2016

Frailty Probit Models for Clustered Interval-Censored Failure Time Data

Haifeng Wu
University of South Carolina

Frailty probit models for clustered interval-censored failure time data

by

Haifeng Wu

Bachelor of Science
Beihang University, 2007

Master of Science
Texas Tech University, 2010

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosphy in

Statistics

College of Arts and Sciences

University of South Carolina

2016

Accepted by:

Lianming Wang, Major Professor

Bo Cai, Committee Member

John Grego, Committee Member

Dewei Wang, Committee Member

Lacy Ford, Senior Vice Provost and Dean of Graduate Studies

Acknowledgments

I would like to express my appreciation to my advisor, Dr. Lianming Wang, who provides tremendous support to my study and research throughout years. Without his guidance I would not have completed this dissertation. Besides how to conduct statistical research, I also learned from my advisor to be diligent, always being positive and dedicated to work.

I would like to thank my dissertation committee members, Dr. John Grego, Dr. Bo Cai and Dr. Dewei Wang, for reviewing and providing insights on my dissertation. Their questions, comments and suggestions help me a lot with completing and polishing this dissertation.

In addition, I am very grateful that I was offered the opportunity to study as a graduate student in the Department of Statistics at University of South Carolina. I really appreciate all the faculty, staff and friends who helped me and showed support over the years.

Finally, I want to thank my wife, my parents. They have been unbelievably supportive and understandable for these years. Their patience, love and encouragement are what took me through this journey and will always be priceless for me in my future endeavor.

Abstract

Survival analysis is an important branch of statistics that deals with time to event data or survival data. An important feature of such data is that the survival time of interest is usually not completely known but is censored due to the design of the study or an early dropout. In this dissertation we focus on studying clustered interval-censored data, a special type of survival data. Interval-censored data arise in many epidemiological, social science, and medical studies, in which subjects are examined at periodical follow-up visits. The survival (or failure) time of interest is never exactly observed but is known to fall within an interval formed by two examination times with changed status of the event of interest. Clustered interval-censored data contributes another complication that the failure times within the same cluster are not independent.

Chapter 1 of this dissertation provides a detailed description of interval-censored data with several real data examples and reviews existing regression models and approaches for clustered interval-censored data.

Chapter 2 proposes a novel frailty Probit model for analyzing clustered interval-censored data. The proposed model has several appealing properties: (1) the marginal covariate effects are proportional to the conditional effect and (2) the intra-cluster association can be quantified in terms of several nonparametric association measures in closed form. the proposed Bayesian estimation approach is easy to implement because all parameters and latent variables have their full conditionals in standard form. The approach has excellent performance in estimating the regression parameters and the baseline survival function and is also robust to misspecification of the

frailty distribution.

Chapter 3 extends the frailty Probit model in Chapter 2 to allow modeling both clustered and independent data through the adoption of a mixture distribution for the frailty. The proposed approach provides tests of the existence of intra-cluster association for each cluster via Bayes factors and can identify clusters with strong (weak) correlation. Two different prior structures are considered in our approach, and both lead to good estimation and testing results.

Chapter 4 studies a joint modeling of clustered interval-censored failure times and the sizes of the clusters. The cluster size is modeled as an ordinal response using a parametric Probit model, and a separate frailty semiparametric Probit model is used to model the clustered failure times. The two submodels are connected through a shared random effect. The performance of the proposed model is evaluated through a simulation study.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Interval-censored data

Survival analysis is a branch of statistics dealing with time to event data or failure time data. When the failure time of interest is not observed exactly but known to fall within an interval, interval-censored data arise. Interval-censored data occur naturally when there are periodic follow-ups in clinical studies or medical studies. Below we provide three real life examples to illustrate the structure of interval-censored data and discuss the common research goals in analyzing interval-censored data.

1.1.1 Breast cosmesis data

Adjuvant chemotherapy improves the relapse-free and overall survival time in at least some subgroups of patients treated by mastectomy. However, there is a concern that acute skin reactions may be worse when patients are treated with adjuvant chemotherapy with postoperative radiation or primary radiation for breast cancer. A study was conducted to compare two types of treatments: radiotherapy alone and radiotherapy with adjuvant chemotherapy. The study was interested in the cosmetic effects of the treatments (Finkelstein 1985). The patients in the study had been treated with either radiation alone or with the combined radiation and adjuvant chemotherapy. Depending on the geographical location of the patients, they were followed-up for every 4 to 6 months. Physicians recorded the cosmetic appearances of the patients with respect to the overall cosmetic result. Breast retraction was one

of the least subjective endpoints, therefore the time until breast retraction was used to compare the effect of the two treatment groups. Table 1.1 displays a subset of the data from the study. Each interval represents the time period the breast retraction first appeared. For example, if an observed interval is $(0, 5]$, the retraction was presented at the first test at month 5. Since the retraction appeared before the first test, this observation is left censored. Similarly, if an observation is $(8, 12]$, the retraction did not appear at month 8 but appeared by month 12. The breast retraction time is interval censored. If an observation is $(13, \_]$, the retraction did not appear at the last test at 13 months, the breast retraction time is called right censored.

Table 1.1: Interval of cosmetic deterioration for early breast cancer patients treated with radiotherapy and chemotherapy vs radiotherapy alone.

Radiotherapy			Radiotherapy and Chemotherapy		
$(45, \_]$	$(25, 37]$	$(37, \_]$	$(8, 12]$	$(0, 5]$	$(30, 34]$
$(6, 10]$	$(46, \_]$	$(0, 5]$	$(0, 22]$	$(5, 8]$	$(13, \_]$
$(0, 7]$	$(26, 40]$	$(18, \_]$	$(24, 31]$	$(12, 20]$	$(10, 17]$
$(17, \_]$	$(46, \_]$	$(24, \_]$	$(17, 27]$	$(11, \_]$	$(8, 21]$
$(46, \_]$	$(27, 34]$	$(36, \_]$	$(17, 23]$	$(33, 40]$	$(4, 9]$
$(7, 16]$	$(36, 44]$	$(5, 11]$	$(24, 30]$	$(31, \_]$	$(11, \_]$
$(17, \_]$	$(46, \_]$	$(19, 35]$	$(16, 24]$	$(13, 39]$	$(14, 19]$
$(7, 14]$	$(36, 48]$	$(25, 17]$	$(13, \_]$	$(19, 32]$	$(4, 8]$

1.1.2 HIV data

The HIV data (De Wolf et al., 2001; Van Sighem et al., 2003) is another example of interval censored data, and it has been used frequently in survival analysis literature. The ATHENA database contains over 9000 HIV-infected patients in the Netherlands. The data include information related to demographic information, antiretrovial therapy (ART) information, and clinical measurements. Since 1987, antiretrovial therapy (ART) was used as the treatment of HIV (Mocroft, 1998). The primary goal of the treatment was to reduce the plasma viral load which results in substantial clinical

benefits. The goal of this study was to identify the factors affecting the time to suppression of plasma viral load after antiretrovial therapy (ART). Viral suppression was defined to be less than 500 copies/ml. Patients were measured one month after the ART treatment. Afterward, they were checked every three to four months until the viral suppression was detected or the end of the study. About 50% of all patients had their suppression of plasma viral load at the first measurement. Therefore, the time to suppression is known to be less than the first check time, which results in left-censored observations. About 35% of the patients had viral suppression between two adjacent examinations, which results in interval-censored observations, and the remaining 15% of the patients did not have viral suppression by the last scheduled measurement, and their suppression times are greater than the time from the treatment to their last check times, resulting in right-censored observations.

1.1.3 Dental data

The third example is a dental study conducted in Hong Kong. Children from eight kindergartens participated in the study. Children with dentin caries in at least 1 primary anterior tooth were included in the study and were treated with 5 different treatments. Follow up examinations were conducted every 6 months after the treatment. The purpose of the study is to determine the effect of different treatments in arresting dentin caries. A total of 375 children were included in the study, and 1483 surfaces with dentin caries from those children were included in the study. The study was interested in the effect of different treatments on the time of arresting dentin caries, which cannot be observed exactly. Therefore, the resulting data is interval censored data. However, the data structure is slightly different from the previous two examples since a hierarchical structure is presented. The subjects in the study are the dentin caries. However, the dentin caries cannot be treated independently. Those dentin caries come from the same child and share some common character-

istics, which induces natural correlation among those times to dentin caries. Thus, this study yields clustered interval-censored data, where the time to dentin caries are correlated within the same cluster.

1.2 Popular statistical models for interval censored data

In most studies, people are interested in estimating the survival function as well as the covariate effects. Below we review a few popular statistical models in the literature for modeling failure time or survival data. Let $\boldsymbol{x}$ denote the vector of covariates, and T is the failure time of interest in the study. The most popular model is the proportional hazards (PH) model introduced by Cox (1972). The PH model specifies its hazard function as follows:

$$\lambda(t|\boldsymbol{x}) = \exp(\boldsymbol{x}'\boldsymbol{\beta})\lambda_0(t)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function. The partial likelihood approach introduced by Cox (1975) allows one to estimate the regression parameter without the need of estimating the unknown baseline hazard function, which makes the PH model widely used in the literature.

Another popular model is the accelerated failure time (AFT) model. Under the AFT model, the failure time can be modeled as follows:

$$\log(T) = \boldsymbol{x}'\boldsymbol{\beta} + e$$

Unlike the PH model, the AFT model assumes that the effect of the covariate is to accelerate or decelerate the failure time by some constant. AFT model is a parametric model when the distribution of the error term is chosen, and the model is fully specified. The AFT model can also be semiparametric if assuming unknown distribution for e. Maximum likelihood can be applied for the estimation procedure.

Another popular model for modeling survival data is the proportional odds (PO) model introduced by Bennett (1983). Instead of assuming the covariate effect is to

increase or decrease the hazard, the PO model assumes the effect of covariate is to increase or decrease the odds of the event of interest:

$$\frac{1 - S(t, x)}{S(t, x)} = \frac{1 - S_0(t)}{S_0(t)} \exp(x'\beta)$$

where $S(t|x)$ is the survival function given covariate x and $S_0(t)$ is the baseline survival function.

Another semiparametric regression model is the Probit model. The Probit model has gained its popularity recently in survival data literature. It is a good alternative to the previous three models, and it is very easy to use. Probit model specifies its cumulative distribution function of the failure time as follows:

$$F(t|x) = \Phi(\alpha(t) + x'\beta),$$

where Φ is the CDF of a standard normal random variable, and $\alpha(\cdot)$ is a unknown nondecreasing function.

1.3 Existing approaches

Semiparametric regression models are really popular for modeling survival data since they usually do not specify the functional form of the baseline survival function. The extra flexibility it provided compared to a fully specified parametric model is desirable because in real life data can be complex and does not follow a pre-specified distribution.

Under the PH model, Finkelstein (1986) treated the interval-censored observations as incomplete observations from grouped response time data and fitted the PH model with maximum likelihood estimation, and a generalization of log-rank test for testing the covariate effects was provided. Huang (1996) considered simultaneous estimation of regression parameters and a nonparametric maximum likelihood estimator of the baseline survival distribution. Satten (1998) developed an imputation approach to

convert the problem to a current status data problem. Cai (2003) provided a piecewise linear spline for estimating the baseline hazard function. Pan (2000) provided a two step imputation approach under PH model with the unspecified baseline survival function.

Under the AFT model, Rabinowitz (1995) developed a class of score statistics for estimating the regression coefficients and an estimator of the covariance of the scores is derived. Xue (2006) proposed a sieve maximum likelihood method to simultaneously estimate all the parameters. Tian (2006) used an efficient MCMC based resampling method for obtaining simultaneously the point estimator and a consistent estimator of its variance -covariance matrix.

Under the PO model, Huang (1997) used sieve maximum likelihood estimator to estimate the finite dimensional regression parameters. Rabinowitz (2000) allowed one to use conditional logistic regression routines in standard statistical packages to fit the PO model.

Lin (2009) proposed a semiparametric probit model with monotone splines. Probit model is well established for dealing with continuous data. But it is not commonly used for modeling failure time data. In this dissertation, we show it can provide very fast and efficient estimation from a Bayesian perspective due to the structure of the model with the specific prior specifications. This dissertation studies three important topics related to interval-censored failure time data using the extended semiparametric probit models.

1.4 Topics to be studied

When the subjects in the study are hierarchically structured, with observation units grouped in clusters, it is important to accommodate this feature into the analysis, otherwise the estimation will be biased. Take the cow udder quarter infection data (Goethals 2009) for an example: the interest of the study was to investigate the time

to infection for each udder quarter of a cow. Undeniably, four udder quarters from the same cow share common characteristics. Ignoring such feature and treating all failure times as independent will lead to biased estimation results. Goethals (2009) adapted the frailty PH model with shared gamma frailty and used frenquetist approach to find the MLE. Bellamy (2004) used weibull frailty model and obtaining the MLE with Newton-Raphson technique. Wong (2005) adapted a Bayesian approach under the PH model. Henschel (2009) provided a semiparametric PH model for interval censored data with log-normal frailty terms. In Chapter 2, we provide an alternative approach under a frailty Probit model with a normal frailty term to incorporate the cluster effects. Our model enjoys several good features. First, the marginal distribution of failure time is also a semiparametric Probit model under the structure of our model. Second, the conditional covariate effects are proportional to the marginal covariate effects. Third, the intra cluster association can be measured by two nonparametric measures and can be easily calculated from the MCMC result.

The second problem of interest here is to provide a statistical test for the cluster effect when the hierarchy structure is presented in the data. Bellamy (2004) developed a score test for the variance of the frailty term under the weibull model with added frailty term in order to determine over-dispersion from a frequentist perspective. Wong (2005) used the same model and provided a formula to estimate the intra-cluster correlation between the items within same cluster using a Bayesian methodology. However, all those tests provide a global result with no implication for a specific cluster if interested. In Chapter 3, we developed a Bayesian hypothesis testing via Bayes Factor that can test cluster effects globally and locally. That is, our model and approach can be used to identify which clusters are "real" clusters with strong correlation and which clusters actually have weak or no correlation.

The third topic in this dissertation is motivated by a more complex setting when the cluster size may have an influence on the outcomes, or vice versa, or possibly

they both are influenced by a third, unobserved latent variable. This is called the informative cluster size problem in the literature. The problem is very common in continous and binary response data in the literature and has been well addressed. In volume-outcome studies (Panageas et al. 2007) specialized surgeons treating many patients may have better outcomes than those treating few patients; in periodontal studies (Williamson et al. 2003; Wang et al. 2011) patients with fewer teeth tend to have a poorer condition for the still remaining teeth; in radiation toxicity studies (Datta and Satten 2008), the number of measurements on successive measurement on an individual depends on the number of radiation therapies, which in turn depend on the underlying severity of cancer. When the cluster size is informative, Hoffman (2001) proposed a within-cluster resampling method for unbiased estimation of the covariate effect. Later Williamson (2003) developed a cluster weighted generalized estimating equation by using the inverse weighting based on the cluster size. Chen (2011) proposed a joint modeling of outcome and cluster size. The informative cluster size problem can also occur for interval censored failure time data. One example is the lymphatic filariasis study discussed by Williamson (2008). However, very few works in the literature. Fan (2011) used marginal AFT model to adapt the informative cluster size feature and proposed a simple adjustment through inverse cluster size reweighting. Williamson (2008) also used cluster weighted approach under weibull and Cox model. Zhang (2010) gave two procedures under parametric framework, one being a weighted score function and the other making use of the within-cluster resampling (WCR) idea. Kim (2010) proposed a joint modeling of failure times and cluster size from frenquentist's prospective. No research has been found on this topic using Bayesian methods. In Chapter 4, we propose a joint modeling of failure times and cluster size from Bayesian perspective. We proposed to use a semiparametric Probit model for the interval censored failure time and an ordinal model for the cluster size, and the two models are linked by a shared frailty. An efficient Gibbs

sampler is proposed based on a data augmentation. This project is on-going and the performance of our approach will be reported in the near future.

Chapter 2

Frailty Probit model for clustered interval-censored failure time data

Clustered interval-censored data commonly arise in many studies of biomedical research where the failure time of interest is subject to interval-censoring and subjects are correlated for being in the same cluster. In this chapter, we propose a new frailty semiparametric Probit regression model to study covariate effects on the failure time by accounting for the intra-cluster dependence. The proposed normal frailty Probit model enjoys several nice properties: (1) the marginal distribution of the failure time is a semiparametric Probit model, (2) the regression parameters can be interpreted as the conditional covariate effects given frailty or the marginal covariate effects up to a multiplicative constant, and (3) the intra-cluster association can be summarized by two nonparametric measures in simple and explicit form. A fully Bayesian estimation approach is developed based on the use of monotone spline for the unknown nondecreasing function and a data augmentation using normal latent variables. The Gibbs sampler is straightforward to implement since all unknowns have standard form of full conditional distributions. The proposed method performs well in estimating the regression parameters and is robust to frailty distribution misspecification in our simulation studies. Two real-life data sets are analyzed for illustration.

2.1 Introduction

Interval-censored data arise naturally in many epidemiological and biomedical studies in which subjects undergo examinations periodically. As a consequence, the failure time of interest cannot be observed exactly but is known to fall within some time interval (Sun, 2006). Furthermore, in many of such studies, subjects are correlated because of sharing some common characteristics for being in the same cluster, leading to clustered interval-censored data. Examples of clustered interval-censored data studied in the literature include the diabetic retinopathy data (Ross, 1999; Lam et al., 2010), the asthma data (Bellamy et al., 2004), the dental data (Wong et al., 2005), the lymphatic filariasis data (Williamson et al., 2008), and the cow udder infection data (Goethals et al., 2009). In those examples, clusters are naturally formed when some subjects are coming from the same animal, person, or family, etc. The intra-cluster correlation contributes additional complication to the analysis of interval-censored data. Ignoring such intra-cluster association may lead to biased estimation (Bellamy et al., 2004). In this chapter, we focus on estimating covariate effects on the failure time subject to interval-censoring by taking into account of the intra-cluster association among subjects.

There are some existing approaches for analyzing clustered interval-censored data. Bellamy et al. (2004) proposed a normal frailty Weibull model and developed a Newton-Raphson algorithm based on a numerically approximated likelihood with Gaussian quadrature. Wong et al. (2005) developed a Bayesian method under the normal frailty Weibull model and Wong et al. (2006) extended it to a frailty Cox model and implemented the new method with WinBUGs. Goethals et al. (2009) proposed a shared gamma frailty Weibull model and developed a Newton-Raphson algorithm with all derivatives in explicit form. Zhang and Sun (2010) studied a marginal Weibull survival model and developed an estimating equation-based ap-

proach and a resampling-based approach to take into account the informative cluster size. Lam et al. (2010) proposed a multiple imputation method based on gamma frailty Weibull model. Kim (2010) proposed a joint modeling of the survival time and the informative cluster size and used a frailty Cox model for the survival time. All these approaches are within the proportional hazards model framework.

In this chapter, we propose a novel frailty semiparametric Probit model to analyze clustered interval-censored data. The frailty Probit model specifies the conditional cumulative distribution function (CDF) of failure time T given frailty ξ in the following form,

$$F(t|\boldsymbol{x},\xi) = \Phi\{\alpha(t) + \boldsymbol{x}'\boldsymbol{\beta} + \xi\}, \tag{2.1}$$

where $\Phi(\cdot)$ is the CDF of a standard normal random variable, $\boldsymbol{x}$ is the covariate vector, $\xi \sim N(0,\sigma^2)$ is the frailty term, and $\alpha(\cdot)$ is an unknown increasing function with $\alpha(0) = -\infty$ and $\alpha(\infty) = \infty$. The unspecified function α makes model (1) a semiparametric regression model. An equivalent form of model (1) is

$$\alpha(T) = -\boldsymbol{x}'\boldsymbol{\beta} - \xi + \epsilon,$$

where $\xi \sim N(0,\sigma^2)$ and $\epsilon \sim N(0,1)$. To see the equivalence clearly, we have $\Pr(T \leq t|\boldsymbol{x},\xi) = \Pr\{\alpha(T) \leq \alpha(t)|\boldsymbol{x},\xi\} = \Pr(\epsilon \leq \alpha(t) + \boldsymbol{x}'\boldsymbol{\beta} + \xi|\boldsymbol{x},\xi) = \Phi\{\alpha(t) + \boldsymbol{x}'\boldsymbol{\beta} + \xi\}$.

While the proposed frailty Probit model has a simple form, surprisingly it is not studied in the survival literature. The proposed normal frailty Probit model has several appealing properties. First, the marginal distribution of T is a semiparametric Probit model of Lin and Wang (2010), which indicates that marginal and conditional CDF of T belong to the same family. Second, the conditional covariate effects given the frailty are proportional to the marginal covariate effects. This allows one to estimate the marginal covariate effects by fitting the frailty model (1) directly. Third, the intra-cluster association can be simply summarized by two nonparametric association measures in simple and explicit form. Details of these properties will be described

in Section 2.2. It is worth noting that such nice properties are not pertained under existing survival models in the literature.

Although the proposed frailty Probit model can be used to model other types of survival data, we focus on its application to clustered interval-censored data and develop an efficient Bayesian estimation approach in this chapter. We develop an efficient Bayesian method for analyzing clustered interval-censored data under the normal frailty Probit model. Specifically, we model the unknown increasing function $\alpha(\cdot)$ with monotone splines of Ramsay (1988) and estimate the regression parameters and spline coefficients jointly. The proposed Gibbs sampler is promising and straightforward to implement because the full conditional distributions for all unknowns are in standard form. Simulation results suggest that the proposed method works very well in estimating the regression parameters as well as the intra-cluster association and that its performance is robust to frailty distribution misspecification.

The remainder of the chapter is organized as follows. Section 2.2 discusses the theoretic properties of the normal frailty Probit model. Section 2.3 presents our estimation procedure in detail. Section 2.4 evaluates the proposed method using simulation. Section 2.5 gives two real-life data applications. Section 2.6 concludes with discussions.

2.2 The normal frailty Probit model

2.2.1 Marginal distribution and marginal effect

The frailty Probit model (1) is an extension to the semiparametric Probit model of Lin and Wang (2010) by incorporating a normal frailty term that induces the correlation among the subjects in the same cluster. The regression coefficients $\boldsymbol{\beta}$ in (1) can be interpreted as the conditional covariate effects on the transformed failure time given the frailty ξ. While this interpretation is appealing, it is conditioning on the unobserved frailty. In this situation, marginal covariate effects are usually

preferred when they are tractable.

Integrating out the normal frailty in the conditional CDF (1), we obtain the marginal CDF of the failure time T in the following form,

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

where $\alpha^*(t) = c\alpha(t)$, $\boldsymbol{\beta}^* = c\boldsymbol{\beta}$, and $c = (1+\sigma^2)^{-1/2}$. This result first implies that the failure time T follows a marginal semiparametric Probit model (Lin and Wang, 2010). Second, there is a multiplicative relationship between the conditional covariate effects $\boldsymbol{\beta}$ and marginal covariate effects $\boldsymbol{\beta}^*$. The multiplicative constant is a deterministic function of the normal frailty variance σ^2. Due to this relationship, the regression parameters $\boldsymbol{\beta}$ can be informally interpreted as marginal covariate effects up to a constant, and the inferences based on $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ will lead to the same conclusion. This relationship also allows us to obtain the exact marginal covariate effects on the failure time by just fitting the frailty model.

2.2.2 Intra-cluster association

The dependence among the subjects in the same cluster is induced by sharing a common frailty in the cluster. The variance of frailties σ^2 measures the correlation among the subjects in the same cluster. In this following, we quantify the intra-cluster association for clustered data in terms of two commonly used nonparametric association measures: Spearman's rank correlation coefficient ρ_s and median concordance κ (Kruskal, 1958; Hougaard, 2000).

For illustration, let T_1 and T_2 denote the two correlated failure times for two subjects in the same cluster. The two subjects can potentially have different covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Spearman's rank correlation coefficient (Kruskal, 1958) is defined as

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

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, and S_1^{-1} and S_2^{-1} are the inverse functions of S_1 and S_2, respectively.

Median concordance, also called quadrant measure by Kruskal (1958), is another nonparametric measure of association between correlated random variables and is defined as

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

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

Spearman's rank correlation coefficient ρ_s and median concordance κ (Kruskal, 1958; Hougaard, 2000) have been widely used for modeling correlated responses, especially for non-survival data, due to the following nice properties. First, they are nonparametric in the sense that their definitions do not rely on specific forms of the distributions of the correlated responses. Second, both of them have a good interpretation of the correlation. They take values 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. They both take 0 when the responses are independent. Third, both of the two measures are invariant to marginal monotone transformations. Here by saying ρ_s is invariant to monotone transformation, we mean $\rho_s(T_1, T_2) = \rho_s(g_1(T_1), g_2(T_2))$ for any two increasing (decreasing) transformations g_1 and g_2. The following theorem summarizes the intra-cluster association explicitly under the normal frailty Probit model (1) for clustered survival or non-survival data.

Theorem 1: The intra-cluster association for clustered data under the normal frailty Probit model (1) is characterized by Spearman's correlation coefficient ρ_s and median

concordance κ as follows,

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

where $\rho = \sigma^2/(1+\sigma^2)$.

The proof of this theorem is sketched in the appendix. This theorem is promising as it provides explicit expression of measures to quantify the intra-cluster association. From Theorem 1, it agrees with the common sense that the association is determined by the frailty variance and that a larger variance σ^2 will lead to a stronger dependence among the subjects within the same cluster. It is also clear that the intra-cluster association is positive under model (1) since ρ is positive. This is expected because ξ is a shared frailty in model (1) by all subjects in the same cluster. Another interesting observation is that both ρ_s and κ are free of covariates, indicating that the intra-cluster association does not depend on the covariates of the subjects.

Pearson's correlation coefficient is widely used for describing linear correlation for bivariate distributions, especially bivariate (multivariate) normal distribution. However, it seldom appears in survival literature because it usually does not have a simple or explicit form under commonly used survival models. It is also the case under the normal frailty model (1). The major reason of that is that Pearson's correlation coefficient is not invariant under marginal monotone transformations. Kendall's concordance τ is commonly used to measure the correlation between bivariate random variables in the survival literature. It is a rank-based nonparametric measure and is invariant under marginal transformation. However, Kendall's τ does not seem to be well defined for the clustered data because its definition requires two independent pairs of correlated random variables.

2.3 The proposed approach

2.3.1 Data and likelihood

Suppose there are n clusters in a survival study. Let T_{ij} denote the failure time of interest for the jth subject in the ith cluster and $(L_{ij}, R_{ij}]$ the observed interval for T_{ij}, $j = 1, ..., m_i$ and $i = 1, ..., n$. The total number of subjects is $M = \sum_{i=1}^{n} m_i$. We rewrite the normal frailty Probit model at the subject level as follows,

$$\alpha(T_{ij}) = -\boldsymbol{x}'_{ij}\boldsymbol{\beta} - \xi_i + \epsilon_{ij}, \tag{2.2}$$

where ϵ_{ij}s are independent standard normal random variables, $\boldsymbol{x}_{ij}$ is the covariate vector associated with the jth subject in cluster i, and $\xi_i \sim N(0, \sigma^2)$ is the shared frailty among all the subjects in cluster i. The observed data for clustered interval-censored data are thus $\{(L_{ij}, R_{ij}), \boldsymbol{x}_{ij}\}, j = 1, ..., m_i, i = 1, ..., n$.

In this chapter, we assume that the failure time is independent of the observation process that produces the observed interval conditioning on the covariates. This non-informative censoring assumption is quite general and is commonly adopted in the literature for studying interval-censored data. Under this assumption, the observed likelihood can be written as

$$L_{obs} = \prod_{i=1}^{n} \int \sigma^{-1}\phi(\sigma^{-1}\xi_i) \prod_{j=1}^{m_i} \{F(R_{ij}|\boldsymbol{x}_{ij}, \xi_i) - F(L_{ij}|\boldsymbol{x}_{ij}, \xi_i)\} d\xi_i, \tag{2.3}$$

where $\phi(\cdot)$ is the density function of a standard normal random variable. The integrals in the observed likelihood (3.1) do not have an explicit form when $m_i \geq 2$, and this makes the observed likelihood difficult to use directly for estimating the unknown parameters $(\boldsymbol{\beta}, \sigma^2, \alpha)$ using Bayesian methods. To alleviate such difficulty, we consider the following conditional likelihood L_{con} by treating all ξ_is as latent variables,

$$\begin{aligned} L_{con} = & \prod_{i=1}^{n} \prod_{j=1}^{m_i} F(R_{ij}|\boldsymbol{x}_{ij}, \xi_i)^{\delta_{ij1}} \{F(R_{ij}|\boldsymbol{x}_{ij}, \xi_i) - F(L_{ij}|\boldsymbol{x}_{ij}, \xi_i)\}^{\delta_{ij2}} \\ & \{1 - F(L_{ij}|\boldsymbol{x}_{ij}, \xi_i)\}^{\delta_{ij3}} \sigma^{-1}\phi(\sigma^{-1}\xi_i) \end{aligned} \tag{2.4}$$

where δ_{ij1}, δ_{ij2} and δ_{ij3} are the censoring indicators for the jth subject in cluster i indicating left, interval, and right censoring, respectively. The introduction of these censoring indicators is to distinguish the three censoring types and to help make clear of our estimation procedure below.

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

Estimating under the normal frailty semiparametric Probit model (1) is challenging due to the existence of the infinite-dimensional parameter α. The same problem occurs when analyzing interval-censored data under all other semiparametric regression models because there does not seem to exist an appropriate partial likelihood that does not contain the unspecified function. To reduce the number of parameters in α while also maintain adequate modeling flexibility, we model $\alpha(t)$ with monotone splines of Ramsay (1988) following the idea in Lin and Wang (2010),

$$\alpha(t) = \gamma_0 + \sum_{l=1}^{k} \gamma_l b_l(t), \tag{2.5}$$

where b_l's are monotone spline basis functions and $\{\gamma_l\}_{l=1}^{k}$ are nonnegative coefficients to ensure the monotonicity of α. The spline basis functions b_l's are essentially piecewise polynomials, and each of the basis function has three stages: equal to 0 at the first stage, increasing from 0 to 1 at the second stages, and then staying plateau at the third stage. These stages and shape are determined by the knot placement and the degree of the splines. Knots are usually a sequence of increasing time points in an interval, within which one wish to estimate the unknown function. The degree controls the overall smoothness of the basis functions, taking value 1 for piecewise linear, 2 for quadratic functions, and 3 for cubic functions, etc. Once the knots and degree are specified, the spline basis functions are deterministic and can be obtained using iterative formula. Our R function for calculating such basis functions is available upon request.

Under the monotone spline representation (2.5), the only unknown parameters involved in α are the spline coefficients γ_l's. The number k of spline coefficients, i.e., the number of basis functions, is equal to the number of interior knots plus the degree (Ramsay, 1988). In general, the more knots taken, the more modeling flexibility, when using splines. However, using too many knots requires much additional computational time and may cause over-fitting problems. Following Wang and Lin (2011) and Wang and Dunson (2011) among others using the monotone splines, we take a moderate number ($10 \sim 30$) of equal-spaced knots to balance the computational burden and modeling flexibility. A shrinkage prior will be used for all spline coefficients functioning to penalize large coefficients and shrink the coefficients for unnecessary basis functions towards zero.

2.3.3 Prior specification and posterior computation

The conditional likelihood (3.2) is still complicated for sampling unknown parameters using Bayesian methods with any prior specifications. Although one can use Metropolis-Hastings or adaptive rejection Metropolis sampling algorithms in this case, we aim to develop a more efficient method that allows to sample all unknowns from standard distributions. To this end, we adopt the following data augmentation motivated by Lin and Wang (2010),

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

where $t_{ij} = R_{ij}1_{(\delta_{ij1}=1)} + L_{ij}1_{(\delta_{ij1}=0)}$, i.e., t_{ij} takes the right end point of the observed interval in the case of right censoring and takes the left end point otherwise for all i and j. The augmented data likelihood function is,

$$L_{aug} = \prod_{i=1}^{n} \left[\prod_{j=1}^{m_i} \phi\{z_{ij} - \alpha(t_{ij}) - \boldsymbol{x}'_{ij}\boldsymbol{\beta} - \xi_i\} 1_{C_{ij}}(z_{ij}) \right] \sigma^{-1}\phi(\sigma^{-1}\xi_i), \tag{2.6}$$

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

augmented likelihood leads to the conditional likelihood (3.2). The augmented likelihood (3.3) is very appealing because it leads to a normal distribution form for each of the unknown parameters and latent variables.

We take the following prior specifications: a multivariate normal prior $N(\boldsymbol{\beta}_0, \Sigma_0)$ for $\boldsymbol{\beta}$, a gamma prior $\mathcal{G}a(a_\sigma, b_\sigma)$ for frailty precision σ^{-2}, a normal prior $N(m_0, \nu_0^{-1})$ for the unconstrained γ_0, and independent exponential priors $Exp(\eta)$ for all $\{\gamma_l\}_{l=1}^k$. We further assign a gamma prior $\mathcal{G}a(a_\eta, b_\eta)$ for η. The independent Exponential priors for γ_l's and Gamma prior for the hyperparameter η have been proved successful to shrink the spline coefficients towards zero and prevent overfitting problems in Lin and Wang (2010), Cai et al. (2011), Wang and Dunson (2011) among others. These priors are natural and allow easy sampling of all the parameters from their full conditional distributions in standard form. Combining these priors and the augmented likelihood (3.3), we develop the following Gibbs sampler.

1. Sample z_{ij} from a truncated normal, $N(\alpha(t_{ij}) + \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \xi_i, 1)1_{C_{ij}}(z_{ij})$, for each j and i.

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

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

3. Sample all γ_l's for $l = 1, ..., k$. For each $l \geq 1$, let $W_l = \sum_{i=1}^{n}\sum_{j=1}^{m_i} b_l^2(t_{ij})$.

   (a) If $W_l = 0$, sample γ_l from the prior $Exp(\eta)$.

   (b) If $W_l > 0$, sample γ_l from $N(E_l, W_l^{-1})1_{(\gamma_l > d_l^*)}$, where

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

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

4. Sample $\boldsymbol{\beta}$ from $N(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$, where $\hat{\Sigma} = (\Sigma_0^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{m_i} \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}')^{-1}$ and

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

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

$$\mu_i = \sigma_i^2\left[\sum_{j=1}^{m_i}\{z_{ij} - \alpha(t_{ij}) - \boldsymbol{x}_{ij}'\boldsymbol{\beta}\}\right].$$

6. Sample η from $\mathcal{G}a(a_\eta + k, b_\eta + \sum_{l=1}^{k}\gamma_l)$.

7. Sample σ^{-2} from $\mathcal{G}a(a_\sigma + 0.5n, b_\sigma + 0.5\sum_{i=1}^{n}\xi_i^2)$.

This Gibbs sampler is very appealing in that all the full conditional distributions are standard distributions and easy to sample from. This property is not pertained in most of the Bayesian methods for analyzing survival data. The Gibbs sampler enjoys good mixing and fast convergence from our observation.

2.4 Simulation studies

Simulation studies were conducted to evaluate the performance of our method. We considered the following model for failure time T_{ij} involving both discrete and continuous covariates,

$$F(t|x_{ij1}, x_{ij2}, \xi_i) = \Phi\{\alpha(t) + x_{ij1}\beta_1 + x_{ij2}\beta_2 + \xi_i\},$$

where x_{ij1} was generated from a standard normal distribution, x_{ij2} was generated from a Bernoulli distribution with probability of success 0.5, and ξ_i was generated from a normal distribution with mean 0 and standard deviation 1. We took true $\alpha(t) = 1 + t + 2\log(t)$, true β_1 equal to 1 or 0, and true β_2 equal to 1, 0, or -1, yielding 6 simulation setups. We generated failure time T_{ij} by solving equation $F(T_{ij}|x_{ij1}, x_{ij2}, \xi_i) = u_{ij}$ numerically, where u_{ij} is a random number from uniform distribution $U(0, 1)$ for each i and j. We generated the observed interval $(L_{ij}, R_{ij}]$ for the failure time T_{ij}

as follows. First we generated an observation process for each subject. We took a random number of observational times for each subject so that subjects can have different numbers of observational times. The random number was taken to be 1 plus a Poisson random variable with mean 3. The observation times were obtained by generating the gap times between adjacent observation times independently from an exponential distribution with mean 0.3. Then the observed interval for T_{ij} was determined by the two adjacent observation times (may include 0 or ∞) that contains T_{ij}. The specification of the observation process was chosen so that none of the censoring types dominates the others. For example, in the case of no covariates (i.e., $\beta_1 = \beta_2 = 0$), there are on average 23.58% left-censored observations, 54.69% interval-censored observations, and 21.73% right-censored observations across all simulated data sets. We generated 500 data sets for each setup and each data set contains 50 clusters with 4 subjects in each cluster.

To specify monotone splines, we used 2 for the degree to ensure adequate smoothness of the splines and took 14 equally spaced interior knots between the minimum and maximum values of the end points of the observed intervals excluding 0 and a ∞ for each data set. This leads to 16 basis spline functions in use throughout the simulation. We adopted the following specifications of the priors for the unknown parameters. We took $m_0 = 1$ and $\nu_0 = 0.1$ leading to a normal prior for γ_0 with a large variance, $a_\eta = b_\eta = 1$ leading to a $\mathcal{G}a(1,1)$ prior for η, $a_\sigma = b_\sigma = 1$ leading to a $\mathcal{G}a(1,1)$ prior for σ^{-2}, and $\boldsymbol{\beta}_0 = 0$ and $\Sigma_0 = M(\sum_{i,j} \boldsymbol{x}_{ij}\boldsymbol{x}'_{ij})^{-1}$ in the bivariate normal prior for $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, where $M = 200$ is the total number of subjects in each simulated data. The prior distribution of $\boldsymbol{\beta}$ is a g-prior with a unit information variance (Zellner, 1986). Fast convergence of the MCMC was observed in our simulation and this is probably due to the fact that all the full conditional distributions are of standard form in the proposed Gibbs sampler. We summarized results based on 5000 iterations of MCMC after discarding a first 1000-iteration as a burn-in. Convergence

of MCMC was checked by using various convergence criteria in the R package CODA (Plummer et al., 2006).

Table 2.1 shows the frequentist operating characteristics of the key parameter estimates from the proposed Bayesian method: BIAS the difference between the average of the 100 point estimates (posterior means) and the true value, ESD the average of the estimated standard deviations of their posterior distributions across the 500 data sets, SSD the sample standard deviation of the 500 point estimates, and CP95 the 95% coverage probability, i.e., the proportion of the 95% credible intervals from 500 data sets that include the true value of the parameter. As seen from Table 2.1, the proposed method works very well in estimating the regression parameters and the standard deviation of frailty distribution σ 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 1 also provides the estimation results of the intra-cluster association in terms of Spearman's correlation coefficient ρ_s and median concordance κ using Theorem 1. These results suggest that our method can estimate the intra-cluster association very accurately.

Our method is developed under an assumption that the frailty has a normal distribution. To investigate how sensitive our method is to this assumption, we conducted additional simulations to apply our method to some cases of non-normal frailties. Two scenarios were considered when generating data: (1) a mixture of normal with $\xi_i \sim 0.45N(0.5, 0.4^2) + 0.55N(-0.5, 0.18^2)$ and (2) a log-Gamma distribution, i.e., $\exp(\xi_i) \sim \mathcal{G}a(1, 1)$. Everything else was kept the same as in the above simulations. Table 2.2 shows the estimation results of regression parameters under the two scenarios based on 100 data sets, with each data set containing 50 clusters and each cluster containing 4 subjects. The results are very promising with small bias in the point estimates, ESD close to SSD, and CP95 close to 0.95. This suggests that our method is robust to frailty distribution misspecifications. This robustness makes our

Table 2.1: Performance of the proposed method in the case of using 50 clusters. 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 CP95 the 95% coverage probability.

True	BIAS	SSD	ESD	CP95
β_1=0	0.0035	0.1210	0.1122	0.922
β_2=-1	-0.0540	0.2380	0.2400	0.948
$\sigma = 1$	0.0426	0.1935	0.1959	0.954
$\rho = 0.48$	0.0072	0.0839	0.0878	0.954
$\kappa = 0.33$	0.0100	0.0641	0.0666	0.954
β_1=1	0.0072	0.0839	0.0878	0.954
β_2=-1	-0.0550	0.2408	0.2469	0.956
$\sigma = 1$	0.0448	0.1985	0.1998	0.958
$\rho = 0.48$	0.0075	0.0863	0.0893	0.958
$\kappa = 0.33$	0.0104	0.0659	0.0677	0.958
β_1=0	0.0016	0.1121	0.1123	0.952
β_2=0	-0.0300	0.2228	0.2221	0.954
$\sigma = 1$	0.0365	0.1896	0.1961	0.968
$\rho = 0.48$	0.0045	0.0832	0.0884	0.968
$\kappa = 0.33$	0.0080	0.0634	0.0669	0.968
β_1=0	-0.0018	0.1152	0.1152	0.948
β_2=1	-0.0019	0.2628	0.2455	0.936
$\sigma = 1$	0.0305	0.1862	0.1992	0.962
$\rho = 0.48$	0.0016	0.0825	0.0902	0.962
$\kappa = 0.33$	0.0058	0.0627	0.0681	0.962
β_1=1	0.0329	0.1612	0.1524	0.932
β_2=0	-0.0226	0.2285	0.2278	0.958
$\sigma = 1$	0.0350	0.1911	0.1987	0.960
$\rho = 0.48$	0.0034	0.0839	0.0896	0.960
$\kappa = 0.33$	0.0072	0.0639	0.0678	0.960
β_1=1	0.0311	0.1593	0.1560	0.944
β_2=1	0.0044	0.2716	0.2512	0.920
$\sigma = 1$	0.0317	0.1928	0.2027	0.974
$\rho = 0.48$	0.0014	0.0856	0.0915	0.974
$\kappa = 0.33$	0.0058	0.0650	0.0692	0.974

method more appealing and more applicable to handle complicated real life clustered interval-censored data.

Table 2.2: Simulation results when the frailty distribution is misspecified. Scenario 1: $\xi_i \sim 0.45N(0.5, 0.4^2) + 0.55N(-0.5, 0.18^2)$ and scenario 2: $\exp(\xi_i) \sim \mathcal{G}a(1, 1)$.

	Scenario 1				Scenario 2			
True	BIAS	ESD	SSD	CP95	BIAS	ESD	SSD	CP95
β_1=0	0.0014	0.1080	0.1144	0.934	-0.0068	0.1154	0.1146	0.960
β_2=0	-0.0119	0.2142	0.2103	0.958	-0.0078	0.2272	0.2292	0.944
β_1=0	0.0018	0.1078	0.1107	0.944	0.0079	0.1161	0.1157	0.950
β_2=-1	-0.0690	0.2297	0.2272	0.944	-0.0359	0.2477	0.2449	0.948
β_1=0	0.0024	0.1116	0.1133	0.948	0.0057	0.1172	0.1115	0.958
β_2=1	0.0578	0.2380	0.2361	0.954	0.0079	0.2485	0.2461	0.962
β_1=1	0.0787	0.1468	0.1542	0.926	0.0130	0.1541	0.1506	0.958
β_2=0	0.0165	0.2201	0.2201	0.946	-0.0096	0.2319	0.2310	0.950
β_1=1	0.0802	0.1461	0.1559	0.914	0.0285	0.1573	0.1634	0.956
β_2=-1	-0.0942	0.2379	0.2453	0.934	-0.0409	0.2548	0.2587	0.950
β_1=1	0.0766	0.1510	0.1654	0.914	0.0286	0.1583	0.1677	0.956
β_2=1	0.0457	0.2420	0.2453	0.952	0.0050	0.2529	0.2627	0.940

2.5 Two real-life data applications

2.5.1 Mastitis data

Udder infections are known to be closely associated with reduced milk yield and poor milk quality (Seegers et al., 2003). In a recent study of infectious mastitis, a total of 100 cows were screened roughly monthly from the time of parturition until the lactation period, although the gap between two adjacent screenings was longer in the summer due to lack of personnel. The response of interest is the time to udder infection for each udder quarter and is typically interval-censored due to the periodic examinations. If no infection was found at a specific udder quarter at the end of lactation or at the last examination of the cow's follow-up, then the infection

time was right-censored. Among a total of 400 udder quarters, 26 had left-censored infection times, 291 had interval-censored infection times, and 83 had right-censored infection times. In this study, a cow forms a natural cluster with four udder quarters in each cluster, yielding clustered interval-censored data for the udder quarter infection times.

There are two covariates: the number of calvings and the position of udder quarter. The number of calvings is a cow level covariate and is shared by all udder quarters on the same cow. Following Goethals et al. (2009), we categorize this variable into three groups: cows with only one calving, cows with between two and four calvings, and cows with more than four calvings. Two dummy variables x_1 and x_2 are introduced for each udder quarter to distinguish these three groups, with $x_1 = x_2 = 0$ if the cow has only one calving, $x_1 = 1$ and $x_2 = 0$ if the cow has two to four calvings, and $x_1 = 0$ and $x_2 = 1$ if the cow only has more than four calvings. We define binary variable x_3 to denote the position of each udder quarter with 0 for front and 1 for rear position.

We apply our method to this data set using degree 2 and 14 interior knots for monotone splines, 20000 iterations were run with the first 5000 discarded as burn in, Table 2.3 shows the posterior means and the corresponding 95% credible intervals of the covariate effects. From Table 3, it seems that neither the position of the udder quarter nor the number of calvings has a significant effect on the infection time because both of their 95% credible intervals contain 0. Table 2.3 also listed the the posterior means and 95% credible intervals for Spearman's correlation coefficient ρ_s and for median concordance κ by using Theorem 1. These results suggest that there is a low to medium intra-cluster association among the udder quarter infections from the same cow.

Table 2.3: Estimation results for the mastitis data: posterior mean and 95% credible interval.

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

2.5.2 Lymphatic filariasis data

Lymphatic filariasis is a disease caused by the parasite Wuchereria bancrofti and transmitted by infectious mosquitoes. When people are bitten by infectious mosquitoes, W. bancrofti larvae enter the skin of people and later grow to adult worms in the lymphatic vessels of people (Williamson et al., 2008). Ultrasound is used to visualize the move of those adult worms and determine their live status. To compare the effectiveness of two treatments, DEC/ALB and DEC alone, in killing the adult worms, a study involving 47 men with lymphatic filariasis was conducted in Brazil (Dreye et al., 2006). These participants were randomized to receive either DEC/ALB combination or DEC alone for their treatment. A total of 78 adult worms were detected among the 47 participants before treatment. Ultrasound examinations were then taken at 7, 14, 30, 45, 60, 90, 180, 270, and 365 days on each participant after treatment. The failure time of interest is the time to clearance of a worm nest, which is not exactly observed but is known to be larger than the last examination time of observing the live worm and be smaller than the first examination time of observing the dead worm. Among the 78 worms, 9 of their clearance times are left censored, 41 are interval-censored, and 28 are right-censored. Also, the clearance times of those worms within the same participant are naturally correlated, leading to clustered interval-censored data.

Effects of two cluster level covariates on the clearance time are of interest: the

difference between treatments and age of participant. We use $x_1 = 1$ for DEC treatment group and 0 for DEC/ALB group and use x_2 for standardized age in the model. We apply our method to this data set using degree 2 and 14 interior knots for monotone splines, 20000 iterations were run with the first 5000 discarded as burn in, and the results are shown in Table 2.4. From Table 2.4, neither age or treatment group shows a significant effect on the clearance time based on the conventional 95% credible intervals. This conclusion agrees with the findings in (Williamson et al., 2008). However, there seems to be substantial evidence that DEC treatment is more effective than DEC/ALB treatment in clearing the worms because the posterior probability $P(\beta_1 > 0|Data) = 0.9554$. This can be also observed from the estimated posterior density function of β_1 in Figure 2.1. Table 2.4 also lists the posterior means and the corresponding 95% credible intervals for Spearman's correlation coefficient ρ_s and median concordance κ. These estimation results suggest that the clearance times of the worms from the same person are strongly correlated.

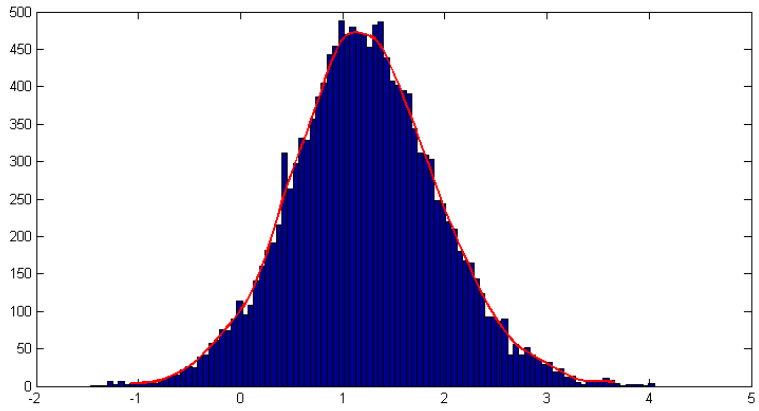
Table 2.4: Estimation results for the lymphatic filariasis data: posterior mean and 95% credible interval.

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

2.6 Discussion

In this chapter, we propose a normal frailty Probit model for analyzing clustered interval-censored data. The proposed model enjoys several appealing properties. First, this model is semiparametric since the nondecreasing function $\alpha(\cdot)$ is unspecified. Second, the conditional CDF and the marginal CDF of the failure time belong

Figure 2.1: Estimated posterior density of the treatment effect in Lymphatic filariasis data.



to the same family. Third, the conditional covariate effects given frailty are proportional to the marginal covariate effects. Forth, the intra-cluster association can be summarized by two nonparametric association measures in simple and explicit form.

It is worth noting that the frailty Probit model can be used to model multivariate survival data and that all those nice properties remain. In particular, Spearman's correlation coefficient ρ_s and median concordance κ have the same form as in Theorem 1 for describing the correlation between two correlated failure times. In this case, Kendall's τ is well defined and takes the same form as κ.

We develop a fully Bayesian method for analyzing clustered interval-censored data. Our Gibbs sampler is straightforward to implement and enjoys fast convergence mainly due to the fact that the full conditional distributions of all unknowns are of standard form. Our method shows a good performance in estimating the regression parameters and the intra-cluster association and is robust to the frailty distribution misspecification as observed in our simulation studies.

Chapter 3

A Bayesian approach to test cluster effects for clustered interval-censored failure time data under Probit model

In chapter 2 of this dissertation, it is shown that ignoring cluster effect may lead to biased estimation when applying a marginal probit model to clustered interval-censored data, where the cluster effect exists due to subjects coming from the same family, animal, clinic center, or region of living. However, in many cases the cluster effect may be very weak or even not exist. Applying the proposed method in chapter 2 may be problematic. To solve this issue, we propose two new frailty Probit models for analyzing potentially clustered interval-censored data. Both models allow us to test whether there is a global cluster effect. Our second model further allows to test local tests in order to identify which clusters have strong local effects.

Efficient Gibbs samplers are developed under both models for posteriors computation. Bayes factor is used for evaluating the global and local tests. Simulation results suggest both model perform well for the global test. Our second model also performs well for the local tests.

Keywords: Clustered interval-censored data, Probit model, semiparametric regression, Gibbs sampler, Bayes factor, Bayesian hypothesis testing, global/local test.

3.1 Introduction

In clinical studies, it is very common to have subjects coming from different groups or locations. Although the effect of groups or locations is not of interest in the study in most cases, ignoring such effect may lead to biased estimation and misleading conclusion. This has been addressed in literature for both binary and continuous response data (Agresti 2000, Localio 2001, Gould 1998.) as well as time to event data (Hougaard 1994, Hunter 1989, Klein 1992, Lancaster 1979, Lancaster 1980, Pickles 1994). In this chapter we focus on the latter when the response variable is time to event data. Two well established approaches are commonly employed to deal with time to event data with cluster effects. One way is to fit a proportional hazards model with a fixed effect for each group (Klein 2003). Another way is to introduce a frailty term into the model to incorporate the group effect (Klein 1992, Vaupel 1979, Nielsen 1992). To test the group effects, in most studies, the proportional hazard model are employed, and a Wald or likelihood ratio test is used to test the group effects. In other cases, a score test of homogeneity is derived from the marginal partial likelihood (Commenges 1995).

A more complicated situation in survival, which is not uncommon at all, is when the time to event or time to failure is not observed exactly, but is known to fall within some interval (Kalbfleisch 2002, Sun 2006). This type of data is called clustered interval-censored data. In this situation, it is just as important as in the regular survival data case to be able to test the cluster effects. However, very limited literature addressed this issue and provided testing procedures for the cluster effect for interval-censored data. Bellamy et al. (2004) developed a score test for the variance of the frailty term under weibull model with added frailty term in order to determine over-dispersion from a frequentist prospective. Wong et al. (2005) used the same model and derived a way to estimate the intra-cluster correlation between the items within

the same cluster using a Bayesian methodology.

The literature listed above solved the problem under a fully specified parametric model. This does not provide enough flexibility, as in the real-life data the underlying distribution is usually unknown. Furthermore, there is no literature available to test cluster effect for individual clusters. In this chapter, we propose a semiparametric Probit model to solve this problem and provide both a global test and local tests simultaneously. To incorporate the potential cluster effects, a frailty term is added to the survival function. Furthermore, we develop an efficient Gibbs sampler algorithm for the posterior computation. To tackle the problem of testing the cluster effect, we develop a Bayesian hypothesis testing approach via Bayes factor.

The remainder of this chapter is organized as follows. In section 3.2 we introduce the proposed model to test the frailties globally. In section 3.3, we layout the second model for testing the frailties locally for individual clusters. We evaluate our approach through a simulation study in section 3.4. In section 3.5, our method is applied to a real life application: lymphatic filariasis data. We conclude the chapter in section 3.6.

3.2 The proposed approach

3.2.1 Data and Likelihood

Suppose there are n clusters in the study. Let T_{ij} denote the failure time of interest for the jth subject in the ith cluster. The observed interval for T_{ij} is (L_{ij}, R_{ij}), $j = 1, ...m_i$ and $i = 1, ..., n$. Therefore, the total number of subjects is $M = \sum_{i=1}^{n} m_i$. We propose the following frailty semiparametric Probit model as follows:

$$\alpha(T_{ij}) = -\boldsymbol{x}'_{ij}\boldsymbol{\beta} - \xi_i + \epsilon_{ij}$$

where $\alpha(\cdot)$ is an unknown increasing function with $\alpha(0) = -\infty$ and $\alpha(\infty) = \infty$, ξ_i is the frailty term for the ith cluster, $\boldsymbol{x}_{ij}$ is the vector of covariates for the jth subject

in the ith cluster, and $\boldsymbol{\beta}$ is the vector of regression parameters.

The purpose of this study is to be able to detect whether the cluster effects exist overall, and furthermore to assess how strong the specific cluster effect is. To accomplish this goal, we consider the following prior setup with ξ_i follows a normal distribution with mean 0 and variance σ^2 if $\sigma^2 > 0$, and $\xi_i = 0$ if $\sigma^2 = 0$. We propose the model the frailty variance σ^2 with a mixture prior probability distribution $\pi(\sigma^2) = p\delta_0 + (1-p)IG(1,1)$. The variance σ^2 with probability p equals 0, and otherwise follows a inverse gamma distribution. This will help to test the existence of cluster effect globally, we layout the Bayesian hypothesis testing procedure via Bayes factor. For the globally test, we need to test if the cluster effects exist in at least one of the clusters, therefore, the hypothesis is setup as follows:

$$Ho : \sigma^2 = 0$$

,

$$Ha : \sigma^2 > 0$$

The null hypothesis indicates no frailties exists, i.e. $\xi_i = 0$ for all i, in the model. First we need to derive the likelihood function. As mentioned in many literature, we assume that the failure time is independent of the observation process that produces the observed interval conditioning on the covariates. Under this assumption, the observed likelihood can be expressed as

$$L_{obs} = \prod_{i=1}^{n} \int \pi(\xi_i) \prod_{j=1}^{m_i} \{F(R_{ij}|\boldsymbol{x}_{ij}, \xi_i) - F(L_{ij}|\boldsymbol{x}_{ij}, \xi_i)\} d\xi_i, \tag{3.1}$$

where $\phi(\cdot)$ is the density function of a standard normal random variable. The above likelihood does not have a closed form, therefore it is hard to estimate the unknown parameter $(\boldsymbol{\beta}, \sigma^2, \alpha, p)$ directly. To overcome this difficulty, we treat all the $\xi_i's$ as

latent variables and work with the following conditional likelihood

$$L_{con} = \prod_{i=1}^{n}\prod_{j=1}^{m_i} F(R_{ij}|\boldsymbol{x}_{ij},\xi_i)^{\delta_{ij1}}\{F(R_{ij}|\boldsymbol{x}_{ij},\xi_i) - F(L_{ij}|\boldsymbol{x}_{ij},\xi_i)\}^{\delta_{ij2}}$$
$$\{1 - F(L_{ij}|\boldsymbol{x}_{ij},\xi_i)\}^{\delta_{ij3}}\pi(\xi_i) \tag{3.2}$$

where δ_{ij1}, δ_{ij2} and δ_{ij3} are the censoring indicators for the jth subject in cluster i indicating left, interval, and right censoring, respectively. The purpose of introducing these censoring indicators is to distinguish the three censoring types and to help making clear of our estimation procedure later on.

3.2.2 Prior specification and posterior computation

The conditional likelihood (3.2) is still complicated for sampling unknown parameters using Bayesian methods with any prior specification. Although one can use Metropolis-Hastings or adaptive rejection Metropolis sampling method in this case, we developed a more efficient way that allows us to sample all unknowns from standard distributions. We adopted the following data augmentation motivated by Lin and Wang (2010),

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

where $t_{ij} = R_{ij}1_{(\delta_{ij1}=1)} + L_{ij}1_{(\delta_{ij1}=0)}$, i.e., t_{ij} takes the right end point of the observed interval in the case of left censoring and takes the left end point otherwise for all i and j. The augmented data likelihood function is,

$$L_{aug} = \prod_{i=1}^{n}\left[\prod_{j=1}^{m_i}\phi\{z_{ij} - \alpha(t_{ij}) - \boldsymbol{x}'_{ij}\boldsymbol{\beta} - \xi_i\}1_{C_{ij}}(z_{ij})\right]\pi(\xi_i), \tag{3.3}$$

where C_{ij} is the constrained space of z_{ij} and takes $(0,\infty)$ if $\delta_{ij1} = 1$, $(\alpha(L_{ij}) - \alpha(R_{ij}), 0)$ if $\delta_{ij2} = 1$, and $(-\infty, 0)$ if $\delta_{ij3} = 1$. Integrating out all z_{ij}'s in the above augmented likelihood leads to the conditional likelihood (3.2). The augmented likelihood (3.3) is very appealing because it leads to a normal distribution form for each of the unknown parameters and latent variables.

We take the following prior specifications: a multivariate normal prior $N(\boldsymbol{\beta}_0, \Sigma_0)$ for $\boldsymbol{\beta}$, a normal prior $N(m_0, \nu_0^{-1})$ for the unconstrained γ_0, and independent exponential priors $Exp(\eta)$ for all $\{\gamma_l\}_{l=1}^k$. We further assign a gamma prior $\mathcal{G}a(a_\eta, b_\eta)$ for η. The independent Exponential priors for γ_l's and Gamma prior for the hyperparameter η have been proved successful to shrink the spline coefficients towards zero and prevent overfitting problems in Lin and Wang (2010), Cai et al. (2011), Wang and Dunson (2011) among others. To avoid MCMC sample of σ^2 being stuck at 0, we introduce latent variable to reparameterize ξ_i and σ^2 following the idea in Dunson and Chen (2011) and Pan et. at. (2015). Let $\rho = 1(\sigma = 0)$ be an indicator that equals 1 if $\sigma = 0$ and 0 otherwise. For computational purpose, σ^2 and ξ_i can be expressed as follows:

$$\sigma = (1-\rho)\tilde{\sigma} \quad and \quad \xi_i = (1-\rho)\tilde{\xi}_i$$

where $\tilde{\xi}_i$ follows normal distribution with mean 0 and variance $\tilde{\sigma^2}$ and $\tilde{\sigma^2}$ follows IG(1,1).

These priors are natural and allow easy sampling of all the parameters from their full conditional distributions in standard form. Combining these priors and the augmented likelihood (3.3), we develop the following Gibbs sampler.

1. Sample z_{ij} from a truncated normal, $N(\alpha(t_{ij}) + \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \xi_i, 1)1_{C_{ij}}(z_{ij})$, for each j and i.

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

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

3. Sample all γ_l's for $l = 1, ..., k$. For each $l \geq 1$, let $W_l = \sum_{i=1}^{n}\sum_{j=1}^{m_i} b_l^2(t_{ij})$.

   (a) If $W_l = 0$, sample γ_l from the prior $Exp(\eta)$.

(b) If $W_l > 0$, sample γ_l from $N(E_l, W_l^{-1})1_{(\gamma_l > d_l^*)}$, where

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

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

4. Sample $\boldsymbol{\beta}$ from $N(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$, where $\hat{\Sigma} = (\Sigma_0^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{x}_{ij}\boldsymbol{x}'_{ij})^{-1}$ and

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

5. Sample $\tilde{\xi}_i$ from $N(0, \tilde{\sigma}^2)$ if $\rho = 1$, otherwise , sample $\tilde{\xi}_i$ from $N(\mu_i, \sigma_i^2)$ for $i = 1, ..., n$, where $\sigma_i^2 = (m_i(1-\rho)^2 + \tilde{\sigma}^{-2})^{-1}$ and

$$\mu_i = \sigma_i^2\left[\sum_{j=1}^{m_i}\{z_{ij} - \alpha(t_{ij}) - \boldsymbol{x}'_{ij}\boldsymbol{\beta}\}\right].$$

6. sample ρ from Bernoulli $(\tilde{\pi_0})$, where

$$\tilde{\pi_0} = \frac{\pi_0}{\pi_0 + (1-\pi_0)C}$$

where $C = L(\xi_i = \tilde{\xi}_i)/L(\xi_i = 0)$ and $L(\xi_i = \tilde{\xi}_i) = exp(-1/2\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{z_{ij} - \alpha(t_{ij}) - \boldsymbol{x}'_{ij}\boldsymbol{\beta} - \tilde{\xi}_i\}^2)$ and $L(\xi_i = 0) = exp(-1/2\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{z_{ij} - \alpha(t_{ij}) - \boldsymbol{x}'_{ij}\boldsymbol{\beta}\}^2)$

7. Sample η from $\mathcal{G}a(a_\eta + k, b_\eta + \sum_{l=1}^{k}\gamma_l)$.

8. Sample $\tilde{\sigma}^{-2}$ from $\mathcal{G}a(a_\sigma + 0.5n, b_\sigma + 0.5\sum_{i=1}^{n}\tilde{\xi}_i^{\,2})$.

9. updating σ^{-2} and ξ_i for $i = 1, ..., n$. according to the following:

$$\sigma = (1-\rho)\tilde{\sigma} \quad and \quad \xi_i = (1-\rho)\tilde{\xi}_i$$

To test the existence of frailty at the global level, the posterior probability of $\rho = 1$ is calculated by averaging the value of ρ from the MCMC iterations. This is equivalent to the posterior probability of the null hypothesis. Using a pre-determined confidence level, one can make conclusion from the result of the test.

3.3 Alternative method for global test and local test

3.3.1 proposed model

In this section, we introduce the second model we proposed. The first model in section 3.2 can test the existence of the frailty term globally. One may also be interested in testing the existence of frailty for the individual clusters. The second model is proposed with the hope of solving this exact problem. To accomplish this goal, we need to assign a prior probability distribution to the frailty term $\xi_i's$ from $\alpha(T_{ij}) = -\boldsymbol{x}'_{ij}\boldsymbol{\beta} - \xi_i + \epsilon_{ij}$, so that for each individual cluster, the frailty term can be zero or non zero. To allow such flexibility, we propose the frailty term with the following mix probability distribution, $\pi(\xi_i) = p\delta_0(\xi_i) + (1-p)N(\xi_i; \mu, \sigma^2)$. This set up is very flexible since cluster effects can exist in some or all of the clusters or they may not exist in any of the clusters at all. For the local test for cluster i, we need to test $\mathrm{H}_0 : \xi = 0$ vs H$a : \xi_i \neq 0$ conditioning on p from the semiparametric Probit model:

3.3.2 Prior specification and posterior computation

We take the following prior specifications: a multivariate normal prior $N(\boldsymbol{\beta}_0, \Sigma_0)$ for $\boldsymbol{\beta}$, a gamma prior $\mathcal{G}a(a_\sigma, b_\sigma)$ for frailty precision σ^{-2}, a normal prior $N(m_0, \nu_0^{-1})$ for the unconstrained γ_0, and independent exponential priors $Exp(\eta)$ for all $\{\gamma_l\}_{l=1}^k$. We further assign a gamma prior $\mathcal{G}a(a_\eta, b_\eta)$ for η. The independent Exponential priors for γ_l's and Gamma prior for the hyperparameter η have been proved successful to shrink the spline coefficients towards zero and prevent overfitting problems in Lin and Wang (2010), Cai et al. (2011), Wang and Dunson (2011) among others. A beta prior $Beta(a, b)$ for p. These priors are natural and allow easy sampling of all the parameters from their full conditional distributions in standard form. Combining these priors and the augmented likelihood (3.3), we develop the following Gibbs sampler.

1. Sample z_{ij} from a truncated normal, $N(\alpha(t_{ij}) + \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \xi_i, 1)1_{C_{ij}}(z_{ij})$, for each j and i.

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

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

3. Sample all γ_l's for $l = 1, ..., k$. For each $l \geq 1$, let $W_l = \sum_{i=1}^{n}\sum_{j=1}^{m_i} b_l^2(t_{ij})$.

   (a) If $W_l = 0$, sample γ_l from the prior $Exp(\eta)$.

   (b) If $W_l > 0$, sample γ_l from $N(E_l, W_l^{-1})1_{(\gamma_l > d_l^*)}$, where

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

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

4. Sample $\boldsymbol{\beta}$ from $N(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$, where $\hat{\Sigma} = (\Sigma_0^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{x}_{ij}\boldsymbol{x}'_{ij})^{-1}$ and

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

5. Sample ξ_i from $p^*\delta_0 + (1-p^*)N(\mu_i, \sigma_i^2)$ for $i = 1, ..., n$, where $\sigma_i^2 = (m_i + \sigma^{-2})^{-1}$ and

$$\mu_i = \sigma_i^2\left[\sum_{j=1}^{m_i}\{z_{ij} - \alpha(t_{ij}) - \boldsymbol{x}'_{ij}\boldsymbol{\beta}\}\right].$$

   and

$$p^* = \frac{pL(\xi_i = 0)}{pL(\xi_i = 0) + (1-p)(m_i + \sigma^{-2})^{-1}\sigma^{-1}}$$

6. Sample p from $beta\{a + \sum_i 1(\xi_i = 0), b + \sum_i 1(\xi_i \neq 0)\}$

7. Sample η from $\mathcal{G}a(a_\eta + k, b_\eta + \sum_{l=1}^{k}\gamma_l)$.

8. Sample σ^{-2} from $\mathcal{G}a(a_\sigma + 0.5n, b_\sigma + 0.5\sum_{i=1}^{n}\xi_i^2)$.

This Gibbs sampler is very appealing in that all the full conditional distributions are standard distributions and easy to sample from. This is not common in Bayesian methods for interval censored survival data. The Gibbs sampler enjoys good mixing and fast convergence from our observation.

3.3.3 Bayes factor calculation

To calculate bayes factor (BF) for the global test, we need to find the posterior odds and the prior odds respectively. Prior odds is defined the ratio of the probability of the alternative hypothesis and the probability of null hypothesis i.e.,

$$Prior\ Odds = \frac{P(H_a)}{P(H_0)},$$

where $P(H_0) = P(\cap_{i=1}^{n}\{\xi_i = 0\}) = E(p^n)$, $P(Ha) = 1 - P(\cap_{i=1}^{n}\{\xi_i = 0\}) = 1 - E(p^n)$. The value of prior odds is completely determined once the prior distribution is given. In this study, we consider two different prior probability distributions for parameter p, an equal prior probability setup and an unequal prior probability setup. In the equal probability setup, we let the prior probability of the null hypothesis and the alternative hypothesis of the global test to be the same, i.e. $P(H_0|prior) = 0.5$. The resulting prior odds is 1. To find a distribution $beta(a, b)$ that will satisfy the above condition, we need to solve the equation $E(p^n) = 0.5$. By fixing $a = 1$, the equation can be simplified as $2\Gamma(1 + b)\Gamma(n + 1) - \Gamma(b + n + 1) = 0$, and b can be solved numerically once n is determined. In the unequal probability setup, we let the prior probability for p to be $beta(1, 1)$. Therefore, the resulting prior odds equals n after a simple calculation. The way of finding posterior odds is the same in both setups, the key is to find the conditional expectation of p^n given data which can be estimated using posterior mean from MCMC. For the local test, to find the Bayes Factor for cluster i, we take the MCMC chain for ξ_i from step 5, $\frac{1}{J}\sum_1^J 1(\xi_i \neq 0)/\frac{1}{J}\sum_i^J 1(\xi_i = 0)$ provide an estimate of posterior odds. Therefore, the BF for cluster i can be found easily.

3.4 Simulation studies

The two proposed methods provide ways to test the existence of cluster effects, i.e. whether the failure times within a cluster are independent or correlated.

To evaluate our method, a simulation study is conducted. Three simulation scenarios are considered: (1) all of the cluster effects are nonzero (2) all of the cluster effects are zero (3) cluster effects are generated in the following way: $\xi_i \sim p\delta_0 + (1-p)N(0,\sigma^2)$ where $0 < p < 1$. It is clear that the subject within clusters are not independent in the first case, independent in the second case and partially independent in the third case. The purpose of having three different simulation scenarios is to show that our method can be applied to very general cases in real life. Simulation performance will be evaluated in scenario 1 and scenario 2 for both models. The second model will also be evaluated in scenario 3. We also consider the two different prior probability setups discussed in section 3.3.3 in an effort to demonstrate the effectiveness of our method, we considered the following model for failure time T_{ij} involving both discrete and continuous covariates,

$$F(t|x_{ij1}, x_{ij2}, \xi_i) = \Phi\{\alpha(t) + x_{ij1}\beta_1 + x_{ij2}\beta_2 + \xi_i\},$$

where x_{ij1} was generated from a standard normal distribution, x_{ij2} was generated from a Bernoulli distribution with probability of success equaling 0.5, $\xi_i's$ are generated according to the three scenarios mentioned above. In scenario 1, all of the cluster effects are not zero, ξ_i is generated from a normal distribution with mean 0 and standard deviation 1. In scenario 2, all of the cluster effects are zero. ξ_i is set to zero. For scenario 3, 20% of the cluster effects $\xi_i's$ are generated from standard normal distribution and the rest were set to zero. Under each of the three scenarios, we take true $\alpha(t) = 1 + t + 2\log(t)$, true $\beta_1 = 1$ or 0, and true $\beta_2 = 1$, 0, or -1, yielding 6 simulation setups in each of the three scenarios.

We generated failure time T_{ij} by solving equation $F(T_{ij}|x_{ij1}, x_{ij2}, \xi_i) = u_{ij}$ numerically, where u_{ij} was a random number from uniform distribution $U(0, 1)$ for each i and j. We generated the observed interval $(L_{ij}, R_{ij}]$ for the failure time T_{ij} as follows. First we generated an observation process for each subject. We took a random number of observational times for each subject so that subjects could have different numbers of observational times. The random number was taken to be 1 plus a Poisson random variable with mean 3. The observation times were obtained by generating the gap times between adjacent observation times independently from an exponential distribution with mean 0.3. Then the observed interval for T_{ij} was determined by the two adjacent observation times (may include 0 or ∞) that contains T_{ij}. The way we generated the observation times for each subject was very general that subjects were not required to have the same number of observation times and the same observation intervals. We generated 100 data set for each setup and each data set contained 50 clusters with 4 subjects in each cluster.

To specify monotone splines, we used 3 for the degree to ensure adequate smoothness of the spline basis functions and took 14 equally spaced interior knots between the minimum and maximum values of the end points of the observed intervals excluding 0 and a ∞ for each data set. This leads to 16 basis spline functions in use throughout the simulation.

We adopted the following specifications of the priors for the unknown parameters. We take $m_0 = 1$ and $\nu_0 = 0.1$ leading to a normal prior for γ_0 with a large variance, $a_\eta = b_\eta = 1$ leading to a $\mathcal{G}a(1, 1)$ prior for η, $a_\sigma = b_\sigma = 1$ leading to a $\mathcal{G}a(1, 1)$ prior for σ^{-2}, and $\boldsymbol{\beta}_0 = 0$ and $\Sigma_0 = M(\sum_{i,j} \boldsymbol{x}_{ij}\boldsymbol{x}'_{ij})^{-1}$ in the bivariate normal prior for $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. The prior distribution of $\boldsymbol{\beta}$ is a g-prior with a unit information variance (Zeller 1986). As discussed in the previous section, we use prior distribution $beta(1, 0.1582)$ for p in the equal prior probability setup, and we use $beta(1, 1)$ for p in the unequal prior probability setup,. Fast convergence of the MCMC was observed

in our simulation and this is probably due to the fact that all the full conditional distributions are of standard form in the proposed Gibbs sampler. We summarized results based on 5000 iterations of MCMC after discarding a first 1000-iteration as a burn-in. Convergence of MCMC was checked by using various convergence criteria in the R package CODA (Plummer 2006).

The result of the regression parameter estimates are shown in Table 3.1, 3.2 and 3.3 which compares the result of the first and the second model. Both our models performed equally well in terms of the regression parameter estimates. The point estimates of the covariate coefficients are very close to the true 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. CP95 is the 95% coverage probability, i.e. the proportion of the 95% credible intervals from each of the 100 data sets that include the true value of the parameter. The results shows that all the parameter in all setups, CP95 are very close to 0.95. The results in table 3.1 and 3.2 provide strong evidence that our method performs very well in estimating the regression parameters.

Table 3.1: Scenario 1: all of the cluster effects are not zero.

	First model				Second model			
True	$\overline{POINT}$	SSD	ESD	CP95	$\overline{POINT}$	SSD	ESD	CP95
β_1=0	0.0067	0.0867	0.0866	0.96	0.0086	0.0843	0.0851	0.95
β_2=1	0.9941	0.1769	0.1842	0.96	0.0.9094	0.1962	0.1867	0.89
β_1=0	0.0013	0.0863	0.0851	0.94	0.0109	0.0895	0.0837	0.94
β_2=-1	-0.9874	0.1833	0.1786	0.93	-0.9642	0.2238	0.1839	0.94
β_1=0	-0.0014	0.0861	0.0852	0.94	-0.0086	0.0846	0.0834	0.95
β_2=0	-0.0149	0.1661	0.1680	0.95	-0.0001	0.1761	0.16450	0.96
β_1=1	0.9970	0.1013	0.1090	0.96	0.9901	0.1437	0.1127	0.95
β_2=0	-0.0164	0.1748	0.1718	0.94	-0.0164	0.1740	0.1713	0.96
β_1=1	1.0007	0.1076	0.1113	0.96	0.9798	0.1416	0.1130	0.96
β_2=1	0.9741	0.1721	0.1854	0.96	0.9537	0.2011	0.1880	0.94
β_1=1	1.0086	0.1102	0.1093	0.97	0.9719	0.1472	0.1112	0.97
β_2=-1	-0.1048	0.1825	0.1839	0.96	-1.0001	0.1803	0.1827	0.95

To evaluate the performance of the models in terms of the hypothesis testing, we

summarize the result in table 3.4 and 3.5 respectively. For model 1, the posterior probability of ρ are summarized in table 3.4. In scenario 1, for all 6 setups in at least 91 out of 100 datasets, we reject the null hypothesis and conclude the alternative hypothesis is true which is the correct test results. A small probability $P(\rho = 1)$ indicates the probability of null hypothesis being true is low. In scenario 2, non of the 100 datasets provide significant evidence to reject the null hypothesis. For model 2, Bayes factor is computed for each data sets and are used to evaluate the effectiveness of our methods in terms of successfully testing the existence of cluster effects. Bayes factor that is less than 1 shows negative support of the alternative hypothesis. Bayes factor that is greater than 100 shows decisive evidence against null hypothesis (Jeffrey 1961). Results are shown in Table 3.4. In scenario 1, data were generated with cluster effects for all the clusters. The bayes factor estimates successfully confirms that with at least 92 and as high as 99 data sets having BF greater than 100 in all simulation setups. In scenario 2, at least 93 out of 100 data sets have BF less than 1 which provide very strong evidence that the cluster effects does not exist. The purpose of having scenario 3 is to find out how our method perform when the cluster effect exist in some clusters but not all. The results show that when 80% true cluster effect does not exist, the BF that are greater than 100 is very rare. It is plausible to say our method provide a strong evidence when the cluster effects is very strong or very weak.

Local tests are performed in scenario 3. For the convenience of tracking the performance, we generate true cluster effect from $N(0, 2)$ for cluster $i = 1, ..., 10$ and assign it to be 0 for the rest of the cluster effects. Our results show that for cluster 11 to 50, those that does not have within cluster association, our test gives conclusion correctly at least 98 out of 100 times. But for cluster 1 to 10, our tests are correct only between 50% to 60% of the time. We suspect this is due to the fact that when the cluster effect is too close to 0, the test is not able to differentiate it from 0, thus

gives higher type II error rate.

3.5 Lymphatic filariasis data

Lymphatic filariasis is a disease caused by the parasite Wuchereria bancrofti and transmitted by infectious mosquitoes. When people are bited by infectious mosquitoes, W. bancrofti larvae enter the skin of people and later grow to adult worms in the lymphatic vessels of people (Willianmson 2003). Ultrasound is used to visualize the move of those adult worms and determine their live status. To compare the effectiveness of two treatments, DEC/ALB and DEC alone, in killing the adult worms, a study involving 47 men with lymphatic filariasis was conducted in Brazil (Dreyer 2006). These participants were randomized to receive either DEC/ALB combination or DEC alone for their treatment. A total of 78 adult worms were detected among the 47 participants before treatment. Ultrasound examinations were then taken at 7, 14, 30, 45, 60, 90, 180, 270, and 365 days on each participant after treatment. The failure time of interest is the time to clearance of a worm nest, which is not exactly observed but is known to be larger than the last examination time of observing the live worm and be smaller than the first examination time of observing the dead worm. Among the 78 worms, 9 of their clearance times are left censored, 41 are interval-censored, and 28 are right-censored. Also, the clearance times of those worms within the same participant are naturally correlated, leading to clustered interval-censored data.

Effects of two cluster level covariates on the clearance time are of interest: the difference between treatments and age of participant. We use $x_1 = 1$ for DEC treatment group and 0 for DEC/ALB group and use x_2 for standardized age in the model. We apply our method to this data set using degree 3 and 14 interior knots for monotone splines. Table 3.7 shows the result. We compared result in both prior setups. Neither age or treatment group show a significant effect on the clearance time based on the conventional 95% credible intervals. This conclusion agrees with the findings

Table 3.2: Scenario 2: all of the cluster effects are zero.

	First model				Second model			
True	$\overline{POINT}$	SSD	ESD	CP95	$\overline{POINT}$	SSD	ESD	CP95
β_1=0	0.0067	0.0789	0.0787	0.95	0.0041	0.0796	0.0768	0.94
β_2=1	0.9941	0.1720	0.1648	0.93	0.9666	0.1699	0.1605	0.95
β_1=0	0.0074	0.0769	0.0788	0.94	-0.0075	0.0679	0.0753	0.97
β_2=-1	-0.9790	0.2171	0.1614	0.92	-0.9017	0.1571	0.1531	0.99
β_1=0	0.0021	0.0717	0.0785	0.97	0.0044	0.0698	0.0756	0.96
β_2=0	-0.0070	0.1485	0.1518	0.97	0.0007	0.1470	0.1466	0.96
β_1=1	1.0211	0.1093	0.0990	0.93	1.0113	0.0938	0.0953	0.94
β_2=0	0.0030	0.1608	0.1557	0.94	-0.0130	0.1628	0.1519	0.93
β_1=1	1.0263	0.1059	0.1013	0.94	1.0045	0.1094	0.0982	0.94
β_2=1	1.0269	0.1748	0.1698	0.93	1.0115	0.1558	0.1653	0.99
β_1=1	1.0308	0.1008	0.0991	0.94	1.0115	0.1064	0.0954	0.93
β_2=-1	-1.0310	0.1723	0.1661	0.94	-1.0062	0.1458	0.1611	0.96

Table 3.3: Scenario 3: 20% of the cluster effects are generated from standard normal distribution and the rest were set to zero.

	First model				Second model			
True	$\overline{POINT}$	SSD	ESD	CP95	$\overline{POINT}$	SSD	ESD	CP95
β_1=0	0.0065	0.0862	0.0857	0.94	-0.0024	0.0853	0.0848	0.96
β_2=1	0.9755	0.1922	0.1809	0.94	0.9577	0.1882	0.1805	0.94
β_1=0	0.0002	0.0868	0.0845	0.94	-0.0094	0.0877	0.0836	0.95
β_2=-1	-1.0050	0.1827	0.1769	0.95	-1.0239	0.1850	0.1785	0.93
β_1=0	0.0012	0.0776	0.0840	0.97	-0.0079	0.654	0.0827	0.99
β_2=0	0.0064	0.1616	0.1655	0.95	0.0038	0.1535	0.1623	0.97
β_1=1	1.0011	0.1157	0.1078	0.94	0.9863	0.1192	0.1071	0.93
β_2=0	-0.0142	0.1762	0.1694	0.95	-0.0271	0.1581	0.1689	0.96
β_1=1	1.0050	0.1195	0.1096	0.93	0.9933	0.1117	0.1105	0.93
β_2=1	0.9813	0.1806	0.1828	0.92	0.9986	0.1836	0.1841	0.96
β_1=1	1.0223	0.1184	0.1089	0.94	1.0043	0.1265	0.1081	0.94
β_2=-1	-1.0250	0.2084	0.1818	0.92	-0.9888	0.1876	0.1809	0.93

Table 3.4: Posterior probability of $\rho = 1$

	True value of β_1, β_2	$(1,1)$	$(1,-1)$	$(1,0)$	$(0,1)$	$(1,-1)$	$(1,0)$
Scenario 1	# of dataset with $P(\rho = 1) < 0.05$	92	91	93	92	94	96
Scenario 2	# of dataset with $P(\rho = 1) < 0.05$	0	0	0	0	0	0

Table 3.5: Bayes factor estimates

	True value of (β_1, β_2)	$(0,1)$	$(0,-1)$	$(0,0)$	$(1,0)$	$(1,1)$	$(1,-1)$
unequal prior	Scenario 1 $BF > 100$	99	97	92	98	97	99
	Scenario 2 $BF < 1$	98	97	99	97	97	98
	Scenario 3 $BF > 100$	2	0	1	4	4	6
	Scenario 3 $10 < BF < 100$	6	8	4	13	15	6
equal prior	Scenario 1 $BF > 100$	96	98	93	99	97	99
	Scenario 2 $BF < 1$	93	96	98	94	96	90
	Scenario 3 $BF > 100$	11	8	3	6	9	9
	Scenario 3 $10 < BF < 100$	14	18	9	15	23	25

in (Willianmson 2003). The results in Table 3.6 confirms that having different prior setups has little effect on the parameter estimates. The log of Bayes Factor estimate is 85 and 82 respectively, which is strongly in favor of the alternative hypothesis. There is strong evidence to support the existence of cluster effects. Furthermore, we rerun the model without the group effect and compare the result to ones from our model. It is very clear that the model tend to underestimate the parameter, in fact, the effect of treatment group become significant compare to our method. This result suggest ignoring cluster effect lead to seriously biased estimates.

Local test has also been performed for the data. For the total of 47 clusters, the Bayes factors are valued between 2 to 6. The result suggest the local cluster effects are not very strong. This is due to the small sizes in this data set.

Table 3.6: Filariasis data: compare covariate effect estimates and CI for different setups

		unequal prior setup	equal prior setup	No group effects
β_1	Mean	1.1352	1.1450	0.7210
	95% CI	(-0.2165, 2.4445)	(-0.1451, 2.5002)	(0.1852, 1.2692)
β_2	Mean	0.2268	0.2068	0.0740
	95% CI	(-0.4946, 0.9896)	(-0.5516, 0.9899)	(-0.2009, 0.3425)
$log(BF)$		85.09	82.57	

3.6 Discussion

In this chapter, we proposed two Bayesian estimation approaches which allow to test the cluster effects for clustered interval censored failure time data. In most literature, statistical test for cluster effect can only provide the global test result, our method can test the existence of individual cluster effect as well. The test result is given by a Bayes factor which can be easily computed from the MCMC sample. Simulation result show our test approach works well for the global test. For the local test, one may encounter higher Type II error rate when the cluster effect is not strong enough. Further research is required to improve the accuracy of the test for all cases.

Chapter 4

Joint modeling of informative cluster size and clustered interval-censored failure time data

4.1 Introduction

Clustered Interval-censored data arises commonly in medical studies. In most cases, the cluster structure induces correlation among subjects within the same cluster. Some of the examples are listed: the patients treated at the same medical center, the bacterias come from the same host. When such data structure is presented, it is very important to in cooperate it in the statistical models. For clustered interval-censored data, common methods for analyzing such data do not assume correlation between the failure time and the cluster size. However, in some cases the cluster size is statistically correlated with the failure times which is called informative cluster size in the literature. One famous example in the literature is the Lymphatic Filariasis (LF) study discussed by Williamson et al. (2008) among others. The goal of the study is to compare the effect of two different treatments in clearing the worms: co-administration of diethylcarbamazine and albendazole (DEC/ALB) versus DEC alone. There are 47 patients included in the study. One or more nests of adult filarial worms were detected for the patients which result in a total of 78 worm counts. It is shown in the study, in either treatment, it took longer to clear the worms in men with multiple worms than in men with a single worm. In another example of a dental study (Wong et al., 2005), the times to arrest dentin caries could be correlated with the number of the active dentin caries from the same child.

A common approach for this scenario is to use within cluster resampling (WCR) technique that assigns equal sampling probabilities for all subunits (Williamson et al., 2008; Zhang et al., 2010). Catalano and Ryan (1992) considered cluster size as a covariate in the model for discrete and continuous response. Dunson et al. (2003) proposed a Bayesian approach to joint modeling the cluster size and failure time data. Kim (2010) proposed to joint modeling the cluster size and failure time for the interval-censored data.

To deal with the informative cluster size problem in the context of the clustered interval-censored data, we propose to joint modeling the cluster size and failure time under the Probit model framework. The cluster size is modeled as an ordinal response using a parametric Probit model, and a separate frailty semiparametric Probit model is used to model the clustered failure times. The two submodels are connected through a shared random effect. The idea is similar to Kim (2010), but both the proposed models and the methods are quite different. The remaining of this chapter is organized as follows: In Section 4.2, the notation and the proposed model are introduced. Prior specification and the posterior computation is discussed in Section 4.3. A simulation study is conducted to evaluate the proposed model in Section 4.4. In Section 4.5, the proposed approach is illustrated with Lymphatic Filariasis (LF) data analysis. Section 4.6 provides a brief discussion.

4.2 Proposed model

Consider a study that involves the clustered interval-censored data, there are a total of n independent clusters. For each cluster i, there are m_i subjects. Therefore, the total number of subject in the study is $N = \sum_{i=1}^{n} m_i$. The failure time, T_{ij}, cannot be observed exactly but is known to fall within an interval $(L_{ij}, R_{ij}]$. When $L_{ij} = 0$, T_{ij} is left censored, when $R_{ij} = \infty$, it is right censored; otherwise, it is interval-censored data. Let δ_{ij1}, δ_{ij2} and δ_{ij3} be censoring indicator variables for the j^{th} subject in

cluster i indicating left-censoring, interval-censoring, and right-censoring respectively. Let $\boldsymbol{x}_i$ be a q-vector of cluster specific covariates and $\boldsymbol{z}_{ij}$ be a p-dimensional vector of covariates for subject j in cluster i, $i = 1, ..., n$, $j = 1, ..., m_i$. Therefore the observed data for the study are $\{(L_{ij}, R_{ij}], \boldsymbol{x}_i, m_i, \boldsymbol{z}_{ij}, \boldsymbol{\delta}_{ij}\}$, where $\boldsymbol{\delta}_{ij} = (\delta_{ij1}, \delta_{ij2}, \delta_{ij3})'$ $j = 1, ..., m_i, i = 1, ..., n$

The proposed joint model includes two parts. A model for cluster size and a model for the subunit failure time, and the two models are connected with a shared frailty. It is assumed that the clustered size and the failure time s are independent given the shared frailty. To model the cluster size, we adopt an ordinal regression model regarding the cluster size. We treat $m_1, \cdots, m_n$ as K category ordinal responses with $P(1 \leq m_i \leq K) = 1$, $i = 1, ..., n$. Given a random effect u_i and cluster-specific covariate $\boldsymbol{x}_i$, the model for the cluster size m_i is given by:

$$P(m_i \leq k|\boldsymbol{x}_i, u_i) = \Phi(s_k + \boldsymbol{x}_i'\boldsymbol{\gamma} + u_i), \ k = 1, ..., K \tag{4.1}$$

where $\boldsymbol{\gamma}$ is a vector of regression coefficients and $S = (s_1, ..., s_K)$ are ordered, category-specific cutoff points or thresholds with $s_K = \infty$. The random effects u_i's represent the heterogeneity among clusters and are assumed to follow a normal distribution with mean 0 and variance σ_u^2.

The failure times model is specified as,

$$F(t|u_i, v_i) = \Phi\{\alpha(t) + \boldsymbol{z}_{ij}'\boldsymbol{\beta} + \eta u_i + v_i\} \tag{4.2}$$

. where $\boldsymbol{\beta}$ is a vector of regression coefficients. Random effect v_i is introduced to induce the correlation among failure time of the subjects within the same cluster. We assume v_i follows a normal distribution with mean 0 and variance σ_v^2. $\alpha(t)$ is a unknown monotone increasing function with $\alpha(0) = -\infty$ and $\alpha(\infty) = \infty$. (4.1) and (4.2) accommodate dependency between the cluster size m_i and the failure time T_{ij} in cluster i through the shared latent variable u_i. The relationship of between failure

time and the cluster size is depending on the parameter η. We assume η follows a two point probability distribution with probability p equals 1 and probability $1-p$ equals -1. The sign of *eta* indicates the relationship between the cluster size and the failure times. For example, $\eta = 1$ indicates a positive relationship between a small cluster size and earlier failure times, and $\eta = -1$ indicates that a larger cluster size results in earlier failure times. The joint observed likelihood is given by:

$$L_{obs} = \prod_{i=1}^{n} \int\int \pi(u_i)\pi(v_i) \prod_{j=1}^{m_i} P(L_{ij} \leq t_{ij} \leq R_{ij}|\boldsymbol{z}_{ij}, u_i, v_i) P(m_i|\boldsymbol{x}_i, u_i) du_i dv_i \quad (4.3)$$

where $P(L_{ij} \leq t_{ij} \leq R_{ij}|\boldsymbol{z}_{ij}, u_i, v_i) = \Phi\{\alpha(R_{ij}) + \boldsymbol{z}'_{ij}\boldsymbol{\beta} + \eta u_i + v_i\} - \Phi\{\alpha(L_{ij}) + \boldsymbol{z}'_{ij}\boldsymbol{\beta} + \eta u_i + v_i\}$, and $P(m_i|\boldsymbol{x}_i, u_i) = \Phi(m_i + \boldsymbol{x}'_i\boldsymbol{\gamma} + u_i) - \Phi(m_i - 1 + \boldsymbol{x}'_i\boldsymbol{\gamma} + u_i)$.

The above integral does not have a closed form, therefore it is difficult to estimate the parameters directly. To overcome this, we treat all the random effects u_i's and v_i's as latent variables and work with the conditional likelihood below instead:

$$L_{con} = \prod_{i=1}^{n} \prod_{j=1}^{m_i} P(L_{ij} \leq t_{ij} \leq R_{ij}|\boldsymbol{z}_{ij}, u_i, v_i) P(m_i|\boldsymbol{x}_i, u_i) \pi(u_i)\pi(v_i) \quad (4.4)$$

Finally, to deal with the unknown monotone increasing function $\alpha(t)$, we employ the monotone splines for the modeling as in the previous chapters.

4.3 Prior specification and posterior computation

A Bayesian Gibbs sampler is employed for the estimation procedure. The above conditional likelihood would not lead to closed form posterior distributions no matter what prior distributions are used. Although Metropolis-Hastings or adaptive rejection sampling algorithm can be used, great computation complexity and inefficiency result. To facilitate efficient computation, we introduce two data augmentations to allow all the unknowns to be sampled from standard distributions. Motivated by Lin and Wang (2010), we adopt the following:

$$w_{ij} \sim N(\alpha(t_{ij}) + \boldsymbol{z}'_{ij}\boldsymbol{\beta} + \eta u_i + v_i, 1),$$

where $t_{ij} = R_{ij}1_{(\delta_{ij1}=1)} + L_{ij}1_{(\delta_{ij1}=0)}$, the right end point of the observed interval in the case of right censoring and the left end point otherwise, for all i and j. The latent variable w_{ij} is subjected to a constraint C_{ij} which defines the space of w_{ij}. w_{ij} take $(0, \infty)$ if $\delta_{ij1} = 1$, $(\alpha(L_{ij}) - \alpha(R_{ij}), 0)$ if $\delta_{ij2} = 1$, and $(-\infty, 0)$ if $\delta_{ij3} = 1$. Introduce a normal latent variable for the cluster size model as follows,

$$q_i \sim N(-\boldsymbol{x}_i'\boldsymbol{\gamma} - u_i, 1)$$

, $i = 1, ..., n$, where q_i is subject to a constraint: $\alpha_{m_i-1} < q_i < \alpha_{m_i}$. The resulting augmented likelihood can be written as

$$\prod_{i=1}^{n}\prod_{j=1}^{m_i} \phi(w_{ij} - \alpha(t_{ij}) - \boldsymbol{z}_{ij}'\boldsymbol{\beta} - \eta u_i - v_i)1_{C_{ij}(w_{ij})}\phi(q_i + \boldsymbol{x}_i'\boldsymbol{\gamma} + u_i)1_{(}\alpha_{m_i-1} \leq q_i \leq \alpha_{m_i})\pi(u_i)\pi(v_i)$$

Integrating out all the w_{ij} and q_i will lead to conditional likelihood (4.4). To utilize Gibbs sampler algorithm, we assign conjugate prior to each of the unknown parameters from the augmented likelihood. We specify the following prior distribution: assign a normal prior $N(\tilde{\gamma_0}; m_0, \nu_0^{-1})$ for the unconstrained $\tilde{\gamma_0}$. We assign independent exponential prior $Exp(\tilde{\eta})$ for all $\{\tilde{\gamma_l}\}_{l=1}^{h}$. A gamma prior $\mathcal{G}a(a_{\tilde{\eta}}, b_{\tilde{\eta}})$ for the hyperparameter $\tilde{\eta}$. For the parameter coefficient $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, assign multivariate normal prior $N(\boldsymbol{\beta}_0, \Sigma_{\boldsymbol{\beta}_0})$ and $N(\boldsymbol{\gamma}_0, \Sigma_{\boldsymbol{\gamma}_0})$ respectively. Assign gamma prior $\mathcal{G}a(a_v, b_v)$ for σ_v^{-2} and a gamma prior $\mathcal{G}a(a_u, b_u)$ for σ_u^{-2}. Finally, assign normal prior $N(m_\eta, \nu_\eta^{-1})$ for η. Based on the above prior specifications and the augmented likelihood. The Gibbs sampler algorithm is given as the following:

1. Sample latent variables w_{ij} for $i = 1, ..., n$, $j = 1, ..., m_i$.

(a) if $\delta_{ij1} = 1$, sample w_{ij} from $N(\alpha(t_{ij}) + \boldsymbol{z}_{ij}'\boldsymbol{\beta} + \eta u_i + v_i, 1)1_{(w_{ij}>0)}$.

(b) if $\delta_{ij2} = 1$, sample w_{ij} from $N(\alpha(t_{ij}) + \boldsymbol{z}_{ij}'\boldsymbol{\beta} + \eta u_i + v_i, 1)1_{(\alpha(L_{ij})-\alpha(R_{ij})<w_{ij}<0)}$.

(c) if $\delta_{ij3} = 1$, sample w_{ij} from $N(\alpha(t_{ij}) + \boldsymbol{z}_{ij}'\boldsymbol{\beta} + \eta u_i + v_i, 1)1_{(w_{ij}<0)}$.

2. Sample $\tilde{\gamma_0}$ from $N(E_0, W_0^{-1})$ where $W_0 = \nu_0 + N$ and

$$E_0 = W_0^{-1}\left[\nu_0 m_0 + \sum_{i=1}^{n}\sum_{j=1}^{m_i}[w_{ij} - \sum_{l=1}^{k}\tilde{\gamma_l}b_l(t_{ij}) - \boldsymbol{z}_{ij}'\boldsymbol{\beta} - \eta u_i - v_i]\right].$$

3. Sample $\tilde{\gamma_l}$ for $l = 1, ..., h$. For each $l \geq 1$, let $W_l = \sum_{i=1}^{n} \sum_{j=1}^{m_i} b_l^2(t_{ij})$.

(a) If $W_l = 0$, sample $\tilde{\gamma_l}$ from the prior $Exp(\tilde{\eta})$.

(b) If $W_l > 0$, sample $\tilde{\gamma_l}$ from $N(E_l, W_l^{-1})1_{(\tilde{\gamma_l} > d_l^*)}$, where

$$E_l = W_l^{-1} \left[\sum_{i=1}^{n} \sum_{j=1}^{m_i} b_l(t_{ij})[w_{ij} - \tilde{\gamma_0} - \sum_{l' \neq l} \tilde{\gamma_{l'}}' b_{l'}(t_{ij}) - \boldsymbol{z}_{ij}' \boldsymbol{\beta} - \eta u_i - v_i] - \tilde{\eta} \right]$$

$$d_l^* = max(c_l^*, 0)$$

and

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

4. Sample $\tilde{\eta}$ from $ga(a_{\tilde{\eta}} + k, b_{\tilde{\eta}} + \sum_{l=1}^{h} \tilde{\gamma_l})$

5. Sample $\boldsymbol{\beta}$ from$N(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$, where $\hat{\Sigma} = (\Sigma_0^{-1} + \sum_{i=1}^{n} \sum_{j=1}^{m_i} \boldsymbol{z}_{ij} \boldsymbol{z}_{ij}')^{-1}$ and

$$\hat{\boldsymbol{\beta}} = \hat{\Sigma} \left\{ \Sigma_0^{-1} \boldsymbol{\beta} + \sum_{i=1}^{n} \sum_{j=1}^{n_i} (w_{ij} - \tilde{\alpha}(t_{ij}) - \eta u_i - v_i) \boldsymbol{z}_{ij} \right\}$$

6. Sample v_i, for $i = 1, ..., n$, from $N(\mu_{v_i}, \sigma_{v_i}^2)$ where $\sigma_{v_i}^2 = (m_i + \sigma_v^{-2})^{-1}$ and

$$\mu_{v_i} = \sigma_{v_i}^2 \left\{ \sum_{j=1}^{m_i} (w_{ij} - \tilde{\gamma_0} - \sum_{l=1}^{h} \tilde{\gamma_l} b_l(t_{ij}) - \boldsymbol{z}_{ij}' \boldsymbol{\beta} - \eta u_i) \right\}.$$

7. Sample σ_v^{-2} from $\mathcal{G}a(a_v + n/2, b_v + 1/2 \sum_{i=1}^{n} v_i^2)$.

8. Sample q_i from $N(-\boldsymbol{x}_i' \boldsymbol{\gamma} - u_i, 1)1(\alpha_{m_i - 1} \leq q_i \leq \alpha_{m_i})$ for $i = 1, ..., n$.

9. Sample u_i from $N(\mu_{u_i}, \sigma_{u_i}^2)$ where $\sigma_{u_i}^2 = (m_i \eta^2 + 1 + \sigma_u^{-2})^{-1}$ and

$\mu_{u_i} = \sigma_{u_i}^2 \left[\eta \sum_{j=1}^{m_i} (w_{ij} - \tilde{\alpha}(t_{ij}) - \boldsymbol{z}_{ij}' \boldsymbol{\beta} - v_i) - q_i - \boldsymbol{x}_i' \boldsymbol{\gamma} \right]$ for $i = 1, ..., n$.

10. Sample σ_u^{-2} from $ga(a_u + n/2, b_u + 1/2 \sum_{i=1}^{n} u_i^2)$.

11. Sample $\boldsymbol{\gamma}$ from $N(\hat{\boldsymbol{\gamma}}, \hat{\Sigma}_{\boldsymbol{\gamma}})$ where $\hat{\Sigma}_{\boldsymbol{\gamma}} = (\Sigma_{\boldsymbol{\gamma}_0}^{-1} + \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i')^{-1}$ and

$\hat{\boldsymbol{\gamma}} = \hat{\Sigma}_{\boldsymbol{\gamma}} (\Sigma_{\boldsymbol{\gamma}_0}^{-1} \gamma_0 + \sum_{i=1}^{n} (q_i - u_i) \boldsymbol{x}_i)$

12. Sample η from $N(E_\eta, W_\eta^{-1})$ where $W_\eta = (\sum_{i=1}^{n} m_i u_i^2 + \nu_\eta)$ and

$E_\eta = W_\eta^{-1} \left[m_\eta \nu_\eta + \sum_{i=1}^{n} \sum_{j=1}^{m_i} (w_{ij} - \tilde{\alpha}(t_{ij}) - \boldsymbol{z}_{ij}' \boldsymbol{\beta} - v_i) u_i \right]$

13. Sample $\alpha_j, j = 1, ..., K-1$, from $uniform(a, b)$ where $a = max\{max\{q_{ij} : m_i = j\}, \alpha_{j-1}\}$ and $b = min\{min\{q_{ij} : m_i = j+1\}, \alpha_{j+1}\})$

4.4 Simulation

In this section, we evaluate the proposed method via a simulation study. The goal is to assess the performance of our approach primarily on parameter estimation and to compare our method with a naive approach where the cluster size effect is ignored. We generated 100 datasets. In each dataset, there are 100 clusters. We generated the cluster specific covariate $\boldsymbol{x} = (x_1, x_2)$ in the following way: x_1 was generate from $N(0, 1)$, and x_2 was generated from Bernoulli distribution with probability 0.5.This specification was chosen so that we can evaluate the proposed method with both discrete and continuous covariates. Similarly, we generated the sub-unit specific covariate $\boldsymbol{z} = (z_1, z_2)$ from a standard normal distribution and a Bernoulli distribution. The frailties $u_i's$ and $v_i's$ are generated from identically independent standard normal distribution $N(0, 1)$. For the covariate coefficients, true β_1 was taken to be 1, and true β_2 was taken to be 1 or -1. The true γ_1 was taken to be 1, and true γ_2 was 1 or -1, which result in 4 different combination configurations for covariates.

To generate the cluster size for each cluster, we specified the thresholds $\alpha = (\alpha_1, \alpha_2, \alpha_3) = (-2, -0.5, 1)$ for the ordinal model (4.1). Given the above setup of parameter specification, we generate the cluster size from (4.1) the following way. For cluster i, we evaluate $\Phi(s_k + \boldsymbol{x}_i'\boldsymbol{\gamma} + u_i),\ k = 1, ..., 3$. As a result, the interval $(0, 1)$ can be partitioned into 4 sub-intervals. A random number was generated from uniform distribution, and the cluster size was determined by the order of the interval that contained the random number. Cluster sizes ranged from 1 to 4 for all clusters.

Once the cluster size was generated, the total number of the subjects in the dataset was known. The next step was to generate failure time for each subject. We specified the true $\alpha(t) = 1 + t + 2\log(t)$, which is a non-linear function. We generated failure time T_{ij} by solving equation $F(T_{ij}|x_{ij1}, x_{ij2}, \xi_i) = r_{ij}$ numerically, where r_{ij} is a random number from uniform distribution $U(0, 1)$ for each i and j. We then

generated the observed interval $(L_{ij}, R_{ij}]$ for the failure time T_{ij} as follows. First we generate an observation process for each subject. We took a random number of observational times for each subject so that subjects can have different numbers of observational times. The random number was taken to be 1 plus a Poisson random variable with mean 3. The observation times are obtained by generating the gap times between adjacent observation times independently from an exponential distribution with mean 0.3. Then the observed interval for T_{ij} was determined by the two adjacent observation times (may include 0 or ∞) that contains T_{ij}. The way we generate the observation times was very general that subjects are not required to have the same number of observation times and the same observation intervals.

We examines the estimates of the covariate coefficients as well as the variance of the frailty terms and the thresholds for the ordinal model. The result are listed in table 4.1. We summarized the frequentist operating characteristics of the parameter estimates from the proposed Bayesian method. For each parameter configuration, POINT is the average of the 100 point estimates (posterior mean). ESD is the average of the estimated standard deviations of their posterior distributions across the 100 data sets, SSD is the sample standard deviation of the 100 point estimates, and CP95 is the 95% coverage probability, i.e. the proportion of the 95% credible intervals from 100 data sets that included the true value of the parameter. In all 4 cases, the parameter estimates are all very close to the true value, the SSD and ESD are very close which indicates a good convergence of the MCMC chain. The CP95 are close to 0.95. The results indicate that our model works very well in different scenarios.

Furthermore, for comparison purpose, we consider the reduced model where the dependence structure of the data between failure time and cluster size is ignored. The reduced model is essentially two models without the shared frailty term u_i. A Probit model for the failure time as the one described in Chapter 2, and a Ordinal model

Table 4.1: Performance of the proposed method in the case of using 100 clusters, POINT denotes the average of 100 point estimates, ESD is the average of the estimated standard deviation, SSD is the sample standard deviation of the 100 point estimates, and CP95 is the 95% coverage probability.

True	$\overline{POINT}$	SSD	ESD	CP95
β_1=1	0.9977	0.2129	0.2044	0.94
β_2=1	0.9742	0.3560	0.3625	0.96
γ_1=0.5	0.5155	0.1854	0.1670	0.91
γ_2=0.5	0.5420	0.3074	0.3034	0.92
α_1=-2	-2.1610	0.3323	0.3301	0.92
α_2=-0.5	-0.5342	0.2435	0.2326	0.91
α_3=1	1.0533	0.2872	0.2603	0.92
σ_u=1	1.0559	0.2001	0.1867	0.96
σ_v=1	1.0476	0.2287	0.2291	0.94
β_1=1	1.0375	0.1965	0.2076	0.97
β_2=-1	-1.0864	0.3721	0.3752	0.94
γ_1=0.5	0.5279	0.1565	0.1688	0.97
γ_2=0.5	0.5677	0.3034	0.3115	0.93
α_1=-2	-2.1636	0.3793	0.3353	0.90
α_2=-0.5	-0.5422	0.2288	0.2378	0.95
α_3=1	1.0772	0.2776	0.2674	0.93
σ_u=1	1.0691	0.1564	0.1880	1
σ_v=1	1.0714	0.2176	0.2390	0.98
β_1=1	1.0328	0.2178	0.2006	0.94
β_2=1	0.9631	0.3237	0.3510	0.94
γ_1=0.5	0.5426	0.1879	0.1732	0.92
γ_2=-0.5	-5101	0.3150	0.3154	0.94
α_1=-2	-2.2011	0.4765	0.3642	0.89
α_2=-0.5	-0.5287	0.2856	0.2387	0.91
α_3=1	1.0849	0.2896	0.2639	0.88
σ_u=1	1.0862	0.2087	0.1864	0.93
σ_v=1	1.0186	0.2122	0.2185	0.96
β_1=1	1.0688	0.1960	0.2040	0.94
β_2=-1	-1.1021	0.3408	0.3692	0.97
γ_1=0.5	0.5410	0.2245	0.1742	0.88
γ_2=-0.5	-0.5820	0.3288	0.3217	0.93
α_1=-2	-2.2934	0.4327	0.3944	0.92
α_2=-0.5	-0.5347	0.2639	0.2504	0.92
α_3=1	1.0602	0.2710	0.2623	0.91
σ_u=1	1.0993	0.1620	0.1900	0.97
σ_v=1	1.0777	0.2217	0.2236	0.94

with cluster size. The result is shown in table 4.2. The parameter estimates for the failure time model is close to the true value. This reinforce the finding in Chapter 2 regarding the robustness of the Probit model. However, the parameter estimate of the Ordinal model for cluster size are bias.

Table 4.2: Reduced model: Performance of the proposed method in the case of using 100 clusters, POINT denotes the average of 100 point estimates, ESD is the average of the estimated standard deviation, SSD is the sample standard deviation of the 100 point estimates, and CP95 is the 95% coverage probability.

True	$\overline{POINT}$	SSD	ESD	CP95
β_1=1	1.0024	0.2025	0.1997	0.92
β_2=1	0.9484	0.3398	0.3451	0.94
γ_1=0.5	1.2308	0.3640	0.5045	0.65
γ_2=0.5	1.2173	0.6365	0.7644	0.82
α_1=-2	-6.4176	3.3400	1.8519	0.22
α_2=-0.5	-1.4501	0.8725	0.7619	0.68
α_3=1	3.4657	2.4202	1.2390	0.51
σ_u=1				
σ_v=1	1.370	0.2523	0.2235	0.58
β_1=1	1.0472	0.2034	0.1923	0.93
β_2=-1	-1.0932	0.3246	0.3497	0.91
γ_1=0.5	1.4356	0.3872	0.5219	0.59
γ_2=0.5	1.2342	0.6382	0.7362	0.79
α_1=-2	-6.5212	3.7543	1.9342	0.20
α_2=-0.5	-1.3879	0.8462	0.7983	0.65
α_3=1	3.2019	2.1093	1.9876	0.52
σ_u=1				
σ_v=1	1.4024	0.2534	0.2345	0.55
β_1=1	1.0123	0.2134	0.2018	0.94
β_2=1	0.9694	0.3285	0.3489	0.92
γ_1=0.5	1.5367	0.4027	0.5328	0.60
γ_2=-0.5	-1.4256	0.6886	0.7352	0.82
α_1=-2	-6.3893	3.0665	1.9078	0.20
α_2=-0.5	-1.3984	0.8289	0.8034	0.68
α_3=1	3.4272	2.1039	1.9983	0.49
σ_u=1				
σ_v=1	1.3980	0.2234	0.2434	0.52
β_1=1	1.0634	0.2234	0.2134	0.94
β_2=-1	-1.1034	0.3343	0.3445	0.91
γ_1=0.5	1.6033	0.3984	0.5324	0.57
γ_2=-0.5	-1.4454	0.6984	0.7634	0.79
α_1=-2	-6.4326	3.2355	2.0456	0.19
α_2=-0.5	-1.4246	0.8046	0.8428	0.65
α_3=1	3.3785	2.2894	2.1034	0.50
σ_u=1				
σ_v=1	1.4022	0.2356	0.2455	0.50

4.5 Data analysis

In this section, we apply the proposed approach to the Lymphatic Filariasis (LF) data. The main purpose of the analysis is to study the effect of the co-administration of DEC and ALB against DEC alone for the treatment of Lymphatic Filariasis worms. Filariasis worms can be easily detected by ultrasound. The patients in this study were followed for a year and periodically examined by ultrasound to determine the clearance of the worms. The response variable or the outcome in the study is the time from treatment administration to clearance of the worms. The worms from the same study subject can be viewed from the same cluster, and therefore the cluster size is the number of filariasis worms in each patient.In this study,it took longer to clear a nest of worms with multiple worms than a nest with a single worm which indicates the cluster size is informative. (shown in table 4.3) Among the total of 47 men included

Table 4.3: Percentages of nests cleared during 360 days in the lymphatic filariasis study.

Number of nests	Percentage cleared
1	81.8
2	62.5
3	50
4 or 5	33.3

in the study, 25 were in the DEC group and 22 were in the DEC/ALB group. The number of worm nests detected from all individuals ranged from 1 to 5 which yeilded a total of 78 adult worm nests. Effects of two cluster level covariates on the clearance time are of interest: the treatment group x_1 (DEC/ALB=0, DEC=1) and the age x_2. We apply the proposed model to this data. The result is shown in table 4.4. Even though based on the 95% credible interval, the treatment effects is not significant. However, we calculated the posterior probability of $P(\beta_1 > 0|Data) = 0.9453$ which provide substantial evidence that DEC treatment is more effective than DEC/ALB treatment. The effect of age is not significant. The estimate of parameter $\eta = 1$

which indicate the bigger the nest is, the smaller the clearance time is. This finding is consistence with literature results (Kim, 2010).

Table 4.4: Filariasis data: compare covariate effect estimates and CI

	Mean	95% CI
β_1	1.0985	(-0.1761, 2.5191)
β_2	0.1910	(-0.4817, 0.9282)
γ_1	-0.4472	(-1.1226, 0.2201)
γ_2	0.0718	(-0.2851, 0.4294)
α_1	-0.8001	(-1.3036, -0.2931)
α_2	0.6517	(0.0437, 1.3000)
α_3	1.1211	(0.4272, 1.9186)
α_4	1.6828	(0.7894, 2.7694)

4.6 Conclusion

In this chapter, we considered a more complexed situation than the previous chapter where the cluster size could provide useful information for the failure times. We proposed a joint modeling approach to deal with this problem. We used a ordinal regression model for the cluster size and a Probit model for the failure time. The two models are linked by a shared frailty term. A Bayesian Gibbs samplers is developed for estimation. The results indicate that the join modeling approach works very well in terms of estimating the covariate effects.

Bibliography

[1] Agresti A., Hartzel J. (2000) TUTORIAL IN BIOSTATISTICS Strategies for comparing treatments on a binary response with multi-centre data. *Statistics in Medicine* **19**: 1115-1139.

[2] Bellamy, S.L., Li, Y., Ryan, L.M., Lipsitz, S., Canner, M.J., and Wright, R. (2004) Analysis of clustered and interval censored data from a community-based study in asthma. *Statistics in Medicine.* **23** 3607–3621.

[3] Cai, T. and Betensky, R.A. (2003) Hazard Regression for Interval-Censored Data with Penalized Spline. *Biometrics.* **59** 570–579.

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

[5] Chen, Z., Zhang, B., and Albert, P.S. (2011) A joint modeling approach to data with informative cluster size: Robustness to the cluster size model. *Statistics in Medicine.* **30** 1825–1836.

[6] Commenges D. Andersen PK. (1995) Score test of homogeneity for survival data. *Lifetime data analysis.* **1**: 145-156.

[7] Datta, S., Lee, K.Y. (2008) A signed-rank test for clustered data. *Biometrics.* **64** 501–507.

[8] De Wolf, A., Lange, JMA. (2001) The ATHENA cohort study: implications of the introduction of HAART for the course of HUV-1 disease, public health and health care as well as the economic costs and benefits. *Monitoring of Human Immunodeficiency Birus Type 1 (HIV-1) Infection in the Netherlands.* 18–51.

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

[10] Fan, J., Datta, S. (2011) Fitting marginal accelerated failure time models to clustered survival data with potentially informative cluster size. *Computational Statistics and Data Analysis.* **55** 3295–3303.

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

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

[13] Finkelstein, D.M. (1986). A Proportional Hazards Model for Interval-Censored Failure Time Data. *Biometrics.* **42** 845–854.

[14] Glen, A.S. (1996) Rank-based inference in the proportional hazards model for interval censored data. *Biometrika.* **83** 355–370.

[15] Goethals, K., Ampe, B., Berkvens, D., Laevens, H., Janssen P., and Duchateau, L. (2009) Modeling Interval-Censored, Clustered Cow Udder Quarter Infection Times Through the Shared Gamma Frailty Model. *Journal of Agricultural, Biological, and Environmental Statistics.* **14** 1–14.

[16] Gould AL. (1998) Multi-centre trial analysis revisited. *Statistics in Medicine.* **17**: 1779-1797.

[17] Henschel, V., Engel, J., Holzel, D., and Mansmann. U. (2009) A semiparametric Bayesian proportional hazards model for interval censored data with frailty effects. *BMC Medical Research Methodology.* **9** 9.

[18] Hoffman, E.B., Sen, P.K., Weinberg, C.R. (2001) Within-cluster resampling. *Biometrika* **88** 1121–1134.

[19] Hougaard, P., Myglegaard, P., and Borch-Johnsen, K. (1994) Heterogeneity models of disease susceptibility with application to diabetic nephropathy. *Biometrics* **50** 1178–1188.

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

[21] Huster, W.J., Brookmeyer, R., and Self, S.G. (1989) Modeling paired survival data with covariates. *Biometrics* **45** 145–156.

[22] Jeffreys H. (1935) Some Tests of Significance, Treated by the Theory of Probability. *Proceedings of the Cambrige Philosophy Society.* **31**: 202-222.

[23] Jeffreys H. *Theory of Probability (3rd ed.)* 1961. Oxford, U.K.: Oxford University Press.

[24] Kalbfleisch JD., Prentice RL. *The Statistical Analysis of Failure Time Data.* (2nd ed). Wiley: New York, 2002

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

[26] Kooperberg, C. and Stone, C.J. (1992) Logspline Density Estimation for Censored Data. *Journal of Computational and Graphical Statistics.* **1** 301–328.

[27] Klein, JP., Moeschberger ML. (2003) *Survival analysis: techniques for censored and truncated data.* Springer: New York.

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

[29] Kruskal, W.H. (1958) Ordinal measures of association. *Journal of the American Statistical Association* **53** 814–861.

[30] Lam, K.F., Xu, Y., and Cheung, T.L. (2010) A multiple imputation approach for clustered interval-censored survival data. *Statistics in Medicine.* **29** 680–693.

[31] Lancaster, T. (1979) Econometric methods for the duration of unemployment. *Econometrica* **47** 939–956.

[32] Lancaster T., Nickell S. (1980) The analysis of re-employment probabilities for the unemployed (with discussion). *J.R. Statist. Soc. A.* **143**: 141-165.

[33] Lin, X. and Wang, L. (2009) A semiparametric probit model for case 2 interval-censored failure time data. *Statistics in Medicine.* **29** 972–981.

[34] Localio AR., Berlin JA., Ten Have TR., Kimmel SE. (2001) Adjustments for Center in Multicenter Studies: An Overview. *Annual of Internal Medicine.* **135(2)**: 112-123.

[35] Mocroft, A., Vella, S. and Benfield, T.L. (1998) Changing patterns of mortality across Europe in patients infected with HIV-1. *Lancet* **352** 1725–1730.

[36] Nielsen G., Gill RD., Andersen PK., Sorensen TIA. (1992) A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics.* **19**: 25-44.

[37] Odell, P.M., Anderson, K.M., and DÁgostino, R.B. (1992) Maximum Likelihood Estimation for Interval-Censored Data Using a Weibull- Based Accelerated Failure Time Model. *Biometrics.* **48** 951–959.

[38] Pan, W. A Multiple Imputation Approach to Cox Regression with Interval Censored Data. *Biometrics.* **56** 199–203.

[39] Panageas, K.S., Schrag, D., Russell, L.A., Venkatraman, E.S., and Begg, C.B. (2007) Properties of analysis methods that account for clustering in volume-outcome studies when the primary predictor is cluster size. *Statistics in Medicine.* **26** 2017–2035.

[40] Pickles A., Crouchley R. (1994) Generalizations and applications of frailty models for survival and event data. *Statist. Meth. Med. Res.* **3**:263-278.

[41] Plummer, M., Best, N., Cowles, K., and Vines, K. (2006) CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News.* **6** 7–11.

[42] Rabinowitz, D., Tsiatis, A., and Aragon, J. (1995) Regression with interval-censored data. *Biometrika.* **82** 501–513.

[43] Ramsay, J. (1988) Monotone Regression Splines in Action. *Statistical Science.* **3** 425–461.

[44] Ross, E., and Moore, D. (1999) Modeling clustered, discrete, or grouped time survival data with covariates. *Biometrics.* **55** 813–819.

[45] Sethuraman, J. (1994) A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, **4**, 639-650.

[46] Seegers, H., Fourichon, C., and Beaudeau, F. (2003) Production effects related to Mastitis and mastitis economics in dairy cattle herds. *Veterinary Research.* **34** 475–491.

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

[48] Van Sighem, A., Van de Wiel, M., and Ghani, A. (2003) Mortality and progression to AIDS after starting highly active antiretroviral therapy. *AIDS.* **17** 2227–2236.

[49] Vaupel JW., Manton KG., Stallard E. (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography.* **16**: 439-454.

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

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

[52] Wang, M., Kong, M.K., and Datta, S. (2011) Inference for marginal linear models for clustered longitudinal data with potentially informative cluster sizes. *Stat Methods Med Res.* **20** 347–367.

[53] Williamson, J.M., Datta. S., and Satten. G.A. (2003) marginal analysis of clustered data when cluster size is informative. *Biometrics.* **59** 36–42.

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

[55] Wong, M.C.M., Lam, K.F., Lo, E.C.M. (2005) Bayesian Analysis of Clustered Interval-censored Data. *Journal of Dental Research* **84** 817–821.

[56] Wong, M., Lam, K.F., and Lo, E. (2006) Multilevel modelling of clustered grouped survival data using Cox regression model: an application to ART dental restorations. *Statistics in Medicine.* **25** 447–457.

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

[58] Zellner, A., 1986. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, (eds. P. K. Goel and A. Zellner), 233-243. North-Holland/Elsevier.

Appendix A

Proof of Theorem 1

Proof: Let T_1 and T_2 denote the failure times of two subjects, with covariate $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ respectively, in the same cluster. Under the normal frailty Probit model (2.1), it is equivalent to write

$$\alpha(T_j) = -\boldsymbol{x}_j\boldsymbol{\beta} - \xi + \epsilon_j, \quad j = 1, 2,$$

where $\xi \sim N(0, \sigma^2)$ is the shared frailty, and ϵ_1 and ϵ_2 are independent standard normal random variables. Define $Y_j = \alpha(T_j)$ for $j = 1$ and 2. It is straightforward to obtain that both Y_1 and Y_2 have a marginally normal distribution with variance $1+\sigma^2$ and their joint distribution is a bivariate normal. Pearson's correlation coefficient between Y_1 and Y_2 is thus

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

Theorem 1 follows directly by using the relationship among Pearson's correlation coefficient, Spearman's correlation coefficient, and median concordance for bivariate normal distribution as follows,

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

see Kruskal (1958) and Houggard (2000) among others.

Appendix B

Extension of the normal frailty to nonparametric frailty distribution in Chapter 2

In chapter 2, we proposed to use normal distribution to model the frailty term under the semiparametric Probit model. In real life data, the distribution of the frailty is often unknown. Therefore, a natural extension of the proposed model is to use nonparametric approach to model the frailty term. In this appendix, we will layout the nonparametric approach as well as the estimation procedure along with the simulation result.

We propose to use Dirichlet process (DP) prior for the frailty distribution. Dirichlet process is very popular in nonparametric Bayesian modeling. It was first introduced by Ferguson (1973). DP is specified by a base distribution H and a concentration parameter α which is a positive real number. The DP process draws distribution around the base distribution in a way that a normal distribution draws value from its mean. DP process is equivalent to a stick-breaking construction which is introduced by Sethuraman (1994).

Under the proposed model (2.1) in chapter 2, we assign a Dirichlet process prior for the frailty distribution, i.e.,

$$\xi_i | G \overset{i.i.d.}{\sim} G, \quad \text{and } G \sim DP(mH),$$

where m is the precision parameter and H is the base distribution of the Dirichlet

process. It is equivalent to write the DP using the stick-breaking form:

$$G(\cdot) = \sum_{h=1}^{\infty} p_h \delta_{\theta_h}(\cdot),$$

where p_hs are random weights taking $p_h = v_h \prod_{l<h}(1 - v_l)$ for all h with v_h being independent $Beta(1, m)$ random variables, θ_hs are random atoms sampled from H, and p_hs and θ_hs are independent. Given the above setup, the marginal distribution of the failure time is given by

$$F(t|\boldsymbol{\xi}) = \sum_{h=1}^{\infty} p_h \Phi\{\alpha(t) + \boldsymbol{\xi}'\beta + \theta_h\}.$$

This distribution is extremely flexible as a mixture of normals is used to model the transformed failure time.

We assign a normal prior $N(m_0, v_0^{-1})$ for μ_H, a Gamma prior $\mathcal{G}a(a_0, b_0)$ for σ_H^{-2}, and a Gamma prior $\mathcal{G}a(a_m, b_m)$ for m. Let N denote the number of clusters used in the Dirichlet process and denote K_i as labeling variable for ξ_i with $K_i = h$ if $\xi_i = \theta_h$ for all i. It is difficult to allow the mean constraint using the Dirichlet process, here we take the unconstrained Dirichlet process and force $\gamma_0 = 0$ for identifiability purpose.

Gibbs sampler based on the augmented likelihood. With all the parameters being sampled initially from their prior, the Gibbs samlper proceeds with the following steps at each iteration:

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

(a) if $\delta_{ij1} = 1$, sample z_{ij} from $N(\alpha(t_{ij}) + \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i, 1)1_{(z_{ij}>0)}$.

(b) if $\delta_{ij2} = 1$, sample z_{ij} from $N(\alpha(t_{ij}) + \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i, 1)1_{(\alpha(L_{ij})-\alpha(R_{ij})<z_{ij}<0)}$.

(c) if $\delta_{ij3} = 1$, sample z_{ij} from $N(\alpha(t_{ij}) + \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i, 1)1_{(z_{ij}<0)}$.

2. Sample $\gamma'_l s$ for $l = 1, ..., k$. For each $l \geq 1$, let $W_l = \sum_{i=1}^{n} \sum_{j=1}^{n_i} b_l^2(t_{ij})$.

(a) If $W_l = 0$, sample γ_l from the prior $Exp(\eta)$.

(b) If $W_l > 0$, sample γ_l from $N(E_l, W_l^{-1})1_{(\gamma_l > d_l^*)}$, where

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

$$d_l^* = max(c_l^*, 0)$$

and

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

3. Sample $\boldsymbol{\beta}$ from$N(\hat{\boldsymbol{\beta}}, \hat{\Sigma})$, where $\hat{\Sigma} = (\Sigma_0^{-1} + \sum_{i=1}^{n} \sum_{j=1}^{n_i} \boldsymbol{x}'_{ij} \boldsymbol{x}_{ij})^{-1}$ and

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

4. Sample η from $ga(a_\eta + k, b_\eta + \sum_{l=1}^{k} \gamma_l)$

5. Sample Sample V_h from $Beta(1 + \sum_{i=1}^{n} 1_{(K_i=h)}, m + \sum_{i=1}^{n} 1_{(K_i>h)})$

6. Sample θ_h from θ_h from $N(\tilde{\mu}_h, \tilde{\sigma}_h^2)$ for $h = 1, ..., N$, where $\tilde{\sigma}_h^2 = (\sigma_H^{-2} + \sum_{(i:K_i=k)} n_i)^{-1}$ and $\tilde{\mu}_h = \tilde{\sigma}_h^2 \{\mu_H \sigma_H^{-2} + \sum_{(i:K_i=h)} \sum_{j=1}^{n_i} (z_{ij} - \alpha(t_{ij}) - x'_{ij}\beta)\}$.

7. Sample u_i from uniform distribution $U(0, p_{K_i})$ for $i = 1, ..., n$.

8. Sample K_i for $i = 1, ..., n$. Let $U^* = \max(u_1, ..., u_n)$.

(1) If $\sum_{h=1}^{N} p_h > 1 - U^*$, Sample K_i from a Multinormial distribution $\mathcal{M}\{1, (q_{i1}, \cdots, q_{iN})\}$ for $i = 1, ..., n$, where

$$q_{ih} = \frac{\exp[-1/2 \sum_{j=1}^{n_i} \{z_{ij} - \alpha(t_{ij}) - x'_{ij}\beta - \theta_h\}^2] 1(p_h > u_i)}{\sum_{h=1}^{N} \exp[-1/2 \sum_{j=1}^{n_i} \{z_{ij} - \alpha(t_{ij}) - x'_{ij}\beta - \theta_h\}^2] 1(p_h > u_i)}$$

(2) Otherwise, keep updating $N = N + 1$ and sampling v_N, θ_N from their prior distributions until $\sum_{h=1}^{N} p_h > 1 - U^*$. Then sample K_i as in (1).

9. Update $\xi_i = \theta_{K_i}$ for $i = 1, \cdots, n$

10. Sample μ_H from

$$N\{(v_0 + n\sigma_H^{-2})^{-1}(v_0 m_0 + \sigma_H^{-2}\sum_{i=1}^{n}\xi_i), (v_0 + n\sigma_H^{-2})^{-1}\},$$

11. Sample σ_H^{-2} from $\mathcal{G}a(a_0 + n/2, b_0 + 1/2\sum_{i=1}^{n}(\xi_i - \mu_H)^2)$

12.Sample m from $\mathcal{G}a(a_m + N, b_m - \sum_{h=1}^{N}\log(1 - V_h))$

To evaluate the performance of the proposed approach, we follow the simulation setup discussed in chapter 2 and layout the result in the tables below. The result is very similar to the result of normal frailty model. All the parameter estimates are very close to the true value. ESE and SSD are very close to each other. CP95 are close to 95%. The nonparametric approach works well for any true distribution of frailties.

Table B.1: Simulation result: normal setup

True	$\overline{POINT}$	ESE	SSD	CP95
β_1=0	-0.0021	0.0972	0.1025	0.94
β_2=0	-0.0168	0.1915	0.1644	0.95
β_1=0	0.0044	0.0976	0.1044	0.92
β_2=-1	-0.9725	0.2112	0.2394	0.99
β_1=0	0.0028	0.0984	0.1001	0.96
β_2=1	0.9818	0.2115	0.2242	0.95
β_1=1	0.9800	0.1314	0.1430	0.95
β_2=0	0.0509	0.1978	0.2055	0.91
β_1=1	1.0095	0.1348	0.1198	0.96
β_2=-1	-1.0027	0.2178	0.2158	0.97
β_1=1	0.9839	0.1331	0.1381	0.91
β_2=1	0.9492	0.2158	0.2389	0.94

Table B.2: Simulation result: $\xi_i \sim 0.45N(0.5, 0.4^2) + 0.55N(-0.5, 0.18^2)$

True	$\overline{POINT}$	ESE	SSD	CP95
β_1=0	0.0037	0.1024	0.1158	0.91
β_2=0	0.0300	0.2000	0.1989	0.97
β_1=0	0.0007	0.1028	0.0988	0.97
β_2=-1	-0.9860	0.2214	0.2225	0.97
β_1=0	0.0041	0.1019	0.1088	0.93
β_2=1	0.9825	0.2184	0.2317	0.91
β_1=1	0.9846	0.1398	0.1519	0.92
β_2=0	0.0515	0.2036	0.2111	0.91
β_1=1	0.9638	0.1358	0.1358	0.92
β_2=-1	-0.9669	0.2212	0.2176	0.99
β_1=1	0.9927	0.1392	0.1516	0.95
β_2=1	0.9811	0.2220	0.2363	0.92

Table B.3: Simulation result: $\exp(\xi_i) \sim \mathcal{G}a(1, 1)$

True	$\overline{POINT}$	ESE	SSD	CP95
β_1=0	0.0006	0.1027	0.0946	0.98
β_2=0	0.0139	0.2041	0.2179	0.97
β_1=0	-0.0123	0.1053	0.1090	0.93
β_2=-1	-1.0021	0.2287	0.2739	0.98
β_1=0	-0.0003	0.1038	0.1124	0.93
β_2=1	0.9916	0.2245	0.2209	0.97
β_1=1	0.9716	0.1450	0.1407	0.96
β_2=0	-0.0262	0.2070	0.1988	0.97
β_1=1	0.9821	0.1453	0.1542	0.94
β_2=-1	-0.9910	0.2309	0.2020	0.99
β_1=1	0.9890	0.1466	0.1285	0.97
β_2=1	0.9742	0.2297	0.2504	0.92

The nonparametric approach we proposed is very flexible. However, the flexibility comes at the cost of computational intensity. It requires many more parameters being samples from the Gibbs sampler algorithm compared to the normal frailty algorithm. The performance from the simulation result shows that the normal frailty model is very robust even when the frailty distribution is misspecified. Therefore, the normal frailty model is preferred between the two due to the less complexity.

Appendix C

Simulation result for the equal prior probabilities setup in Chapter 3

Table C.1: Performance of the proposed method in the cases of using 50 clusters under **equal prior probabilities setup**. POINT denotes the average of the 100 point estimates, ESD the average of the estimated standard deviations, SSD the sample standard deviation of the 100 point estimates, and CP95 the 95% coverage probability

	Scenario 1				Scenario 2				Scenario 3			
True	$\overline{POINT}$	SSD	ESD	CP95	$\overline{POINT}$	SSD	ESD	CP95	$\overline{POINT}$	SSD	ESD	CP95
β_1=0	0.009	0.086	0.086	0.96	0.002	0.064	0.077	0.98	-0.002	0.085	0.084	0.96
β_2=1	0.938	0.154	0.182	0.95	0.937	0.162	0.161	0.92	0.957	0.188	0.180	0.94
β_1=0	0.004	0.077	0.085	0.99	-0.004	0.064	0.075	0.99	-0.009	0.087	0.083	0.95
β_2=-1	-0.994	0.208	0.178	0.92	-0.947	0.175	0.154	0.90	-1.023	0.185	0.178	0.93
β_1=0	-0.005	0.074	0.084	0.98	0.002	0.055	0.077	1	-0.007	0.654	0.082	0.99
β_2=0	-0.028	0.168	0.168	0.94	-0.007	0.149	0.150	0.96	0.003	0.1535	0.162	0.97
β_1=1	1.005	0.095	0.109	0.99	1.004	0.094	0.095	0.95	0.986	0.119	0.107	0.93
β_2=0	0.022	0.168	0.172	0.94	-0.015	0.169	0.152	0.92	-0.027	0.158	0.168	0.96
β_1=1	0.982	0.111	0.110	0.93	1.020	0.108	0.098	0.93	0.993	0.111	0.110	0.93
β_2=1	0.984	0.177	0.185	0.94	1.021	0.18	0.166	0.94	0.998	0.183	0.181	0.96
β_1=1	0.983	0.100	0.107	0.95	1.023	0.101	0.096	0.94	1.0043	0.126	0.108	0.94
β_2=-1	-0.973	0.181	0.183	0.97	-1.013	0.173	0.162	0.94	-0.988	0.187	0.180	0.93

Table C.2: Bayes factor estimates

	True value of (β_1, β_2)	$(0,1)$	$(0,-1)$	$(0,0)$	$(1,0)$	$(1,1)$	$(1,-1)$
equal prior	Scenario 1 $BF > 100$	96	98	93	99	97	99
	Scenario 2 $BF < 1$	93	96	98	94	96	90
	Scenario 3 $BF > 100$	11	8	3	6	9	9
	Scenario 3 $10 < BF < 100$	14	18	9	15	23	25