Summer 2019

Multivariate Probit Models for Interval-Censored Failure Time Data

Yifan Zhang

Multivariate Probit Models for Interval-Censored Failure Time Data

by

Yifan Zhang

Bachelor of Science
Nankai University, 2012

Master of Science
Rutgers University, the State University of New Jersey, 2014

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

2019

Accepted by:

Lianming Wang, Major Professor

Yen-Yi Ho, Committee Member

Xiaoyan Lin, Committee Member

Bo Cai, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

I consider myself fortunate indeed to have had the opportunity to pursue research work toward to the PhD with Dr. Lianming Wang. The statistical foundations of Dr. Wang's work in survival analysis are both powerful and elegant. His unwavering enthusiasm for statistics kept me constantly engaged with my research. There are so many times when I felt discouraged, it was him who never failed to offer advice and encouragement. Without his guidance and persistent help, this dissertation would not have been possible.

I also want to thank my committee members Dr. Yen-Yi Ho, Dr. Xiaoyan Lin and Dr. Bo Cai. Their mentoring and encouragement have been especially valuable. The brilliant comments and suggestions given by the committee made my defense an enjoyable moment in my life. They has provided me extensive personal and professional guidance and taught me a great deal about both scientific research and life in general.

Finally, I would like to give my special thanks to my parents, Zhenwu Zhang and Hong Zhang and my boyfriend, Tuan Do, for always believing in me and encouraging me to follow my dreams. My dearest family, I know they are always there whenever for whatever.

Abstract

Survival analysis is an important branch of statistics that analyzes the time to event data. The events of interest can be death, disease occurrence, the failure of a machine part, etc.. One important feature of this type of data is censoring: information on time to event is not observed exactly due to loss to follow-up or non-occurrence of interested event before the trial ends. Censored data are commonly observed in clinical trials and epidemiological studies, since monitoring a person's health over time after treatment is often required in medical or health studies. In this dissertation we focus on studying multivariate interval-censored data, a special type of survival data. By saying multivariate interval-censored data, we mean that there are multiple failure time events of interest, and these failure times are known only to lie within certain intervals instead of being observed exactly. These events of interest can be associated because of sharing some common characteristics. Multivariate interval-censored data draw more and more attention in epidemiological, social-behavioral and medical studies, in which subjects are examined multiple times and several events of interest are tested at the observation times.

There are some existing methods available in literatures for analyzing multivariate interval-censored failure time data. Various models were developed for regression analysis. However, due to the complicated correlation structure between events, analyzing such type of survival data is much more difficult and new efficient methodologies are needed.

Chapter 1 of this dissertation illustrates the important concepts of interval-censored data with several real data examples. A literature review of existing regression models

and approaches is included as well. Chapter 2 introduces a new normal-frailty multivariate probit model for regression analysis of interval-censored failure time data and proposes an efficient Bayesian approach to get parameter estimates. Simulations and an analysis on a real data set are conducted to evaluate and illustrate the performance of this new method. This new approach is proved efficient and has accurate estimations on both the regression parameters and the baseline survival function. Several appealing properties of the model are discussed here. Chapter 3 proposes a more general multivariate probit model for multivariate interval-censored data. This new model allows arbitrary correlation among the correlated survival times. A new Gibbs sampler is proposed for the joint estimation of the regression parameters, the baseline CDF, and the correlation parameters. Chapter 4 extends the normal frailty multivariate probit model to allow arbitrary pairwise correlations. Simulation studies are conducted to explore the underlying relationship between the normal frailty multivariate probit model and the general multivariate probit model.

Table of Contents

List of Tables

Chapter 1

Introduction

1.1 Data Structure

Survival analysis is a branch of statistics that specifically deal with data with outcome variable being the time until the occurrence of an event of interest. For example, if the event of interest is heart attack, then the survival time can be the time in years until a person experiences a heart attack. In survival analysis, subjects are usually followed up over a specified time period and the focus is on the time at which the event of interest occurs. However, it is often not observed directly. Interval censoring occurs when subjects in the study are examined only at discrete observational times and the status of event of interest is known at those observational times. Consequently, the failure time of interest is not observed exactly but is known only to take place within some time interval. Data of this type usually consist of left-, interval-, and right-censored observations. In many clinical trails, patients are scheduled several visits during the study period and they are examined whether a certain event of interests (e.g. failure) is occurred at the observation time. However, the event may occur before the first visit, which results in a left-censored observation. Or in between visits, or we only know that the true event time is greater than last observation time at which the event hasn't appeared yet and less than the first observation time at which the event status has changed, which contributes to interval-censored data. Or the event after the last visit, which is ended of being a right-censored observation. Unavailable exact observed time of event caused difficulty in analyzing interval-censored data.

1.1.1 Case I interval-Censoring and Case II interval-Censoring

If each subject is observed only once in a study, the failure time of interest is known only to fall before or after the observation time point. This type of data are referred to as current status data and only contain left-censored or right-censored observations. Current status data is a special case of general interval censored data.

This dissertation studies general type of interval-censored data, which contain left-, interval-, and right-censored observations. Below we provide three real life examples of general interval-censored data.

1.1.2 Interval-Censored data examples

Breast Cancer Study

A dataset on breast cancer was studied by Finkelstein and Wolfe (1985) from a study on breast cancer patients at the Joint Center for Radiation in Boston between 1976 and 1980. Two types of treatments were compared in this study: radiotherapy (RT) alone and radiation therapy plus adjuvant chemotherapy (RCT). Ninety-four breast cancer patients were involved, with 46 of them treated by radiotherapy only and the rest of them treated by the combined treatments of radiation and adjuvant chemotherapy. In this study, patients were scheduled to visit the clinicians every 4 to 6 months. However, actual visit times varied from patient to patient. Some of the patients missed the scheduled visits. During each visit, physicians recorded the cosmetic appearance such as breast retraction, a response which was highly correlated with a negative impact on overall cosmetic appearance. The failure time of interest was the time until breast retraction.

The data for the interval-censored event time of breast retraction are shown in Table 1.1. There are 38 patients with intervals without a right endpoint. These patients did not experience breast retraction during the study period and they represented right censored observations. For those with starting point at 0, such as $(0, 8]$, it means

Table 1.1: Event intervals for breast cancer patients treated by radiotherapy vs radiotherapy and chemotherapy

Therapy	**Event Intervals**								
RT	(45,]	(25, 37]	(37,]	(6,10]	(46,]	(0,5]	(0,7]	(26,40]	(18,]
	(46,]	(46,]	(24,]	(46,]	(27,34]	(36,]	(7,16]	(36,44]	(5,11]
	(17,]	(46,]	(19, 35]	(7, 14]	(36, 48]	(17, 25]	(37, 44]	(37,]	(24,]
	(0, 8]	(40,]	(32,]	(4, 11]	(17, 25]	(33,]	(15,]	(46,]	(19, 26]
	(11, 15]	(11, 18]	(37,]	(22,]	(38,]	(34,]	(46,]	(5, 12]	(36,]
	(46,]								
RCT	(8, 12]	(0, 5]	(30, 34]	(0, 22]	(5, 8]	(13,]	(24, 31]	(12, 20]	(10, 17]
	(17, 27]	(11,]	(8, 21]	(17, 23]	(33, 40]	(4, 9]	(24, 30]	(31,]	(11,]
	(16, 24]	(13, 39]	(14, 19]	(13,]	(19, 32]	(4, 8]	(11, 13]	(34,]	(34,]
	(16, 20]	(13,]	(30, 36]	(18, 25]	(16, 24]	(18, 24]	(17, 26]	(35,]	(16, 60]
	(32,]	(15, 22]	(35, 39]	(23,]	(11, 17]	(21,]	(44, 48]	(22, 32]	(11, 20]
	(14, 17]	(10, 35]	(48,]						

that the breast retraction happened before the first time examination and the failure time is left censored. For the rest of the data, taking the observation $(25, 37]$ as an example, it represents that the breast retraction did not appear at month 25 but is shown up by month 37, so the exact breast retraction time lies between 25 and 37 months.

Respiratory Symptoms in Aluminum Potroom Workers Study

A longitudinal study of respiratory symptoms among 1301 aluminum potroom workers was conducted in the Nordic countries between 1986 and 1989. The workers were scheduled for at least two health examinations and asked to report in questionnaires about respiratory symptoms. If the workers reported wheezing and dyspnea, then they were considered symptomatic. Investigators were interested in analyzing the time from employment to the development of asthmatic symptoms (wheezing and dyspnoea).

In this study, the workers leaving the potroom or ending the survey without respiratory symptoms have right-censored observations. The workers that developed asthmatic symptoms between two consecutive health examinations have interval-censored observations. More details can be found in Samuelsen and Kongerud (1994).

Hemophilia Data

The Hemophilia data were collected from 262 patients with Type A or B hemophilia at the Hospitals Kremlin Bicetres and Coeur des Yvelines in France between 1978 and 1988. Twenty-five of the persons were detected to be infected with HIV on their first lab test. By August 1988, 197 of the hemophiliacs had become infected and 43 of these showed clinical symptoms (AIDS, lymphadenopathy or leukopenia). All these individuals in this study were believed to have become infected from infusions of contaminated blood factor they received periodically to treat their hemophilia. Blood samples were periodically collected and stored to decide a time interval during which the infection occurred. The infection times are censored into the interval between the last negative and first positive lab result. More information about this data can be found in De Gruttola and Lagakos (1989).

1.2 Motivating Examples

Multivariate failure time data are commonly encountered in biomedical areas when one is interested in several failure time events. For example, the study subject may experience multiple events. This type of data can also arise when the failure times are clustered, such as in family studies. The key feature of this type of data is that the failure times are related to each other. These multivariate events can be interval-censored, the exact times of events are not known since the events could have happened any time during two adjunct visits. This dissertation will focus on analyzing multivariate interval-censored data. Two real life data examples are presented below to illustrate multivariate interval censored data.

1.2.1 AIDS Clinical Trail Data

One goal of the ACTG 181 study is to determine the natural history of the opportunistic infection cytomeglovirus (CMV) in an HIV-infected individual. CMV virus

(shedding of the virus), were tested during scheduled clinic visits in the blood and in the urine. The question of interest in this study is whether the stage of HIV disease at study entry contributed to an increased risk for CMV shedding in either the blood or the urine. Samples from urine and blood were collected every 4 weeks and 12 weeks respectively. Since the samples come from the same patient, the outcomes are correlated. The real sample collection time differed from patient to patient. Some patients were observed missed their visits and came back with changed CMV shedding status. The failure time was only known to be between the times specified by the last negative and the first positive assessment, yielding, hence, interval-censored observations. Left-censored shedding times resulted from those patients who were already shedding at the time of the study. Right-censored times occured where some patients had not yet started shedding by the end of the follow-up period. More details can be referred to Goggins and Finkelstein (2000).

1.2.2 Sexually Transmitted Infection (STI) data

STI data collected on young women as a part of the Young Women's Project (YWP) is analyzed as an illustration for the proposed model. The details for study design and folllow-up protocol were previously described, see Tu et al. (2009) and Tu et al. (2011). In this study, infections with Chlamydia trachomatis (CT), Neisseria gonorrhoeae (GC) and Trichomonas vaginalis (TV) are the three outcomes of interest. This analysis focuses on the time to first STI infection for each of these three types. Three hundred and eighty seven adolescent young women aged 14 to 17 years were observed between 1999 and 2007 in this observational study. At enrollment, participants were interviewed and asked to complete a detailed questionnaire about their sexual behaviors such as the number of sex partners, age of first sex, etc.. Patients were examined every three months and actual examination times differed from patient to patient since some of them missed their visits. As a result, the exact times

of infections were not directly observable since the infections could have happened at any time in the interval between the last visit with negative result and the first visit with a positive diagnosis. In this case, the time to infection was interval-censored. The failure times were right-censored at the last visit time if no infection was detected throughout the follow-up, or left-censored at the beginning of the study if a patient is detected positive at the time of her first testing.

1.3 Commonly Used Models

Let T denotes a non-negative continuous random variable, representing the survival time until the occurrence of an event. Its probability density function (p.d.f.) is denoted by $f(t)$ and cumulative distribution function (c.d.f.) is $F(t) = Pr\{T \leq t\}$. Then the survival function of T is defined as the probability that T exceeds a time t, given by

$$S(t) = Pr\{T > t\} = 1 - F(t) = \int_t^\infty f(s)ds,\ 0 < t < \infty.$$

The hazard function, or instantaneous rate of occurrence of the event, is defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt | T \geq t)}{dt}.$$

The relationships between the survival function and the hazard function can be written as

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d \log S(t)}{dt},$$

and correspondingly,

$$S(t) = e^{-\int_0^t \lambda(s)ds} = e^{-\wedge(t)},$$

and

$$f(t) = \lambda(t) e^{-\wedge(t)},$$

with $\wedge(t) = \int_0^t \lambda(s)ds,\ 0 < t < \infty$, as the cumulative hazard function of T.

In survival analysis study, scientists are interested in investigating the association between the survival time of patients and predictor variables. Below is a short review

of a few popular statistical models that are widely known for analyzing failure time data.

1.3.1 The Proportional Hazards Model

The Cox proportional-hazards (PH) model Cox (1972) specifies the hazard function $h(t)$, which can be interpreted as the risk of dying at time t. It is defined as below,

$$h(t|x) = \exp(x'\beta)h_0(t),$$

where x represents the vector of covariates, $h_0(t)$ is baseline hazard function and β measures the impact of covariates.

The quantities $\exp(\beta)$ is called hazard ratio (HR). If the hazard ratio for the ith covariate is greater than one, then it indicates that as the value of the ith covariate increases, the event hazard increases and the length of survival decreases. Correspondingly, if the hazard ratio is smaller than 1, then the event hazard decreases and the length of survival increases with increasing value of ith covariate. And if the hazard ratio is 1, then there's no effect in that covariate. In the case of 2-sample problem, i.e., the ratio of hazards between the treatment group ($x = 1$) and the control group ($x = 0$), the ratio of the hazard functions has the form:

$$\frac{h(t; x = 1)}{h(t; x = 0)} = e^{\beta}.$$

Thus the ratio indicates that the covariates has multiplicative effects on the hazard function under the PH model.

The availability of the partial likelihood approach proposed by Cox (1975) made the proportional hazards model the most popular model for analyzing right-censored data in survival analysis. The approach is efficient since the estimator of β is asymptotically equivalent to the estimator of β from the full likelihood method, and β can be estimated without specifying the unknown baseline hazard function. The partial

likelihood is defined as below,

$$L(\beta) = \prod_{j=1}^{k} \left(\frac{h_0(t_j) \exp(x_j'\beta)}{\sum_{l \in R(t_j)} h_0(t_j) \exp(x_l'\beta)} \right)^{\delta_j}$$
$$= \prod_{j=1}^{k} \left(\frac{\exp(x_j'\beta)}{\sum_{l \in R(t_j)} \exp(x_l'\beta)} \right)^{\delta_j},$$

where $R(t) = \{j : t_j \geq t\}$ is the set of individuals who are "at risk" for failure at time t and $\delta_j = 0$ if t_j is a censoring time, 1 otherwise.

1.3.2 The Accelerated Failure Time model

The Accelerated Failure Time model (AFT model) assumes that the effect of a covariate is to accelerate or decelerate the life course of a disease by some constant. The failure time T can be modeled as follows:

$$\log(T) = x'\beta + \epsilon,$$

where x is covariate vector and ϵ is the disturbance term. By assuming different distributions for ϵ, the failure time T has different parametric distributions. Table 1.2 gives some of these distributions. Maximum likelihood approach can be applied for estimation purpose.

Table 1.2: Distributions for T with different ϵ under AFT models

Distribution of ϵ	Distribution of T
extreme values (2 parameters)	Weibull
extreme values (1 parameter)	exponential
log-gamma	gamma
logistic	log-logistic
normal	log-normal

1.3.3 The Proportional Odds Model

The proportional odds (PO) model was proposed by Bennett (1983). It specifies

$$\frac{F(t;x)}{1-F(t;x)} = \frac{F_0(t;x)}{1-F_0(t;x)} e^{x'\beta},$$

or

$$\text{logit}F(t;x) = \text{logit}F_0(t) + x'\beta \ ,$$

where $F_0(t)$ is an unknown baseline cumulative distribution function. Thus, the PO model assumes that each explanatory variable exerts the same effect on each cumulative logit. The ratio of the hazards changes with time t under PO model.

1.3.4 The Additive Hazards Model

The additive hazards model is given by

$$h(t;x) = h_0(t) + x'\beta,$$

where $h_0(t)$ is an arbitrary unspecified baseline hazard function. This model specifies that the effects of the covariates are additive rather than multiplicative as in the Cox model. The model was developed first by Aalen (1989).

1.4 Existing Regression Analysis approaches

1.4.1 Regression analysis of Interval-Censored Data

Many methods have been developed for analyzing interval-censored data in the past two decades. The primary goal in these regression analyses is to estimate the covariate effects on the failure time. Semiparametric regression models are popular since they enjoy great flexibility as compared to parametric models by allowing the baseline survival function to be unspecified.

Finkelstein (1986) proposed a maximum likelihood estimation method under the proportional hazards model for interval censored data. His method based on a Newton-Raphson algorithm provides estimates for covariate effects that are compatible with those from the Cox PH model, and the score test based on Finkelstein's method can be used for hypothesis testing. Among others Satten (1996) proposed

a marginal likehood approach to fit the proportional hazards model; Betensky et al. (2002) adopted local likelihood methods for the proportional hazards regression analysis; Cai and Betensky (2003) developed a new approach for estimating the hazard function for interval censored survival data by applying a piecewise linear spline and maximizing the penalized likelihood by a mixed model-based approach; Goggins et al. (1998) proposed a Markov Chain Monte Carlo expectation-maximization (EM) algorithm for fitting the proportional hazards model, Goeteghebeur and Ryan (2000) developed an EM algorithm estimating the covariate effects and baseline hazard function by maximizing a Cox partial likelihood and using the Breslow estimator. Shao et al. (2014) incorporated a semiparametric varying-coefficient model for interval censored data with a cured proportion. Wang et al. (2016) presented a novel EM algorithm relied on a two-stage data augmentation for analyzing interval-censored data under the PH model.

The proportional odds model were studied by Huang and Rossini (1997) and Shen (1998), both of which applied sieve estimation procedures. The former took use of a piecewise linear function, while the latter employed a monotone spline to approximate the baseline log odds function. Rabinowitz et al. (2000) applied conditional logistic regression by assuming that all examination times, even after the event, are recorded. Under the accelerated failure time model framework, Rabinowitz et al. (1995) and Betensky et al. (2001) explored estimating equation approaches and score statistics. Li and Pu (2003) applied a U-statistic based on ranks to estimate covariate coefficients and Xue et al. (2006) adopted the sieve estimation idea. Zeng et al. (2006) proposed a maximum likelihood approach under the additive hazards model. More recent research includes Zeng et al. (2016), who devised an EM-type algorithm through semiparametric transformation models, and Zhang and Zhao (2013) studied rank-based estimation methods for linear transformation models.

1.4.2 Regression Analysis of Multivariate Interval-Censored Data

There are some existing methods available in literatures for analysing multivariate interval-censored data. Wei et al. (1989) proposed a marginal proportional hazards model. This method adopted the working interdependence assumption among the multivariate failure times, and it inherits the advantages of Cox model. Kim and Xue (2002) extended this marginal approach by assuming the marginal distribution for each event is based on a discrete analogue of the proportional hazards model. Goggins and Finkelstein (2000) also outlined a method based on the discrete proportional hazards model. Chen et al. (2007) developed the proportional odds model for multivariate interval censored failure time data. Shen (2015) considered a general class of additive transformation model. Another popular approach for handling correlated failure time data is through frailty models. Models in this class assume that, conditional on some unobserved quantity, which is called 'frailty', the lifetimes are independent. When the unknown random effect is integrated out, the lifetimes become dependent; the frailty terms are introduced into models for survival data to represent the dependence. See more details in Hougaard (2000), Ibrahim et al. (2008) and Wienke (2012). For example, Oakes (1989) considered the class of bivariate survival distributions by inducing frailties. Komarek and Lessaffre (2007) proposed a Bayesian accelerated failure time model with frailty. Lin and Wang (2011) developed a Bayesian proportional odds model with a gamma frailty. Zuma (2007) explored the Gamma-frailty Weibull model. Wang et al. (2015) studied the gamma-frailty proportional hazards model by using the EM algorithm for bivariate current status data. Gamage et al. (2018) generalized this method for correlated bivariate interval-censored data. Chen et al. (2009). studied the PH model with a normal frailty and a probit model with normal frailty was proposed by Dunson and Dinse (2002) for analyzing multivariate case I interval-censored data.

1.5 Outline of The Dissertation

The remainder of this dissertation contains three main parts about statistical analysis of multivariate interval censored failure time data. In chapter 2, we discuss the regression analysis under the normal frailty probit model from Bayesian perspectives. This new model assumes that there exists a common but unobserved frailty, and the correlated failure times are independent given the frailty. The frailty term induces correlations among the multiple survival times. A Bayesian approach is developed for estimating the covariate effects under this normal frailty probit model for multivariate interval censored data. Simulation studies and analysis of a real data set are conducted to evaluate the performance of the proposed method.

In chapter 3, a Bayesian estimation approach for regression analysis of multivariate interval censored data under a semiparametric multivariate probit model is developed. The association structure between multiple failure times is modeled through the covariance matrix of correlated random errors. This model allows arbitrary correlations among multiple failure times. An efficient Gibbs sampling technique is developed for the estimations on covariate effects and the correlations. Extensive simulation study is conducted to assess the performance of the proposed method.

Some future work are discussed in chapter 4.

Chapter 2

Bayesian regression analysis of multivariate interval-censored failure time data under the normal frailty probit model

Summary: Interval-censored data naturally arise in many epidemiological, social- behavioral, and medical studies, in which subjects are examined multiple times and the failure times of interest are not observed exactly, but fall within some intervals. Correlated survival times arise when the subject experiences several events, and the events are potentially correlated. Frailty modeling is a popular approach for this type of data since it acknowledges this data specialty and directly models the correlation structure through frailty terms. In this chapter, a new frailty probit model is proposed for the regression analysis of multivariate interval-censored data, and this model allows explicit form of the pairwise statistical associations among the failure times. Monotone splines are applied for the purpose of approximating the unknown functions, significantly reducing the number of unknown parameters while retaining modeling flexibility. An efficient Bayesian estimation approach is proposed under this model and allows joint estimation of regression parameters and other secondary parameters. The proposed method is evaluated by extensive simulation studies and illustrated by a real-life application.

2.1 Introduction

Interval-censored failure time data are commonly observed when subjects undergo periodic follow-ups in clinical studies. The failure time of interest is not observed exactly but is known to fall within some interval. (Kalbfleisch and Prentice, 2002; Sun, 2006). Multivariate interval-censored data commonly arise when there are multiple failure time events and only interval-censored data are available for each failure time. Furthermore, these events of interest are typically correlated because of sharing some common characteristics. One example of multivariate interval-censored data in the literature is the AIDS clinical trial data, where the occurrences of bacterial and viral infections were the correlated interval-censored events of interest (Betensky and Finkelstein, 1999). Another example is sexually transmitted infection (STI) data, where the times to first infection with Chlamydia trachomatis (CT), Neisseriagonorrhoeae (NG), and Trichomonas vaginalis (TV) are the endpoints of interest (Tu et al., 2009). These concurrent infections with multiple organisms are correlated, resulting in multivariate interval-censored data. Ignoring such correlation will lead to biased estimation. With the correlation structure among multiple failure events, the analysis for the interval-censored data analysis becomes very complicated. Therefore, a model that can efficiently estimate the covariate effects under the association structure for multiple failure times is in need, and method on estimating the correlations for multivariate interval-censored data is desirable.

There is a substantial literature on the use of frailty for multivariate failure time data. For multivariate right-censored data, existing work includes Oakes (1989), Klein (1992), Andersen et al. (1997), Cui and Sun (2004), Rondeau et al. (2003) and Yin and Ibrahim (2005) among many others. For multivariate interval-censored data, existing work includes Wen and Chen (2011), Dunson and Dinse (2002), Chen et al. (2009), Lam et al. (2010), Henschel et al. (2009), Yavuz and Lambert (2016), Wen and Chen (2013), Lin and Wang (2011), Komarek and Lessaffre (2007), Gamage et al.

(2018), Chang et al. (2007), Zuma (2007), Hens et al. (2009), and Wang et al. (2015).

This chapter discusses the fitting of a multivariate probit model with a normal frailty term to interval censored data. The remaining of this chapter is organized as follows. Section 2.2 presents the proposed model and its properties. Section 2.3 gives the details of the proposed approach, including the use of monotone splines to approximate the unknown function, a data augmentation procedure, and the proposed Gibbs sampler for posterior computation. Section 2.4 evaluates the proposed method via a simulation study, and Section 2.5 provides a real-life application as an illustration. Discussions are given in section 2.6.

2.2 Models and Properties

Let $F_j(\cdot|\boldsymbol{x})$ denotes the cumulative distribution function (CDF) of the failure time of interest T_j given covariate vector $\boldsymbol{x}$. The new normal-frailty multivariate probit model specifies the conditional cumulative distribution function in the following form:

$$F_j(t|\boldsymbol{x},\zeta) = \Phi\{\alpha_j(t) + \boldsymbol{x'\beta_j} + \zeta\}, \forall t \in (0,\infty), \tag{2.1}$$

where $\Phi(\cdot)$ is the CDF of a standard normal random variable, $\boldsymbol{\beta_j}$ is a vector of regression coefficients, $\zeta \sim N(0,\sigma^2)$ is the frailty term, and α_j is an unknown nondecreasing function with $\alpha_j(0) = -\infty$ and $\alpha_j(\infty) = \infty$. The common frailty induces correlation among those T_j's. We can rewrite the model at the subject level as follows, with T_{ij} denoting the jth failure time for the ith subject, $i = 1, ..., n$ and $j = 1, ..., k$.

$$\alpha_j(T_{ij}) = -\boldsymbol{x_i'\beta_j} - \zeta_i + \epsilon_{ij}, \tag{2.2}$$

where ϵ_{ij}s are independent standard normal random variables, and ζ_i's are frailties that are normally distributed with mean 0 and variance σ^2. These ζ_i's show the heterogeneity among the subjects.

The proposed normal frailty Probit model has a simple form but enjoys several appealing properties. First, the marginal distribution of T_j is a semiparametric Probit

model of Lin and Wang (2010), and thus the marginal and conditional distributions of T_j belong to the same family. Second, the conditional covariate effects given the frailty are proportional to the marginal covariate effects. Thus, we can estimate the marginal covariate effects easily through this model and give marginal interpretations for regression parameters. Third, the association between different failure times of interest can be explicitly quantified via three nonparametric association measures in simple form. Details of these properties are presented in section 2.2.2.

2.2.1 Marginal covariate effect

The proposed frailty multivariate probit model (2.1) essentially is an extension to the semiparametric probit model of Lin and Wang (2010), with an extra frailty term that adjusts the correlation between the multiple events. Given the frailty term ζ, the coefficients $\boldsymbol{\beta_j}$ can be interpreted as the conditional covariate effects on the transformed failure time T_j. However, since the frailty is unknown, the use of marginal covariate effects are preferred for interpretation purpose.

The marginal CDF of the failure time T_j can be obtained by integrating out ζ from the conditional CDF (2.1):

$$\begin{aligned} F_j^*(t|\boldsymbol{x}) = Pr(T_j \leq t|\boldsymbol{x}) &= \int \Phi\Big(\alpha_j(t) + \boldsymbol{x'}\boldsymbol{\beta}_j + \zeta\Big)\pi(\zeta)d\zeta \\ &= \Phi\{\alpha_j^*(t) + \boldsymbol{x'}\boldsymbol{\beta_j^*}\}, \end{aligned} \tag{2.3}$$

where $\alpha_j^*(t) = \theta\alpha_j(t)$, $\boldsymbol{\beta_j^*} = \theta\boldsymbol{\beta}_j$ and $\theta = (1+\sigma^2)^{-1/2}$. From (2.3), we observe that the failure time T_j follows a marginal semiparametric probit model. See Lin and Wang (2010). The regression coefficients $\boldsymbol{\beta_j}$ can be interpreted as the marginal covariate effects $\boldsymbol{\beta}^*$ up to a multiplicative constant θ. This relationship implies that the inferences based on the conditional covariate effects $\boldsymbol{\beta}$ and the marginal covariate effects $\boldsymbol{\beta}^*$ will lead to the same conclusion. We can easily estimate the marginal covariate effects through the conditional covariate effects obtained from this multivariate probit model.

2.2.2 Multiple Events Association

Measures of association are well studied and often applied to data that are completely observed. Some methods also applied on right censored data (e.g. Clayton (1978), Dabrowska (1986), Oakes (1982)). In this section, we applied three most widely used statistical methods for modeling correlated responses under our proposed model: Spearman's rank correlation coefficient ρ_s, median concordance κ and kendall's τ (Kruskal (1958), Hougaard (2000)).

Now suppose we have two correlated failure times T_1 and T_2, and these two events of interest have the same set of covariates $\boldsymbol{x}$. The Spearman's rank correlation coefficient is defined as

$$\rho_s = 12 \int_0^1 \int_0^1 S(S_1^{-1}(u), S_2^{-1}(v))\, du\, dv - 3,$$

where $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$ is the joint survival function, S_1 and S_2 are the marginal survival functions of T_1 and T_2 respectively. S_1^{-1} and S_2^{-1} are the inverse functions of S_1 and S_2.

Median concordance by Kruskal (1958) is a nonparametric measure of association between correlated random variables and is defined as below:

$$\kappa = E\Big[\text{sign}\Big\{(T_1 - M_1)(T_2 - M_2)\Big\}\Big],$$

where sign$(\cdot)$ is the sign function taking 1 for positive values, 0 for zero and -1 for negative values. M_1 and M_2 are the population medians of T_1 and T_2, respectively.

Kendall's τ is another rank-based nonparametric measure which 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 sign function.

Spearman's rank correlation coefficient, median concordance and Kendall's τ are nonparametric methods and do not require specific forms of the distributions for the

correlated failure times. And they are invariant to marginal monotone transformations, which means $\rho_s(T_1, T_2) = \rho_s(g_1(T_1), g_2(T_2))$ for any two increasing (decreasing) transformations g_1 and g_2. The normal-frailty multivariate probit model provides closed-form expressions for the statistical association between correlated failure times in terms of the Spearman's correlation coefficient ρ_s, median concordance κ and kendall's τ, defined as follows,

$$\rho_s = 6\pi^{-1} \sin^{-1}(\rho/2) = \frac{6}{\pi} \arcsin\left(\frac{\sigma^2}{2(1+\sigma^2)}\right), \tag{2.4}$$

$$\kappa = 2\pi^{-1} \sin^{-1}(\rho) = \frac{2}{\pi} \arcsin\left(\frac{\sigma^2}{1+\sigma^2}\right), \tag{2.5}$$

$$\tau = 2\pi^{-1} \sin^{-1}(\rho) = \frac{2}{\pi} \arcsin\left(\frac{\sigma^2}{1+\sigma^2}\right). \tag{2.6}$$

The proof is sketched in the Appendix A. The association among the failure times is explicitly quantified. As seen in (2.4), (2.5) and (2.6), the magnitude of correlation depends only on the frailty variance, and a large variance will lead to a strong dependence among these failure times. Another interesting finding is that ρ_s, κ and τ are all free of covariates, indicating that the association does not rely on the covariates. The values for Spearman's rank correlation coefficient, median concordance and Kendall's τ both range from -1 and 1. A positive value indicates a positive relationship between the responses while a negative value expresses a negative association. A value of zero indicates that no association exists between the events.

2.3 Estimation Method

2.3.1 Modeling α_j's with monotone splines

From model (2.1), the unknown parameters involve the regression parameters $\boldsymbol{\beta_j}$, the function α_j's and the frailty variance parameter σ^2. For the jth event, the marginal baseline CDF is represented by $F_{j0}(t) = \Phi(\alpha_j(t))$ under (2.3), and thus α_j can be regarded as the transformed baseline CDF for the jth failure event with probit link.

The well established partial likelihood method for Cox PH model allow one to consistently estimate the covariate effects $\boldsymbol{\beta}$, without the need of estimating the baseline hazard function for right-censored data. However, those techniques no longer work for interval censored data due to its complex data structure, nor those techniques exist under the probit model. Recall that the unknown function $\alpha(\cdot)$ is a nondecreasing function with an infinite dimension, the estimation under this model is challenging. Inspired by Lin and Wang (2010), Wang and Dunson (2011), Wang et al. (2016), and Cai et al. (2011), we approximate the unknown hazard function α_j, $j = 1, 2 \cdots k$, through the use of monotone splines of Ramsay (1988) as follows,

$$\alpha_j(t) = \gamma_{j0} + \sum_{l=1}^{m} \gamma_{jl} b_l(t), \ j = 1, \cdots k \tag{2.7}$$

where b_l's are monotone integrated spline basis functions, each of which is nondecreasing from 0 to 1, γ_0 is an unconstrained intercept of a monotone spline, and γ_{jl}'s are the corresponding unknown spline coefficients, which are constrained to be nonnegative to ensure the monotonicity of α_j. The spline basis functions $b_l(t)$'s are piecewise polynomials, taking 0 at the very beginning stage, increasing from 0 to 1 in the middle stage and staying plateau at the last stage. Here we adopt the same set of monotone spline basis function for all events. This is reasonable because the same observational process is available for all events for each subject. The two key factors in determining these spline functions include the knot placement and the degree of the splines. The placement of knots determines the shape of basis splines, with more knots introducing greater modeling flexibility. The degree controls the smoothness of the basis functions. For example, the degree of 1, 2, 3 represents linear, quadratic and cubic basis functions, respectively. The spline basis functions will be fully determined after the placement of knots and degree are specified, with the number of spline basis functions equals the number of interior knots plus the degree of the splines. More details about splines can be referred to Ramsay (1988).

The new presentation (2.7) for α_j is very flexible as it can approximate any non-

decreasing continuous function by using I spline basis functions. The infinite dimensional parameter in α_j is reduced to a finite number of parameters γ_{jl}'s. The number of unknown parameters is significantly reduced while retaining modeling flexibility. In general, even though using more knots leads to greater model flexibility, having too many basis functions may cause over-fitting problem and increase computation burden. Following the conclusions from Lin and Wang (2010) and Wang and Dunson (2011), we adopt a moderate number (10-30) of knots to allow for efficient computation while maintaining modeling flexibility. A degree of 2 or 3 usually guarantees adequate smoothness. As discussed in Cai et al. (2011), Lin and Wang (2010), Wang and Lin (2011), and Wang and Dunson (2011), there are two common ways to design knot placement: use equally spaced knots or place the knots based on the quantiles of the observed interval within the data range.

2.3.2 Bayesian method

Let $(L_{ij},\ R_{ij})$ denote the observed interval for T_{ij}, for $i = 1, \ldots, n,\ j = 1, \ldots, k$. Here, L_{ij} and R_{ij} denote the left and right bounds of the observed interval for the jth event of the ith subject respecitively, with $L_{ij} < R_{ij}$. In our study, we consider a case II interval-censored data, which includes left, interval, and right-censored observations. To further illustrate, when $L_{ij} = 0$, the failure time T_{ij} is left-censored; for those intervals with $R_{ij} = \infty$, T_{ij}'s are right-censored; the failure time T_{ij} is interval-censored otherwise. Define δ_{ij1}, δ_{ij2} and δ_{ij3} to be the censoring indicators representing left-, interval-, and right-censoring, respectively, with the constraint $\delta_{ij1} + \delta_{ij2} + \delta_{ij3} = 1$. Then the observed data are $\Delta = \{(L_{ij}, R_{ij}, \boldsymbol{x_i}, \delta_{ij1},\ \delta_{ij2},\ \delta_{ij3}); i = 1, 2, ...n; j = 1, 2, ...k\}$. In this chapter, we adopt non-informative censoring assumption, which suggests that the failure time and the observation process that generates the observed interval are independent, given the covariates information. This assumption is common in survival literatures; see, e.g., Zhang and Sun (2010)

among others. Therefore, the observed likelihood can be expressed as

$$L_{obs} = \prod_{i=1}^{n} \int \sigma^{-1}\phi(\sigma^{-1}\zeta_i) \prod_{j=1}^{k} \left[F_j(R_{ij}|\boldsymbol{x_i}, \zeta_i) - F_j(L_{ij}|\boldsymbol{x_i}, \zeta_i)\right] \, d\zeta_i,$$

where $\phi(\cdot)$ is the density function of a standard normal random variable. We can further write down the likelihood as

$$\begin{aligned} L_{obs} = \prod_{i=1}^{n} \int \sigma^{-1}\phi(\sigma^{-1}\zeta_i) \prod_{j=1}^{k} & \left[F_j(R_{ij}|\boldsymbol{x_i}, \zeta_i)\right]^{\delta_{ij1}} \left[F_j(R_{ij}|\boldsymbol{x_i}, \zeta_i) - F_j(L_{ij}|\boldsymbol{x_i}, \zeta_i)\right]^{\delta_{ij2}} \\ & \left[1 - F_j(L_{ij}|\boldsymbol{x_i}, \zeta_i)\right]^{\delta_{ij3}} \, d\zeta_i, \end{aligned} \tag{2.8}$$

We observe that for multiple failure times when $k > 1$, the integrals in the observed likelihood (2.8) do not have an explicit form, and this makes the observed likelihood impossible to use directly for estimating the unknown parameters through Bayesian methods. In order to facilitate the posterior computation, we consider the following conditional likelihood L_{con} by treating all frailties ζ_i's as latent variables.

$$\begin{aligned} L_{con} = \prod_{i=1}^{n} \sigma^{-1}\phi(\sigma^{-1}\zeta_i) \prod_{j=1}^{k} & \left[F_j(R_{ij}|\boldsymbol{x_i}, \zeta_i)\right]^{\delta_{ij1}} \left[F_j(R_{ij}|\boldsymbol{x_i}, \zeta_i) - F_j(L_{ij}|\boldsymbol{x_i}, \zeta_i)\right]^{\delta_{ij2}} \\ & \left[1 - F_j(L_{ij}|\boldsymbol{x_i}, \zeta_i)\right]^{\delta_{ij3}}. \end{aligned} \tag{2.9}$$

Notice that by integrating the conditional likelihood (2.9) over ζ_i, we will obtain the observed data likelihood (2.8).

In general, Bayesian methods require sampling all the unknown parameters from their posterior distributions generated by combining the likelihood function and the prior distributions. From the discussions before, $\Theta = (\boldsymbol{\beta_j'}s, \gamma_j's, \sigma^2)$ is the set of unknown parameters, where $\gamma_j = (\gamma_{j0}, \gamma_{j1}, \ldots, \gamma_{jm})'$. However, the conditional likelihood is still too complicated for estimating the unknown parameters with any priors. Motivated by Lin and Wang (2010), an additional data augmentation layer is added by introducing normal latent variables,

$$z_{ij} \sim N(\alpha_j(t_{ij}) + \boldsymbol{x_i}\boldsymbol{\beta_j} + \zeta_i, 1), \; i = 1, \cdots n, \; j = 1, \cdots k.$$

where $t_{ij} = R_{ij}I_{(\delta_{ij1}=1)} + L_{ij}I_{(\delta_{ij1}=0)}$, i.e., t_{ij} takes the right end point of the observed interval when the failure time is left censoring and takes the left end point otherwise for all i and j. Then the augmented data likelihood function can be written as:

$$L_{aug} = \prod_{i=1}^{n} \sigma_\zeta^{-1}\phi(\sigma_\zeta^{-1}\zeta_i) \left\{ \prod_{j=1}^{k} \phi(z_{ij} - \alpha_j(t_{ij}) - \boldsymbol{x_i}\boldsymbol{\beta_j} - \zeta_i) I_{C_{ij}}(z_{ij}) \right\}, \tag{2.10}$$

where C_{ij} is the constrained space of z_{ij},

$$C_{ij} = \begin{cases} (0, \infty) & \text{if} \quad \delta_{ij1} = 1 \\ \Big(\alpha_j(L_{ij}) - \alpha_j(R_{ij}), 0\Big) & \text{if} \quad \delta_{ij2} = 1 \\ (-\infty, 0) & \text{if} \quad \delta_{ij3} = 1 \end{cases}$$

The likelihood (2.9) can be obtained after integrating out all z_{ij}s in (2.10). This augmented data likelihood has a nice form for sampling.

The following priors are assigned in order to promote the posterior computation: a multivariate normal prior $N(\boldsymbol{\beta_0}, \Sigma_0)$ for regression coefficients $\boldsymbol{\beta}$ and a normal prior $N(m_{j0}, \upsilon_{j0}^{-1})$ for the unconstrained γ_{j0}. Independent exponential priors $Exp(\eta_j)$ are adopted for all spline basis coefficients γ_{jl}'s for each j, and further a $\mathcal{G}a(a_{j\eta}, b_{j\eta})$ hyper prior is assigned for η_j. These prior specifications are appealing since it gives conjugate forms for the conditional posterior distributions of γ_{jl}'s and η_j, and they can shrink those small unnecessary spline coefficients to zero, serving to penalize the large nonzero spline coefficients, and thus resulting in basis function selection, see Cai et al. (2011). This nice property can alleviate the overfitting problems. A gamma prior $\mathcal{G}a(a_\zeta, b_\zeta)$ is given for frailty precision σ_ζ^{-2}.

Gibbs sampling is a popular Markov chain Monte Carlo (MCMC) algorithm for Bayesian computation (Geman and Geman, 1984). The idea is to generate posterior samples by sweeping through each variable to sample from its conditional distribution with the remaining variables fixed to their current values. We adopt Gibbs sampling for our posterior computation. Based on the above assigned priors and the augmented likelihood (2.10), the following Gibbs sampler is developed.

1. Sample latent variables z_{ij} for $i = 1, ..., n$ and $j = 1, ..., k$.

   - if $\delta_{ij1} = 1$, sample z_{ij} from $N(\alpha_j(t_{ij}) + \boldsymbol{x_i}\boldsymbol{\beta_j} + \zeta_i, 1)I_{(z_{ij}>0)}$;
   - if $\delta_{ij2} = 1$, sample z_{ij} from $N(\alpha_j(t_{ij}) + \boldsymbol{x_i}\boldsymbol{\beta_j} + \zeta_i, 1)I_{(\alpha_j(L_{ij})-\alpha_j(R_{ij})<z_{ij}<0)}$;
   - if $\delta_{ij3} = 1$, sample z_{ij} from $N(\alpha_j(t_{ij}) + \boldsymbol{x_i}\boldsymbol{\beta_j} + \zeta_i, 1)I_{(z_{ij}<0)}$.

2. Sample γ_{j0} from $N(E_{j0}, W_{j0}^{-1})$ where $W_{j0} = \upsilon_{j0} + n$ and

$$E_{j0} = W_{j0}^{-1}\left[m_{j0}\upsilon_{j0} + \sum_{i=1}^{n}\{z_{ij} - \sum_{l=1}^{m}\gamma_{jl}b_l(t_{ij}) - \boldsymbol{x_i}\boldsymbol{\beta_j} - \zeta_i\}\right].$$

3. Sample all γ_{jl}'s for $l = 1, 2..., m$ and $j = 1, ..., k$. For each l, let $W_{jl} = \sum_{i=1}^{n} b_l^2(t_{ij})$.

   - If $W_{jl} = 0$, sample γ_{jl} from $Exp(\eta_j)$.
   - If $W_{jl} > 0$, sample γ_{jl} from $N(E_{jl}, W_{jl}^{-1})1(\gamma_{jl} > d_{jl}^*)$, where

$$E_{jl} = W_{jl}^{-1}\left[\sum_{i=1}^{n} b_l(t_{ij})\{z_{ij} - \gamma_{j0} - \sum_{jl'\neq jl}\gamma_{jl'}b_{l'}(t_{ij}) - \boldsymbol{x_i}\boldsymbol{\beta_j} - \zeta_i\} - \eta_j\right],$$

$$d_l^* = \max(c_l^*, 0) \text{ and } c_l^* = \max_{\{(i,j):\delta_{ij2}=1\}}\left[\frac{-z_{ij}-\sum_{jl'\neq jl}\gamma_{jl'}\{b_{l'}(R_{ij})-b_{l'}(L_{ij})\}}{b_l(R_{ij})-b_l(L_{ij})}\right].$$

4. Sample $\boldsymbol{\beta_j}$ from $N(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \hat{\Sigma}_j)$, where $\hat{\Sigma}_j = (\Sigma_{j0}^{-1} + \sum_{i=1}^{n}\boldsymbol{x_i}\boldsymbol{x_i'})^{-1}$ and

$$\hat{\boldsymbol{\beta}}_j = \hat{\Sigma}_j\left[\Sigma_{j0}^{-1}\beta_{j0} + \sum_{i=1}^{n}\{z_{ij} - \alpha_j(t_{ij}) - \zeta_i\}\boldsymbol{x_i}\right].$$

5. Sample ζ_i from $N(\mu_i, \sigma_i^2)$ for $i = 1, ..., n$ where $\sigma_i^2 = (k + \sigma_\zeta^{-2})^{-1}$ and

$$\mu_i = \sigma_i^2\sum_{j=1}^{k}\{z_{ij} - \alpha_j(t_{ij}) - \boldsymbol{x_i'}\beta\}.$$

6. Sample η_j for $j = 1, ..., k$ from $Gamma(a_{j\eta} + m, b_{j\eta} + \sum_{l=1}^{m}\gamma_{jl})$.

7. Sample σ_ζ^{-2} from $Gamma(a_\zeta + 0.5n, b_\zeta + 0.5\sum_{i=1}^{n}\zeta_i^2)$.

The above Gibbs sampler is fast and easy to implement since all the parameters and latent variables can be updated through standard distributions. Based on simulation studies below, we observe that the proposed method enjoys fast convergence.

2.4 Simulation Studies

In this section, we evaluate the performance of the proposed model through simulation studies. First we generated 100 data sets with sample size $n = 500$ under the following model involving both discrete and continuous covariates,

$$F_j(t|x_{i1}, x_{i2}, \zeta_i) = \Phi\{\alpha_j(t) + x_{i1}\beta_{j1} + x_{i2}\beta_{j2} + \zeta_i\},$$

where x_{i1} is a normal random variable follows $N(0, .25)$ and x_{i2} is a Bernoulli random variable with probability of success 0.5, and ζ_i was generated from $N(0, \sigma^2)$ with $\sigma = .5$, 1 and 2 respectively. We considered three events of interest and took true $\alpha_1(t) = 1 + t + \log(t)$, $\alpha_2(t) = t^2 + \log(t)$ and $\alpha_3(t) = 1 + t + \log(t)$, $\boldsymbol{\beta_1} = (1, 0)$, $\boldsymbol{\beta_2} = (0, 1)$, $\boldsymbol{\beta_3} = (-1, 1)$. Below is how we obtain the observed interval (L_{ij}, R_{ij}) for T_{ij}.

- For subject i, we first generate the number k_i of observation times from a Poisson distribution plus 1. This guarantees that k_i is at least one and that different subjects can have different number of observation times.
- Generate k_i gap times for subject i independently from an Exponential family with mean 4. Denote these gap time by $g_{i1}, \ldots, g_{ik_i}$.
- Obtain the observed times by $O_{id} = \sum_{l=1}^{d} g_{il}$, for $d = 1, \ldots, k_i$.
- For each j, we calculate $F_j(O_{id} \mid x_i, \delta_i)$ for $d = 1, \ldots k_i$ and generate u_{ij} from $U(0, 1)$.
- The observed interval (L_{ij}, R_{ij}) will be taken as interval (O_{ic}, O_{ic+1}), where $F_j(O_{ic} \mid x_i, \delta_i) \leq u_i < F_j(O_{ic+1} \mid x_i, \delta_i)$.

The specification of the observation process was chosen so that none of the censoring types dominates the others. For example, in the case of $\sigma^2 = 4$, there are on average 51.08% left-censored observations, 36.63% interval-censored observations, and

12.29% right-censored observations for the first event; 28.03% left-censored observations, 48.39% interval-censored observations, and 23.59% right-censored observations for the second event; 36.56% left-censored observations, 41.90% interval-censored observations, and 21.54% right-censored observations for the third event across all simulated data sets.

In specifying the monotone splines, we chose 2 for the degree to guarantee adequate smoothness of the splines. For each generated data set, knots were spaced equally within the minimum and the maximum value of the finite endpoints of the observation times. The distance between two adjacent knots is equal to 0.3. Consequently, the number of knots varies from data set to data set and ranges from 12 to 24. For the Bayesian computation, the following specifications were given: $m_{10} = -3$, $m_{20} = -4$ and $m_{30} = -3$; $\upsilon_{10} = \upsilon_{20} = \upsilon_{30} = 0.1$, which will results in a normal prior for γ_{j0} with a large variance, a $\mathcal{G}a(1,1)$ prior for η_j with $a_{j\eta} = b_{j\eta} = 1$, a $\mathcal{G}a(1,1)$ prior with $a_\zeta = b_\zeta = 1$ for σ^{-2}, and $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\Sigma_0 = n(\boldsymbol{X'X})^{-1}$, where $\boldsymbol{X}$ is the covariate matrix. For each data set, we implemented the Gibbs sampler and summarized results based on 4000 iterations of MCMC after discarding first 1000 iterations as a burn-in. This was observed to be sufficient due to good mixing observed in the sample chains.

Table 2.1 shows the performance of the proposed method in the case of using 100 data sets under three scenarios: true $\zeta \sim N(0, .25)$, $\zeta \sim N(0, 1)$, $\zeta \sim N(0, 4)$. The Bias is calculated as the difference between the average of the 100 point estimates (posterior means) and the true value, ESD denotes the average of the estimated standard deviations of their posterior distributions, SSD is the sample standard deviation of the 100 point estimates, and the CP95 represents the 95% coverage probability.

From the results in Table 2.1, we can tell the proposed method works very well. The biases for all point estimates are small, the ESDs are close to corresponding SSDs, and the 95% coverage probabilities are close to 0.95 for all the regression parameters and the frailty variance parameter σ^2 under all simulation settings. In addition to

Table 2.1: Simulation results of the proposed method under three scenarios: $\zeta_i \sim N(0, .25)$, $\zeta_i \sim N(0, 1)$ and $\zeta_i \sim N(0, 4)$ based on 100 datasets. Presented results include the bias, the average of the estimated standard deviations, the sample standard deviation of the 100 point estimates, and the 95% coverage probability for the conditional covariate effects.

	Scenario I					Scenario II					Scenario III				
	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95
β_{11}	1	-0.012	0.134	0.140	0.96	1	0.004	0.162	0.160	0.96	1	-0.020	0.235	0.235	0.93
β_{12}	0	0.022	0.142	0.133	0.94	0	0.014	0.174	0.155	0.93	0	0.034	0.247	0.234	0.92
β_{21}	0	-0.014	0.127	0.118	0.92	0	-0.007	0.140	0.142	0.93	0	-0.039	0.244	0.226	0.95
β_{22}	1	0.045	0.132	0.129	0.94	1	0.056	0.161	0.151	0.93	1	0.010	0.244	0.231	0.97
β_{31}	-1	-0.037	0.122	0.127	0.92	-1	-0.046	0.161	0.148	0.93	-1	-0.010	0.247	0.229	0.94
β_{32}	1	0.012	0.128	0.130	0.96	1	0.031	0.145	0.152	0.95	1	0.001	0.243	0.232	0.96
σ^2	.25	-0.020	0.060	0.059	0.92	1	0.080	0.120	0.136	0.94	4	-0.060	0.523	0.491	0.95
ρ	.191	-0.009	0.035	0.034	0.92	.483	0.025	0.032	0.036	0.94	.786	-0.001	0.021	0.020	0.95
κ	.128	-0.006	0.024	0.023	0.92	.333	0.018	0.023	0.027	0.94	.590	-0.002	0.021	0.020	0.95
τ	.128	-0.006	0.024	0.023	0.92	.333	0.018	0.023	0.027	0.94	.590	-0.002	0.021	0.020	0.95

the excellent estimation accuracy in regression coefficients and frailty variance, the proposed method also provides precise estimation results for the association in terms of Spearman's correlation coefficient ρ_s, median concordance κ and the Kendall's τ, as seen in Table 2.1.

From the discussions in section 2.2, the covariate coefficients $\boldsymbol{\beta}$ from Table 2.1 can be interpreted as the conditional covariate effects on the transformed failure time. Though this interpretation is appealing, it is conditioning on the unknown frailty term. The marginal covariate effects are preferred since there exits a multiplicative relationship between the conditional covariate effects $\boldsymbol{\beta}$ and the marginal covariate effects $\boldsymbol{\beta}^*$ under the proposed multivariate probit model. As a further illustration and comparison, the semiparametric probit model by Lin and Wang, 2010, which from henceforth will be referred as the univariate approach was considered. This competing approach uses the idea of modeling each of failure times separately under the semiparametric probit model while ignoring the underlying correlated structure. Table 2.2 presents a summary of the regression parameter estimates obtained by the univariate approach and the corresponding marginal covariate coefficients from our proposed methodology for the same simulation configurations as were considered in

Table 2.1. The estimations for marginal covariate coefficients are given in Table 2.2 with different values of σ^2. As seen from Table 2.2, we note that the univariate approach also performs well (this is not surprising since the probit model is the true marginal model), but differences are obvious when comparisons are made. In particular, the bias obtained from univariate approach are bigger than those obtained from our proposed model. Moreover, the empirical coverage probabilities for the univariate approach were not at their nominal level around 95%, especially with the estimates for $\boldsymbol{\beta_2}$ and $\boldsymbol{\beta_3}$ tending to under cover. These losses in estimation precision are likely attributable to the fact that the univariate approach ignores the dependence between the failure times during estimation. In summary, Table 2.2 shows that the proposed method is capable of accurately estimating the unknown model parameters and delivers reliable inference.

2.5 Real Data Analysis

2.5.1 Sexually transmitted infection (STI) data

In this section, we apply our method to STI data, which were collected on young women as a part of the Young Women's Project (YWP). The details for study design and folllow-up protocol can be found in Tu et al., 2009 and Tu et al., 2011. In this study, infections with Chlamydia trachomatis (CT), Neisseria gonorrhoeae (GC) and Trichomonas vaginalis (TV) are the three outcomes of interest. This analysis focuses on the time to initial STI infections for each of these three organisms. Three hundred and eighty seven adolescent young women aged 14 to 17 years were observed between 1999 and 2007 in this observational study. At enrollment, participants were interviewed and asked to complete detailed questionnaire about their sexual behaviors such as the number of sex partners, age of first sex, infection history, etc.. Patients were examined every three months and actual examination times differed from patient to patient since some of them missed their visits. As a result, the precise times of

Table 2.2: Marginal covariate effects comparison between univariate probit model and normal-frailty multivariate probit model based on 100 datasets under the scenarios $\sigma^2 = 0.25$, $\sigma^2 = 1$ and $\sigma^2 = 4$. Bias denotes the difference between the average of the 100 point estimates and the true value, SSD the sample standard deviation of the 100 point estimates, ESD is the average of the estimated standard deviations, and CP95 the 95% coverage probability

		Univariate Probit Model				Multivariate Probit Model			
	True	Bias	SSD	ESD	CP95	Bias	SSD	ESD	CP95
$\sigma^2 = 0.25$	$\beta_{11} = 1$	-0.007	0.119	0.128	0.97	-0.004	0.118	0.123	0.96
	$\beta_{12} = 0$	0.014	0.128	0.118	0.95	0.019	0.126	0.118	0.94
	$\beta_{21} = 0$	-0.015	0.111	0.105	0.92	-0.013	0.113	0.105	0.92
	$\beta_{22} = 1$	0.067	0.111	0.112	0.90	0.047	0.113	0.113	0.93
	$\beta_{31} = -1$	-0.061	0.107	0.110	0.90	-0.040	0.108	0.111	0.93
	$\beta_{32} = 1$	0.032	0.113	0.113	0.95	0.018	0.111	0.114	0.96
$\sigma^2 = 1$	$\beta_{11} = 1$	-0.009	0.122	0.115	0.95	-0.013	0.119	0.114	0.95
	$\beta_{12} = 0$	0.006	0.130	0.113	0.94	0.010	0.126	0.112	0.93
	$\beta_{21} = 0$	-0.010	0.094	0.102	0.97	-0.005	0.101	0.102	0.93
	$\beta_{22} = 1$	0.073	0.110	0.107	0.85	0.025	0.114	0.108	0.94
	$\beta_{31} = -1$	-0.070	0.106	0.104	0.88	-0.018	0.111	0.106	0.93
	$\beta_{32} = 1$	0.049	0.100	0.108	0.94	0.006	0.105	0.108	0.96
$\sigma^2 = 4$	$\beta_{11} = 1$	0.023	0.103	0.106	0.92	-0.006	0.104	0.104	0.94
	$\beta_{12} = 0$	0.009	0.109	0.107	0.93	0.015	0.111	0.103	0.92
	$\beta_{21} = 0$	-0.021	0.101	0.099	0.93	-0.017	0.108	0.100	0.95
	$\beta_{22} = 1$	0.062	0.101	0.103	0.90	0.008	0.106	0.102	0.93
	$\beta_{31} = -1$	-0.063	0.102	0.100	0.89	-0.008	0.108	0.101	0.94
	$\beta_{32} = 1$	0.049	0.096	0.104	0.95	0.004	0.107	0.102	0.95

infections were not directly observable since the infections could have happened at any time during the interval between the last visit with a negative result and the first visit with a positive result. In other words, the time to each type of infection is interval-censored. The event times for a subject are right-censored at the last visit time if no any infection was detected throughout the follow-up. The event times are left-censored at the beginning of the study if detected positive at the time of first testing.

Chalamydia, gonorrhea and trichomoniasis are the three most common bacterial infections for sexually transmitted diseases that often co-exist. Times to the initial infections within the same individual are correlated due to the same physiological environment and sexual behavior. This multivariate interval-censored data analy-

sis jointly model the infection times, studies several participant characteristics and examines the associations between these infections.

There are five covariates of interest: age when enter the study, the number of partners, age at first intercourse, race and the infection status at the beginning of the study. Twenty-seven patients were excluded from study due to missing data or data discrepancies. After data cleaning, a subset of 360 patients was included in the analysis. Among these individuals, 10.28%, 44.72%, and 45% were left-, interval-, and right-censored, respectively for C. trachomatis; 1.67%, 28.89%, and 69.44% were left-, interval-, and right-censored, respectively for N.gonorrhea and 1.67%, 28.89%, and 69.44% were left-, interval-, and right-censored, respectively for T.vaginalis.

2.5.2 Data analysis results

We applied our proposed method to this data set with 16 knots for monotone quadratic splines. The knots are assigned according to the quantiles of the observation intervals. The same prior specifications as simulation study are used here. A total of 20000 iterations were run in our Gibbs sampler and the first 5000 iterations were discarded as a burn-in. A summary of the posterior mean estimates and the corresponding 95% credible intervals for the regression parameters on the 15000 iterations of the Markov chain is presented in Table 2.3. This analysis indicates that the infection status at the beginning of the study has a big impact on the first infection time in the study for all the three infections since their 95% credible intervals are all outside 0. Having an infection history was associated with an increased risk of early infection acquisition. While the age when entering the study, age at first intercourse and race seems irrelevant to the time to first infection with N.gonorrhoeae and C. trachomatis, these characteristics contributes to the time to first infection with T. vaginalis. An earlier age at first sexual intercourse, older age when entering the study, and being African American were associated with an increased risk with T. vaginalis. Moreover, the

Table 2.3: STI data: Conditional covariate effects estimations, posterior mean and 95% credible interval are provided

	CT	GC	TV
Age when enter the study	0.0275 (-0.1255, 0.1769)	-0.0185 (-0.1778, 0.1442)	0.1774 (0.061, 0.3519)
number of partners	-0.0183 (-0.1762, 0.1414)	0.0744 (-0.0864, 0.2327)	0.1269 (-0.0360, 0.2882)
Age at first intercourse	-0.0751 (-0.2279, 0.0703)	-0.2136 (-0.1918, 0.1423)	-0.2843 (-0.4639, -0.1130)
Race	0.1139 (-0.0230, 0.2511)	0.0533 (-0.1013, 0.2148)	0.2890 (0.1048, 0.4794)
Infection history	0.2315 (0.0889, 0.3766)	0.3017 (0.1384, 0.4715)	0.1977 (0.0408, 0.3523)

Table 2.4: STI data: Estimation results for posterior mean and 95% credible interval of ρ, κ and τ are provided

	Mean	Std.	95%CI
σ^2	0.259	0.066	(0.1531, 0.3878)
ρ	0.195	0.039	(0.1269, 0.2677)
κ	0.131	0.027	(0.0848, 0.1803)
τ	0.131	0.027	(0.0848, 0.1803)

proposed method is capable to quantify the statistical association among these first infection times. An estimate of the posterior means and 95% credible intervals for Spearman's correlation coefficient, median concordance and Kendall's concordance τ between the three infection times is listed in table 2.4. The estimated $\rho_s = 0.195$, $\kappa = 0.131$ and $\tau = 0.131$ suggests that there is a weak association between the three failure times.

2.6 Discussion

In this chapter, a novel normal-frailty multivariate probit model is proposed for analyzing multivariate interval-censored survival data. This semiparametric probit model provides an attractive alternative to the proportional hazards or the proportional odds

model. The proposed model enjoys several appealing properties. First, this model is semiparametric since the nondecreasing function α is unspecified. Second, the conditional CDF and the marginal CDF of the failure time belong to the same family. Third, the conditional covariate effects given frailty are proportional to the marginal covariate effects. Fourth, the association among multiple correlated failure times can be summarized by three nonparametric association measures in simple and explicit form.

We developed a fully Bayesian method for analyzing the multivariate interval-censored data. Our approach makes use of monotone splines representation to approximate the unknown conditional cumulative baseline hazard function and allows one to estimate the regression parameters and spline coefficients jointly. The derivation of the algorithm is based on data augmentation and all the parameters can be updated in standard formulas. The proposed Gibbs sampler has great computational advantages over the existing Bayesian methods in that it does not require imputing the unobserved failure times or contain complicated Metropolis-Hastings or adaptive rejection Metropolis sampling steps, and all the sampling steps are straightforward and enjoys fast convergence. Through simulation studies, it has been shown that the proposed method can robustly and efficiently estimate all the regression parameters, spline coefficients and the normal frailty variance parameter.

Chapter 3

Regression analysis of interval-censored failure time data under multivariate probit model with arbitrary correlations

Summary: Multivariate interval-censored data are frequently encountered in many applications. In medical studies, several infections caused by a certain disease may co-exist and these infections are correlated. Sometimes multiple diseases are observed within an individual and the responses are correlated. A joint analysis of this multivariate data with the consideration of the association structure is in great demand, and ignoring such underlying correlated data structure can lead to inaccurate estimations. Multivariate probit model, which was introduced by Ashford and Snowden (1970), is most widely accepted for studying multivariate binary responses. Inspired by the applications of this model, we proposed a new semiparametric multivariate probit model to study correlated failure time data. This new model allows the correlations between different failure times to be arbitrary. One of the challenges with multivariate probit model is the difficulty on likelihood computation, as it is obtained by intergrating over a multidimensional constrained space of latent variables. Another difficulty arises on estimating the covariance matrix efficiently. In this chapter, we develop a parameter-extended data augmentation Gibbs sampling algorithm under multivariate probit model, which can be applied for estimating the covariate effects and correlation matrix jointly.

3.1 Introduction

3.1.1 Literature Review and Computational Challenges

Motivated by the data gathered in STI study by Tu et al. (2009), we try to build model on correlated time-to-event data, when the responses are not observed exactly, but can only be determined to lie in an interval obtained from a sequence of observation times. Considerable amount of studies has been developed on estimation of regression coefficients and survival functions for multivariate censored data for the past decade (See, Hougaard, 2000). In these studies, multiple models such as the proportional hazards (PH), the proportional odds (PO), additive hazards (AH) and accelerated failure time (AFT) models (Hanson and Johnson, 2004) have been discussed either from a marginal approach or a frailty approach.

The probit model has been widely studied in generalized linear models, but it is rarely seen in the field of survival analysis. Dunson and Dinse (2002) developed a normal frailty probit model for case I interval-censored data. Lin and Wang (2010) proposed a semiparametric probit model, which serves as an alternative to the PH, PO, AH, and AFT models, and developed a novel Bayesian approach for analyzing univariate interval-censored data. Liu and Qin (2018) studied the maximum likelihood estimations for univariate and bivariate current-status data under the semiparametric probit models. Chin et al. (2018) applied multivariate probit models with panel data. Du and Sun (2019) developed a semiparametric probit model for informative current status data. Wu and Wang (2019) studied clustered interval-censored data by proposing a semiparametric frailty probit regression model. Shiboski (1998) pointed out that the PH, PO and probit models are special cases of the generalized linear models. Motivated by Lin and Wang (2010), we developed a new semiparametric multivariate probit model onto the multivariate interval-censored failure time data. The transformed baseline cumulative distribution function is approximated

through linear combinations of monotone splines. There are several computational challenges under this model. As discussed in Chapter 2, the first computational challenge of estimating the multivariate probit model, given that its likelihood involves integrating over a multidimensional constrained space of latent variables, significantly limits its application in practice. Thus, the MVP models are not as popular as the Cox PH models, the PO mdels and AFT models, etc. in published survival analysis literatures, and its potentials have not been fully studied yet.

In many medical studies, scientists are not only interested in estimating the covariate effects on the correlated outcomes, but would also like to explore and study the association structure between the correlated responses. In our multivariate probit model, this association is modeled through the correlation matrix of $\boldsymbol{\epsilon}$. However, sampling the correlation matrix is an another challenge under the multivaraite probit model, because the scale parameters, i.e., the diagonal elements of the matrix, are fixed. From the Bayesian respective, a prior needs to be placed on correlation matrix R directly to calculate the posterior distributions. However, there does not exist a conjugate prior for the correlation matrix. Therefore, the posterior computation is very challenging.

There are some existing literature on studying how to sample from the correlation matrix under multivariate probit model for binary data. Barnard et al. (2000) adopted the Griddy Gibbs sampler (Ritter and Tanner, 1992) under a hierarchical shrinkage model to sample the components of the correlation matrix one by one at each time. This approach is time consuming, especially when the dimension is high. Chib and Greenberg (1998) developed a more efficient Metropolis-Hastings Random Walk algorithm to sample the correlation matrix by drawing the correlation coefficients in blocks. However, this algorithm has the problem of slow mixing in high dimensions and it cannot guarantee the resulting correlation matrix to be positive definite.

The method of parameter expansion together with data augmentation (PX-DA) enjoys a high popularity in recent years in the MVP model literature. As discussed by Liu and Wu (1999), PX-DA can be selected as an alternative approach when estimating the correlation matrix. This type of parameter expanded data augmentation algorithm is proved to be useful for accelerating Gibbs sampling algorithms and is closely related to reparameterization techniques. The idea behind this approach is to expand R into a less constrained covariance matrix, say $\Sigma = DRD$, and then update this covariance matrix before projecting it back to a correlation matrix. Liu (2001) discussed simulation of correlation matrix through PX-DA under the multivariate probit model by relaxing the correlation matrix R back to an unconstrained covariance matrix and borrowing the scales from the latent variables. One restriction for his method is the prior for R has to be Jeffrey's prior. Instead of using a marginal prior for R, Zhang et al. (2006) demonstrated a parameter-extended Metropolis-Hastings algorithm for sampling from the posterior distribution of a correlation matrix by applying a joint prior derived from an inverse Wishart distribuion of $\Sigma = DRD$. Liu and Daniels (2006) extended this approach by adopting a two-stage parameter expanded re-parameterization and Metropolis-Hastings algorithm. Under their algorithm, they first draw all elements of R simultaneously by drawing a covariance matrix from an inverse Wishart distribution, and then translating it back to a correlation matrix through a reduction function and accepting it based on a Metropolis-Hastings acceptance probability. Talhouk et al. (2012) proposed an efficient Markov Chain Monte Carlo algorithm relying on a parameter expansion scheme to sample from the resulting posterior distributions. Their method allows one to update the correlation matrix within a simple Gibbs sampling framework and make inference in the multivariate probit model. Knowing that the inverse Wishart distribution on Σ is a conjugate prior in the parameter expanded model, posterior sampling of Σ can be accomplished by Gibbs sampling from the full conditional easily. A realisation of R is then achieved

by normalizing Σ. Similar discussions by using different priors on R can also be found in Lawrence et al. (2008). Chin et al. (2018) proposed an efficient data augmentation for multivariate probit models with panel data. A comprehensive discussion about data augmentation methods can be found in Dyk and Meng (2001).

In this chapter, we propose a new multivariate probit model for analyzing correlated interval censored failure time data. Our proposed model is better than those shared frailty proportional hazard or proportional odds models, and also the normal frailty multivariate probit model in chapter 2. The new multivariate probit model here can provide marginal covariate effect estimates directly, and it enjoys the advantage of allowing different pairs of failure times to have different correlations. Under this model, the pairwise statistical associations can be quantified by three nonparametric measures in explicit simple forms. An efficient Bayesian approach is developed to estimate the regression parameters, the baseline survival function and the pairwise Pearson's correlation matrix jointly.

3.1.2 Outline

In this chapter, we will provide details for the proposed Bayesian approach under the new multivariate probit model. The research goal is to find a general framework that can estimate the covariate effects, the baseline survival function and the correlations jointly.

This chapter will be structured as follows:

Section 3.2 introduces the notations, the proposed model and its properties. Section 3.3 presents the details of the proposed approach, including the application of monotone splines to approximate the unknown function, a parameter expansion data augmentation procedure, and a fully developed Gibbs sampler for posterior computation. A summary of the algorithm is included at the end of this section. Section 3.4 evaluates the performance of the proposed approach through extensive simula-

tion studies. Two illustrative real life examples are provided in section 3.5. Some discussions and concluding remarks are made in section 3.6.

3.2 Model and Properties

The multivariate probit model (MVP) in generalized linear models is a generalization of the probit model for studying multiple correlated binary outcomes jointly. Suppose we have n subjects. For each subject i, $i = 1, \ldots, n$, let $\boldsymbol{x}_i = (x_{i1} x_{i2} \ldots x_{ip})'$ denotes the $p \times 1$ vector of covariates. And the covariate matrix can be represented by $\boldsymbol{X}$. Define $\boldsymbol{\beta}_j = (\beta_{j1} \ldots \beta_{jp})'$ as the corresponding unknown covariate coefficient vector for the jth event of interest, $j = 1, 2, \ldots, J$. The covariate coefficient matrix is denoted by $\boldsymbol{\beta}$.

$$\underset{n\times p}{\boldsymbol{X}} = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix} \text{ and } \underset{p\times J}{\boldsymbol{\beta}} = \begin{pmatrix} \beta_{11} & \ldots & \beta_{1J} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \ldots & \beta_{pJ} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1' \\ \vdots \\ \boldsymbol{\beta}_J' \end{pmatrix}.$$

Usually in the multivariate probit model, the response variables are binary. Let $\boldsymbol{Y_i} = (Y_{i1}, \ldots Y_{iJ})'$ denotes the collection of observed binary 0/1 responses on the ith subject. The MVP model assumes that given a set of explanatory variables, the multivariate response is an observed indicator of some underlying Gaussian latent variables fall within certain intervals. And the Gaussian latent variables are correlated with a covariance matrix Σ. Then the probability that $\boldsymbol{Y_i} = y_i$, is given by

$$P(\boldsymbol{Y_i} = y_i | \boldsymbol{\beta}, \Sigma) = \int_{A_{iJ}} \cdots \int_{A_{i1}} \phi_J(t|0, \Sigma)\, dt,$$

where $\phi_J(t|0, \Sigma)$ is the density of a J-variate normal distribution with mean vector 0 and covariance matrix Σ, and A_{ij} is the interval taking the following form:

$$A_{ij} = \begin{cases} (-\infty, x_i'\beta_j) & \text{if} \quad y_{ij} = 1 \\ (x_i'\beta_j, \infty) & \text{if} \quad y_{ij} = 0 \end{cases}$$

Thus, the marginal model for Y_{ij} for $i = 1, \dots n,\ j = 1, \dots J$ has the following form,

$$P(Y_{ij} = y_{ij}) = \begin{cases} \Phi(x_i'\beta_j) & \text{if} \quad y_{ij} = 1 \\ \Phi(-x_i'\beta_j) & \text{if} \quad y_{ij} = 0 \end{cases}$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution.

3.2.1 The proposed model

In the multivariate time to event analysis, the binary response vector $\boldsymbol{Y_i}$ can be interpreted as whether the set of events of interest happened or not on the ith subject, and the time to the events of interest is our response outcome. The new proposed model has the following form:

$$\underset{J\times 1}{\boldsymbol{\alpha}(T_i)} = \begin{pmatrix} \alpha_1(T_{i1}) \\ \alpha_2(T_{i2}) \\ \vdots \\ \alpha_J(T_{iJ}) \end{pmatrix} = -\boldsymbol{\beta}'\boldsymbol{x}_i + \boldsymbol{\epsilon}_i\ , \tag{3.1}$$

where $\alpha_j(\cdot)$ is an unknown nondecreasing function with $\alpha_j(0) = -\infty$ and $\alpha_j(\infty) = \infty$, for $j = 1, \dots J$. It indicates that our proposed model is a semiparametric model. To further illustrate the correlated structure under this model, the error terms $\boldsymbol{\epsilon}_i = (\epsilon_{i1} \dots \epsilon_{iJ})' \sim N_J(0, \Sigma)$. The distribution $N_J(0, \Sigma)$ is a J-variate normal distribution with mean vector 0 and $J \times J$ covariance matrix Σ.

$$\underset{J\times 1}{\boldsymbol{\epsilon}_i} = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{iJ} \end{pmatrix} \sim N_J(0, \Sigma), \text{ where } \underset{J\times J}{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1J}\sigma_1\sigma_J \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \dots & \rho_{2J}\sigma_2\sigma_J \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{J1}\sigma_J\sigma_1 & \rho_{J2}\sigma_J\sigma_2 & \dots & \sigma_J^2 \end{pmatrix},$$

with ρ_{jl}, $j, l = 1 \dots J$ as the Pearson's correlation coefficient between different pairs of failure time, σ_j is the standard deviation of transformed failure time j.

As specified by Chib and Greenberg (1998), it is important to restrict the covariance matrix Σ to be a correlation matrix for identifiability purpose. In the multivariate probit model, the unknown parameters $(\boldsymbol{\beta}, \Sigma)$ are not identifiable. Let's say, we have an alternative parameterisation as $(\boldsymbol{\theta}, \Omega)$, then it could be easily seen that the likelihood of $\mathbf{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \Sigma$ is the same as $\mathbf{Y}|\boldsymbol{X}, \boldsymbol{\theta}, \Omega$, with $\beta_j = c_{jj}^{-1/2}\theta_j$, $\Sigma = C\Omega C'$ and $C = \text{diag}\{c_{11}^{-1/2}, \ldots, c_{JJ}^{-1/2}\}$. Therefore, to avoid the problem of identifiability, we followed the idea of Chib and Greenberg (1998) by imposing the restriction of using correlation matrix instead of covariance matrix. The correlation matrix R is shown below:

$$\underset{J\times J}{R} = \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1J} \\ \rho_{21} & 1 & \ldots & \rho_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{J1} & \rho_{J2} & \ldots & 1 \end{pmatrix}.$$

Therefore, we have $\boldsymbol{\epsilon_i} \sim N_J(0, R)$ instead. ρ_{jl} in R for $j, l = 1 \ldots J$ is the Pearson's correlation coefficients between different pairs of failure time. Our proposed model allows the Pearson's correlations to be different from each other and it can take both positive and negative values. However, this correlation matrix also introduces more constraints since it requires that the diagonal elements to be fixed and can only take the value 1, and the off-diagonal elements are between -1 and 1.

3.2.2 Model Properties

Marginal distribution and marginal effect

Now let $F_j(\cdot|\boldsymbol{x_i})$ denotes the marginal cumulative distribution function (CDF) of the failure time of interest for the jth event given the covariate vector $\boldsymbol{x_i}$. The multivariate probit model specifies the marginal cumulative distribution function of T_{ij}, where T_{ij} is the jth failure time for the ith subject, in the following form:

$$F_j(t|\boldsymbol{x_i}) = \Phi\{\alpha_j(t) + \boldsymbol{x_i'}\boldsymbol{\beta}_j\}, \ \forall t \in (0, \infty). \tag{3.2}$$

where $\Phi(\cdot)$ is the CDF of a standard normal random variable, $\alpha_j(\cdot)$ is an unknown nondecreasing function with $\alpha_j(0) = -\infty$ and $\alpha_j(\infty) = \infty$, $j = 1, \ldots J$. This result implies that the failure time T_{ij} follows a marginal semiparametric Probit model (Lin and Wang, 2010). We can rewrite the model at the subject level as follows,

$$\alpha_j(T_{ij}) = -\boldsymbol{x_i'\beta_j} + \epsilon_{ij}, \text{ with } \epsilon_{ij} \sim N(0,1). \tag{3.3}$$

To see the equivalence between (3.2) and (3.3), we have $P(T_{ij} \leq t|\boldsymbol{x}_i) = P\{\alpha_j(T_{ij}) \leq \alpha_j(t)|\boldsymbol{x}_i\} = P(\epsilon_{ij} \leq \alpha_j(t) + \boldsymbol{x_i'\beta}_j|\boldsymbol{x}_i) = \Phi\{\alpha_j(t) + \boldsymbol{x_i'\beta}_j\}$. Based on equation (3.2), one can write $\alpha_j(t) + \boldsymbol{x_i'\beta}_j = \Phi^{-1}(F_j(t|\boldsymbol{x_i}))$, with the right side being the inverse-probit transformed probability of the failure of interest, where Φ^{-1} is the inverse function of Φ. Thus the interpretation of β_{jp}, the pth element of $\boldsymbol{\beta}_j$ corresponding to the pth covariate x_p, can be given as the change in the inverse-probit transformed probability of the failure of interest due to one unit increase in x_p, while keeping all other covariates at the fixed levels.

Multiple Events Association

The dependence among multiple events of interest is modeled by the Pearson's correlation matrix R in model (3.1). As discussed in section 2.2.2, the three nonparametric measures for quantifying the statistical association between multiple events: Spearman's rank correlation coefficient ρ_s, median concordance κ and kendall's τ (Kruskal, 1958 and Hougaard, 2000) can also be applied under our proposed model (3.1). The multivariate probit model provides closed-form expressions for the pairwise statistical associations between correlated failure times in terms of these three measures in the following theorem,

Theorem 3.2.1. *The pairwise correlations for multivariate survival data under the general MVP model* (3.1) *is characterized by Spearman's correlation coefficient* ρ_s*, median concordance* κ *and Kendall's* τ *as follows, with the Pearson's correlation*

denoted by ρ_{jl}, *for* $j, l = 1, \ldots J$.

$$\rho^s_{jl} = 6\pi^{-1} sin^{-1}(\rho_{jl}/2), \tag{3.4}$$

$$\kappa_{jl} = 2\pi^{-1} sin^{-1}(\rho_{jl}), \tag{3.5}$$

$$\tau_{jl} = 2\pi^{-1} sin^{-1}(\rho_{jl}). \tag{3.6}$$

The proof is similar as the one in the Appendix A, mainly based on the relationship among Pearson's correlation coefficient, Spearman's correlation coefficient, median concordance and Kendall's τ for multivariate normal distribution (Kruskal, 1958). This theorem is promising as it provides explicit expression of measures to quantify the pairwise correlations. They are nonparametric measures so that no specific forms of the correlated failure time distributions are required. The values for these three measures all ranging between -1 and 1, with positive (negative) values representing a positive (negative) relationship. Their magnitude measures the degree of the correlation, a larger magnitude indicating a stronger correlation. A value of zero indicates that no association exists between the failure events.

3.3 The Proposed Method

3.3.1 Data and likelihood

Suppose there are n subjects in our study. It is assumed that conditional on the covariates, the failure times is independent of the observation process. This assumption is quite common in survival analysis literature studying interval-censored data. We consider case II interval-censored data in our model, with the observed data $\mathcal{D} = \{(L_{ij}, R_{ij}], \boldsymbol{x}_i\}$, where $(L_{ij}, R_{ij}]$ is the observed time interval for T_{ij}. To be specific, $L_{ij} = 0$ indicates that the jth failure time for the ith subject is left censored and $R_{ij} = \infty$ indicates the case of right-censoring. We use the indicators δ_{ij1}, δ_{ij2} and δ_{ij3} to denote left-, interval-, and right-censored data respectively. Note that these

censoring indicators subject to the constraint $\delta_{ij1} + \delta_{ij2} + \delta_{ij3} = 1$. From model (3.1), we have $\boldsymbol{\alpha}(T_i) \sim N_J(-\boldsymbol{\beta}'\boldsymbol{x}_i, R)$. It is the same as follows,

$$\underset{J\times 1}{\boldsymbol{\epsilon}_i} = \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{iJ} \end{pmatrix} = \boldsymbol{\beta}'\boldsymbol{x}_i + \boldsymbol{\alpha}(T_i) \sim N_J(0, R).$$

Then the likelihood can be written as

$$\underbrace{P\Big(T_{ij} \in (L_{ij}, R_{ij}) \mid \boldsymbol{x}_i\Big)}_{i=1,\ldots,n \;\; j=1,\ldots J} = \prod_{i=1}^{n} \int \cdots \int_{\boldsymbol{\epsilon}_i \in A_i} \phi_J(\boldsymbol{\epsilon}_i | \mathbf{0}, R)\, d\boldsymbol{\epsilon}_i. \tag{3.7}$$

where $A_i = [\alpha_1(L_{i1}) + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1, \alpha_1(R_{i1}) + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1] \times \cdots \times [\alpha_J(L_{iJ}) + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_J, \alpha_J(R_{iJ}) + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_J]$ and $\phi_J(\cdot)$ is the probability distribution function (pdf) for J-variate normal distribution.

From (3.7), let $\epsilon_{ij}^* = \epsilon_{ij} - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_j - \alpha_j(t_{ij})$ with $t_{ij} = R_{ij}1(\delta_{ij1} = 1) + L_{ij}1(\delta_{ij1} = 0)$ (See, Lin and Wang, 2010). Therefore, when the failure time is left-censored, t_{ij} equals to the right observation time point and when it is interval-censored or right-censored, t_{ij} equals to its left point of observation interval. Note that $\epsilon_{ij}^* \in \Big(\alpha_j(L_{ij}) - \alpha_j(t_{ij}), \alpha_j(R_{ij}) - \alpha_j(t_{ij})\Big)$ and $\boldsymbol{\epsilon}_i^* = (\epsilon_{i1}^*, \cdots, \epsilon_{iJ}^*)' \sim N(-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{\alpha}(t_i), R)$, with $t_i = (t_{i1} \ldots t_{iJ})'$. Thus, equation (3.7) is equvilent to

$$L_{obs} = \prod_{i=1}^{n} \int \cdots \int_{\boldsymbol{\epsilon}_i^* \in A_i^*} \phi_J(\boldsymbol{\epsilon}_i^* | -\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i - \boldsymbol{\alpha}(t_i), R)\, d\boldsymbol{\epsilon}_i^*, \tag{3.8}$$

where $A_i^* = [\alpha_J(L_{iJ}) - \alpha_J(t_{iJ}), \alpha_J(R_{iJ}) - \alpha_J(t_{iJ})] \times \cdots \times [\alpha_1(L_{i1}) - \alpha_1(t_{i1}), \alpha_1(R_{i1}) - \alpha_1(t_{i1})]$.

To facilitate Bayesian computation, we consider a data augmentation by defining latent variables,

$$\boldsymbol{Z_i} = -\boldsymbol{\epsilon}_i^* \sim N_J(\boldsymbol{\alpha}(t_i) + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, R).$$

Then the augmented likelihood function can be written as

$$L_{aug} = \prod_{i=1}^{n} \phi_J\Big(\boldsymbol{z_i}|\boldsymbol{\alpha}(t_i) + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, R\Big)\times \\ \prod_{j=1}^{J}\Big\{\delta_{ij1}1(z_{ij}>0) + \delta_{ij2}1\Big(\alpha_j(L_{ij}) - \alpha_j(R_{ij}) < z_{ij} < 0\Big) + \delta_{ij3}1(z_{ij}<0)\Big\}. \tag{3.9}$$

By intergrating out all z_{ij} from (3.9), one obtains the likelihood function (3.7).

3.3.2 Modeling $\alpha(\cdot)$ with monotone splines

The unknown nondecreasing function α_j, $j = 1, \cdots, J$ is difficult to estimate since α_j is infinite-dimensional. As introduced by chapter 2, using splines to model unknown functions is very common in statistics studies and it provides modeling flexibility. By applying monotone splines of Ramsay (1988) for modeling α_j, we only need to estimate a finite number of parameters. Followed by the idea in Lin and Wang (2010), Cai et al. (2011) and Gamage et al. (2018), α_j can be modeled in the following way:

$$\alpha_j(t) = \gamma_{j0} + \sum_{l=1}^{m}\gamma_{jl}b_l(t), \text{ for } j = 1, \cdots, J \tag{3.10}$$

where $\{b_l\}_{l=1}^m$ are monotone I (integrated) spline basis functions, each of which is nondecreasing from 0 to 1. The basis functions do not depend on j, see my argument in chapter 2.3. Here, γ_{j0} is an unconstrained intercept of a monotone spline. $\{\gamma_{jl}\}_{l=1}^m$ are spline basis coefficients, the values are all taken nonnegative such that α_j is nondecreasing. To specify the I spline basis functions, knots and degree need to be identified first. Even though more knots introduces greater flexibility, Ramsay (1988) recommended that a small number of knots should be chosen as large number of knots is unnecessary and takes more computation time. As claimed by Lin and Wang (2010), a moderate number (10 to 30) of equally spaced knots guarantees modeling flexibility and saves computation time for analyzing interval-censored data. As for the degree of I spline basis, we used quadratic splines.

3.3.3 Bayesian Inference in Multivariate Probit model

Inspired by the work on PX-DA method, we extended this idea to multivariate probit model on time to event data analysis. We propose a parameter expanded data augmentation Gibbs sampling algorithm to sample the unknown parameters and the correlation matrix for multivariate interval-censored data jointly. Let $\Omega = (\boldsymbol{\beta}, R, \boldsymbol{\gamma_0}, \boldsymbol{\gamma})$ be the full parameter set, with $\boldsymbol{\gamma_0} = (\gamma_{10} \dots \gamma_{J0})'$ and $\boldsymbol{\gamma} = (\gamma_{jl})_{J\times m}$, $j = 1, \dots J$ and $l = 1, \dots m$. The first step of the algorithm involves sampling $\mathbf{Z} = (\boldsymbol{z_1}, \cdots, \boldsymbol{z_n})'$. For simpler notations later, let $\boldsymbol{x_i^*} = \left(1\ b_1(t_{ij}) \cdots\ b_m(t_{ij})\ \boldsymbol{x_i'}\right)$ and $\boldsymbol{\beta_j^*} = (\gamma_{j0}\ \gamma_{j1} \cdots, \gamma_{jm}\ \boldsymbol{\beta_j})'$, then $\boldsymbol{\alpha}(t) + \boldsymbol{X\beta} = (\boldsymbol{X^*\beta^*})_{n\times J}$, with

$$\boldsymbol{X}^* = \begin{bmatrix} 1 & b_1(t_{11}) & \dots & b_m(t_{11}) & \boldsymbol{x_1'} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & b_1(t_{nJ}) & \dots & b_m(t_{nJ}) & \boldsymbol{x_n'} \end{bmatrix}_{n\times(1+m+p)} \quad \text{and } \boldsymbol{\beta}^* = \begin{bmatrix} \gamma_{10} & \cdots & \gamma_{J0} \\ \vdots & \ddots & \vdots \\ \gamma_{1m} & \cdots & \gamma_{Jm} \\ \boldsymbol{\beta}_1 & \dots & \boldsymbol{\beta}_J \end{bmatrix}_{(1+m+p)\times J}.$$

From section 3.3.1, the latent variable $\boldsymbol{z_i}$ has the distribution:

$$\pi(\boldsymbol{z_i}|\boldsymbol{\beta}^*, R) \sim N_J(\boldsymbol{z_i}|(\boldsymbol{X^*\beta^*})_i', R)1_{C_i}, \tag{3.11}$$

where $(\boldsymbol{X^*\beta^*})_i'$ is the ith row of the matrix $(\boldsymbol{X^*\beta^*})'$ and $C_i = \prod_{j=1}^{J} C_{ij}$, with C_{ij} as the constrained space of z_{ij},

$$C_{ij} = \begin{cases} (0, \infty) & \text{if} \quad \delta_{ij1} = 1 \\ (\alpha_j(L_{ij}) - \alpha_j(R_{ij}), 0) & \text{if} \quad \delta_{ij2} = 1 \\ (-\infty, 0) & \text{if} \quad \delta_{ij3} = 1 \end{cases}$$

Therefore, we can draw $\mathbf{Z}$ from a truncated multivariate Gaussian distribution as in equation (3.11). To sample from this distribution, one can use the method by Geweke (1991) by composing a cycle of Gibbs steps through univariate truncated normal distributions. In each step of this cycle, z_{ij} is drawn from $z_{ij}|\{z_{i,-j}, \Omega\}$, which is a univariate normal distribution truncated to $(0, \infty)$ if $\delta_{ij1} = 1$, $(\alpha_j(L_{ij}) - \alpha_j(R_{ij}), 0)$ if $\delta_{ij2} = 1$ and $(-\infty, 0)$ if $\delta_{ij3} = 1$. The details are given in Appendix B (B.1).

Given the latent variables $\mathbf{Z}$ sampled from the truncated multivariate Gaussian distribution, we would like to study how to sample the correlation matrix next. Unfortunately, sampling the correlation matrix in MCMC algorithms can be problematic. First, the correlation matrix has to be positive definite and it has the restriction that the diagonal elements need to be fixed at 1. In addition, the number of unknown elements in the correlation matrix increases quadratically with the dimension J. These facts make simulating a correlation matrix difficult.

An instinctive way to solve this problem is to relax R into a less constrained space, say $\Sigma = DRD$ and update Σ instead. We will adopt the method of data augmentation parameter expansion and follow the idea by Talhouk et al. (2012) to propose a new approach for sampling correlation matrix under multivariate probit model.

Let $\boldsymbol{W} = \mathbf{Z}D$, where D is the expansion parameter and it is a $J \times J$ diagonal matrix with $d_{jj} > 0$. Then $\pi(\boldsymbol{W}|\boldsymbol{\beta}^*, R, D) \sim N_{n,J}(W; \boldsymbol{X}^*\boldsymbol{\beta}^* D, DRD)$. Now define a latent parameter $\theta = (\theta_1, \cdots, \theta_J)$ with $\theta_j = \frac{r^{jj}}{2d_j^2}$, where r^{jj} is the jth diagonal element of R^{-1} and d_j is the jth diagonal element of D. The latent parameter θ is defined in such a way that the resulting posterior distribution will be easily to sample from. More details will be discussed later.

The basic procedure of PX-DA algorithm for sampling can be described as in algorithm 3.1. Here $|J : \mathbf{Z} \to \boldsymbol{W}|$ is the Jacobian transformation from $\mathbf{Z}$ to $\boldsymbol{W}$. Let $\boldsymbol{W_i}$ and $(\boldsymbol{X}^*\boldsymbol{\beta}^* D)_i$ represent the ith row of $\boldsymbol{W}$ and $(\boldsymbol{X}^*\boldsymbol{\beta}^* D)$ respectively. Then,

$$
\begin{aligned}
p(\mathbf{Z}|\boldsymbol{\beta}^*, R)|J : \mathbf{Z} \to \boldsymbol{W}| &\sim N_{n,J}(W; \boldsymbol{X}^*\boldsymbol{\beta}^* D, DRD) \\
&\propto |DRD|^{-\frac{n}{2}} \exp\Bigg[-\frac{1}{2} \sum_{i=1}^{n} \Big(\boldsymbol{W_i} - (\boldsymbol{X}^*\boldsymbol{\beta}^* D)_i\Big)^{\mathrm{T}} \\
&\qquad (DRD)^{-1}\Big(\boldsymbol{W_i} - (\boldsymbol{X}^*\boldsymbol{\beta}^* D)_i\Big)\Bigg].
\end{aligned}
$$

Algorithm 3.1 PX-DA Algorithm

At iteration i,

1: Draw $\mathbf{Z}$ from the truncated multivariate normal distribution (3.11) and compute $\boldsymbol{W}$;

2: Draw $(\boldsymbol{\beta}^*, R, \theta)$ jointly conditional on the latent data $\mathbf{Z}$,

$$\boldsymbol{\beta}^*, R, \theta | \boldsymbol{W} \sim \underbrace{p(\mathbf{Z}|\boldsymbol{\beta}^*, R)|J : \mathbf{Z} \to \boldsymbol{W}|}_{p(\boldsymbol{W}|\boldsymbol{\beta}^*, R, \theta)} \underbrace{p(\theta|\boldsymbol{\beta}^*, R)\pi(R, \boldsymbol{\beta}^*)}_{\text{prior}}.$$

Now define

$$\Sigma = DRD, \tag{3.12}$$

$$E = (\mathbf{Z} - \boldsymbol{X}^*\boldsymbol{\beta}^*)D. \tag{3.13}$$

Therefore, the transformed likelihood under the parameter expansion can be written as:

$$p(\mathbf{Z}|\boldsymbol{\beta}^*, R)|J : \mathbf{Z} \to \boldsymbol{W}| \propto |\Sigma|^{-n/2} \exp\{tr(\Sigma^{-1}E^{\mathrm{T}}E)\}. \tag{3.14}$$

This transformed likelihood 3.14 enjoys the convenience of deriving posterior distributions that are easily sampled from.

Prior Specifications

Assume that the priors for R and $\boldsymbol{\beta}^*$ are independent, i.e., $\pi(R, \boldsymbol{\beta}^*) = \pi(R)\pi(\boldsymbol{\beta}^*)$. It is not easy and strightforward to find a joint prior on $\boldsymbol{\beta}^*$, which is a combination of $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. However, it is equivalent to consider $\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma}_0)\pi(\boldsymbol{\gamma})$ since $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are independent. Hence, we introduce the priors for $\Omega = (\boldsymbol{\beta}, R, \boldsymbol{\gamma_0}, \boldsymbol{\gamma})$ one by one.

1. *Prior for $\boldsymbol{\beta}$*

   We adopt a multivariate Gaussian distribution prior for $\boldsymbol{\beta}$ as

$$\pi(\boldsymbol{\beta}) \propto exp\Big[-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^{\mathrm{T}}\Psi_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\Big]. \tag{3.15}$$

We would like to choose large values for the diagonal elements of Ψ_0 such that the prior on $\boldsymbol{\beta}$ is uninformative. In addition, let $\vec{\beta} = vec(\boldsymbol{\beta})$, representing the $pJ \times 1$ vector that stacks the columns of the $J \times p$ regression coefficient matrix $\boldsymbol{\beta}$, then we have

$$\pi(\vec{\beta}) \sim N_{pJ}(\vec{\beta_0}, \Psi_0 \otimes I_J). \tag{3.16}$$

2. *Prior for R*

Since there's no conjugate prior available for sampling R, the Bayesian inference on correlation matrix can be difficult. There are some discussions on the choices of prior for the correlation matrix R:

- (P1) Multivariate truncated normal distribution prior.
- (P2) Jeffrey's prior: $\pi(R) \propto |R|^{-\frac{p+1}{2}}$.
- (P3) Jointly uniform prior.
- (P4) Marginally uniform prior.
- (P5) Hierarchical prior on the partial correlation matrix.

The application of (P1) was studied by Chib and Greenberg (1998), and they proposed a random walk Metropolis-Hastings algorithm with a multivariate t proposal density to sample each r_{ij} of R in blocks. The resulting proposal cannot be guaranteed to be a correlation matrix. Moreover, as with random walk algorithms in general, this approach has slow exploration of parameter space, and tuning the parameters of proposal distribution requires finding a mode of the posterior distribution and the observed Fisher information for each iteration, leading to high computation burden. Barnard et al. (2000) used the Griddy Gibbs approach based on this prior, and by solving an equation to decide the support for r_{ij} first, this approach guarantees the resulting correlation matrix to be valid. However, the authors also pointed out that this prior is

inefficient due to its tendency to push marginal correlations to zero in high dimensions. Liu and Daniels (2006) developed an MH accept-reject algorithm for the proposed correlation matrix by using this prior as well. The posterior analysis resulting from this multivariate truncated normal prior is difficult to conduct.

Liu (2001) and Lawrence et al. (2008) avoided this sampling difficulty by using (P2). However, this prior has the problem of being improper. The posterior distribution may not be well-defined and it has been proved that improper priors on covariance matrices is informative and tends to push marginal correlations towards the bounds (-1 and 1). See, Rossi et al., 2005.

Barnard et al. (2000) suggested a uniform prior (P3) over all correlation matrices in R^J, where R^J is the correlation matrix space. The uniform prior on R^J is:

$$p(R) \propto 1, R \in R^J.$$

This prior is also a special case of the LKJ prior of Lewandowski et al. (2009) with unit shape parameter. It has a greater density around zero for each r_{jl} in high dimensions and thus is highly informative.

(P4) was also proposed by Barnard et al. (2000), by decomposing a covariance matrix into diagonal matrices of standard deviations and correlation matrix to obtain a prior distribution on R as

$$\pi(R) \propto |R|^{\frac{J(J-1)}{2}-1}\Big(\Pi_j|R_{jj}|\Big)^{-\frac{(J+1)}{2}}. \tag{3.17}$$

where J is the number of events and R_{jj} denotes the j-th principal submatrix of R. Even though (3.17) is not easy to sample from directly, when combined with PX-DA strategy, it can be proved that it is equivalently to sample from a standard inverse Wishart distribution and project it back to a correlation matrix. The proof can be found in Appendix B.2. This marginally uniform

prior enjoys several nice properties: First, it is a proper prior and we can get the normalizing constant. Second, The marginal densities for each r_{jl} follows a uniform distribution on $[-1, 1]$. Despite a jointly uniform prior tends to favor correlations to 0, the marginally uniform prior is uninformative. Third, further studies on model selections can be conducted.

Wong and Kohn (2003) proposed a prior for the covariance matrix of Gaussian data that allows the off-diagonal elements of its inverse to be identically zero. A hierarchical prior (P5) was built for the partial correlation matrix. Pitt et al. (2006) used this prior to conduct Bayesian inference for Gaussian copula regression models. This prior is more complicated to conduct posterior analysis.

Due to the nice properties of marginally uniform prior and the possibility of sampling the correlation matrix by applying parameter expansion data augmentation strategy, we will use (P4) as our prior for the correlation matrix R.

3. *Prior for γ's*

   A normal prior $N(m_{j0}, v_{j0}^{-1})$ is assigned for the unconstrained γ_{j0}'s. Indepedent exponential priors $Exp(\eta_j)$ are assigned for all the nonnegative $\{\gamma_{jl}\}_{l=1}^{m}$. A further prior $\mathcal{G}a(a_{j\eta}, b_{j\eta})$ is given for the hyper parameter η_j. These prior assignments for basis coefficients have the advantage of selecting basis functions by shrinking small spline coefficients towards zero and punishing large spline coefficients. In this way, it can help prevent overfitting problems. See, Lin and Wang (2010), Wang and Dunson (2011), Cai et al. (2011).

Data Transformation and Posterior Sampling

With all the above prior distributions and the expanded likelihood specified, an efficient Gibbs sampling is developed. For the joint prior $\pi(\boldsymbol{\beta}, R, \theta, \boldsymbol{\gamma}_0, \boldsymbol{\gamma})$, it is equal

to $\pi(\theta|R)\pi(R)\pi(\boldsymbol{\beta})\pi(\boldsymbol{\gamma}_0)\pi(\boldsymbol{\gamma})$. Barnard et al. (2000) proved that sampling Σ from a standard inverse Wishart distribution with the degree of freedom $d = J + 1$ is equivalent to sampling from the prior $\pi(\theta|R)\pi(R)$. There are several ways of writing the probability distribution function for inverse Wishart distribution, we adopt the one used by Barnard et al. (2000). Let $\Sigma \sim IW(d, I_J)$, where d is the degree of freedom, then:

$$f_J(\Sigma|d) \propto |\Sigma|^{-\frac{1}{2}(d+J+1)} \exp\Big(-\frac{1}{2} tr(\Sigma^{-1})\Big), \tag{3.18}$$

which is the probability distribution function for inverse Wishart distribution used in our approach. By the transformation (3.12) and Jacobian transformation formulas, we have the following:

$$\pi(\Sigma) = \pi(\theta, R) \times |J : \Sigma \to (D, R)| = \pi(\theta|R)\pi(R), \tag{3.19}$$

where $\pi(R)$ is taken as the marginally uniform prior (3.17), and

$$\pi(\theta|R) \sim \mathcal{G}a(\frac{J+1}{2}, 1). \tag{3.20}$$

The pdf of $\mathcal{G}a\Big(\frac{J+1}{2}, 1\Big)$ is given as:

$$f_\theta\Big(\theta \mid \frac{J+1}{2}, 1\Big) = \frac{1}{\Gamma(\frac{J+1}{2})} \theta^{\frac{J+1}{2}-1} \exp(-\theta).$$

Based on the above discussions, we give the following theorem:

Theorem 3.3.1. *By specifying the prior for R as* (3.17) *and a gamma prior* (3.20) *for θ, under the transformation* (3.12)*, simulating R is equivalent to sample Σ first from a standard inverse Wishart distribution with degree of freedom $d = J + 1$ and project it back to the correlation matrix R through $R = D^{-1}\Sigma D^{-1}$.*

The proof of this theorem can be found in the Appendix B.2.

From step 2 of algorithm 3.1, by assuming that $\boldsymbol{\beta}$, R, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\gamma}$ are independent, we combine the transformed likelihood (3.14), the marginally uniform prior on R in

(3.17), the gamma prior on θ, the prior on $\boldsymbol{\beta}$ in (3.16), the normal prior on $\boldsymbol{\gamma}_0$ and the exponential prior on $\boldsymbol{\gamma}$, we have the joint posterior distribution:

$$\begin{aligned}\pi(R,\theta,\boldsymbol{\beta},\boldsymbol{\gamma}_0,\boldsymbol{\gamma}|\boldsymbol{W}) \propto & |\Sigma|^{-\frac{n}{2}} \exp tr(\Sigma^{-1}E^{\mathrm{T}}E) \\ & \times |R|^{\frac{J(J-1)}{2}-1}\Big(\prod_j |R_{jj}|\Big)^{-\frac{(J+1)}{2}} \\ & \times \mathcal{G}a(\frac{J+1}{2},1) \\ & \times N_{pJ}(\vec{\beta_0},\Psi_0 \otimes I_J) \\ & \times \pi(\boldsymbol{\gamma}_0,\boldsymbol{\gamma}). \end{aligned} \tag{3.21}$$

In order to sample from the joint posterior distribution in (3.21), a Gibbs sampling framework is conducted. To sample R, from (3.21) we have

$$\begin{aligned}\pi(R,\theta|\boldsymbol{W},\boldsymbol{\beta},\boldsymbol{\gamma}_0,\boldsymbol{\gamma}) \propto & |\Sigma|^{-\frac{n}{2}} \exp tr(\Sigma^{-1}E^{\mathrm{T}}E) \\ & \times |R|^{\frac{J(J-1)}{2}-1}\Big(\prod_j |R_{jj}|\Big)^{-\frac{(J+1)}{2}} \\ & \times \mathcal{G}a(\frac{J+1}{2},1). \end{aligned} \tag{3.22}$$

By the transformation $\Sigma = DRD$, (3.22) is equivalent to:

$$\begin{aligned}\pi(\Sigma|\boldsymbol{W},\boldsymbol{\beta},\boldsymbol{\gamma}_0,\boldsymbol{\gamma}) &\propto \pi(R,\theta|\boldsymbol{W},\boldsymbol{\beta},\boldsymbol{\gamma}_0,\boldsymbol{\gamma}) \times |J:(D,R)\to\Sigma| \\ &= |\Sigma|^{-\frac{n}{2}} \exp tr(\Sigma^{-1}E^{\mathrm{T}}E) \times |\Sigma|^{-\frac{1}{2}2(J+1)} \times \exp\Big(-\frac{1}{2}tr(\Sigma^{-1})\Big) \\ &= |\Sigma|^{-\frac{1}{2}(d+J+1)} \exp\Big(-\frac{1}{2}tr(\Sigma^{-1}S)\Big). \end{aligned} \tag{3.23}$$

It's clearly to see that Σ comes from an inverse Wishart distribution with $d = n+J+1$ and $S = E^{\mathrm{T}}E$. Therefore, the Gibbs steps to sample R can be summarized as below:

- Draw θ_j from gamma distribution $\mathcal{G}a(\frac{J+1}{2},1)$.
- Compute diagonal matrix D, with the jth element of D as $d_j = \sqrt{\frac{r^{jj}}{2\theta_j}}$, where r^{jj} is the jth diagonal element of R^{-1}.
- Compute $E = (\boldsymbol{Z} - \boldsymbol{X}^*\boldsymbol{\beta}^*)D$.

- Draw Σ from an inverse Wishart distribution, with $\Sigma \sim IW(V, S)$ where $V = n + J + 1$ and $S = E'E$.
- Compute $R = D^{-1}\Sigma D^{-1}$.

Next, we want to derive the posterior distribution for $\boldsymbol{\beta}$. Let $\vec{\mathbf{Z}} = vec(\mathbf{Z}')$, $\vec{\boldsymbol{\alpha}}(t) = vec\Big(\boldsymbol{\alpha}'(t)\Big)$ by stacking the columns of $\mathbf{Z}'$ and $\boldsymbol{\alpha}'(t)$ respectively,

$$\vec{\mathbf{Z}} = vec(\mathbf{Z}') = \begin{bmatrix} z_{11} \\ \vdots \\ z_{1J} \\ z_{21} \\ \vdots \\ z_{2J} \\ \vdots \\ z_{n1} \\ \vdots \\ z_{nJ} \end{bmatrix}_{nJ \times 1}$$

and

$$\vec{\boldsymbol{\alpha}}(t) = vec\Big(\boldsymbol{\alpha}'(t)\Big) = \begin{bmatrix} \alpha_1(t_{11}) \\ \vdots \\ \alpha_J(t_{1J}) \\ \alpha_1(t_{21}) \\ \vdots \\ \alpha_J(t_{2J}) \\ \vdots \\ \alpha_1(t_{n1}) \\ \vdots \\ \alpha_J(t_{nJ}) \end{bmatrix}_{nJ \times 1} .$$

Then the computation for the posterior distribution of $\boldsymbol{\beta}$ can be derived as below:

$$
\begin{aligned}
\pi(\vec{\beta} \mid \vec{\mathbf{Z}}, R) &\propto \pi(\vec{\mathbf{Z}} \mid \vec{\beta}, R) \times \pi(\vec{\beta}) \\
&\propto N_{nJ}\Big(\vec{\mathbf{Z}} \mid (\boldsymbol{X} \otimes I_J)\vec{\beta} + \vec{\boldsymbol{\alpha}}(t), I_n \otimes R\Big) \times N_{pJ}(\vec{\beta} \mid \vec{\beta_0}, \Psi_0 \otimes I_J) \\
&\propto \exp\Big\{ -\frac{1}{2}\Big[\vec{\mathbf{Z}} - \vec{\boldsymbol{\alpha}}(t) - (\boldsymbol{X} \otimes I_J)\vec{\beta}\Big]'(I_n \otimes R)^{-1}\Big[\vec{\mathbf{Z}} - \vec{\boldsymbol{\alpha}}(t) - (\boldsymbol{X} \otimes I_J)\vec{\beta}\Big]\Big\} \\
&\quad \times \exp\Big\{ -\frac{1}{2}(\vec{\beta} - \vec{\beta_0})'(\Psi_0 \otimes I_J)^{-1}(\vec{\beta} - \vec{\beta_0})\Big\} \\
&\propto \exp\Big\{ -\frac{1}{2}\Big[\Big(\vec{\mathbf{Z}} - \vec{\boldsymbol{\alpha}}(t)\Big)'(I_n \otimes R)^{-1}\Big(\vec{\mathbf{Z}} - \vec{\boldsymbol{\alpha}}(t)\Big) + \vec{\beta_0'}(\Psi_0 \otimes I_J)^{-1}\vec{\beta_0} \\
&\quad + \vec{\beta}'(\boldsymbol{X} \otimes I_J)'(I_n \otimes R)^{-1}(\boldsymbol{X} \otimes I_J)\vec{\beta} + \vec{\beta}'(\Psi_0 \otimes I_J)^{-1}\vec{\beta} \\
&\quad - \vec{\beta}'(\boldsymbol{X} \otimes I_J)'(I_n \otimes R)^{-1}\Big(\vec{\mathbf{Z}} - \vec{\boldsymbol{\alpha}}(t)\Big) - \vec{\beta}'(\Psi_0 \otimes I_J)^{-1}\vec{\beta_0} \\
&\quad - \Big(\vec{\mathbf{Z}} - \vec{\boldsymbol{\alpha}}(t)\Big)'(I_n \otimes R)^{-1}(\boldsymbol{X} \otimes I_J)\vec{\beta} - \vec{\beta_0'}(\Psi_0 \otimes I_J)^{-1}\vec{\beta}\Big]\Big\} \\
&\propto \exp\Big\{ -\frac{1}{2}\Big[\vec{\beta}'\Big((\boldsymbol{X} \otimes I_J)'(I_n \otimes R)^{-1}(\boldsymbol{X} \otimes I_J) + (\Psi_0 \otimes I_J)^{-1}\Big)\vec{\beta} \\
&\quad - \vec{\beta}'\Big((\boldsymbol{X} \otimes I_J)'(I_n \otimes R)^{-1}(\vec{\mathbf{Z}} - \vec{\boldsymbol{\alpha}}(t)) + (\Psi_0 \otimes I_J)^{-1}\vec{\beta_0}\Big) \\
&\quad - \Big((\vec{\mathbf{Z}} - \vec{\boldsymbol{\alpha}}(t))'(I_n \otimes R)^{-1}(\boldsymbol{X} \otimes I_J) - \vec{\beta_0'}(\Psi_0 \otimes I_J)^{-1}\Big)\vec{\beta}\Big]\Big\}. \qquad (3.24)
\end{aligned}
$$

From Kronecker product algebra, we have:

$$(\boldsymbol{X} \otimes I_J)'(I_n \otimes R^{-1}) = \boldsymbol{X}' \otimes R^{-1}. \qquad (3.25)$$

and

$$(\boldsymbol{X} \otimes I_J)'(I_n \otimes R^{-1})(\boldsymbol{X} \otimes I_J) = \boldsymbol{X}'\boldsymbol{X} \otimes R^{-1}. \qquad (3.26)$$

By applying (3.25) and (3.26), we have

$$\pi(\vec{\beta} \mid \vec{\mathbf{Z}}, R, \vec{\boldsymbol{\alpha}}) \propto N_{pJ}(\vec{\beta} \mid \tilde{\beta}, \tilde{\psi}). \qquad (3.27)$$

where

$$\tilde{\beta} = \tilde{\psi}\Big[(\boldsymbol{X}' \otimes R^{-1})\Big(\vec{\mathbf{Z}} - \vec{\boldsymbol{\alpha}}(t)\Big) + (\psi_0 \otimes I_J)^{-1}\beta_0\Big]$$

and

$$\tilde{\psi} = \Big[(\boldsymbol{X}'\boldsymbol{X} \otimes R^{-1}) + (\psi_0 \otimes I_J)^{-1}\Big]^{-1}.$$

Next we will derive the posterior distributions for the spline coefficients. The Gibbs sampling steps for γ_{jl} and γ_{j0} can be summarized from their posterior distributions easily.

3.3.4 Algorithm Summary

From the discussions above, the full algorithm is summarized in Algorithm 3.2. The Gibbs sampler is very appealing in that all of the full conditional distributions are standard distributions and are easy to sample from. This property is rarely seen in existing Bayesian methods for analyzing multivariate survival data in the literature. The proposed Gibbs sampler is observed good mixing and fast convergence from our observation.

3.4 Simulation Studies

In this section, we use simulation studies to evaluate the performance of our proposed multivariate probit model. We assume that the covariate x_{i1} is a normal variable with mean 0 and variance 0.25, and covariate x_{i2} is a Bernoulli random variable with the success probability of 0.5. We considered three events of interest and take true $\alpha_1(t) = 1 + t + \log(t)$, $\alpha_2(t) = t^2 + \log(t)$ and $\alpha_3(t) = 1 + t + \log(t)$, respectively. The ture covariate coefficients are $\boldsymbol{\beta_1} = (1, 0)$, $\boldsymbol{\beta_2} = (0, 1)$, $\boldsymbol{\beta_3} = (-1, 1)$. The true correlation matrix is set as

$$R = \begin{bmatrix} 1 & 0.3 & 0.6 \\ 0.3 & 1 & 0 \\ 0.6 & 0 & 1 \end{bmatrix}.$$

The procedure about how we obtain the observed interval (L_{ij}, R_{ij}) for each T_{ij} is the same as the one in section 2.4. The specification of the observation process was chosen so that none of the censoring types dominates the others. Quadratic splines were applied to ensure adequate smoothness of the splines. Equally spaced knots

Algorithm 3.2 Full PX-DA Sampling Scheme in Multivariate Probit for interval-censored data

At iteration i,

1: **Sample latent variables $\boldsymbol{z_i}$** for $i = 1, ..., n$ from a truncated multivariate normal distribution $\pi(\boldsymbol{z_i}|\boldsymbol{\beta^*}, R) \sim N_J(\boldsymbol{z_i}|(\boldsymbol{x^*\beta^*})'_i, R)I_{C_{ij}}$, where

$$X^* = \begin{pmatrix} 1 & b_1(t_{11}) & \dots & b_m(t_{11}) & X_1^T \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & b_1(t_{nJ}) & \dots & b_m(t_{nJ}) & X_n^T \end{pmatrix}$$

and

$$\boldsymbol{\beta^*} = \begin{pmatrix} \gamma_{10} & \gamma_{20} & \dots & \gamma_{J0} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{1m} & \gamma_{2m} & \dots & \gamma_{Jm} \\ \beta_1 & \beta_2 & \dots & \beta_J \end{pmatrix}$$

with $t_{ij} = R_{ij}I_{(\delta_{ij1}=1)} + L_{ij}I_{(\delta_{ij1}=0)}$. And $(\boldsymbol{x^*\beta^*})'_i$ is the ith row of the matrix $(\boldsymbol{x^*\beta^*})'$.

2: **Sample** R,

- Draw θ_j from gamma distribution $\mathcal{G}a(\frac{J+1}{2}, 1)$.
- Compute diagonal matrix D, with the jth element of D as $d_j = \sqrt{\frac{r^{jj}}{2\theta_j}}$, where r^{jj} is the jth diagonal element of R^{-1}.
- Compute $E = (\boldsymbol{Z} - \boldsymbol{X^*\beta^*})D$.
- Draw Σ from an inverse Wishart distribution, with $\Sigma \sim IW(V, S)$ where $V = n + J + 1$ and $S = E'E$.
- Compute $R = D^{-1}\Sigma D^{-1}$.

3: **Sample** γ_{j0} from $N(M_{j0}, W_{j0}^{-1})$, where $W_{j0} = \upsilon_{j0} + nr_{jj}$ and

$$M_{j0} = W_{j0}^{-1}\Bigg[m_{j0}\upsilon_{j0} + \sum_{i=1}^{n}\Bigg[(z_{ij} - \sum_{l=1}^{m}\gamma_{jl}b_l(t_{ij}) - \boldsymbol{x_i\beta_j}) * r_{jj} - \sum_{j'\neq j} r_{j'j}\Big(\gamma_{j'0} - (z_{ij'} - x'_i\beta_{j'} - \sum_{l=1}^{m}\gamma_{j'l}b_l(t_{ij'}))\Big)\Bigg]\Bigg].$$

Here r_{jj} is the jth diagonal element of the correlation matrix R.

4: **Sample all γ_{jl}'s** for $l = 1, 2..., m$ and $j = 1, ..., k$. For each l, let $W_{jl} = \sum_{i=1}^{n} r_{jj} b_l^2(t_{ij})$, where r_{jj} is the j-th diagonal element of R.

- If $W_{jl} = 0$, sample γ_{jl} from $Exp(\eta_j)$.
- If $W_{jl} > 0$, sample γ_{jl} from $N(H_{jl}, W_{jl}^{-1})1(\gamma_{jl} > d_{jl}^*)$, where

$$H_{jl} = W_{jl}^{-1}\left[\sum_{i=1}^{n} b_l(t_{ij})\Big\{r_{jj}[z_{ij} - \gamma_{j0} - \sum_{l' \neq l} \gamma_{jl'} b_{l'}(t_{ij}) - \boldsymbol{x_i'\beta_j}] + \sum_{j' \neq j} r_{j'j}[z_{ij'} - \boldsymbol{x_i'\beta_{j'}} - \gamma_{j'0} - \sum_{l=1}^{m} \gamma_{j'l} b_l(t_{ij'})]\Big\} - \eta_j\right]$$

with $d_{jl}^* = \max(c_{jl}^*, 0)$ and $c_{jl}^* = \max_{\{(i,j):\delta_{ij2}=1\}} \left[\frac{-z_{ij} - \sum_{jl' \neq jl} \gamma_{jl'}\{b_{l'}(R_{ij}) - b_{l'}(L_{ij})\}}{b_l(R_{ij}) - b_l(L_{ij})}\right]$.

5: **Sample** η_j for $j = 1, ..., J$ from $\mathcal{G}a(a_{j\eta} + m, b_{j\eta} + \sum_{l=1}^{m} \gamma_{jl})$.

6: **Sample** $\vec{\beta}$ from $N_{pJ}(\tilde{\boldsymbol{\beta}}, \tilde{\Psi})$, where $\tilde{\Psi} = [(\boldsymbol{X'X} \otimes R^{-1}) + (\Psi_0 \otimes I_J)^{-1}]^{-1}$ and

$$\tilde{\boldsymbol{\beta}} = \tilde{\Psi}\left[(\Psi_0 \otimes I_J)^{-1}\beta_0 + (\boldsymbol{X'X} \otimes R^{-1})(\vec{\boldsymbol{Z}} - \vec{\boldsymbol{\alpha}}(\boldsymbol{t}))\right]$$

with

$$\vec{\beta} = vec(\boldsymbol{\beta'}) = (\beta_{11} \ldots \beta_{J1}\, \beta_{12} \ldots \beta_{J2}\, \ldots \beta_{1p} \ldots \beta_{Jp})'$$
$$\vec{Z} = vec(\boldsymbol{Z'}) = (z_{11} \ldots z_{1J}\, z_{21} \ldots z_{2J}\, \ldots, z_{n1} \ldots z_{nJ})'$$

and

$$\vec{\boldsymbol{\alpha}}(\boldsymbol{t}) = vec(\boldsymbol{\alpha'(t)}) = (\alpha_1(t_{11}) \ldots \alpha_J(t_{1J})\, \alpha_1(t_{21}) \ldots \alpha_J(t_{2J})\, \ldots \alpha_1(t_{n1}) \ldots \alpha_J(t_{nJ}))'$$

Repeat until convergence.

were assigned within the range of the finite endpoints of the observation times for each generated data set. The gap distance between two adjacent knots is set to 0.3. The number of knots is different from data set to data set and takes values from 12 to 26.

We adopted the following prior specifications for the unknown parameters. $m_{10} = -3$, $m_{20} = -4$, $m_{30} = -3$ and $\upsilon_{10} = \upsilon_{20} = \upsilon_{30} = 0.1$, resulting in a normal prior for γ_{j0} with a large variance; a $\mathcal{G}a(1, 1)$ prior for η_j with $a_{j\eta} = b_{j\eta} = 1$; $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\Psi_0 = n(\boldsymbol{X'X})^{-1}$, where $\boldsymbol{X}$ is the covariate matrix. Fast convergence of the proposed Gibbs sampler was observed in the simulation study as the fact that all the

Table 3.1: Performance of the proposed method in the case of using 100 datasets. BIAS denotes the difference between the average of the 100 point estimates and the true value, ESD the average of the estimated standard deviations, SSD the sample standard deviation of the 100 point estimates, and the CP95 the 95% coverage probability.

True	Bias	SSD	ESD	CP95
$\beta_{11} = 1$	0.033	0.155	0.220	0.99
$\beta_{12} = 0$	0.022	0.151	0.159	0.98
$\beta_{21} = 0$	0.006	0.145	0.162	0.97
$\beta_{22} = 1$	0.056	0.164	0.157	0.92
$\beta_{31} = -1$	-0.104	0.162	0.172	0.90
$\beta_{32} = 1$	0.099	0.132	0.155	0.92
$\rho_{12} = .3$	0.022	0.070	0.091	0.97
$\rho_{13} = .6$	-0.072	0.053	0.075	0.93
$\rho_{23} = 0$	-0.011	0.074	0.085	0.98

parameters can be updated by their full conditional distributions in standard forms. A total number of 4000 iterations of MCMC were ran, and the first 1000 iterations were discarded as a burn-in period. The estimation results include the estimated bias (Bias) given by the average of the estimates minus the true value, the sample standard deviation (SSD) of the estimates, the average of the estimated standard deviation (ESD), and the 95% empirical coverage probability (CP). The estimation results for regression coefficients and correlation coefficients are presented in Table 3.1. The estimations for the three nonparametric measures are shown in Table 3.2. The results in these two tables are based on 100 datasets, each with a sample size 300. Table 3.3 and Table 3.4 using the same simulation set ups as above give the results based on 500 datasets, each with a sample size $n = 100$.

The results in Table 3.1 and Table 3.3 indicate that the estimates for the covariate effects from our proposed method are accurate since the bias is small. It is observed that the sample standard deviation and the estimated standard error are quite close. The 95% coverage probability are close to the nominal level 0.95 in all parameter configurations. Beyond the precise estimations in covariate effects, the estimations for pairwise correlations are also accurate, which suggests that our model performs very

Table 3.2: Estimates of associations in the case of using 100 datasets. BIAS denotes the difference between the average of the 100 point estimates and the true value, ESD the average of the estimated standard deviations, SSD the sample standard deviation of the 100 point estimates, and the CP95 the 95% coverage probability.

True	Bias	SSD	ESD	CP95
$\rho^s_{12} = 0.288$	0.020	0.068	0.088	0.97
$\rho^s_{13} = 0.582$	0.072	0.053	0.075	0.93
$\rho^s_{23} = 0$	-0.010	0.071	0.082	0.98
$\kappa_{12} = 0.194$	0.013	0.047	0.061	0.97
$\kappa_{13} = 0.410$	0.054	0.041	0.056	0.93
$\kappa_{23} = 0$	-0.007	0.048	0.054	0.98
$\tau_{12} = 0.194$	0.013	0.047	0.061	0.97
$\tau_{13} = 0.410$	0.054	0.041	0.056	0.93
$\tau_{23} = 0$	-0.007	0.048	0.054	0.98

Table 3.3: Performance of the proposed method in the case of using 500 datasets. BIAS denotes the difference between the average of the 500 point estimates and the true value, ESD the average of the estimated standard deviations, SSD the sample standard deviation of the 500 point estimates, and the CP95 the 95% coverage probability.

True	Bias	SSD	ESD	CP95
$\beta_{11} = 1$	-0.042	0.370	0.385	0.960
$\beta_{12} = 0$	-0.060	0.472	0.296	0.932
$\beta_{21} = 0$	-0.015	0.261	0.275	0.956
$\beta_{22} = 1$	-0.036	0.386	0.284	0.918
$\beta_{31} = -1$	-0.010	0.452	0.314	0.954
$\beta_{32} = 1$	-0.019	0.325	0.282	0.932
$\rho_{12} = .3$	-0.007	0.166	0.152	0.94
$\rho_{13} = .6$	0.085	0.127	0.126	0.91
$\rho_{23} = 0$	-0.022	0.152	0.142	0.95

well in estimations with arbitrary correlations between different events of interest. Table 3.2 and Table 3.4 proved that the model has a good performance in estimating the three nonparametric measures for statistical associations as well.

3.5 Real Data Analysis

3.5.1 STI Data

In this section, we apply the proposed methodology to the Sexually Transmitted Infection (STI) data set. STIs are prevalent in the US population, especially among

Table 3.4: Estimates of associations in the case of using 500 datasets. BIAS denotes the difference between the average of the 500 point estimates and the true value, ESD the average of the estimated standard deviations, SSD the sample standard deviation of the 500 point estimates, and the CP95 the 95% coverage probability.

True	Bias	SSD	ESD	CP95
$\rho^s_{12} = 0.288$	0.046	0.288	0.218	0.94
$\rho^s_{13} = 0.582$	0.092	0.355	0.371	0.91
$\rho^s_{23} = 0$	-0.034	0.277	0.321	0.96
$\kappa_{12} = 0.194$	0.056	0.329	0.356	0.94
$\kappa_{13} = 0.410$	-0.088	0.441	0.410	0.91
$\kappa_{23} = 0$	0.027	0.244	0.214	0.96
$\tau_{12} = 0.194$	0.056	0.329	0.356	0.94
$\tau_{13} = 0.410$	0.088	0.441	0.410	0.91
$\tau_{23} = 0$	0.027	0.244	0.214	0.96

young people aged 15-24. STIs can cause many serious problems such as pelvic inflammatory disease, ectopic pregnancy, tubal infertility, preterm birth, and increased susceptibility to human immunodeficiency virus infection (HIV). As introduced in section 2.5.1, three types of infections, Chlamydia trachomatis (CT), Neisseria gonorrhoeae (GC) and Trichomonas vaginalis (TV) are of interest. The data structure for this data set is shown in Table 3.5.

Table 3.5: STI data structure for infection times: sample size n=360.

	CT	NG	TV
left-censored	10.28%	1.67%	1.67%
interval-censored	44.72%	28.89%	28.89%
right-censored	45%	69.44%	69.44%

Monotone quadratic splines with 16 knots were adopted for this data set. The same prior specifications as the ones in section 2.5 were applied. A total of 20000 iterations were run in our Gibbs sampler and the first 5000 iterations were discarded as a burn-in. A summary of the posterior mean estimates and the corresponding 95% credible intervals for the regression parameters on the 15000 iterations of the Markov chain is presented in Table 3.6.

As demonstrated in Table 3.6, one can see that there was a generally positive

Table 3.6: Covariate effects estimations for STI data: posterior mean and 95% credible interval are provided.

	CT	GC	TV
age when enter the study	0.0372 (-0.0740 0.1515)	-0.0033 (-0.1417 0.1456)	0.0838 (-0.0655 0.2261)
number of partners	-0.0072 (-0.0468 0.0320)	0.0147 (-0.0254 0.0549)	0.0288 (-0.0120 0.0693)
age at first intercourse	-0.0316 (-0.1071 0.0454)	-0.0075 (-0.0966 0.0811)	-0.1584 (-0.2515 -0.0661)
race	0.1746 (-0.0903 0.4371)	0.0834 (-0.2096 0.3896)	0.6256 (0.2471 1.0531)
initial infection status	0.3726 (0.1022 0.6479)	0.5345 (0.2422 0.8294)	0.3731 (0.0762 0.6636)

association between infection history and STI acquisition. Being infected before the study contributes to a higher risk of early infection acquisition. The STI infection risk with Trichomonas vaginalis for subjects with a younger age at first intercourse was higher than those with an older age. African American adolescents tended to have a higher STI risk with T.vaginalis than white Amercians. The infections with C. trachomatis and T. vaginalis are not related to the age when enter the study, number of partners, age at first intercourse or race. The findings here proved the complexity of the STI risk in adolescents and a more careful evaluation of the behavioral markers for STI screening is needed.

Table 3.7 provides the statistical association estimations between these initial infection times. The estimates of the posterior means and 95% credible intervals for the piecewise Pearson's correlations ρ, Spearman's correlation coefficients ρ_s, median concordances κ and Kendall's concordances τ are listed. This analysis points to a generally positive association between these three infections. The small numbers observed from ρ, ρ_s, κ and τ in Table 3.7 indicate that the associations between CT, GC and TV infections are weak. However, one can see that the correlation between CT infection and GC infection is the strongest among all the pairwise correlations, and CT infection is less correlated with TV infection compared with the correlation with

Table 3.7: STI data: Estimation results for posterior mean and 95% credible interval of ρ, ρ_s, κ and τ are provided

	Mean	Std.	95%CI
ρ_{12}	0.2841	0.0733	(0.1387 0.4253)
ρ_{13}	0.2058	0.0773	(0.0531 0.3559)
ρ_{23}	0.2581	0.0832	(0.0904 0.4188)
ρ^s_{12}	0.2724	0.0707	(0.1325 0.4093)
ρ^s_{13}	0.1971	0.0743	(0.0507 0.3417)
ρ^s_{23}	0.2473	0.0802	(0.0864 0.4029)
κ_{12}	0.1839	0.0488	(0.0886 0.2797)
κ_{13}	0.1324	0.0505	(0.0338 0.2317)
κ_{23}	0.1668	0.0550	(0.0576 0.2751)
τ_{12}	0.1839	0.0488	(0.0886 0.2797)
τ_{13}	0.1324	0.0505	(0.0338 0.2317)
τ_{23}	0.1668	0.0550	(0.0576 0.2751)

GC infection. In comparison, it is impossible to estimate the dependence between the infection times of chlamydia, gonorrhea and trichomonas through the use of univariate modeling techniques, or to see the pairwise correlation relationships between these infections by the use of normal frailty multivariate probit model in Chapter 2.

3.5.2 AIDs Clinical Trail Data

Now we apply the proposed method to the bivariate interval-censored AIDS data discussed before in section 1.2.1. The data comes from an observational study usually referred as ACTG 181, collected from an AIDS clinical trial on human immunodeficiency virus (HIV)-infected individuals. In this study, 204 patients provided urine and blood samples at their clinical visits every 4 weeks and every 12 weeks, respectively. At each visit, the presence of the opportunistic infection cytomegalovirus (CMV) was tested. Two questions are of interest in this study: What is the covariate effect? The covariate is CD4 cell counts at study entry, which is an indicator of disease stage. If the CD4 cells/$\mu l < 75$, then the patient was in late stage. The CD4 cell counts effect is important, since physicians want to know the optimum timing for initiating prophylaxis for CMV disease. Another question of interest is the correlation between the

two shedding times in blood and urine. The reason behind this question is that the correlation provides an estimate of the two infection processes. If they are perfectly correlated, then the scientists only need to collect sample from one of them in the future. It will save resources and money. If the two are independent, then it indicates that the infection processes in urine and blood are different. The data stucture for AIDS clinical trail data is shown in Table 3.8.

Table 3.8: AIDS data structure for infection times: sample size n=204

	Blood	Urine
left-censored	3.43%	24.02%
interval-censored	11.28%	32.84%
right-censored	85.29%	43.14%

We adopted 19 knots for the monotone quadratic splines to guarantee the flexibility of the model. The knots are assigned according to the quantiles of the observation intervals. The same prior specifications as the ones in simulation study are used here. A total of 20000 iterations were ran in our Gibbs sampler and the first 5000 iterations were discarded as a burn-in. A summary of the posterior mean estimates and the corresponding 95% credible intervals for the regression parameters on the 15000 iterations of the Markov chain is presented in Table 3.9.

Table 3.9: The covariate effect estimation for the AIDs Data: posterior mean and 95% credible interval.

	Blood	Urine
cd4ind	0.6479	0.5104
	(0.0122, 1.3869)	(0.1071, 0.9132)

The results in Table 3.9 indicate that patients with baseline CD4 cell counts lower than 75/μl are at a considerable increased risk of CMV shedding in the urine and blood, since the 95% credible intervals are all beyond 0. To evaluate the statistical association that exists between these two CMV shedding times in blood and urine, we can obtain estimates from the Pearson's correlation ρ, Spearman's rank correlation

coefficient ρ_s, median concordance κ and kendall's τ between the two times from our joint analysis. The results are shown in Table 3.10, which suggest that there is a moderate positive association between the two failure times.

Table 3.10: AIDS data: Estimation results for posterior mean and 95% credible interval of ρ, κ and τ are provided.

	Mean	Std.	95%CI
ρ	0.5257	0.1025	(0.3130, 0.7087)
ρ_s	0.5087	0.1015	(0.3001, 0.6918)
κ	0.3553	0.0773	(0.2027, 0.5015)
τ	0.3553	0.0773	(0.2027, 0.5015)

3.6 Discussion

We presented a Bayesian approach for regression analysis of arbitrarily correlated failure time data under the semiparametric multivariate probit model. Monotone splines are adopted for approximating the unspecified nonparametric transformation functions. Maximum likelihood based methods are not feasible in closed form in the multivariate probit models, due to the intractability of the high dimensional integral in the likelihood function. As a comparison, Bayesian approach is preferred as it provides a full posterior distribution on all unknown parameters. The Gibbs sampler in section 3.3.4 is based on the idea of parameter expansion data augmentation, which gives full conditional posterior distributions in closed form. The proposed approach has many nice properties: It avoids the identifiability problem in the multivariate probit model by constraining the covariance to be a correlation matrix, and conjugate prior for the covariance matrix is applicable in deriving the posterior distribution through parameter expansion. The marginally uniform prior for the correlation matrix R is a proper prior and is uninformative, not favoring marginal correlations close to 0 or the bounds even in high dimensions. Based on this, a straightforward Gibbs sampler was proposed and the simulation study proved that our approach allows one to estimate the regression coefficients and pairwise correlations jointly. This new pro-

posed model enjoys the advantage of allowing different pairs of failure times to have different correlations, which is more flexible compared with the shared frailty models.

Our application looked at the STI data from Tu et al., 2009 and the AIDS data from Goggins and Finkelstein, 2000. An examination of the correlation matrix for the STI data revealed a complex dependence structure between the infections with Chlamydia trachomatis, Neisseria gonorrhoeae and Trichomonas vaginalis, hence indicating the plausibility of our formulation to model these infection times in a multivariate setting. Compared with the normal frailty multivariate probit model in chapter 2, the multivariate probit model in chapter 3 is a more general model, it allows the correlations between different events of interest to be arbitrary.

Chapter 4

Semiparametric regression analysis of multivariate interval-censored failure time data under frailty probit model allowing for arbitrary pairwise correlations

Summary: Correlated data arise when pairs or clusters of observations are related and thus are more similar to each other than to other observations in the dataset. In multivariate interval-censored data set, the multiple events are correlated. When more than two events of interest are investigated, the strength of association between different pairs of events can be different. For example, for the study of a disease's impact to multiple body parts, the infections with arms and legs can be more related compared with the eyes infection. Observations from different subjects can also be related differently. For example, when study family disease, the disease impact on twin pairs can be more correlated compared with the other members within this family. Therefore, models that allow arbitrary correlations are preferred. In this chapter, we extended the normal frailty multivariate probit model (normal frailty MVP) in chapter 2 to allow arbitrary pairwise correlations. This extended study makes the new normal frailty MVP model comparable to the MVP model in chapter 3. The underlying relationship between the two models is explored. Simulation results suggest that both models have good performance for estimating the regression parameters and the correlation coefficients. Our analysis suggests that the extended

normal frailty multivariate probit model is equivalent to the general multivariate probit model, and it enjoys simpler Gibbs algorithm formulas and fast computation.

4.1 Motivation

Recall that in chapter 2, we introduced the normal frailty multivariate probit model for estimating the covariate effects and statistical associations jointly. We refer to this model as normal frailty MVP model. The magnitude of association between multiple events depends on the shared frailty variance σ^2. A more general multivariate probit model is proposed in chapter 3, which is referred to as MVP model. Under this model, the association structure is evaluated by the correlation matrix. As a comparison, we notice that one limitation of normal frailty MVP model is that it assumes the correlations among multiple events of interest are the same. However, it may not be realistic, considering that, for example, the results in 3.5.1 indicate that the infection with Chalamydia trachomatis is more related to the infection with Neisseria gonorrhoeae compared with Trichomonas vaginalis in the STI data. In this case, the normal frailty MVP model in Chapter 2 failed to capture the difference in strength between different pairs of correlations. The MVP model in chapter 3, however, enjoys the advantage of allowing arbitrary correlations. Motivated by this observation, we want to explore the relationship between this two MVP models. Furthermore, an extension work is conducted on the normal frailty MVP model, which allows one to estimate arbitrary correlations.

4.2 Extended Normal Frailty MVP Model

Back to section 2.2, we know that the correlation among T_j's in model (2.1) is induced by the common frailty ζ. In order to allow model (2.1) to cooperate arbitrary correlations, an adjustment parameter c_j is introduced. The extended Normal Frailty

MVP model has the following form:

$$F_j(t \mid \boldsymbol{x}, \xi) = \Phi\{\alpha_j(t) + \boldsymbol{x}'\boldsymbol{\beta}_j + c_j\zeta\}, j = 1, \ldots J. \tag{4.1}$$

The parameters c_j's are unknown constants, except $c_1 = 1$ for identifiability purpose. Having c_j's in the model allows different pairs of event to have different correlations. We can rewrite the model (4.1) at the subject level:

$$\alpha_j(T_{ij}) = -\boldsymbol{x_i'\beta_j} - c_j\zeta_i + \epsilon_{ij}, i = 1, ...n, j = 1, ..., J. \tag{4.2}$$

The frailty term ζ_i follows normal distribution $N(0, \sigma^2)$, and the random variable ϵ_{ij} follows $N(0, 1)$. Now let $T^*_{ij} = \alpha_j(T_{ij})$ for $j = 1, 2, ...J$. Then $T^*_{ij} \sim N(-\boldsymbol{x_i'\beta_j}, \Sigma)$, where Σ is the covariance matrix of T^*_{ij},

$$\Sigma = \Big(a_{jl}\Big) \text{ with } a_{jl} = \begin{cases} a_{jj} = 1 + c_j^2\sigma^2 \\ a_{jl} = c_jc_l\sigma^2 \end{cases} \quad j = 1, \ldots J \text{ and } l = 1, \ldots J.$$

Under the multivarite probit model (4.2), the unknown parameters $(\boldsymbol{\beta}, \Sigma)$ are not identifiable. The reasons are discussed in Chapter 3. Therefore, marginal covariate coefficients $\boldsymbol{\beta}^*$ are estimated instead. By intergrating out the frailty term ζ from (4.1), we have the marginal coefficient $\beta^*_j = \frac{\beta_j}{\sqrt{1+c_j^2\sigma^2}}$. Now we decompose the covariance matrix into correlation matrix to avoid the identifiability problem when estimating the associations.

$$\text{Let } \underset{k\times k}{\Sigma} = \begin{pmatrix} 1 + c_1^2\sigma^2 & c_1c_2\sigma^2 & \ldots & c_1c_J\sigma^2 \\ c_2c_1\sigma^2 & 1 + c_2^2\sigma^2 & \ldots & c_2c_J\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ c_Jc_1\sigma^2 & c_Jc_2\sigma^2 & \ldots & 1 + c_J^2\sigma^2 \end{pmatrix} = DRD,$$

where

$$D = \begin{pmatrix} (1 + c_1^2\sigma^2)^{\frac{1}{2}} & 0 & \ldots & 0 \\ 0 & (1 + c_2^2\sigma^2)^{\frac{1}{2}} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & (1 + c_J^2\sigma^2)^{\frac{1}{2}} \end{pmatrix}$$

and

$$R = \begin{pmatrix} 1 & \frac{c_1 c_2 \sigma^2}{\sqrt{(1+c_1^2\sigma^2)(1+c_2^2\sigma^2)}} & \cdots & \frac{c_1 c_J \sigma^2}{\sqrt{(1+c_1^2\sigma^2)(1+c_J^2\sigma^2)}} \\ \frac{c_2 c_1 \sigma^2}{\sqrt{(1+c_2^2\sigma^2)(1+c_1^2\sigma^2)}} & 1 & \cdots & \frac{c_2 c_J \sigma^2}{\sqrt{(1+c_2^2\sigma^2)(1+c_J^2\sigma^2)}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{c_J c_1 \sigma^2}{\sqrt{(1+c_J^2\sigma^2)(1+c_1^2\sigma^2)}} & \frac{c_J c_2 \sigma^2}{\sqrt{(1+c_J^2\sigma^2)(1+c_2^2\sigma^2)}} & \cdots & 1 \end{pmatrix}.$$

With this decomposition, We see that (4.1) is essentially equivalent to the MVP model in Chapter 3, with the Pearson's correlation $\rho_{jl} = \rho(T_{ij}^*, T_{il}^*) = \frac{c_j c_l \sigma^2}{\sqrt{1+c_j^2\sigma^2}\sqrt{1+c_l^2\sigma^2}}$, for $j, l = 1, \ldots J$. From this relationship, we conclude that the extended Normal Frailty MVP model is a special case of the general MVP model. The good properties discussed in Chapter 2 of model (2.1) remain in the extended model (4.1). The new expressions for the pairwise statistical associations between correlated failure times in terms of the Spearman's rank correlation coefficient ρ_s, median concordance κ and kendall's τ are updated in the following theorem:

Theorem 4.2.1. *The pairwise statistical associations for multivariate survival data under the extended normal frailty MVP model is characterized by Spearman's correlation coefficient ρ_s, median concordance κ and Kendall's τ as follows, with the Pearson's correlation $\rho_{jl} = \rho(T_{ij}^*, T_{il}^*) = \frac{c_j c_l \sigma^2}{\sqrt{1+c_j^2\sigma^2}\sqrt{1+c_l^2\sigma^2}}$, for $j, l = 1, \ldots J$.*

$$\rho_s = 6\pi^{-1} sin^{-1}(\rho_{jl}/2), \tag{4.3}$$

$$\kappa = 2\pi^{-1} sin^{-1}(\rho_{jl}), \tag{4.4}$$

$$\tau = 2\pi^{-1} sin^{-1}(\rho_{jl}). \tag{4.5}$$

In order to facilitate the Bayesian approach for estimating covariate effects and correlations, we notice that c_j is an additional unknown parameter based on the Gibbs sampler in section 2.3.2 and need to be updated. A normal prior $N(0, \sigma_c^2)$ is assigned to c_j and one more step (step 8) is added to the proposed Gibbs sampler.

The new Gibbs sampler under the extended normal frailty MVP model is summarized in Algorithm 4.2.

Algorithm 4.2 Gibbs Sampler for extended normal frailty MVP model

At iteration i,

1: **Sample latent variables** z_{ij} for $i = 1, ..., n$ and $j = 1, ..., J$.

- if $\delta_{ij1} = 1$, sample z_{ij} from $N(\alpha_j(t_{ij}) + \boldsymbol{x_i\beta_j} + c_j\zeta_i, 1)I_{(z_{ij}>0)}$,
- if $\delta_{ij2} = 1$, sample z_{ij} from $N(\alpha_j(t_{ij}) + \boldsymbol{x_i\beta_j} + c_j\zeta_i, 1)I_{(\alpha_j(L_{ij})-\alpha_j(R_{ij})<z_{ij}<0)}$,
- if $\delta_{ij3} = 1$, sample z_{ij} from $N(\alpha_j(t_{ij}) + \boldsymbol{x_i\beta_j} + c_j\zeta_i, 1)I_{(z_{ij}<0)}$,

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

2: **Sample** γ_{j0}, from $N(E_{j0}, W_{j0}^{-1})$, where $W_{j0} = \upsilon_{j0} + n$ and

$$E_{j0} = W_{j0}^{-1}\left[m_{j0}\upsilon_{j0} + \sum_{i=1}^{n}\{z_{ij} - \sum_{l=1}^{m}\gamma_{jl}b_l(t_{ij}) - \boldsymbol{x_i\beta_j} - c_j\zeta_i\}\right].$$

3: **Sample all** γ_{jl}**'s** for $l = 1, 2..., m$ and $j = 1, ..., J$. For each l, let $W_{jl} = \sum_{i=1}^{n} b_l^2(t_{ij})$.

- If $W_{jl} = 0$, sample γ_{jl} from $Exp(\eta_j)$,
- If $W_{jl} > 0$, sample γ_{jl} from $N(E_{jl}, W_{jl}^{-1})1(\gamma_{jl} > d_{jl}^*)$, where

$$E_{jl} = W_{jl}^{-1}\left[\sum_{i=1}^{n} b_l(t_{ij})\{z_{ij} - \gamma_{j0} - \sum_{jl'\neq jl}\gamma_{jl'}b_{l'}(t_{ij}) - \boldsymbol{x_i\beta_j} - c_j\zeta_i\} - \eta_j\right],$$

$$d_l^* = \max(c_l^*, 0) \text{ and } c_l^* = \max_{\{(i,j):\delta_{ij2}=1\}}\left[\frac{-z_{ij}-\sum_{jl'\neq jl}\gamma_{jl'}\{b_{l'}(R_{ij})-b_{l'}(L_{ij})\}}{b_l(R_{ij})-b_l(L_{ij})}\right].$$

4: **Sample** $\boldsymbol{\beta_j}$ from $N(\hat{\boldsymbol{\beta}}_{\boldsymbol{j}}, \hat{\Sigma}_j)$, where $\hat{\Sigma}_j = (\Sigma_{j0}^{-1} + \sum_{i=1}^{n}\boldsymbol{x_i x_i'})^{-1}$ and

$$\hat{\boldsymbol{\beta}}_j = \hat{\Sigma}_j\left[\Sigma_{j0}^{-1}\beta_{j0} + \sum_{i=1}^{n}\{z_{ij} - \alpha_j(t_{ij}) - c_j\zeta_i\}\boldsymbol{x_i}\right].$$

5: **Sample** ζ_i from $N(\mu_i, \sigma_i^2)$ for $i = 1, ..., n$ where $\sigma_i^2 = (\sum_{j=1}^{J} c_j^2 + \sigma_\zeta^{-2})^{-1}$ and

$$\mu_i = \sigma_i^2\sum_{j=1}^{J} c_j\{z_{ij} - \alpha_j(t_{ij}) - \boldsymbol{x_i'}\beta\}.$$

6: **Sample** η_j for $j = 1, ..., J$ from $\mathcal{G}a(a_{j\eta} + m, b_{j\eta} + \sum_{l=1}^{m}\gamma_{jl})$.

7: **Sample** σ_ζ^{-2} from $\mathcal{G}a(a_\zeta + 0.5n, b_\zeta + 0.5\sum_{i=1}^{n}\zeta_i^2)$.

8: **Sample** c_j from $N(M_{j0}, V_{j0}^{-1})$, where $V_{j0} = \sigma_c^{-2} + \sum_{i=1}^{n}\zeta_i^2$ and $M_{j0} = V_{j0}^{-1}\left[\sum_{i=1}^{n}\zeta_i\Big(z_{ij} - \alpha_j(t_{ij}) - \boldsymbol{x_i'\beta_j}\Big)\right]$, $j = 2, \ldots, J$.

4.3 Simulation Studies

In this section, the extended normal frailty MVP model is evaluated through simulation study. The results are summarized on the marginal covariate effects $\boldsymbol{\beta}_j^*$. Same simulation settings as the ones used to generate Table 2.1 are applied. In addiction, the true values for c are set to be $c_1 = 1$, $c_2 = 1$ and $c_3 = 2$. The prior for c is $N(0, 10)$. The results for marginal covariates effects and Pearson's correlation coefficients from the extend normal frailty MVP model are given in Table 4.1, based on the results from 100 data sets, each with a sample size 200. Table 4.2 provides the pairwise results for the three nonparametric measures: Spearman's rank correlation coefficient ρ^s, median concordance κ and Kendall's τ.

Table 4.1: Simulation results of the Extended Normal Frailty MVP model with pairwise correlations. Presented results include the bias, the average of the estimated standard deviations, the sample standard deviation of the 100 point estimates, and the 95% coverage probability for the marginal covariate effects and Pearson's correlation coefficients.

	$\xi_i \sim N(0, .25)$					$\xi_i \sim N(0, 1)$					$\xi_i \sim N(0, 4)$				
	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95
β_{11}^*	0.894	0.019	0.172	0.194	0.94	0.707	0.017	0.171	0.178	0.97	0.447	0.004	0.163	0.160	0.94
β_{12}^*	0	-0.041	0.186	0.187	0.96	0	-0.041	0.159	0.175	0.98	0	-0.004	0.142	0.160	0.98
β_{21}^*	0	-0.019	0.178	0.169	0.94	0	-0.009	0.156	0.163	0.97	0	0.006	0.144	0.152	0.97
β_{22}^*	0.894	-0.012	0.173	0.181	0.95	0.707	-0.022	0.174	0.171	0.92	0.447	-0.005	0.146	0.157	0.96
β_{31}^*	-0.707	0.023	0.174	0.171	0.93	-0.447	0.030	0.164	0.160	0.91	-0.243	0.010	0.164	0.151	0.91
β_{32}^*	0.707	-0.029	0.191	0.174	0.90	0.447	-0.005	0.167	0.163	0.93	0.243	0.007	0.139	0.155	0.98
ρ_{12}	0.200	-0.038	0.087	0.075	0.94	0.500	0.015	0.067	0.063	0.98	0.800	0.000	0.039	0.035	0.94
ρ_{13}	0.316	-0.041	0.100	0.079	0.90	0.632	-0.002	0.066	0.060	0.97	0.868	-0.001	0.032	0.029	0.93
ρ_{23}	0.316	0.009	0.090	0.088	0.93	0.632	-0.008	0.062	0.058	0.95	0.868	-0.013	0.026	0.025	0.93

As seen from Table 4.1 and Table 4.2, the extended method works very well in estimating the regression parameters, with small bias in the point estimates, ESDs being close to SSDs, and the 95% coverage probabilities being close to 0.95 for all of the parameters. Table 4.2 also provides the estimation results of the association in terms of Spearman's correlation coefficient ρ_s, median concordance κ and Kendall's τ using Theorem 4.2.1. These results suggest that our extended method can estimate the pairwise association very accurately.

To further illustrate the relationship between the MVP model in chapter 3 and

Table 4.2: Simulation results of the extended normal frailty MVP model with pairwise correlations. Presented results include the bias, the average of the estimated standard deviations, the sample standard deviation of the 100 point estimates, and the 95% coverage probability for the Spearman's rank correlation coefficients, Median concordance and Kendall's τ.

	$\xi_i \sim N(0, .25)$					$\xi_i \sim N(0, 1)$					$\xi_i \sim N(0, 4)$				
	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95
ρ^s_{12}	.191	-0.036	0.084	0.072	0.94	.483	0.014	0.066	0.062	0.93	.786	0.000	0.041	0.036	0.94
ρ^s_{13}	.303	-0.040	0.097	0.077	0.90	.614	-0.003	0.067	0.060	0.97	.857	-0.001	0.032	0.031	0.91
ρ^s_{23}	.303	0.008	0.089	0.086	0.92	.614	-0.008	0.062	0.058	0.94	.857	-0.014	0.027	0.027	0.93
κ_{12}	.128	-0.026	0.057	0.050	0.94	.333	0.009	0.049	0.046	0.93	.590	-0.003	0.044	0.037	0.94
κ_{13}	.205	-0.030	0.067	0.054	0.90	.436	-0.005	0.055	0.049	0.97	.669	-0.005	0.040	0.037	0.91
κ_{23}	.205	0.004	0.062	0.060	0.92	.436	-0.009	0.051	0.048	0.94	.669	-0.020	0.037	0.034	0.93
τ_{12}	.128	-0.026	0.057	0.050	0.94	.333	0.009	0.049	0.046	0.93	.590	-0.003	0.044	0.037	0.94
τ_{13}	.205	-0.030	0.067	0.054	0.90	.436	-0.005	0.055	0.049	0.97	.669	-0.005	0.040	0.037	0.91
τ_{23}	.205	0.004	0.062	0.060	0.92	.436	-0.009	0.051	0.048	0.94	.669	-0.020	0.037	0.034	0.93

the extended normal frailty MVP model, a comparison simulation study on the MVP model is conducted. In this simulation study, the true values of the correlation coefficients for MVP model are set to be the ones got from the extended normal frailty MVP model in Table 4.1. Table 4.3 presents the results from MVP model with the same simulation settings as the one used by Table 4.1. Scenario I, II and III in this table are in correspondence with the extended normal frailty MVP model when the frailty variance $\sigma^2 = 0.25$, 1 and 2, respectively. The estimates for Spearman's correlation coefficient, median concordance and Kendall's τ are given in Table 4.4.

$$\underbrace{R = \begin{bmatrix} 1 & 0.200 & 0.316 \\ 0.200 & 1 & 0.316 \\ 0.316 & 0.316 & 1 \end{bmatrix}}_{\text{Scenario I}} \quad \underbrace{R = \begin{bmatrix} 1 & 0.500 & 0.632 \\ 0.500 & 1 & 0.632 \\ 0.632 & 0.632 & 1 \end{bmatrix}}_{\text{Scenario II}} \quad \underbrace{R = \begin{bmatrix} 1 & 0.800 & 0.868 \\ 0.800 & 1 & 0.868 \\ 0.868 & 0.868 & 1 \end{bmatrix}}_{\text{Scenario III}}$$

The results in Table 4.3 and 4.4 demonstrate that the MVP model performs well. The mean estimates are very close to the true value of the parameters, and the averaged standard errors are in close agreement to the standard deviations. Estimated coverage probabilities for 95% confidence intervals are at nominal level. Through Table 4.1 to Table 4.4, it is proved that the extended normal frailty MVP model

Table 4.3: Simulation results of the MVP model by using the true pairwise Pearson's correlation coefficients from the extended normal frailty MVP model. Presented results include the bias, the average of the estimated standard deviations, the sample standard deviation of the 100 point estimates, and the 95% coverage probability for the marginal covariate effects and Pearson's correlation coefficients.

	Scenario I					Scenario II					Scenario III				
	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95
β^*_{11}	0.894	0.024	0.241	0.198	0.95	0.707	0.032	0.208	0.183	0.98	0.447	0.013	0.212	0.198	0.96
β^*_{12}	0	-0.036	0.203	0.184	0.93	0	-0.027	0.178	0.169	0.94	0	-0.012	0.172	0.164	0.92
β^*_{21}	0	-0.013	0.184	0.176	0.93	0	-0.012	0.166	0.161	0.97	0	0.010	0.152	0.144	0.95
β^*_{22}	0.894	-0.008	0.182	0.174	0.91	0.707	-0.012	0.167	0.159	0.92	0.447	-0.009	0.150	0.152	0.97
β^*_{31}	-0.707	0.013	0.184	0.176	0.97	-0.447	0.023	0.172	0.176	0.91	-0.243	0.008	0.144	0.150	0.95
β^*_{32}	0.707	-0.032	0.242	0.235	0.92	0.447	-0.010	0.189	0.177	0.92	0.243	0.006	0.162	0.156	0.94
ρ_{12}	0.200	0.006	0.111	0.107	0.93	0.500	0.060	0.129	0.115	0.96	0.800	0.008	0.129	0.104	0.94
ρ_{13}	0.316	0.035	0.097	0.104	0.95	0.632	0.082	0.168	0.102	0.94	0.868	-0.009	0.118	0.094	0.96
ρ_{23}	0.316	0.010	0.086	0.093	0.98	0.632	0.090	0.171	0.145	0.97	0.868	-0.019	0.128	0.102	0.91

Table 4.4: Simulation results of the MVP model by using the true pairwise Pearson's correlation coefficients from the extended normal frailty MVP model. Presented results include the bias, the average of the estimated standard deviations, the sample standard deviation of the 100 point estimates, and the 95% coverage probability for the Spearman's rank correlation coefficients, Median concordance and Kendall's τ.

	Scenario I					Scenario II					Scenario III				
	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95
ρ^s_{12}	.191	0.007	0.115	0.111	0.98	.483	0.051	0.128	0.113	0.96	.786	0.006	0.127	0.095	0.94
ρ^s_{13}	.303	0.037	0.099	0.107	0.95	.614	0.083	0.168	0.101	0.94	.857	-0.008	0.116	0.104	0.92
ρ^s_{23}	.303	0.011	0.088	0.096	0.98	.614	0.080	0.171	0.144	0.97	.857	-0.016	0.127	0.100	0.92
κ_{12}	.128	0.002	0.078	0.073	0.98	.333	0.019	0.099	0.082	0.96	.590	-0.006	0.104	0.098	0.94
κ_{13}	.205	0.022	0.069	0.072	0.95	.436	-0.008	0.074	0.090	0.94	.669	-0.009	0.092	0.087	0.92
κ_{23}	.205	0.005	0.064	0.065	0.98	.436	-0.114	0.168	0.048	0.97	.669	-0.008	0.110	0.134	0.92
τ_{12}	.128	0.002	0.078	0.073	0.98	.333	0.019	0.099	0.082	0.96	.590	-0.006	0.104	0.098	0.94
τ_{13}	.205	0.022	0.069	0.072	0.95	.436	-0.008	0.074	0.090	0.94	.669	-0.009	0.092	0.087	0.92
τ_{23}	.205	0.005	0.064	0.065	0.98	.436	-0.114	0.168	0.048	0.97	.669	-0.008	0.110	0.134	0.92

is essentially a special case of MVP model. We also noticed that by introducing the unknown constant c_j in the normal frailty MVP model, the extended model can further handle negative correlations, as the value of c_j, $j = 2 \dots J$ can be negative. The general MVP model in Chapter 3 enjoys this property as well since the values of Pearson's correlation in the correlation matrix can be negative.

Table 4.5 and 4.6 presents the results for marginal covariate effects and pairwise statistics associations from the extended normal frailty MVP model. In these two tables, the same simulation settings are adopted, except $c_1 = 1$, $c_2 = -1$ and $c_3 = 2$. By taking $c_2 = -1$, the pairwise correlations between event 1 and event 2, the pairwise

Table 4.5: Simulation results of the Extended Normal Frailty MVP model with negative pairwise correlations. Presented results include the bias, the average of the estimated standard deviations, the sample standard deviation of the 100 point estimates, and the 95% coverage probability for the marginal covariate effects and Pearson's correlation coefficients.

	$\xi_i \sim N(0,.25)$					$\xi_i \sim N(0,1)$					$\xi_i \sim N(0,4)$				
	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95
β^*_{11}	0.894	0.017	0.170	0.194	0.96	0.707	0.011	0.172	0.179	0.96	0.447	0.002	0.162	0.159	0.94
β^*_{12}	0	-0.042	0.185	0.187	0.96	0	-0.043	0.158	0.175	0.98	0	-0.008	0.136	0.159	1.00
β^*_{21}	0	-0.015	0.170	0.169	0.92	0	-0.022	0.160	0.162	0.96	0	-0.012	0.146	0.150	0.93
β^*_{22}	0.894	0.000	0.173	0.180	0.96	0.707	0.004	0.167	0.169	0.94	0.447	0.010	0.144	0.156	0.96
β^*_{31}	-0.707	0.021	0.174	0.170	0.93	-0.447	0.026	0.165	0.160	0.90	-0.243	0.010	0.156	0.149	0.91
β^*_{32}	0.707	-0.029	0.189	0.173	0.92	0.447	-0.010	0.168	0.163	0.94	0.243	0.004	0.134	0.153	0.97
ρ_{12}	-0.200	0.048	0.073	0.076	0.91	-0.500	-0.009	0.070	0.064	0.93	-0.800	-0.004	0.037	0.036	0.92
ρ_{13}	0.316	-0.040	0.082	0.078	0.89	0.632	-0.002	0.066	0.060	0.94	0.868	-0.000	0.033	0.029	0.90
ρ_{23}	-0.316	-0.012	0.093	0.089	0.90	-0.632	0.018	0.067	0.058	0.92	-0.868	-0.010	0.031	0.027	0.91

correlations between event 2 and 3 are negative. As a comparison, the true values of the correlation matrix for MVP model are set to be the ones got from the extended normal frailty MVP model in Table 4.5. Table 4.7 and Table 4.8 give the results for the covariate effects and statistics associations under the general MVP model. Scenario I, II and III in these two tables are in correspondence with the extended normal frailty MVP model in Table 4.5 when the frailty variance $\sigma^2 = 0.25$, 1 and 2, respectively.

$$\underbrace{R = \begin{bmatrix} 1 & -0.200 & 0.316 \\ -0.200 & 1 & -0.316 \\ 0.316 & -0.316 & 1 \end{bmatrix}}_{\text{Scenario I}} \quad \underbrace{R = \begin{bmatrix} 1 & -0.500 & 0.632 \\ -0.500 & 1 & -0.632 \\ 0.632 & -0.632 & 1 \end{bmatrix}}_{\text{Scenario II}} \quad \underbrace{R = \begin{bmatrix} 1 & -0.800 & 0.868 \\ -0.800 & 1 & -0.868 \\ 0.868 & -0.868 & 1 \end{bmatrix}}_{\text{Scenario III}}$$

We observe that both models performed equally well in terms of the regression parameter and statistics association estimates. The point estimates of the covariate coefficients are close to the ture values. ESD is the average of the estimated standard deviations of the posterior distribution of the parameter across the 100 data sets. SSD is the sample standard deviation of the point estimates from the 100 data sets. SSD and ESD are very close in all the setups for all the parameters. 95% coverage probability are also close to 0.95. The results from Table 4.5 to Table 4.8 provide strong evidence that both the models perform very well in estimating the regression

Table 4.6: Simulation results of the extended normal frailty MVP model with negative pairwise correlations. Presented results include the bias, the average of the estimated standard deviations, the sample standard deviation of the 100 point estimates, and the 95% coverage probability for the Spearman's rank correlation coefficients, Median concordance and Kendall's τ.

	$\xi_i \sim N(0, .25)$					$\xi_i \sim N(0, 1)$					$\xi_i \sim N(0, 4)$				
	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95
ρ^s_{12}	- .191	0.047	0.076	0.074	0.92	-.483	-0.009	0.069	0.063	0.95	-.786	-0.004	0.038	0.037	0.95
ρ^s_{13}	.303	-0.040	0.085	0.078	0.90	.614	-0.003	0.066	0.060	0.93	.857	-0.000	0.035	0.031	0.90
ρ^s_{23}	-.303	-0.011	0.090	0.086	0.93	-.614	0.019	0.068	0.059	0.90	-.857	0.011	0.032	0.029	0.90
κ_{12}	-.128	0.033	0.053	0.051	0.92	-.333	-0.005	0.052	0.047	0.95	-.590	-0.001	0.040	0.038	0.95
κ_{13}	.205	-0.032	0.060	0.055	0.90	.436	-0.005	0.054	0.049	0.93	.669	-0.004	0.044	0.037	0.90
κ_{23}	-.205	-0.004	0.063	0.060	0.93	-.436	0.019	0.056	0.049	0.90	-.669	0.018	0.041	0.036	0.90
τ_{12}	-.128	0.033	0.053	0.051	0.92	-.333	-0.005	0.052	0.047	0.95	-.590	-0.001	0.040	0.038	0.95
τ_{13}	.205	-0.032	0.060	0.055	0.90	.436	-0.005	0.054	0.049	0.93	.669	-0.004	0.044	0.037	0.90
τ_{23}	-.205	-0.004	0.063	0.060	0.93	-.436	0.019	0.056	0.049	0.90	-.669	0.018	0.041	0.036	0.90

Table 4.7: Simulation results of the MVP model by using the true pairwise Pearson's correlation coefficients from the extended normal frailty MVP model. Presented results include the bias, the average of the estimated standard deviations, the sample standard deviation of the 100 point estimates, and the 95% coverage probability for the marginal covariate effects and Pearson's correlation coefficients.

	Scenario I					Scenario II					Scenario III				
	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95
β^*_{11}	0.894	-0.031	0.282	0.238	0.92	0.707	0.042	0.178	0.184	0.96	0.447	0.033	0.191	0.208	0.90
β^*_{12}	0	-0.011	0.246	0.196	0.92	0	-0.035	0.204	0.214	0.92	0	-0.012	0.199	0.182	0.97
β^*_{21}	0	0.014	0.192	0.187	0.96	0	-0.082	0.209	0.231	0.93	0	0.024	0.187	0.152	0.92
β^*_{22}	0.894	0.065	0.209	0.186	0.88	0.707	-0.032	0.198	0.187	0.93	0.447	-0.014	0.203	0.187	0.91
β^*_{31}	-0.707	-0.042	0.366	0.197	0.90	-0.447	0.032	0.198	0.176	0.92	-0.243	0.038	0.167	0.152	0.92
β^*_{32}	0.707	0.043	0.285	0.191	0.92	0.447	-0.041	0.198	0.179	0.94	0.243	0.034	0.152	0.172	0.91
ρ_{12}	-0.200	-0.002	0.142	0.108	0.95	-0.500	0.07	0.147	0.134	0.92	-0.800	0.023	0.16	0.154	0.93
ρ_{13}	0.316	0.064	0.127	0.107	0.94	0.632	0.079	0.121	0.112	0.94	0.868	-0.023	0.176	0.054	0.98
ρ_{23}	-0.316	-0.045	0.126	0.097	0.96	-0.632	0.079	0.177	0.162	0.90	-0.868	-0.034	0.182	0.179	0.92

Table 4.8: Simulation results of the MVP model by using the true pairwise Pearson's correlation coefficients from the extended normal frailty MVP model. Presented results include the bias, the average of the estimated standard deviations, the sample standard deviation of the 100 point estimates, and the 95% coverage probability for the Spearman's rank correlation coefficients, Median concordance and Kendall's τ.

	Scenario I					Scenario II					Scenario III				
	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95	True	Bias	SSD	ESD	CP95
ρ^s_{12}	-.191	-0.001	0.139	0.104	0.95	-.483	0.062	0.134	0.123	0.93	-.786	0.023	0.141	0.124	0.93
ρ^s_{13}	.303	0.061	0.124	0.104	0.94	.614	0.079	0.187	0.169	0.94	.857	-0.092	0.166	0.183	0.97
ρ^s_{23}	-.303	-0.043	0.123	0.094	0.96	-.614	0.087	0.168	0.144	0.91	-.857	-0.056	0.165	0.178	0.91
κ_{12}	-.128	0.002	0.107	0.071	0.95	-.333	0.034	0.127	0.145	0.93	-.590	-0.046	0.166	0.188	0.93
κ_{13}	.205	0.041	0.087	0.072	0.94	.436	-0.048	0.174	0.189	0.94	.669	-0.049	0.191	0.183	0.97
κ_{23}	-.205	-0.029	0.087	0.065	0.96	-.436	-0.098	0.145	0.132	0.91	-.669	-0.067	0.162	0.151	0.91
τ_{12}	-.128	0.002	0.107	0.071	0.95	-.333	0.034	0.127	0.145	0.93	-.590	-0.046	0.166	0.188	0.93
τ_{13}	.205	0.041	0.087	0.072	0.94	.436	-0.048	0.174	0.189	0.94	.669	-0.049	0.191	0.183	0.97
τ_{23}	-.205	-0.029	0.087	0.065	0.96	-.436	-0.098	0.145	0.132	0.91	-.669	-0.067	0.162	0.151	0.91

parameters and pairwise statistics associations, and enjoy the flexibility of allowing the correlation to be negative.

The extended normal frailty MVP model estimates the correlations from the frailty variance σ^2 and the constant c, resulting in significant gains in efficiency. For the general MVP model, the correlation is estimated through correlation matrix directly and the number of unknown parameters grows rapidly as the number of dimensions for T increases. It usually takes more time for the general MVP model to get results, especially for high dimension. From simulation studies, we observed that though both models give accurate estimations, the normal frailty MVP model enjoys faster computation, and thus is more efficient than the general MVP model. The computation difficulties for general MVP model arise mainly from the sampling of univariate truncated Normal distribution for latent variable $\boldsymbol{Z}$ (Tabet, 2007). The method of Robert (1995) was adopted (See Appendix B), which is based on an accept and reject algorithm. It may happen that under certain simulations, the accepting values is small and significantly slow down the method.

4.4 Real Data Analysis

In this section, we applied the extended normal frailty MVP model on STI data and ACTG181 data, to re-evaluate the marginal covariate effects and the statistical associations.

4.4.1 STI Data

As introduced in section 3.5.1, STI data is a complicated multivariate interval-censored data set. The extended normal frailty MVP model is applied on this data set. The same specifications for the parameters are adopted as discussed before. The marginal covariate effects estimates are given in Table 4.9. The results for the statistical associations are given in Table 4.10.

Table 4.9: Marginal covariate effects for STI data based on extended normal frailty MVP model

	CT	GC	TV
age when enter the study	0.0281 (-0.1041 0.1604)	-0.0086 (-0.1511 0.1315)	0.1548 (0.0044 0.3047)
number of partners	-0.0158 (-0.1572 0.1215)	0.0705 (-0.0688 0.2121)	0.1089 (-0.0349 0.2530)
age at first intercourse	-0.0640 (-0.1974 0.0704)	-0.0179 (-0.1662 0.1234)	-0.2455 (-0.4010 -0.0885)
race	0.3042 (-0.1107 0.6958)	0.0958 (-0.3440 0.5054)	0.7199 (0.2191 1.2298)
initial infection status	0.4111 (0.1427 0.6744)	0.5386 (0.2442 0.8118)	0.3615 (0.0638 0.6568)

Table 4.10: Statistics associations for STI data based on extended normal frailty MVP model: posterior mean and 95% credible interval are provided

	Mean	Std.	95%CI
σ^2	0.2431	0.0619	(0.1453 0.3817)
$\rho_{s_{12}}$	0.4072	0.0441	(0.3248 0.4964)
$\rho_{s_{13}}$	0.1117	0.0357	(0.0429 0.1829))
$\rho_{s_{23}}$	0.2468	0.0737	(0.0981 0.3861)
κ_{12}	0.2785	0.0318	(0.2198 0.3436)
κ_{13}	0.0746	0.0239	(0.0286 0.1225)
κ_{23}	0.1664	0.0505	(0.0655 0.2631)
τ_{12}	0.2785	0.0318	(0.2198 0.3436)
τ_{13}	0.0746	0.0239	(0.0286 0.1225)
τ_{23}	0.1664	0.0505	(0.0655 0.2631)

From the results in Table 4.9 and Table 4.10, we see that being older when enter the study, at a younger age at first intercourse, being African American and has infection history before the study will lead to an increasing risk of early TV infection acquisition. For CT and GC infection, only the infection history has an impact for the early acquisition. We also see the complex correlation structure behind this data set. As observed in Table 4.10, the correlation between CT and GC infection is the strongest among all the three pairwise correlations.. TV infection is less correlated with CT infection as compared with GC infection. The conclusions form Table 4.9 and Table 4.10 are the same as the ones got from 3.6 and 3.7 in chapter 3.

Table 4.11: AIDS data: marginal covariate effects estimations from normal frailty MVP model, posterior mean and 95% credible interval are provided

	Blood	Urine
cd4ind	0.6259	0.6011
	(0.1787 1.0852)	(0.2647 0.9308)

Table 4.12: AIDS data: Estimation results under normal frailty MVP model for posterior mean and 95% credible interval of ρ, κ and τ are provided

	Mean	Std.	95%CI
ρ_s	0.7098	0.0923	(0.4738 0.8446)
κ	0.5222	0.0813	(0.3268 0.6541)
τ	0.5222	0.0813	(0.3268 0.6541)

4.4.2 AIDS Data

Now we apply the frailty MVP model on the AIDS data. The covariate of interest is CD4 cell counts. The same prior specifications as discussed in section 3.5.2 are adopted. A total of 20000 iterations were ran in the Gibbs sampler and the first 5000 iterations were discarded as burn-in. A summary of the posterior mean estimates and the corresponding 95% credible intervals for the regression parameters on the 15000 iterations of the Markov chain is shown in Table 4.11. The marginal covariate effects estimates are given in Table 4.11. The results for the statistical associations are given in Table 4.12. The estimates for the Pearson's correlation ρ, Spearman's rank correlation coefficient ρ_s, median concordance κ and Kendall's τ between the infection times in urine and blood are shown in Table 4.12.

We can see that patients in the late disease stage (with CD4 cell counts lower than 75/μl has higher risk of CMV shedding in the urine and blood. A moderate to strong association exists between the failure times in urine and blood. The marginal covariate effect estimations from the normal frailty MVP model are close to the ones got from the general MVP model, indicating that both models can analyze the multivariate interval-censored data well.

4.5 Conclusion

In this project, we developed two semi-parametric models under the multivariate probit model framework to estimate covariate effects and statistical association jointly. Both the extended normal frailty MVP model and the general MVP model can allow arbitrary pairwise correlations between different failure times. Efficient Bayesian approaches for regression analysis of multivariate interval-censored data under the two models were presented. Monotone splines are adopted for approximating the unspecified function, which provides computational efficiency and model flexibility. By incorporating a normal frailty in the MVP model, the correlation structure in our joint modeling approach is simplified and the computation is more efficient. Simulations and real data applications showed that both the proposed models work reasonably well.

Bibliography

Aalen, O.O. (1989). "A linear regression model for the analysis of life times". In: *Statistics in Medicine* 8, pp. 907–925.

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". In: *Biometrics* 53, pp. 1475–1484.

Ashford, J. and Snowden, R. (1970). "Multivariate probit analysis". In: *Biometrics* 26, pp. 535–546.

Barnard, J., McCulloch, R., and Meng, X. (2000). "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, With Application to Shrinkage". In: *Statistica Sinica* 10, pp. 1281–1311.

Bennett, S. (1983). "Analysis of survival data by the proportional odds model". In: *Statistics in Medicine* 2, pp. 273–277.

Betensky, R.A. and Finkelstein, D.M. (1999). "A non-parametric maximum likelihood estimator for bivariate interval censored data". In: *Statistics in Medicine* 18, pp. 3089–3100.

Betensky, RA., Rabinowitz, D., and Tsiatis, AA. (2001). "Computationally simple accelerated failure time regression for interval censored data". In: *Biometrika* 88, pp. 703–711.

Betensky, RA., Lindsey, JC., Ryan, LM., and Wand, MP. (2002). "A local likelihood proportional hazards model for interval censored data". In: *Statistics in Medicine* 21, pp. 263–275.

Cai, B., Lin, X., and Wang, L. (2011). "Bayesian proportional hazards model for current status data with monotone splines". In: *Computational Statistics and Data Analysis* 55, pp. 2644–2651.

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

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". In: *Statistica Sinica* 17, pp. 1023–1046.

Chen, M., Tong, X., and Sun, J. (2007). "The proportional odds model for multivariate interval-censored failure time data." In: *Statistics in Medicine* 26, pp. 5147–5161.

Chen, M.-H., Tong, X.W., and Sun, J. (2009). "A frailty model approach for regression analysis of multivariate current status data". In: *Statistics in Medicine* 28, pp. 3424–3436.

Chib, S. and Greenberg, E. (1998). "Bayesian analysis of multivariate probit models". In: *Biometrika* 85, pp. 347–361.

Chin, V., Gunawan, D., Fiebig, D.G., Kohn, R., and Sisson, S.A. (2018). "Efficient data augmentation for multivariate probit models with panel data: An application to general practitioner decision-making about contraceptives". In: *arXiv preprint arXiv:1806.07274*.

Clayton, D. G. (1978). "A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence". In: *Biometrika* 65, pp. 141–152.

Cox, D. R. (1972). "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34, pp. 187–220.

Cox, D. R. (1975). "Partial likelihood". In: *Biometrika* 62, pp. 269–276.

Cui, S. and Sun, Y. (2004). "Checking for the Gamma frailty distribution under the marginal proportional hazards frailty model". In: *Statistica Sinica* 14, pp. 249–267.

Dabrowska, D. (1986). "Rank tests for independence for bivariate censored data". In: *The Annals of Statistics* 14, pp. 250–264.

De Gruttola, Victor. and Lagakos, W. Stephen. (1989). "Analysis of Doubly-Censored Survival Data, with Application to AIDS". In: *Biometrics* 45, pp. 1–11.

Du M., Hu T. and Sun, J. (2019). "Semiparametric probit model for informative current status data". In: *Statistics in Medicine*, pp. 1–9.

Dunson, D.B. and Dinse, G. (2002). "Bayesian models for multivariate current status data with informative censoring". In: *Biometrics* 58, pp. 79–88.

Dyk, D.A. van and Meng, X. (2001). "The Art of Data Augmentation". In: *Journal of Computational and Graphical Statistics* 10, pp. 1–50.

Finkelstein, D. (1986). "A proportional hazards model for interval-censored failure time data". In: *Biometrics* 42, pp. 845–854.

Finkelstein, D.M. and Wolfe, R.A. (1985). "A Semiparametric Model for Regression Analysis of Interval-Censored Failure Time Data". In: *Biometrics* 41, pp. 933–945.

Gamage, Withana Prabhashi W., McMahan, S. Christopher., Wang, L., and Tu, W. (2018). "A gamma-frailty proportional hazards model for bivariate interval-censored data". In: *Computational Statistics and Data Analysis* 128, pp. 354–366.

Geman, S. and Geman, D. (1984). "Stochastic relaxiation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, pp. 721–741.

Geweke, J. (1991). "Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints". In: *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface, Alexandria, VA: American Statistical Association.*

Goeteghebeur, E. and Ryan, L. (2000). "Semiparametric regressionanalysis of interval-censored data". In: *Biometrics* 48, pp. 1139–1144.

Goggins, WB., Finkelstein, DM., Schoenfeld, DA., and Zaslavsky, AM. (1998). "A Markov chain Monte Carlo EM algorithm for analyzing interval censored data under the Cox proportional hazards model". In: *Biometrics* 54, pp. 1498–1507.

Goggins, William B. and Finkelstein, Dianne M. (2000). "A Proportional Hazards Model for Multivariate Interval-Censored Failure Time Data". In: *Biometrics* 56, pp. 940–943.

Hanson, T. and Johnson, W. O. (2004). "A bayesian semiparametric AFT model for interval-censored data". In: *Journal of Computational and Graphical Statistics* 13, pp. 341–361.

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". In: *Statistics in Medicine* 28, pp. 2785–2800.

Henschel, V., Engel, J., Holzel, D., and Mansmann, U. (2009). "A semiparametric bayesian proportional hazards model for interval censored data with frailty effects". In: *BMC Medical Research Methodology* 9, p. 9.

Hougaard, P. (2000). "Analysis of Multivariate Survival Data". In: *Springer; New York*.

Huang, J. and Rossini, AJ. (1997). "Sieve estimation for the proportional odds failure-time regression model with interval censoring". In: *Journal of the American Statistical Association* 92, pp. 960–967.

Ibrahim, J.S., Chen, M.-H., and Sinha, D. (2008). "Bayesian Survival Analysis". In: *Springer; New York*.

Kalbfleisch, J.D. and Prentice, R. L. (2002). "The statistical analysis of failure Time data (2nd edn)". In: *Wiley: New York*.

Kim, Y. Mimi. and Xue, Xiaonan (2002). "The analysis of multivariate interval-censored surviva ldata". In: *Statistics in Medicine* 21, pp. 3715–3726.

Klein, J.P. (1992). "Semiparametric estimation of random effects using the Cox model based on the EM algorithm". In: *Biometrics* 48, pp. 795–806.

Komarek, A. and Lessaffre, E. (2007). "Bayesian accelerated failure time model for correlated interval-censored data with a normal mixture as error distribution". In: *Statistica Sinica* 17, pp. 549–569.

Kruskal, W.H. (1958). "Ordinal measures of association". In: *Journal of the American Statistical Association* 53, pp. 814–861.

Lam, K., Xu, Y., and Cheung, T. (2010). "A multiple imputation approach for clustered interval-censored survival data". In: *Statistics in Medicine* 29, pp. 680–693.

Lawrence, E., Bingham, C., and Nair, V. N. (2008). "Bayesian inference for multivariate ordinal data using parameter expansion". In: *Technometrics* 50(2), pp. 182–191.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). "Generating random correlation matrices based on vines and extended onion method". In: *Journal of Multivariate Analysis* 100, p. 9.

Li, L. and Pu, Z. (2003). "Rank estimation of log-linear regression with interval-censored data". In: *Lifetime Data Analysis* 9, pp. 57–70.

Lin, X. and Wang, L. (2010). "A semiparametric probit model for case 2 interval-censored failure time data". In: *Statistics in Medicine* 29, pp. 972–981.

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

Liu, C. (2001). "Bayesian analysis of multivariate probit models - discussion on the art of data augmentation by van dyk and meng". In: *Journal of Computational and Graphical Statistics* 10, pp. 75–81.

Liu, H. and Qin, J. (2018). "Semiparametric probit models with univariate and bivariate current-status data". In: *Biometrics* 74, pp. 68–76.

Liu, J. and Wu, Y. (1999). "Parameter expansion for data augmentation". In: *Journal of the American Statistical Association* 94, pp. 1264–1274.

Liu, X. and Daniels, M. (2006). "A New Algorithm for Simulating a Correlation Matrix Based on Parameter Expansion and Reparameterization". In: *Journal of Computational and Graphical Statistics* 15, pp. 897–914.

Oakes, D. (1982). "A concordance test for independence in the presence of censoring". In: *Biometrics* 38, pp. 451–455.

Oakes, D. (1989). "Bivariate survival models induced by frailties". In: *Journal of the AmericanStatistical Association* 84, pp. 487–493.

Pitt, M., Chan, D., and Kohn, R. (2006). "Efficient Bayesian inference for Gaussian copula regression models". In: *Biometrika* 93, pp. 537–554.

Rabinowitz, D., Tsiatis, AA., and Aragon, J. (1995). "Regression with interval-censored data". In: *Biometrika* 82, pp. 501–513.

Rabinowitz, D., Betensky, RA., and Tsiatis, AA. (2000). "Using conditional logistic regression to fit proportional odds models to interval censored data". In: *Biometrics* 56, pp. 511–518.

Ramsay, J. (1988). "Monotone Regression Splines in Action". In: *Statistical Science* 3, pp. 425–461.

Rao, C.R. (1965). ": Linear Statistical Inference and Its Applications". In: *New York: Wiley.*

Ritter, C. and Tanner, M. A. (1992). "Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler". In: *Journal of the American Statistical Association* 87, pp. 861–868.

Robert, C. (1995). “Simulation of Truncated Normal Variables”. In: *Statistics and Computing* 5, pp. 121–125.

Rondeau, V., Commenges, D., and Joly, P. (2003). “Maximum penalized likelihood estimation in a Gamma-frailty model”. In: *Lifetime Data Analysis* 9, pp. 139–153.

Rossi, P., Allenby, G., and McCulloch, R. (2005). “Bayesian Statistics and Marketing”. In: *John Wiley and Sons*.

Samuelsen, Sven Ove and Kongerud, Johny (1994). “Interval censoring in longitudinal data of respiratory symptoms in aluminium potroom workers: a comparison of methods”. In: *Statistics in Medicine* 13, pp. 1771–1780.

Satten, Glen A. (1996). “Rank-based inference in the proportional hazards model for interval censored data”. In: *Journal of Applied Statistics* 83, pp. 355–370.

Shao, F., Li, J., Ma, S., and Lee, M. (2014). “Semiparametric varyingâĂŘcoefficient model for interval censored data with a cured proportion”. In: *Statistics in Medicine* 33, pp. 1700–1712.

Shen, P. (2015). “Additive transformation models for multivariate interval-censored data”. In: *Communications in Statistics-Theory and Methods* 44, pp. 1065–1079.

Shen, X. (1998). “Proportional odds regression and sieve maximum likelihood estimation”. In: *Biometrika* 85, pp. 165–177.

Shiboski, S. (1998). “Generalized additive models for current status Data”. In: *Lifetime Data Analysis* 4, pp. 29–50.

Sun, J. (2006). “Statistical analysis of interval-censored data (2nd edn)”. In: *Springer: Berlin*.

Tabet, A. (2007). “Bayesian inference in the multivariate probit model: estimation of the correlation matrix”. In: *Unpublished Masters thesis, The University of British Columbia, British Columbia*.

Talhouk, Aline., Doucet, Arnaud., and Murphy, Kevin. (2012). “Efficient bayesian inference for multivariate probit models with sparse inverse correlation matrices”. In: *Journal of Computational and Graphical Statistics* 21, pp. 739–757.

Tu, W., Batteiger, B., Wiehe, S., Ofner, S., Pol, B.V.D., Katz, B., Orr, D., and Fortenberry, J (2009). “Time from first intercourse to first sexually transmitted infection diagnosis among adolescent women”. In: *Archives of Pediatrics and Adolescent Medicine* 163, pp. 1106–1111.

Tu, W., Ghosh, P., and Katz, B (2011). "A stochastic model for assessing chlamydia trachoma-tis transmission risk by using longitudinal observational data". In: *Journal of the Royal Statistical Society: Series A* 174, pp. 975–989.

Wang, D., Zhou, H., and Kulasekera, K. (2016). "A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data". In: *Biometrics* 72, pp. 222–231.

Wang, L. and Dunson, D. (2011). "Semiparametric Bayes Proportional Odds Models for Current Status Data with Under-reporting". In: *Biometrics* 67, pp. 1111–1118.

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

Wang, N., Wang, L., and McMahan, C. (2015). "Regression analysis of bivariate current status data under the Gamma-frailty proportional hazards model using the EM algorithm". In: *Computational Statistics and Data Analysis* 83, pp. 140–150.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). "Regression analysis of multivariate incomplete failure time data by modelling of marginal distributions". In: *Journal of the American Statistical Association* 84, pp. 1065–1073.

Wen, C. and Chen, Y. (2013). "A frailty model approach for regression analysis of bivariate interval-cenosred survival data". In: *Statistica Sinica* 23, pp. 383–408.

Wen, C.C. and Chen, Y.H. (2011). "Nonparametric maximum likelihood analysis of clustered current status data with the Gamma-frailty Cox model". In: *Computational Statistics and Data Analysis* 55, pp. 1053–1060.

Wienke, A. (2012). "Frailty Models in Survival Analysis". In: *Chapman Hall*.

Wong F., Carter C. K. and Kohn, R. (2003). "Efficient Estimation of Covariance Selection Models". In: *Biometrika* 90, pp. 809–830.

Wu, H. and Wang, L. (2019). "Normal frailty probit model for clustered interval-censored failure time data". In: *Biometrical Journal*, pp. 1–14.

Xue, H., Lam, KF., Ben, C., and de, Wolf F. (2006). "Semiparametric accelerated failure time regression analysis with application to interval-censored HIV/AIDS data". In: *Statistics in Medicine* 25, pp. 3850–3863.

Yavuz, A. C. and Lambert, P. (2016). "Semi-parametric frailty model for clustered interval-censored data". In: *Statistical Modelling* 16, pp. 360–391.

Yin, G. and Ibrahim, J.G. (2005). "A class of Bayesian shared Gamma frailty models with multivariate failure time data". In: *Biometrics* 61, pp. 208–216.

Zeng, D., Cai, J., and Shen, Y. (2006). "Semiparametric additive risks model for interval-censored data". In: *Statistica Sinica* 16, pp. 287–302.

Zeng, D., Mao, L., and Lin, D. (2016). "Maximum likelihood estimation for semiparametric transformation models with interval-censored data". In: *Biometrika* 103, pp. 253–271.

Zhang, X., Boscardin, W. J., and Belin, T. R. (2006). "Sampling correlation matrices in bayesian models with correlated latent variables". In: *Journal of Computational and Graphical Statistics* 15, pp. 880–896.

Zhang, Z. and Sun, J. (2010). "Interval-censoring". In: *Statistical Methods in Medical Research* 19, pp. 53–70.

Zhang, Z. and Zhao, Y. (2013). "Empirical likelihood for linear transformation models with interval-censored failure time data". In: *Journal of Multivariate Analysis* 116, pp. 398–409.

Zuma, K. (2007). "A bayesian analysis of correlated interval-censored data". In: *Communicationsin Statistics - Theory and Methods* 36, pp. 725–730.

Appendix A

Chapter 2 Supplementary Materials

A.1 Proofs from Section 2.2

Let T_{i1}, $T_{i2}...T_{ik}$ denote the k failure times of a subject i, with covariate $\boldsymbol{x_i}$. Under the proposed normal frailty multivariate probit model (3.2), it is equivalent to write as

$$\alpha_j(T_{ij}) = -\boldsymbol{x_i'}\boldsymbol{\beta_j} - \zeta_i + \epsilon_{ij}, i = 1, ...n, j = 1, ..., k$$

where $\zeta_i \sim N(0, \sigma^2)$ is the frailty, and ϵ_{ij}s are independent standard normal random variables. Define $Y_j = \alpha(T_{ij})$ for $j = 1, 2, ...k$. Obviously, Y_j has a marginallly normal distribution with variance $1 + \sigma^2$ and their joint distribution is a multivariate normal distribution. Pearson's correlation coefficient between each two of Y_js can be derived as

$$\rho = \frac{cov(Y_1, Y_2)}{\sqrt{var(Y_1)var(Y_2)}} = \frac{\sigma^2}{1 + \sigma^2}.$$

By applying the relationships among Pearson's correlation coefficient, Spearman's correlation coefficient, median concordance κ and Kendall's concordance τ under multivariate normal distribution, $\rho_s = 6\pi^{-1}\sin^{-1}(\frac{\rho}{2})$, $\kappa = 2\pi^{-1}\sin^{-1}(\rho)$ and $\tau = 2\pi^{-1}\sin^{-1}(\rho)$ (See Kruskal, 1958, Hougaard, 2000 among others), (2.4) and (2.5) and (2.6) are proved.

Appendix B

Chapter 3 Supplementary Materials

B.1 Sampling from Multivariate truncated normal distribution

Under the multivariate probit model, we are interested in drawing samples from a truncated multivariate normal distribution. In this chapter, we provide a detailed algorithm from Geweke (1991). (Referenced in Section 3.3 of Chapter 3).

The construction of samples from a J-dimensional normal distribution subject to linear inequality restrictions,

$$x \sim N(\mu, \Sigma),\ a \leq x \leq b \tag{B.1}$$

where μ is a $J \times 1$ mean vector and Σ is a $J \times J$ covariance matrix. The elements of a and b can take $-\infty$ and ∞ respectively. Sampling x from (B.1) is the same as sampling from

$$z \sim N(0, \Sigma),\ \alpha \leq x \leq \beta \tag{B.2}$$

where $\alpha = a - \mu$ and $\beta = b - \mu$. We take $x = \mu + z$. Based on Geweke (1991), a Gibbs sampler is adopted. From the conditional multivariate normal distribution theory, in the non-truncated distributional $N(0, \Sigma)$,

$$E(z_i \mid z_1, \ldots z_{i-1}, z_{i+1}, \ldots, z_J) = \sum_{j \neq i} c_{ij} z_j\ . \tag{B.3}$$

Then the truncated distribution can be represented by

$$z_i = \sum_{j \neq i} c_{ij} z_j + h_i \epsilon_i, \tag{B.4}$$

$$\text{with } \epsilon_i \sim TN\Big(\frac{\alpha_i - \sum_{j \neq i} c_{ij} z_j}{h_i}, \frac{\beta_i - \sum_{j \neq i} c_{ij} z_j}{h_i}\Big). \tag{B.5}$$

where (B.5) is the univariate truncated normal distribution and the vector of coefficients in the conditional mean is denoted by:

$$c^i = (c_{i1}, \ldots, c_{i,i-1}, c_{i,i+1}, \ldots, c_{iJ}),\ i = 1, \ldots, J \tag{B.6}$$

From the conventional theory for the conditional multivariate normal distribution and based on the inverse of a partitioned symmetric matrix (Rao (1965)),

$$c^i = -(\Sigma^{ii})^{-1}\Sigma^{i,<i},\ \text{and}\ h_i^2 = (\Sigma^{ii})^{-1}, \tag{B.7}$$

where Σ^{ii} is the diagonal element of Σ^{-1} and $\Sigma^{i,<i}$ is row i of Σ^{-1} with Σ^{ii} deleted. These computations need only be performed once, before the sampling begins. Starting by assigning initial values for z and sweep through Gibbs cycles, we compute $x = \mu + z$ at the end of each pass. Therefore, the Gibbs steps are summarized as follows:

- Assign initial values $z^0 = 0$.
- Generate J successive variables from

  $$z_i^{(1)} \mid (z_1^{(1)}, \ldots, z_{i-1}^{(1)}, z_{i+1}^{(0)}, \ldots, z_J^{(0)}) \sim f_i(z_1^{(1)}, \ldots, z_{i-1}^{(1)}, z_{i+1}^{(0)}, \ldots, z_J^{(0)}),\ i = 1, \ldots, J.$$

- Repeat at the jth pass,

  $$z_i^{(j)} \mid (z_1^{(j)}, \ldots, z_{i-1}^{(j)}, z_{i+1}^{(j-1)}, \ldots, z_J^{(j-1)}) \sim f_i(z_1^{(j)}, \ldots, z_{i-1}^{(j)}, z_{i+1}^{(j-1)}, \ldots, z_J^{(j-1)}),$$
  $$i = 1, \ldots, J.$$

- Compute $x^{(j)} = \mu + z^{(j)}$ at the end of each pass.

To sample the univariate truncated normal, we adopt the algorithm from Robert (1995). For the one-sided truncation,

$$x \sim N(\mu, \mu^-, \sigma^2),$$

where $N(\mu, \mu^-, \sigma^2)$ is the truncated normal distribution with left truncation point μ^-, i.e. the distribution is as below:

$$f(x|\mu, \mu^-, \sigma^2) = \frac{\exp(-(x-\mu)^2/2\sigma^2)}{\sqrt{2\pi}\sigma(1-\Phi((\mu^- - \mu)/\sigma))}\mathbf{1}_{\{x \geq \mu^-\}}.$$

Robert (1995) adopts an optimal exponential accept-reject algorithm sampling scheme. Without loss of generality, assume that $\mu = 0$ and $\sigma^2 = 1$.

1. Generate $z \sim \mathcal{E}xp(\alpha^*, \mu^-)$;
2. Compute $\rho(z) = exp(-\frac{(z-\alpha^*)^2}{2})$;
3. Generate $u \sim U(0,1)$ and take $x = z$ if $u \leq \rho(z)$, otherwise go back to step 1.

$\mathcal{E}xp(\alpha^*, \mu^-)$ is the translated exponential distribution with density

$$f(z \mid \alpha^*, \mu^-) = \alpha^* e^{-\alpha^*(z-\mu^-)}\mathbf{1}_{\{z \geq \mu^-\}},$$

and the optimal value of $\alpha^* = \frac{\mu^- + \sqrt{(\mu^-)^2+4}}{2}$. For two-sided truncated normal distribution,

$$x \sim N(\mu, \mu^-, \mu^+, \sigma^2),$$

where $N(\mu, \mu^-, \mu^+, \sigma^2)$ is the truncated normal distribution with left truncation point μ^- and right truncation point μ^+, i.e. the distribution is as below:

$$f(x|\mu, \mu^-, \mu^+, \sigma^2) = \frac{\exp(-(x-\mu)^2/2\sigma^2)}{\sqrt{2\pi}\sigma[\Phi((\mu^+ - \mu)/\sigma) - \Phi(\mu^- - \mu)/\sigma)]}\mathbf{1}_{\{\mu^- \leq x \leq \mu^+\}}.$$

Without loss of generality, $\mu = 0$ and $\sigma^2 = 1$. The accept-reject algorithm based on $U(\mu^-, \mu^+)$ is as follows:

1. Generate $z \sim U(\mu^-, \mu^+)$;
2. Compute

$$\rho(z) = \begin{cases} \exp\{-\frac{z^2}{2}\} & \text{if} \quad 0 \in (\mu^-, \mu^+) \\ \exp\{\frac{(\mu^+)^2 - z^2}{2}\} & \text{if} \quad \mu^+ < 0 \\ \exp(\frac{\{\mu^-)^2 - z^2}{2}\} & \text{if} \quad 0 < \mu^- \end{cases}$$

3. Generate $u \sim U(0,1)$ and take $x = z$ if $u \leq \rho(z)$, otherwise go back to step 1.

This accept-reject algorithm is proved more efficient than rejection sampling or the inverse cdf method (Robert, 1995).

B.2 Marginal uniform prior proof from Barnard et al. (2000)

This part serves as proof of Theorem 4.2.1. Under the construction $\Sigma = (\sigma_{ij}) = DRD$, for $i = 1, \ldots, J$ and $j = 1, \ldots, J$.

$$(\sigma_{ij}) = \begin{cases} \sigma_{ij} = d_i d_j r_{ij} & \text{if} \quad i \neq j \\ \sigma_{ii} = d_i^2 & \text{if} \quad i = j \end{cases}$$

Then,

$$|J : \Sigma \to (D, R)| = \frac{\partial \sigma_{ij}}{\partial r_{ij}}.$$

The Jacobian is given as:

$$\begin{aligned} |J : \Sigma \to (D, R)| &= \frac{\partial \sigma_{ij}}{\partial r_{ij}} \\ &= 2^J (\prod_i d_i)^J. \end{aligned} \tag{B.8}$$

Take Σ as a 3×3 covariance matrix as an example,

$$\underset{3\times 3}{\Sigma} = \begin{pmatrix} d_1^2 & d_1 d_2 & d_1 d_3 \\ d_1 d_2 & d_2^2 & d_2 d_3 \\ d_1 d_3 & d_2 d_3 & d_3^2 \end{pmatrix}.$$

Then the Jacobian is:

$$|J : \Sigma \to (D, R)| = \left| \frac{\partial(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{12}, \sigma_{13}, \sigma_{23})}{\partial(d_1, d_2, d_3, r_{12}, r_{13}, r_{23})} \right| \tag{B.9}$$

$$= \begin{vmatrix} 2d_1 & 0 & 0 & d_2 r_{12} & d_3 r_{13} & 0 \\ 0 & 2d_2 & 0 & d_1 r_{12} & 0 & d_3 r_{23} \\ 0 & 0 & 2d_3 & 0 & d_1 r_{13} & d_2 r_{23} \\ 0 & 0 & 0 & d_1 d_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & d_1 d_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & d_2 d_3 \end{vmatrix} \tag{B.10}$$

$$= 2^3 d_1^3 d_2^3 d_3^3 \ . \tag{B.11}$$

From Barnard et al. (2000), $\Sigma(\upsilon, I_J)$ and the inverse Wishart distribution is defined as in (3.18): $f_J(\Sigma|\upsilon) \propto |\Sigma|^{-\frac{1}{2}(\upsilon+J+1)} \exp\Big(-\frac{1}{2} tr(\Sigma^{-1})\Big)$.

$$\begin{aligned} \pi(R, D \mid \upsilon) &\propto |DRD|^{-\frac{1}{2}(\upsilon+J+1)} \exp\Big(-\frac{1}{2} tr(DRD)^{-1}\Big) \times |J| \\ &\propto |R|^{-\frac{1}{2}(\upsilon+J+1)} (\prod_i d_i)^{-(\upsilon+J+1)} (\prod_i d_i)^J \exp\Big(-\frac{1}{2} tr(DRD)^{-1}\Big) \\ &\propto |R|^{-\frac{1}{2}(\upsilon+J+1)} \prod_i \left(d_i^{-(\upsilon+1)} \exp\Big(-\frac{r^{ii}}{2d_i^2}\Big)\right). \end{aligned} \tag{B.12}$$

where r^{ii} is the ith diagonal element of R^{-1}, and the distribution of R is:

$$\begin{aligned} f(R \mid \upsilon) &= \int_0^\infty \pi(R, D|\epsilon) dD \\ &\propto \int_0^\infty |R|^{-\frac{1}{2}(\upsilon+J+1)} \prod_i \left(d_i^{-(\upsilon+1)} \exp\Big(-\frac{r^{ii}}{2d_i^2}\Big)\right) dD. \end{aligned} \tag{B.13}$$

Now let$\theta_i = \frac{r^{ii}}{2d_i^2}$, then

$$\begin{aligned} f(R \mid \upsilon) &\propto \int_0^\infty |R|^{-\frac{1}{2}(\upsilon+J+1)} \prod_i \left(d_i^{-(\upsilon+1)} \exp\Big(-\frac{r^{ii}}{2d_i^2}\Big)\right) dD \\ &\propto |R|^{-\frac{1}{2}(\upsilon+J+1)} \prod_i \int_0^\infty (d_i)^{-(\upsilon+1)} \exp(-\theta_i) \frac{d_i^3}{r^{ii}} d\theta_i \\ &\propto |R|^{-\frac{1}{2}(\upsilon+J+1)} \prod_i \int_0^\infty \left(\frac{d_i^2}{r^{ii}}\right)^{(-\upsilon+2)/2} \exp(-\theta_i) \frac{(r^{ii})^{(-\upsilon+2)/2}}{r^{ii}} d\theta_i \\ &\propto |R|^{-\frac{1}{2}(\upsilon+J+1)} \Big(\prod_i r_{ii}\Big)^{-\frac{\upsilon}{2}} \prod_i \int_0^\infty (\theta_i)^{(\upsilon-2)/2} \exp(-\theta_i) d\theta_i \ . \end{aligned} \tag{B.14}$$

From (B.14), we could see that

$$\pi(R, D) = \pi(R, \theta) = \pi(\theta \mid R)\pi(R), \tag{B.15}$$

where

$$\pi(\theta_i \mid R) \sim \mathcal{G}a\Big(\frac{J+1}{2}, 1\Big), \tag{B.16}$$

$$\pi(R) \propto |R|^{\frac{J(J-1)}{2}-2}\Big(\prod_i |R_{ii}|\Big)^{-\frac{(J+1)}{2}}. \tag{B.17}$$

(B.17) comes from $r^{ii} = \frac{|R_{ii}|}{|R|}$, where R_{ii} is the principal submatrix of R.

There's a nice property of the inverse Wishart distribution, which is that the principal submatrix of an inverse Wishart distribution is also an inverse Wishart distribution. This property can be used to get the marginal distribution of r_{ij} for $i = 1, \ldots, J$ and $j = 1, \ldots, J$. Note that in the special case when choose a 2×2 sub-covariance matrix, the marginal density is

$$f(r_{ij} \mid \upsilon) = (1 - r_{ij})^{\frac{(\upsilon - J - 1)}{2}}. \tag{B.18}$$

And this is $Beta(\frac{\upsilon-J+1}{2}, \frac{\upsilon-J+1}{2})$ on $[-1, 1]$, and is uniform distribution if $\mu = J + 1$. Therefore, we can see that by given the joint distribution of R as (B.17), the marginal distribution for each element of R follows a uniform distribution on $[-1, 1]$.