival Analysis.* Springer: New York.

[45] Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-Breaking priors. *Journal of the American Statistical Association* **96**, 161-173.

[46] Jewell, N. P. and Van der Laan, M. J. (2003). Current status and right-censored data structures when observing a marker at the censoring time. *Annals of Statistics* **31**, 512-535.

[47] Kim, M. Y. and Xue, X. N. (2002). The analysis of multivariate interval-censored survival data. *Statistics in Medicine* **21**, 3715-3726.

[48] Kim, Y. J., 2014. Regression analysis of bivariate current status data using a multistate model. *Communications in Statistics − Simulation and Computation* **43**, 462-475.

[49] Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the EM algorithm. *Biometrics* **48**, 795-806.

[50] Komarek, A. and Lessafre, E. (2007). Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution. *Statistica Sinica* **17**, 549-569.

[51] Lin, D. Y. (1994). Cox regression-analysis of multivariate failure time data - The marginal approach. *Statistics in Medicine* **13**, 2233-2247.

[52] 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.

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

[54] Lo, A. Y. (1984) On a class of Bayesian nonparametric estimates: I, density estimates. *Annals of Statistics* **12**, 351-357.

[55] McMahan, C., Wang, L., and Tebbs, J. (2013). Regression analysis of current status data using the EM algorithm. *Statistics in Medicine* **32**, 4452-4466.

[56] Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249-265.

[57] Oaks, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association* **84**, 487-493.

[58] Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective markov chain monte carlo methods for Dirichlet process hierarchical models. *Biometrika* **95**, 169-186.

[59] Papaspiliopoulos, O. (2008). *A note on posterior sampling from Dirichlet mixture models*. Technical Report.

[60] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25**, 855-900.

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

[62] Rasmussen, C. E. (2000). The infinite Gaussian mixture model. *in Advances in Neural Information Processing Systems 12 S.A. Solla, T.K. Leen and K.-R. Muller (eds.)* 554-560,

[63] Rondeau, V., Commenge, D. and Joly, P. (2003). Maximum penalized likelihood estimation in a Gamma-frailty model. *Lifetime Data Analysis* **9**, 139-153.

[64] Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics* **51**, 874-887.

[65] Scharfstein, D. O., Tsiatis, A. A. and Gilbert, P. B. (1998). Semiparametric efficient estimation in the generalized odds-Rate class of regression models for right-censored time-to-event data. *Lifetime Data Analysis* **4**, 355-391.

[66] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639-650.

[67] Shih, J. H. and Louis, T. A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51**, 1384-1399.

[68] Sun, L., Wang, L. and Sun, J. (2006) Estimation of the association for bivariate interval-censored failure time data. *Scandinavian Journal of Statistics* **33**, 637-649.

[69] Sun J. (2006). *The Statistical Analysis of Interval-Censored Data*. Springer: New York.

[70] Shen, X. T. (1998). Proportional odds regression and sieve maximum likelihood estimation. *Biometrika* **85**, 165-177.

[71] Tong X. W., Chen M-H. and Sun J. (2008). Regression analysis of multivariate interval-censored failure time data with application to tumorigenicity experiments. *Biometrical Journal* **50**, 364-374.

[72] Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistical Society, Series B* **28**, 290-295.

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

[74] Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation* **36**, 45-54.

[75] Wang, W. and Ding, A. A. (2000). On assessing the association for bivariate current status data. *Bimotrika* **87**, 879-893.

[76] Wang, W. and Wells, M. T. (2000). Estimation of Kendall's tau under censoring. *Statistica Sinica* **10**, 1199-1215.

[77] Wang, L., Sun, J., Tong, X. W. (2008). Efficient estimation for bivariate current status data. Lifetime Data Analysis, **14**, 134-153.

[78] Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**, 1065-1073.

[79] Wen, C-C, and Chen, Y-H. (2011). Nonparametric maximum likelihood analysis of clustered current status data with the gamma-frailty Cox model. *Computational Statistics and Data Analysis* **55**, 1053-1060.

[80] Wienke, A. (2012). *Frailty Models in Survival Analysis*. Chapman & Hall.

[81] Yin, G. and Ibrahim, J. G. (2005). A class of Bayesian shared Gamma frailty models with multivariate failure time data. *Biometrics* **61**, 208-216.

[82] Young, E., Albert, J., Satayathum, S., et al. (2005). Predictors and consequences of altered mineral metabolism: the Dialysis Outcomes and Practice Patterns Study. *Kidney Int.* **67**, 1179-1187.

[83] Yu, Q., Li, L. and Wong, G. (2000). On consistency of self-consistent estimator of survival functions with interval-censored data. *candinavian Journal of Statistics* **27**, 35-44.

[84] Zeng, D. and Lin, D. Y. (2007). Maximum likelihood estimation in semiparametric regression models with censored data. *Journal of the Royal Statistical Society, Series B* **69**, 507-564.

[85] Zuma, K. (2007). A Bayesian analysis of correlated interval-censored Data. *Communications in Statistics-Theory and Methods* **36**, 725-730.

Appendix A

The conditional expectations in Section 2.3

The conditional expectations in Section 2.3 are summarized as follows,

$$\begin{aligned}
E(\eta_i|\mathcal{D},\boldsymbol{\theta}) &= \frac{\delta_{i1}\delta_{i2}-\delta_{i1}\{S_2(c_i|\boldsymbol{x}_i)\}^{(1+\nu^{-1})}-\delta_{i2}\{S_1(c_i|\boldsymbol{x}_i)\}^{(1+\nu^{-1})}}{\delta_{i1}\delta_{i2}-\delta_{i1}S_2(c_i|\boldsymbol{x}_i)-\delta_{i2}S_1(c_i|\boldsymbol{x}_i)+S(c_i,c_i|\boldsymbol{x}_i)} \\
&\quad + \frac{\{S(c_i,c_i|\boldsymbol{x}_i)\}^{(1+\nu^{-1})}}{\delta_{i1}\delta_{i2}-\delta_{i1}S_2(c_i|\boldsymbol{x}_i)-\delta_{i2}S_1(c_i|\boldsymbol{x}_i)+S(c_i,c_i|\boldsymbol{x}_i)}, \\
E\{\log(\eta_i)|\mathcal{D},\boldsymbol{\theta}\} &= \frac{\Gamma'(\nu)}{\Gamma(\nu)}-\frac{B_i(\nu,\boldsymbol{\theta})}{\delta_{i1}\delta_{i2}-\delta_{i1}S_2(c_i|\boldsymbol{x}_i)-\delta_{i2}S_1(c_i|\boldsymbol{x}_i)+S(c_i,c_i|\boldsymbol{x}_i)}, \\
E(z_{ij}|\mathcal{D},\boldsymbol{\theta}) &= \frac{\delta_{ij'}[1-\{S_{j'}(c_i|\boldsymbol{x}_i)\}^{(1+\nu^{-1})}]-(1-\delta_{ij'})\{S_{j'}(c_i|\boldsymbol{x}_i)\}^{(1+\nu^{-1})}}{\delta_{ij'}-S_{j'}(c_i|\boldsymbol{x}_i)-\delta_{ij'}S_j(c_i|\boldsymbol{x}_i)+S(c_i,c_i|\boldsymbol{x}_i)}\rho_{ij}\delta_{ij}, \\
&\quad j'\neq j, \\
E(z_{ijl}|\mathcal{D},\boldsymbol{\theta}) &= \{\Lambda_{0j}(c_i)\}^{-1}\gamma_{jl}I_l(c_i)E(z_{ij}|\mathcal{D},\boldsymbol{\theta}),
\end{aligned}$$

where $B_i(\nu,\boldsymbol{\theta})=\delta_{i1}\delta_{i2}\log(\nu)-\delta_{i1}S_2(c_i|\boldsymbol{x}_i)\log(\nu+\rho_{i2})-\delta_{i2}S_1(c_i|\boldsymbol{x}_i)\log(\nu+\rho_{i1})+S(c_i,c_i|\boldsymbol{x}_i)\log(\nu+\rho_{i1}+\rho_{i2})$ and $\rho_{ij}=\Lambda_{0j}(c_i)\exp(\boldsymbol{x}_i'\boldsymbol{\beta}_j)$ for $j=1,2$ and $i=1,\ldots,n$.

The derivation of $E(\eta_i|\mathcal{D},\boldsymbol{\theta})$ and $E\{\log(\eta_i)|\mathcal{D},\boldsymbol{\theta}\}$ arises directly from the augmented likelihood (2.5). When deriving $E(z_{ij}|\mathcal{D},\boldsymbol{\theta})$, we first use the law of iterative expectation

$$E(z_{ij}|\mathcal{D},\boldsymbol{\theta})=E\{E(z_{ij}|\mathcal{D},\eta_i,\boldsymbol{\theta})\}=E\Big\{\frac{\rho_{ij}\delta_{ij}}{1-\exp(-\rho_{ij}\eta_i)}|\mathcal{D},\boldsymbol{\theta}\Big\}.$$

The last step uses the fact that the conditional distribution of z_{ij}, given η_i and the observed data, is a truncated Poisson with a support of all positive integers when $\delta_{ij}=1$ and is degenerated at 0 when $\delta_{ij}=0$. One can complete the conditional expectation in the above expression based on the augmented likelihood (2.5). The derivation of $E(z_{ijl}|\mathcal{D},\boldsymbol{\theta})$ is straightforward by using the law of iterative expectation and noting that the conditional distribution of $(z_{ij1},\ldots,z_{ijk})$, given z_{ij}, is a multinomial.