Bayesian Threshold Regression for Current Status Data with Informative Censoring

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Tao Xiao, M.S.

Graduate Program in Public Health

The Ohio State University

2015

Dissertation Committee:

Michael L. Pennell, Advisor
Grzegorz A. Rempala
Soledad Fernandez
Abstract

In some biomedical studies, there is interest in making inferences about a time to event distribution but the exact time of the event is unknown. For instance in carcinogenicity studies in animals, tumors are not discovered until the time of examination and hence time to tumor is left censored; this is known as current status data. Sometimes, the examination time is not independent of the event time. For example, in an animal study, an exam may have occurred because the animal died due to a cause related to tumor development. In this case, survival analysis methods which assume independent censoring would result in biased inferences. To address this issue, we propose a Bayesian approach which jointly models time to event and time to censoring using latent Wiener processes which fail once they hit a boundary value. Using data augmentation, we sample the unobserved event time and values of the latent processes for those subjects who do not experience an event. Informative censoring is accounted for also by modeling time to censoring using latent health process.

Sometimes multivariate current status data also arise, e.g., tumors can develop in multiple organ sites in carcinogenicity studies. Examination time occurring at natural death could be affected by these different types of tumors which may intrinsically correlate with each other. We propose a multivariate Bayesian approach to accommodate multiple left censored events driven by different latent Wiener processes. We use
a random effect shared by the drifts of the processes underlying the events of interest to model the correlation of the event times. The censoring process is modeled using a latent Wiener process whose time scale is affected by the occurrence of an event thus accounting for dependent censoring.

Our models are conceptually appealing and do not require the assumption of proportional hazards of some standard methods. In simulation studies, we found that the proposed informative censoring models provide more accurate estimates of regression coefficients than the independent censoring models when the data do not satisfy the assumption of independent censoring. We applied our methods to data from National Toxicology Program studies.
This is dedicated to my parents, brother, wife and daughter.
Acknowledgments

I wish to express my sincerest gratitude for my advisor, Dr. Michael L. Pennell, for valuable advice and guidance toward the completion of my dissertation. I want to thank my dissertation committee members, Dr. Grzegorz A. Rempala and Dr. Soledad Fernandez for their kindness to serve on my committee and help. I thank Dr. Mei-Ling Ting Lee, Dr. George Alex Whitmore, Dr. Xin He, Dr. Bo Lu, Dr. Yoonkyung Lee, Dr. Thomas J. Santner and Dr. Haikady N. Nagaraja for their help and support on my doctoral study.

Last but not least, I owe many thanks to my parents, my brother, my wife and my daughter.
Vita

2002 .............................................. B.E. in Electrical Engineering,  
Yanshan University, China

2008 .............................................. M.S. in Biostatistics,  
The Ohio State University, USA

2013 .............................................. M.S. in Statistics,  
The Ohio State University, USA

2006 .............................................. Graduate Research Associate,  
Department of Computer Science and Engineering,  
The Ohio State University, USA

2006-2008 ...................................... Graduate Research Associate,  
Biostatistics Division,  
The Ohio State University, USA

2008-2011 ...................................... Faculty Research Assistant/Lecturer,  
Department of Epidemiology and Biostatistics,  
University of Maryland - College Park, USA

2012-2013 ...................................... Graduate Teaching Associate,  
Biostatistics Division,  
The Ohio State University, USA

2013-2014 ...................................... Lecturer,  
Department of Epidemiology and Biostatistics,  
University of Maryland - College Park, USA

2014 .............................................. Statistician,  
Merkle Inc., USA
Publications


Fields of Study

Major Field: Public Health
Area of Specialization: Biostatistics
Additional Field: Data Mining and Machine Learning
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. Current Status Data</td>
<td>6</td>
</tr>
<tr>
<td>2.1 Types of Censoring</td>
<td>6</td>
</tr>
<tr>
<td>2.2 Models Assuming Independent Censoring</td>
<td>9</td>
</tr>
<tr>
<td>2.2.1 Cox Proportional Hazards Regression Model</td>
<td>9</td>
</tr>
<tr>
<td>2.2.2 Additive Hazards Regression Model</td>
<td>12</td>
</tr>
<tr>
<td>2.2.3 Accelerated Failure Time Model</td>
<td>13</td>
</tr>
<tr>
<td>2.2.4 Proportional Odds Regression Model</td>
<td>15</td>
</tr>
<tr>
<td>2.3 Models Allowing Informative Censoring</td>
<td>17</td>
</tr>
<tr>
<td>2.3.1 Accounting for Informative Censoring Through Shared Random Effects</td>
<td>17</td>
</tr>
<tr>
<td>2.3.2 Accounting for Informative Censoring Through Strict Lethality Assumptions</td>
<td>21</td>
</tr>
<tr>
<td>2.3.3 Accounting for Informative Censoring Through a Relaxed Lethality Assumption</td>
<td>26</td>
</tr>
</tbody>
</table>
2.3.4 Accounting for Informative Censoring Through a Copula Function ........................................... 34
2.4 Conditions Under Which Informative Censoring Should Be Considered 36
2.5 Summary ............................................................................................................................... 38

3. Threshold Regression ............................................................................................................. 40
   3.1 Threshold Regression for Right Censored Data ................................................................. 40
   3.2 Threshold Regression for Current Status Data ................................................................. 46
   3.3 Bayesian Approaches ....................................................................................................... 48
   3.4 Threshold Regression with Operational Time Scale ......................................................... 51

4. Bayesian threshold regression with informative censoring .................................................. 56
   4.1 Introduction ...................................................................................................................... 56
   4.2 Threshold Regression for Current Status Data ................................................................. 58
      4.2.1 Model Under Independent Censoring ..................................................................... 58
      4.2.2 Model Under Informative Censoring ..................................................................... 61
   4.3 Simulation Studies ............................................................................................................ 65
      4.3.1 Description of Data and Methods ........................................................................... 65
      4.3.2 Results ..................................................................................................................... 67
   4.4 Application to Tumorigenicity Study ................................................................................. 68
      4.4.1 Description of Data and Analysis Methods ............................................................ 68
      4.4.2 Results ..................................................................................................................... 72
   4.5 Concluding Remarks ........................................................................................................ 74

5. Bayesian threshold regression for multivariate current status data with informative censoring .................................................. 78
   5.1 Introduction ...................................................................................................................... 78
   5.2 Multivariate Threshold Regression Model for Current Status Data ............................... 80
      5.2.1 Model Under Independent Censoring ..................................................................... 80
      5.2.2 Model Under Informative Censoring ..................................................................... 84
   5.3 Simulation Studies ............................................................................................................ 87
      5.3.1 Description of Data and Methods ........................................................................... 87
      5.3.2 Results ..................................................................................................................... 90
   5.4 Application to Tumorigenicity Study ................................................................................. 91
      5.4.1 Description of Data and Analysis Methods ............................................................ 91
      5.4.2 Results ..................................................................................................................... 98
   5.5 Concluding Remarks ........................................................................................................ 106
6. Concluding Remarks ................................................. 110

6.1 Overview ......................................................... 110
6.2 Strengths and Limitations of Proposed Models ................. 111
6.3 Future Research .................................................. 113

Appendices ............................................................ 116

A. Chapter 4 Supplemental Material ................................. 116
   A.1 Model Under Independent Censoring: Posterior Computation and
       Data Augmentation ............................................... 116
       A.1.1 MCMC Algorithm ......................................... 116
   A.2 Model Under Informative Censoring: Posterior Computation and
       Data Augmentation ............................................... 118
       A.2.1 Likelihood Function ...................................... 118
       A.2.2 MCMC Algorithm ......................................... 119
       A.2.3 Simulation Results for the Parameters of Death Processes . 126
       A.2.4 Trace Plots of Regression Parameters from the Informative
           Censoring Model in the Analysis of the F344/N Male Rat
           Lung Tumor Data ............................................ 128

B. Chapter 5 Supplemental Material ................................. 131
   B.1 Model Under Independent Censoring: Posterior Computation and
       Data Augmentation ............................................... 131
       B.1.1 Likelihood Function ...................................... 131
       B.1.2 MCMC Algorithm ......................................... 132
   B.2 Model Under Informative Censoring: Posterior Computation and
       Data Augmentation ............................................... 135
       B.2.1 Likelihood Function ...................................... 135
       B.2.2 MCMC Algorithm ......................................... 137
       B.2.3 Simulation Results for the Parameters of Death Processes . 143
       B.2.4 Trace Plots of Regression Parameters from the Informative
           Censoring Model in the Analysis of the B6C3F1/NCTR Male
           Mouse Heart and Liver Hemangiosarcoma Data ............ 146

C. Use of the MatLab packages: TRInform and MultiTRInform .... 150
   C.1 The TRInform Package ....................................... 150
      C.1.1 Syntax ...................................................... 150
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Simulation results for the parameters of the event process by the informative and independent censoring models for different values of $\pi$</td>
<td>69</td>
</tr>
<tr>
<td>4.2 Distributions of number of death and lung tumor occurrence v.s. dose groups for the F344/N male rat lung tumor data (N=50/group)</td>
<td>72</td>
</tr>
<tr>
<td>4.3 Analysis result of the F344/N male rat lung tumor data by the informative and independent censoring models</td>
<td>75</td>
</tr>
<tr>
<td>5.1 Simulation results for the parameters of the event processes by the informative and independent censoring models for different values of $\tau$</td>
<td>92</td>
</tr>
<tr>
<td>5.2 Distribution of number of death v.s. dose groups for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data</td>
<td>96</td>
</tr>
<tr>
<td>5.3 Distribution of number of heart hemangiosarcomas v.s. dose groups for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data</td>
<td>97</td>
</tr>
<tr>
<td>5.4 Distribution of number of liver hemangiosarcomas v.s. dose groups for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data</td>
<td>97</td>
</tr>
<tr>
<td>5.5 Analysis result of the B6C3F1/NCTR male mouse heart and liver hemangiosarcoma data by the informative and independent censoring models, as well as the model with right censored data</td>
<td>100</td>
</tr>
<tr>
<td>A.1 Simulation results for the parameters of the death processes by the informative censoring model for different values of $\pi$</td>
<td>127</td>
</tr>
<tr>
<td>B.1 Simulation results for the parameters of the censoring process in the operational time scale</td>
<td>145</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Log-Log Plot by the “treatment2” Variable for the Leukemia Data</td>
<td>55</td>
</tr>
<tr>
<td>3.2</td>
<td>Kaplan-Meier Plot by the “treatment2” Variable for the Leukemia Data</td>
<td>55</td>
</tr>
<tr>
<td>3.3</td>
<td>Cox Predicted Plot v.s. Kaplan-Meier Plot by the “treatment2” Variable for the Leukemia Data</td>
<td>55</td>
</tr>
<tr>
<td>3.4</td>
<td>TR Predicted Plot v.s. Kaplan-Meier Plot by the “treatment2” Variable for the Leukemia Data</td>
<td>55</td>
</tr>
<tr>
<td>4.1</td>
<td>Comparisons of survival estimates for the event of interest $E$ provided by the informative and independent censoring models for data simulated at $\pi = 0.9, 0.75, 0.5$ and $0.25$.</td>
<td>70</td>
</tr>
<tr>
<td>4.2</td>
<td>Kaplan-Meier survival estimates for the four different dose groups of the F344/N male rat lung tumor data (in days)</td>
<td>71</td>
</tr>
<tr>
<td>4.3</td>
<td>Estimated survival probability curves for lung tumor onset at the mean baseline weight of 121.545g by the informative and independent censoring models for the F344/N male rat lung tumor data (in days)</td>
<td>76</td>
</tr>
<tr>
<td>5.1</td>
<td>Comparisons of survival estimates provided by the informative and independent censoring models for the first event of interest ($E_1$) in the simulation studies</td>
<td>93</td>
</tr>
<tr>
<td>5.2</td>
<td>Comparisons of survival estimates provided by the informative and independent censoring models for the second event of interest ($E_2$) in the simulation studies</td>
<td>94</td>
</tr>
</tbody>
</table>
5.3 Kaplan-Meier survival estimates for the four different dose groups of the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data (in days) ................................................................. 96

5.4 Kaplan-Meier survival estimates and Nelson-Aalen cumulative hazard estimates for subjects with only one of the two hemangiosarcomas and with both of the two hemangiosarcomas versus those without any hemangiosarcoma, for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data (in days) .............................. 102

5.5 Estimated survival probability curves at the mean baseline weight of 14.102g by the informative and independent censoring models, as well as the right censored data model for heart hemangiosarcoma, for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data (in days) 107

5.6 Estimated survival probability curves at the mean baseline weight of 14.102g by the informative and independent censoring models, as well as the right censored data model for liver hemangiosarcoma, for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data (in days) 108

A.1 $\alpha_{E0}$ ............................................................... 128

A.2 $\alpha_{E1}$ ............................................................... 128

A.3 $\alpha_{E2}$ ............................................................... 128

A.4 $y_{E0}$ ............................................................... 128

A.5 $\alpha_{C10}$ ............................................................ 129

A.6 $\alpha_{C11}$ ............................................................ 129

A.7 $\alpha_{C12}$ ............................................................ 129

A.8 $y_{C10}$ ............................................................ 129

A.9 $\alpha_{C20}$ ............................................................ 130

A.10 $\alpha_{C21}$ .......................................................... 130

A.11 $\alpha_{C22}$ .......................................................... 130

xiv
Chapter 1: INTRODUCTION

Current status data are a special case of interval-censored data, where the observation of time to event $S$ is restricted to knowledge of whether or not $S$ exceeds an examination (censoring) time $U$ (Jewell and van der Laan, 2004). For example, in carcinogenicity experiments in animals, investigators are interested in inferences on tumor onset time. However, the exact tumor onset time for an animal usually cannot be observed; whether an animal gets a tumor can only be known at natural death or sacrifice (e.g., Gart et al., 1990; Sun and Kalbfleisch, 1993). In studies examining the distribution of age at weaning ($S$), the exact measurements of the age at weaning are usually inaccurate and biased, and for this reason, investigators often treat $S$ as interval censored over $(0, U)$, where $U$ is the survey time (e.g., Diamond et al., 1986, 1991; Grummer-Strawn, 2008). Other examples of current status data include HIV infection time (e.g., Jewell and Shiboski, 1990; Shiboski, 1998) and time of a non-fatal human disease (e.g., Keiding, 1991; Keiding et al., 1996) collected in cross-sectional studies.

There are many papers addressing the analysis of current status data. For instance, Cox (Huang, 1996), additive hazards (Lin et al., 1998; Martinussen and Scheike, 2002), accelerated failure time (Rabinowitz et al., 1995) and proportional
odds models (Rossini and Tsiatis, 1996) have been proposed. All of the aforementioned methods assume that censoring time \((U)\) and event time \((S)\) are independent. However, this independence assumption is questionable in many situations. For example, in tumorigenicity studies, the examination for tumors can be carried out only when an animal dies or is sacrificed at the end of study. Animals who developed a tumor should die sooner than those without a tumor, and hence, for many animals, examination time is related to tumor onset time.

A few authors have developed methods for current status data which account for dependent censoring. Lin et al. (1998) used an additive hazards model to analyze current status data and assumed a proportional hazards model for the censoring time. Dependent censoring was accounted for solely through covariates shared by these two models, i.e., censoring time and time to event are conditionally independent given covariates. Zhang et al. (2005) extended the model proposed by Lin et al. (1998) by incorporating random effects shared by the censoring and event time models. While these models are reasonable for some settings, the event may not be related to the censoring event for all subjects. For example, in tumorigenicity studies, tumors can be either lethal or non-lethal, and non-lethal tumors should not cause dependence between death time and tumor onset time. Lagakos and Louis (1988) discussed how bias is incurred if a tumor is treated as either of two extremes: non-lethal or rapidly lethal. To address this issue, they proposed a score test for comparing control and exposed groups in tumorigenicity experiments. Their test incorporates a function quantifying the probability that the tumor observed at death was lethal. Although their method
is an improvement over previous methods which assume either rapid lethality or non-lethality, their method still needs to pre-specify the lethality probability and requires a sensitivity analysis.

Also motivated by tumorigenicity studies, Dunson and Dinse (2002) proposed a Bayesian regression model for current status data with informative censoring. The authors accounted for tumor lethality using a latent binary variable which can be augmented for each subject, indicating whether or not death was caused by tumor. Subjects with tumor affected deaths indicated by this latent binary variable would have different likelihood contributions than those whose deaths had nothing to do with tumor, and hence informative censoring is accounted for. While Dunson and Dinse (2002) avoided the assumption that every tumor is lethal and the need to pre-specify the lethality probability, their model assumes constant tumor lethality across subjects. However, since tumor lethality could vary across subjects due to genetic, treatment or environmental factors, the constant tumor lethality assumption might be unreasonable. In addition, the computational burden of their model is high since a discretized time scale is used and a large amount of regression parameters are needed to estimate the event and examination time distributions in the discretized time scale.

To address the limitations in Dunson and Dinse’s method and other previous methods for current status data, we propose Bayesian threshold regression models for the analysis of the current status data subject to informative censoring. Threshold regression models time-to-event as the first time a latent stochastic process reaches a threshold level. This feature is conceptually appealing in biomedical studies. For example, when a subject is exposed to a carcinogen, normal cells are damaged and malignant cells begin to emerge. Once enough of these malignant cells accumulate
in the same organ, a tumor appears. This entire process is latent making threshold regression a natural model for time to tumor. There is an extensive literature on applications of threshold regression to biomedical data (e.g., Lee et al., 2004; Whitmore and Su, 2007; Balka et al., 2009; Yu et al., 2009). Most of the previous work on threshold regression was for right censored survival data. Tong et al. (2008) proposed a frequentist threshold regression model for current status data and used their method to model time to tumor onset following exposure to a carcinogen. However, the authors assumed independent censoring in their model, which is questionable for time to tumor data.

Similar to Tong et al. (2008), we model time to an interval censored event using a Wiener diffusion process. However, to account for informative censoring, we also model time to examination using a diffusion process, which may be dependent of the event of interest, starting at the occurrence of the event of interest. The two informative censoring models proposed in this dissertation are based on this approach and hence assume a shifting of the censoring time distribution following the event of interest.

In Chapter 2, we review previous methods for the analysis of current status data. We review the independent censoring models for current status data, the available dependent censoring models, and the situations that the informative censoring should be considered for current status data.

In Chapter 3, we review the literature on threshold regression modelling. We start with the basic threshold regression model for right censored data, and then we review several extensions of this basic model, including models for independently
censored current status data, random effect models, and the use of operational time in threshold regression.

In Chapter 4, we propose a Bayesian threshold regression for current status data with informative censoring. We conduct a simulation study to assess the performance of our model in comparisons with the model without accounting for informative censoring. To illustrate our method, we apply our model to a NTP lung tumor dataset.

In Chapter 5, we propose a Bayesian threshold regression model for multivariate current status data. We utilize the threshold regression with operational time scale in the censoring process to account for dependent censoring of the multivariate current status data.

In Chapter 6, we summarize our proposed methods and discuss future research directions.
Chapter 2: CURRENT STATUS DATA

2.1 Types of Censoring

In survival analysis, the primary interest is to characterize the failure time \( T \) distribution. One hopes to observe enough exact failure time values and use them to estimate the distribution of \( T \). However, since investigators often cannot follow all subjects until failure, censoring times (denoted by \( U \)) exist. The most common censoring mechanism is right censoring, where the event of interest is observed only if it happens prior to time \( U_r \) (\( U_r \) for “right” censoring time). Another censoring mechanism is left censoring, where the event of interest has already occurred before the subject is observed at time \( U_l \) (\( U_l \) for “left” censoring time). When both left censoring and right censoring occur in a study, such data are called doubly censored data. A more general type of censoring, interval censoring, occurs when the failure time is only known to happen within an interval \( (U_l, U_r) \). Interval censoring is a generalization of left and right censoring since, when the interval is \( (0, U_r) \) we have left censoring and, when the interval is \( (U_l, +\infty) \) we have right censoring. An important case of interval censored data is when each subject \( i \) is observed only once at time \( U_i \) and the failure time is only known to be smaller or greater than \( U_i \), or in other words, falling into either \( (0, U_i) \) or \( (U_i, +\infty) \). We define a binary random variable \( D \)
for such information: $D$ is 1 if $T \leq U$ and 0 if $T > U$. Thus $P(D = 1) = P(T \leq u \mid U = u) = F(u)$. Given a specific model for $F(U)$, the likelihood of these data for a finite sample size $n$ is

$$L = \prod_{i=1}^{n} F(u_i)^{d_i} (1 - F(u_i))^{1-d_i}$$  \hspace{1cm} (2.1)

Such data are commonly referred to as current status data.

For current status data, the relationship between the failure time $T$ and a vector of covariates $Z$ is often in our interest, although $T$ cannot be observed exactly. This analysis can be achieved through generalized linear regression models, with $T$ or a function characterizing the distribution of $T$ (e.g., survival function, hazard function) linked to $Z$.

When the censoring time $U$ is unrelated to the event time $T$ in the study design, we have independent censoring. Written formally, suppose that we have an indicator $I_i$ that takes the value 1 when $U_i = T_i$ (i.e., when we observe the actual event time of interest) and the value 0 when we only observe a censoring time $U$, we have independent censoring when a subject whose event time is right censored has the same risk of experiencing the event of interest in the small time interval $[t, t + dt)$ as would have been the case in the situation without censoring, i.e.,

$$P(t \leq U_i < t + dt, I_i = 1 \mid U_i \geq t, \text{past}) = P(t \leq T_i < t + dt \mid T_i \geq t),$$  \hspace{1cm} (2.2)

where past denotes the event history up to time $U$ for all $n$ subjects (i.e., we know for all $i$ whether $U_i \geq t$ or $U_i \leq t$) and $dt$ is some small increment in time (Aalen et al., 2008). Note that the left hand side of Equation (2.2) is equivalent to

$$P(t \leq T_i < t + dt \mid T_i \geq t, U_i \geq t, \text{past}).$$
To see how 2.2 follows from the usual definition of independence, let's define the following events:

\[ A_1 = \{ t \leq T_i < t + dt \}, \]
\[ B_1 = \{ T_i \geq t \}, \]
\[ C_1 = \{ U_i \geq t, \text{past} \}. \]

Using this event notation we have \( P(A_1 \mid B_1, C_1) = P(A_1 \mid B_1) \), and hence \( A_1 \) is independent of \( C_1 \), conditional on \( B_1 \); or in other words, censoring status at \( t \) is unrelated to future risk of an event.

In terms of left censored data, the independent censoring assumption means that a subject whose event time is left censored has the same risk of experiencing the event of interest in the small time interval \((t - dt, t]\) as would have been the case in the situation without censoring, i.e.,

\[ P(t - dt < U_i \leq t, I_i = 1 \mid U_i \leq t, \text{past}) = P(t - dt < T_i \leq t \mid T_i \leq t), \]  \hspace{1cm} (2.3)

Again, note that the left hand side of Equation (2.3) is equivalent to

\[ P(t - dt < T_i \leq t \mid T_i \leq t, U_i \leq t, \text{past}). \]

Again, if we define the following events:

\[ A_2 = \{ t - dt < T_i \leq t \}, \]
\[ B_2 = \{ T_i \leq t \}, \]
\[ C_2 = \{ U_i \leq t, \text{past} \}; \]

then we have \( P(A_2 \mid B_2, C_2) = P(A_2 \mid B_2) \), and hence \( A_2 \) is independent of \( C_2 \), conditional on \( B_2 \); or in other words, censoring status at \( t \) does not provide information about the risk of an event over \((0, t]\).
With the inclusion of the vector of fixed covariates $x_i$ for subject $i$ in a regression model, Equations (2.2) and (2.3) are slightly modified as

$$P(t \leq U_i < t + dt, I_i = 1 | U_i \geq t, \text{past}^*) = P(t \leq T_i \leq t + dt | T_i \geq t, x_i) \quad (2.4)$$

and

$$P(t - dt < U_i \leq t, I_i = 1 | U_i \leq t, \text{past}^*) = P(t - dt < T_i \leq t | T_i \leq t, x_i) \quad (2.5)$$

where $\text{past}^*$ denotes the event history up to time $U$ and covariate values for all $n$ subjects. The independence between the event of interest and the event of censoring for subject $i$ is now conditional on $x_i$; or in other words, independent censoring for a regression model means that conditional on covariate values, censoring status at $t$ does not provide information about past or future risk of the event.

The analysis of current status data has drawn the attention of many statisticians. In the remainder of this chapter we review different regression models that have been proposed to analyze current status data. We first review current status data models assuming independent censoring, then review models accounting for informative censoring and finally review the conditions under which informative censoring must be accounted for.

2.2 Models Assuming Independent Censoring

2.2.1 Cox Proportional Hazards Regression Model

The Cox proportional hazards regression model (Cox, 1972) serves as the most popular regression model in survival analysis for studying the relationship between event times and covariates. It does not require any assumptions about the baseline hazard function. The Cox model specifies the hazard function of $T$ given the covariate
vector $z$ in the following way:

$$h(t \mid z) = h_0(t) \exp(\beta' z)$$  \hspace{1cm} (2.6)$$

where $h_0(t)$ is an arbitrary baseline hazard function and $\beta$ is a parameter vector. Thus the cumulative hazard and survival functions are respectively:

$$H(t \mid z) = \int_0^t h(u \mid z) du = H_0(t) \exp(\beta' z)$$  \hspace{1cm} (2.7)$$

and

$$S(t \mid z) = \exp\{-H(t \mid z)\} = S_0(t)^{\exp(\beta' z)}$$  \hspace{1cm} (2.8)$$

where $H_0(t)$ and $S_0(t \mid z) = 1 - F_0(t \mid z)$ are, respectively, the baseline cumulative hazard function and survival function of $T$; $H$ and $S$ are assumed to be continuous.

The Cox model is often called a proportional hazards model since, if we look at two different subjects with covariate values $z$ and $z^*$, the ratio of their hazard functions is:

$$\frac{h(t \mid z)}{h(t \mid z^*)} = \frac{h_0(t) \exp(\beta' z)}{h_0(t) \exp(\beta' z^*)} = \exp [\beta'(z - z^*)]$$  \hspace{1cm} (2.9)$$

which is independent of $t$ and hence the hazard functions of these two subjects are proportional throughout $t$. The ratio in (2.9) is often called hazard ratio (or relative instantaneous risk) of a subject with covariate vector $z$ as compared to a subject with covariate vector $z^*$. For a treatment covariate $z_1$ ($z_1 = 1$ if treatment and $z_1 = 0$ if placebo), $\log \left[ h(t \mid z_1 = 1)/h(t \mid z_1 = 0) \right] = \beta_1$, and hence $\beta_1$ can interpreted as the log hazard ratio of subjects who received the treatment as compared to those who received the placebo. Although the interpretability of regression coefficients in the Cox model is easily understood, the proportional hazards assumption in this model may be violated in some applications.
Huang (1996) studied maximum likelihood estimation (MLE) of the Cox proportional hazards model for current status data assuming independent censoring. For a single observation $X = (U, d, Z)$, its likelihood function is proportional to

$$P_{\theta, F}(x) = F(U \mid Z)^d S(U \mid Z)^{1-d} \left[1 - S_0(U)^{\exp(\beta'Z)}\right]^{d} S_0(U)^{(1-d)^{\exp(\beta'Z)}}$$ \tag{2.10}

And hence the log-likelihood function is, up to an additive constant,

$$l(\theta, F) = d \log(1 - S_0(U)^{\exp(\beta'Z)}) + (1 - d) e^{\beta'Z} \log S_0(U)$$ \tag{2.11}

Note that the hazard function $h$ can be calculated as

$$h(x) = \frac{f(x)}{1 - F(x)} = -\frac{d \log[1 - F(x)]}{dx}$$ \tag{2.12}

where $f$ is the probability density function of the event time, and hence

$$H(x) = \int_0^x h(u) du = -\log[1 - F(x)]$$ \tag{2.13}

Let $(U_1, D_1, Z_1), \ldots, (U_n, D_n, Z_n)$ be an i.i.d. sample distributed as $(Y, D, Z)$. The log-likelihood function for the sample can be written in terms of $H$:

$$l_n(\beta, H) = \sum_{i=1}^{n} \left[ d_i \log(1 - \exp(-H_0(U_i) \exp(\beta'Z_i))) - (1 - d_i) \exp(\beta'Z_i) H_0(U_i) \right].$$ \tag{2.14}

The author treated $H$ as a right-continuous increasing step function, and as a result, the maximum likelihood estimator of $\beta$ and $H_0$ was obtained by maximizing

$$\phi(\beta, \bar{x}) = \sum_{i=1}^{n} \left\{ d_{(i)} \log \left[ 1 - \exp \left( -\exp(\beta'Z_{(i)}x_i) \right) \right] - (1 - d_{(i)}) \exp(\beta'Z_{(i)}x_i) \right\}$$ \tag{2.15}

subject to $\beta \in \Theta$ and $0 \leq x_1 \leq x_2 \leq \cdots \leq x_n$, where $\Theta \subset \mathbb{R}^d$ is the finite-dimensional parameter space of $\beta$ and $x_i$ stands for $H_0(U_{(i)})$, where $U_{(i)}$ is the $i$th order statistic of $U_i$'s.
2.2.2 Additive Hazards Regression Model

In Section 2.2.1, we discussed the analysis of current status data by the proportional hazards regression model, in which the effect of covariates act multiplicatively on an unspecified baseline hazard function. In this section, we shall review the analysis of current status data by a so-called additive hazard model in which the effects of covariates act additively on an unspecified baseline hazard function in the following manner (Lin et al., 1998):

\[ h(t \mid Z) = h_0(t) + \beta' \mathbf{Z}(t) \]  \hspace{1cm} (2.16)

where \( h_0(t) \) is an unspecified baseline hazard function, \( \mathbf{Z}(\cdot) \) is a \( p \)-dimensional vector of time-dependent covariates and \( \beta \) is a vector of regression coefficients. Although the multiplicative and additive forms of the hazard functions are quite different, Ling and Ying demonstrated an interesting relationship between the two models which simplifies maximum likelihood estimation for the additive hazards models. Let \( \{U_i, D_i, \mathbf{Z}_i(\cdot)\} (i = 1, \ldots, n) \) be the observations in a dataset. Subject \( i \) is at risk at \( t \) if and only if \( U_i \geq t \). Consider the hazard \( h_i(t) \) for a subject that is examined at time \( t \) and still failure-free. Two things are true for this subject: (i) \( U_i = t \) and (ii) the resulting examining reveals that subject \( i \) has been failure-free up to \( t \). The hazard of event (i) is denoted by \( h_c(t) \). Under model (2.16), event (ii) has the conditional probability

\[ \Pr\{T_i \geq t \mid \mathbf{Z}_i(s), s \leq t\} = e^{-H_0(t) - \beta' \mathbf{Z}_i^*(t)}, \]  \hspace{1cm} (2.17)

where

\[ H_0(t) = \int_0^t h_0(s)ds \]  \hspace{1cm} (2.18)

\[ Z_i^*(t) = \int_0^t Z_i(s)ds \]
Under the assumption of independent censoring, we take the product of the hazard for events (i) and conditional probability of event (ii) to obtain

\[ h_i(t) = e^{-\beta' Z_i(t)} h_0(t), \]  

(2.19)

where \( h_0(t) = e^{-H_0(t)} h_c(t) \). Equation (2.19) is the Cox proportional hazards model. It turns out that one can just input the data into standard software for fitting the Cox proportional hazards model with right-censored data to make inferences about the regression coefficients. However, unlike the Cox model for current status data, in which the regression coefficients are interpreted as log-hazard ratios, the regression coefficients in the additive hazard model are additive effects on the hazard, since hazard is modeled as a linear function of covariates. In addition, the additive hazard model doesn’t assume proportional hazards, which the Cox model does.

2.2.3 Accelerated Failure Time Model

Rabinowitz et al. (1995) proposed an accelerated failure time model for interval censored data. Suppose that data are collected from \( n \) subjects. Let \( T_i \) denote the survival time, and \( Z_i \) a \( p \)-dimensional covariate associated with the \( i \)th subject. Suppose that the natural logarithm of the survival time \( V_i = \log(T_i) \) is related to the covariates through a linear regression

\[ V_i = \beta' Z_i + \varepsilon_i \]  

(2.20)

where \( \beta \) is a \( p \)-dimensional regression coefficient vector and \( \varepsilon_i \) are independent and identically distributed residuals. Let \( S_0(T_i) \) denote the survival function given \( Z_i = 0 \).
The survival function is then

$$\Pr[T_i > t_i \mid Z_i] = \Pr[V_i > \log(t_i) \mid Z_i]$$

$$= \Pr[\epsilon_i > \log(t_i) - \beta' Z_i \mid Z_i]$$

$$= \Pr[\exp(\epsilon_i) > t_i \exp(-\beta' Z_i) \mid Z_i]$$

$$= S_0(t_i \exp(-\beta' Z_i))$$

Thus by (2.21), the predictors $Z_i$ change the time scale by a factor $\exp(-\beta' Z_i)$. If $-\beta' Z_i$ is positive, then $\exp(-\beta' Z_i)$ is greater than one and the time is accelerated; otherwise the time is degraded. Unlike the Cox proportional hazards model and the additive hazards model reviewed in Sections 2.2.1 and 2.2.2, which model the hazard function of the event time, the accelerated failure time model provides an analog of the classical linear model and models the event time directly. However, the use of the accelerated failure time model is restricted by the distribution assumption of the residuals.

Let’s denote the distribution function of the residuals by $F_\epsilon$, and its derivative by $f_\epsilon$. Let $n_i$ denote the number of examinations experienced by the $i$th patient, and let

$$X_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,n_i})$$

denote the $i$th patient’s ordered sequence of log examination times. For convenience, define $X_{i,0} = -\infty$ and $X_{i,n_i+1} = \infty$. Assume that the $\epsilon_i$ are independent of the pairs $(Z_i, X_i)$, which are independent and identically distributed.

Let $X_{i,L}$ be the last of the $i$th subject’s examination times preceding $T_i$, and let $X_{i,U}$ be the first examination time following $T_i$. The authors defined the pair of examination times on the time scale of the residual by

$$X_{i,L}(\beta) = X_{i,L} - \beta' Z_i, X_{i,U}(\beta) = X_{i,U} - \beta' Z_i$$

(2.22)
It can be demonstrated that the score function of the model is

\[
\dot{S}(\beta) = \sum_{i=1}^{n} \left[ \frac{g[F_\varepsilon\{X_{i,U}(\beta)\}] - g[F_\varepsilon\{X_{i,L}(\beta)\}]}{F_\varepsilon\{X_{i,U}(\beta)\} - F_\varepsilon\{X_{i,L}(\beta)\}} Z_i \right]
\]  

(2.23)

If \(F_\varepsilon\) is known, then \(g = f_\varepsilon \circ F_\varepsilon^{-1}\). However, since \(F_\varepsilon\) is often unknown, the authors proposed to substitute an estimate of \(F_\varepsilon\) that maximizes the resulting log-likelihood. A procedure of finding the optimal score function, including the choice of \(g(\cdot)\) and the number of jumps that an estimate of \(F_\varepsilon\) is to take, was explored. An estimation procedure for the regression coefficients using the score statistic was proposed.

For current status data, there is only one examination time \(X_i\), and hence only one \(X_i(\beta)\). Let \(Y_i = 1\{\varepsilon_i < X_i(\beta)\}\). The author proposed the following score statistics for current status data:

\[
\dot{S}(\beta) = \sum_j \sum_{i \in U_j(\beta)} \left[ Y_i - \bar{Y}_j(\beta) \right] \left[ Z_i - \bar{Z}_j(\beta) \right] w_j(\beta)
\]  

(2.24)

where \(U_j(\beta)\) is the subset of subjects whose censoring times are used to calculate \(\tau_j(\beta)\), which is the \(j\)th jump point of the estimate of \(F_\varepsilon\). The terms \(\bar{Y}_j(\beta)\) and \(\bar{Z}_j(\beta)\) are the average of the \(Y_i\) and \(Z_i\), respectively, for \(i \in U_j(\beta)\). Finally,

\[
w_j(\beta) = g[\hat{F}_\varepsilon[\tau_j(\beta)]]/[\hat{F}_\varepsilon[(\tau_j(\beta))]][1 - \hat{F}_\varepsilon[\tau_j(\beta)]]].
\]

2.2.4 Proportional Odds Regression Model

Rossini and Tsiatis (1996) proposed a proportional odds model for the analysis of current status data. Assume that for the \(i\)th subject, \(i = 1, \ldots, n\), there exists a failure time \(T_i\) and a \(p\)-dimensional vector of covariates, \(Z_i\). Denote the conditional distribution function of \(T_i\) given \(Z_i\) by \(F(\cdot | Z_i)\). For current status data, the pair of variables \((U_i, D_i)\) are observed where \(U_i\) is the monitoring time and \(D_i = I(T_i \leq U_i)\).
This model is defined by the relation

\[ \text{logit} F(t \mid Z) = \alpha(t) + \beta' Z \quad (2.25) \]

with the logit function defined by \( \text{logit}(s) = \log(s/(1-s)) \), \( \alpha(t) \) a monotone-increasing function and \( \beta \) a \( p \)-dimensional vector of regression coefficients. The parameter \( \alpha(t) \) is interpreted as the log-odds of a failure at time \( t \), when the associated covariates are \( Z = 0 \), and \( \beta'Z \) is the additive change in the log-odds of failure of an observation with covariates \( Z \). Thus

\[
\Pr(D = d \mid U, Z) = F(U \mid Z)^d(1 - F(U \mid Z))^{1-d} \quad (2.26)
\]

\[
= \frac{\exp(d(\alpha(U) + \beta'Z))}{1 + \exp(\alpha(U) + \beta'Z)}
\]

Inference for \( \beta, \alpha \) is based on the conditional likelihood given \( U \) and \( Z \),

\[
l(\beta, \alpha) = \prod_{i=1}^{n} \left\{ \frac{\exp(d_i(\alpha(U_i) + \beta'Z_i))}{1 + \exp(\alpha(U_i) + \beta'Z_i)} \right\} \quad (2.27)
\]

The author developed an estimation procedure for \( \beta \), with the baseline log-odds function \( \alpha \) treated as a nuisance parameter approximated by a step function. The estimates of \( \beta \) and the step function approximating \( \alpha \) are obtained by maximizing (2.27) with \( \alpha \) substituted by the step function. Treating the baseline log-odds function as a nuisance parameter removes restrictions caused by distribution assumptions. However, the estimation of \( \alpha \) can be biased at small sample sizes leading to biased estimation of \( \beta \). Furthermore, when the sample size is large, the estimation of the nuisance parameter \( \alpha \) could be too computationally intensive since the number of steps increases with the sample size and as does the number of parameters used to characterize the step function. The regression coefficients in this proportional odds model are the log-odds ratios comparing odds of an event occurring by some time \( t \)
between two individuals who differ by one unit in the corresponding covariate; these odds ratios are more intuitive to clinicians than the regression coefficients in the Cox proportional hazards model which correspond to log-hazard ratios. However, similar to the Cox model, this proportional odds model makes the assumption that the odds ratios don’t vary with time, which could be violated in some applications.

2.3 Models Allowing Informative Censoring

In this section, we review models for current status data allowing informative censoring. There have been two different mechanisms of dependency assumed in the literature. The first mechanism assumes a continuous relationship between the processes underlying time to event and time to censoring through shared frailty type models. The second is a “shifting mechanism” in which the dependence between the event time and the censoring time is established when the occurrence of an event causes a change in the censoring distribution. We review models with the first mechanism in Section 2.3.1. We review models with the second mechanism in Sections 2.3.2 and 2.3.3. We also review an approach which models the correlations between the distribution functions of time to event and time to censoring by a copula function in Section 2.3.4.

2.3.1 Accounting for Informative Censoring Through Shared Random Effects

Zhang et al. (2005) extended the additive hazards regression model proposed by Lin et al. (1998) as described in Section 2.2.2 by incorporating random effects shared by the censoring and event time to account for informative censoring. The model is described as follows.
Consider a survival study that involves \( n \) independent subjects and each subject gives rise to the event time of interest \( T_i \) and the censoring time \( U_i \), which may depend on \( T_i, i = 1, \ldots, n \). For subject \( i \), let \( \mathbf{Z}_i(t) \) denote a \( p \)-dimensional vector of covariates that may be time-dependent. The authors assumed the relationship between \( T_i \) and \( U_i \) can be characterized by a random effect \( b_i(t) \), which could be time-dependent. The authors assumed that the hazard function of the survival times \( T_i \)'s at time \( t \) has the form

\[
h_i(t \mid \mathbf{Z}_i(s), b_i(s), s \leq t) = h_1(t) + b_i(t) + \beta' \mathbf{Z}_i(t) \tag{2.28}
\]

given \( \{\mathbf{Z}_i(s), b_i(s), s \leq t\} \), where \( h_1(t) \) is an unknown baseline hazard function and \( \beta \) denotes the vector of \( p \)-dimensional regression coefficients. That is, the \( T_i \)'s follow the additive hazards frailty model.

For the \( U_i \)'s, the authors assumed that given \( \{\mathbf{Z}_i(s), b_i(s), s \leq t\} \), they follow the proportional hazards frailty model

\[
h_i^U(t \mid \mathbf{Z}_i(s), b_i(s), s \leq t) = h_2(t) \exp\{\gamma' \mathbf{Z}_i(t) + b_i(t)\} \tag{2.29}
\]

where \( h_2(t) \) is an unknown baseline hazard function as \( h_1(t) \) and \( \gamma \) denotes the effect of covariates on the \( U_i \)'s. Furthermore, it is assumed that the random effects \( b_i(t) \)'s are arbitrary processes with mean zero. This way, the time of event \( T_i \) and the censoring time \( U_i \) for the \( i \)th subject are modelled as dependent on each other through the shared random effects \( b_i(t) \).

Note that model (2.28) is equivalent to the usual additive hazards model. This can be seen by noting that under model (2.28), we have

\[
\Pr\{T_i \geq t \mid \mathbf{Z}_i(s), s \leq t\} = \exp\{-H_1^*(t) - \beta' \mathbf{Z}_i^*(t)\} \tag{2.30}
\]
where $Z_i^*(t) = \int_0^t Z_i(s) ds$, $H_i^*(t) = \int_0^t h_1(s) ds - \log[E\{e^{-B_i(t)}\}]$ and $B_i(t) = \int_0^t b_i(s) ds, i = 1, \ldots, n$. And hence the hazard function of $T_i$ can be written as the following additive hazards model
\[ h_1(t \mid Z_i(s), s \leq t) = h_1^*(t) + \beta'Z_i(t) \] (2.31)
where $h_1^*(t)$ is the derivative of $H_i^*(t)$. The authors provided the estimation procedure for the regression parameter vector $(\beta, \gamma)$.

Kim (2012) proposed a model for bivariate current status data accounting for informative censoring using shared random effects. This model was also motivated by a tumorigenicity study on two different types of tumors. Let $T_{ij}$ denote the $j$th unobservable tumor occurrence time. The author proposed to use bivariate normal random effects $b_i = (b_{i1}, b_{i2})'$ to account for both the correlation between the two tumor onsets and that between tumor onset and death time. The hazard functions of tumor 1 and tumor 2 were defined as follows:
\[
\begin{align*}
\alpha_{i1}(s \mid b_{i1}) &= \alpha_0(s) \exp(z_{i1}' \beta_1 + b_{i1}), \\
\alpha_{i2}(s \mid b_{i2}) &= \alpha_0(s) \exp(z_{i2}' \beta_2 + b_{i2})
\end{align*}
\]
where $\alpha_0(s)$ is a baseline hazard function, $z_i'$ is the covariate vector, $\beta_1$ and $\beta_2$ are regression coefficient vectors. Note that this method makes a strong assumption that the hazard functions of tumor onset for the two tumors have a same baseline hazard function $\alpha_0(s)$. This assumption also induces a direct dependence between the onset of the two tumors.
Kim assumed that tumor 1 and tumor 2 have different lethalities. The hazard functions of death among animals with each tumor type were defined as follows:

\[ h_{i1}(t \mid b_{i1}) = h_0(t) \exp(z_i' \gamma_1 + b_{i1}), \]
\[ h_{i2}(t \mid b_{i2}) = h_0(t) \exp(z_i' \gamma_2 + b_{i2}) \]

The hazard function of death without tumor was defined as

\[ h_{i3}(t) = h_0(t) \exp(z_i' \gamma_3) \]

where \( h_0(t) \) is the baseline hazard function; \( \gamma_1, \gamma_2 \) and \( \gamma_3 \) are regression coefficient vectors. Again, the author makes the strong assumption that the hazard functions of death due to different causes have the same baseline hazard function \( h_0(t) \). The author adopted piecewise constant baseline hazard functions for \( \alpha_0(s) \) and \( h_0(t) \), and used an EM algorithm to find the maximum likelihood estimates of regression parameters and the baseline hazard functions \( \alpha_0(s) \) and \( h_0(t) \).

Chen et al. (2012) also proposed a class of semiparametric models for current status data accounting for informative censoring through random effects. In their model, the correlation between the event time \( T \) and the censoring time \( U \) was accounted for by a normal random effect \( b_i \). Given \( b_i \), the cumulative hazard function of \( T_i \) at time \( t \) was

\[ H_T(t \mid Z_i, b_i) = e^{b_i} G[\exp(Z_i' \beta) R(t)] \]  \hspace{1cm} (2.32)

where \( G \) is a prespecified strictly increasing function, \( Z_i \) is a covariate vector, \( \beta \) is a regression coefficient vector, and \( R \) is an unspecified increasing real-valued function which can be estimated. When \( G(x) = x \), model 2.32 reduces to the proportional hazards frailty model with log-normal frailty and \( R(t) \) is a baseline cumulative hazard function.
Given $b_i$, the cumulative hazard function of $U_i$ was modeled using a proportional hazards model

$$H_C(u \mid \mathbf{Z}_i, b_i) = \exp(\mathbf{Z}_i' \gamma + b_i)H_0(u)$$

(2.33)

where $\gamma$ is a covariate vector and $H_0$ denotes the unknown baseline cumulative hazard function.

To alleviate computation burden, the authors proposed to use a piecewise constant approximation for $R$. The authors used an EM algorithm to find the maximum likelihood estimates of regression parameters of this model.

The above approaches accounting for informative censoring through shared random effects, account for correlations between the event and censoring times, however, they intrinsically assume that censoring time is correlated with event time for each subject and thus may not be suitable for situations when this assumption violated. For example, in tumorigenicity studies of animals, onset of a mild tumor may only affect death for a portion of animals, not all of them.

### 2.3.2 Accounting for Informative Censoring Through Strict Lethality Assumptions

Lagakos and Louis (1988) specifically developed a statistical test for dealing with current status data in tumorigenicity studies, accounting for informative censoring. Before their method, the two most commonly used statistical tests for comparing control and exposed groups in tumorigenicity experiments were the Hoel-Walburg and logrank tests. The former is appropriate for non-lethal tumor types - that is, tumors that cannot cause death and do not alter the risk of death from other causes. In contrast, the logrank test is appropriate for rapidly lethal tumors (i.e., the time to
death following tumor onset is short). When applied to tumors of intermediate lethality, however, both tests can be biased. Peto et al. (1980) proposed a cause-of-death test combining the log-rank and Hoel-Walburg tests for tumors with intermediate lethality, however, this test requires the information of cause of death which is usually unavailable. To solve this issue, the authors developed a test incorporating a tumor lethality function with which the researchers can specify the probability that a tumor present at death is the cause of death. By replacing the unavailable cause-of-death information with the tumor lethality function, this test is more flexible than the cause-of-death test when sensitivity analysis on a plausible range of the tumor lethality function value is applied.

To illustrate this method, firstly let us see the four states commonly used to describe the sequence of events compromised by historical appearance of a tumor followed by a death for experimental animals in tumorigenicity studies (Kodell et al., 1982; McKnight and Crowley, 2001; Lagakos and Ryan, 1985). Each animal is tumor-free at the beginning (State 1) and then either develops a tumor (State 2) or dies. Animals that develop a tumor either die from the tumor (State 3) or die from other causes (State 4).

Let $U_1$ denote the time until the first event, no matter this event is just tumor onset, or death without tumor; let $d$ be the indicator of whether or not tumor onset occurs before death ($d = 1$: yes; $d = 0$: no); let $U$ denote the time to death (either natural death or terminal sacrifice); and let $c$ be the indicator of whether death is due to the tumor or to other causes ($c = 1$: tumor; $c = 0$: other causes). A Probabilistic model with the four states can be described with the following four intensity functions:
\[
\alpha(t) = \lim_{dt \to 0} \frac{\Pr\{U_1 < t + dt, d = 1 \mid U_1 \geq t\}}{dt},
\]
\[
\beta(t) = \lim_{dt \to 0} \frac{\Pr\{U_1 < t + dt, d = 0 \mid U_1 \geq t\}}{dt},
\]
\[
\gamma(t, x) = \lim_{dt \to 0} \frac{\Pr\{U < t + dt, c = 1 \mid U_1 = x, d = 1, U \geq t\}}{dt},
\]
\[
\lambda(t, x) = \lim_{dt \to 0} \frac{\Pr\{U < t + dt, c = 0 \mid U_1 = x, d = 1, U \geq t\}}{dt}
\]

Note that \(\alpha(t)\) and \(\beta(t)\) are the cause-specific hazard functions corresponding to tumor onset and death without tumor at time \(t\) respectively, while \(\gamma(t, x)\) and \(\lambda(t, x)\) are the cause-specific hazard functions for death from tumor and death from other causes at time \(t\) respectively, given tumor onset at \(x\). In comparing the control and exposed groups, \(\alpha(t)\) and \(\gamma(t, x)\) are regarded as appropriate bases for assessing tumorigenicity, where \(\beta(t)\) and \(\lambda(t, x)\) are often regarded as nuisance parameters.

For occult tumors, the time of tumor onset cannot be directly observed. Instead, each observation only contains time to death (\(U\)) and an indicator \(d\) on tumor presence/absence at death. i.e., we have current status data. In the cases of rapidly lethal tumor and non-lethal tumor, we can use the log-rank and Hoel-Walburg tests respectively to assess the carcinogenic potential of a substance under the null hypothesis that \(\alpha_{\text{exposed}}(t) = \alpha_{\text{controls}}(t)\) for current status data. And note that, for non-lethal tumors, the comparison of \(\alpha_{\text{exposed}}(t) = \alpha_{\text{controls}}(t)\) is mathematically equivalent to a comparison of \(\pi_{\text{exposed}}(t) = \pi_{\text{controls}}(t)\), where \(\pi(t)\) is a prevalence function measuring the probability that an animal living at time \(t\) has a tumor (McKnight and Crowley, 2001).

With animal study as an example, let \(w\) denote the group (\(w = 0\) for controls and \(w = 1\) for exposed) and let \((U_i, d_i, w_i)\) indicate the values of \((U, d, w)\) for the \(i\)th of
n rodents. For the distinct times of death $t_1, t_2, \ldots, t_k$, both the log-rank and Hoel-Walburg tests have a similar form of test statistic, by computing and then summing an observed minus expected tumor count for the exposed group (i.e., $w = 1$) at each time of death, i.e., $(O - E) / \sqrt{V}$, with the definitions of $O$, $E$ and $V$ given as

$$O = \sum_{j=1}^{k} n_{1j}, \quad E = \sum_{j=1}^{k} (n_{0j} + n_{1j}) N_{1j} / (N_{0j} + N_{1j}) \quad \text{and} \quad V = \sum_{j=1}^{k} \frac{n_j(N_j - n_j)N_{0j}N_{1j}}{N_j^2(N_j - 1)}$$

(2.34)

where $n_{wj}$ are defined in both tests as the number of deaths with a tumor in group $w$ at time $t_j$ for $w = 0, 1$ and $j = 1, 2, \ldots, k$:

$$n_{wj} = \sum_{i=1}^{n} I(U_i = t_j, w_i = w) d_i$$

(2.35)

And $N_{wj}$ is defined in log-rank test as the number at risk in group $w$ at time $t_j$:

$$N_{wj} = \sum_{i=1}^{n} I(U_i \geq t_j, w_i = w)$$

(2.36)

while in Hoel-Walburg test, $N_{wj}$ is defined as the number of deaths in group $w$ at time $t_j$:

$$N_{wj} = \sum_{i=1}^{n} I(U_i = t_j, w_i = w)$$

(2.37)

Under the null hypotheses, both statistics are approximately $N(0, 1)$. In practice, the $N_{wj}$ are often small in Hoel-Walburg test so it is common to group deaths into intervals rather than to compute an $O_j$ and $E_j$ at each distinct death time.

For tumors with intermediate lethality, neither log-rank nor Hoel-Walburg tests should be used otherwise the results would be biased. Peto et al. (1980) proposed to combine the log-rank and Hoel-Walburg tests as the cause-of-death test, which, however, requires that the cause-of-death information ($c_i$) is available for each animal with a tumor.
If we distinguish the terms with the superscripts $L$ and $H$ corresponding to logrank test and Hoel-Walburg test respectively, the cause-of-death test is based on the statistic

$$Z = \frac{O^L + O^H - E^L - E^H}{\sqrt{(V^L + V^H)}}$$

where $O^L$, $E^L$, $V^L$, $O^H$, $E^H$, and $V^H$ are defined in the same form as in 2.34, but with the corresponding $n^L_{wj}$, $N^L_{wj}$, $n^H_{wj}$, and $N^H_{wj}$ defined as

$$n^L_{wj} = \sum_{i=1}^{n} I(U_i = t_j, w_i = w)c_i$$

$$N^L_{wj} = \sum_{i=1}^{n} I(U_i \geq t_j, w_i = w)$$

$$n^H_{wj} = \sum_{i=1}^{n} I(U_i = t_j, w_i = w)(1 - c_i)d_i$$

$$N^H_{wj} = \sum_{i=1}^{n} I(U_i \geq t_j, w_i = w)(1 - c_i)$$

Since the cause of death information $(c_i)$ is usually unavailable, the use of the cause-of-death test is restricted. In order to analyze the intermediate tumor data lacking of the cause of death information $(c_i)$, Lagakos and Louis (1988) proposed an incomplete data score test, which is similar in form to the cause-of-death test, but with the (unknown) $c_i$ replaced in equations in (2.38) by

$$L(u_i, w_i) = E(c_i | u_i, w_i, d_i = 1)$$

$L(u, w)$ is referred to as the tumor lethality function since it gives the probability that tumor present at death is the cause of death. In most applications, the lethality function will not be known, however, the test can be evaluated for a range of plausible lethality values in a sensitivity analysis. And this will indicate how strongly the results depend on the assumed lethality and whether an unequivocal message emerges from the available data.
2.3.3 Accounting for Informative Censoring Through a Relaxed Lethality Assumption

Also motivated by tumorigenicity studies, Dunson and Dinse (2002) proposed a Bayesian regression model for current status data with informative censoring. The authors defined a binary parameter $C$ to indicate whether the cause of death is with respect to tumor or not ($C = 1$: the death is caused directly or indirectly by one or more tumors; $C = 0$: the death is not related to any tumors). When $C = 1$, the distribution of the censoring time by natural death is affected by the tumor onset event, introducing informative censoring. Note that, for those animals with a tumor present and died naturally, we may not know whether they died of tumor, or of other non-tumor causes; hence the value of $C$ may not be observed. However, the value of $C$ can be augmented during the Bayesian estimation procedure in this model for each subject even when the information of $C$ is not observed.

The model is established on a discrete time scale. Suppose that there are $J$ distinct death times with ordered values $t_1 < \cdots < t_J$ and set $t_0 = 0, t_{J+1} = \infty$, and $I_j = (t_{j-1}, t_j](j = 1, \cdots , J+1)$. The primary interest is the probability mass function of the tumor onset time on this discrete time scale: $f(t_j), j = 1, \ldots, J$ with $F(t_j)$ as its cumulative distribution function.

Let $t$ denote the death time, $\delta$ denote the type of death for each subject ($\delta = 0$: random or terminal sacrifice; $\delta = 1$: natural death, including moribund sacrifice), $\pi_j$ denote the probability of terminal sacrifice at time $t_j$, $\beta_j$ denote the probability of natural death at time $t_j$ by causes other than tumor, $\gamma_j$ denote the probability of natural death at time $t_j$ by a cause related to tumor, $d$ denote the current status of tumor at death ($d = 0$: tumor not present at death; $d = 1$: tumor present at death),
and $\lambda_j$ denote the probability of developing a tumor during the $j$th time interval $I_j$, conditional on being free of tumor before time $t_j$.

Next let’s introduce the construction of the likelihood function of this model, by detailing the likelihood contributions of the four different types of observations:

1. Subject being terminally sacrificed at $t$ without tumor present ($\delta=0, d=0$).

2. Subject died naturally at $t$ without tumor present ($\delta=1, d=0$).

3. Subject being terminally sacrificed at $t$ with tumor present ($\delta=0, d=1$).

4. Subject died naturally at $t$ with tumor present ($\delta=1, d=1$).

In the first scenario of being terminally sacrificed at $t$ without tumor present ($\delta=0, d=0$), the following events are experienced simultaneously:

- the event of being free of tumor at or before time $t$, which occurs with probability $\prod_{j=1}^{t} \{1 - \lambda_j\}$.

- the event of not dying naturally due to a non-tumor related factor at or before time $t$, which occurs with probability $\prod_{j=1}^{t} (1 - \beta_j)$.

- the event of not being terminally sacrificed at or before time $t$, which occurs with probability $\prod_{j=1}^{t-1} (1 - \pi_j)$.

- the event of being terminally sacrificed instead of natural death at time $t$, which occurs with probability $\pi_t$. 


Hence the likelihood contribution for a subject being sacrificed at time $t$ without a tumor is:

$$L_{\{\delta=0,d=0\}} = \left[ \prod_{j=1}^{t} (1 - \lambda_j) \right] \times \left[ \prod_{j=1}^{t-1} (1 - \beta_j) \right] \times \left[ \prod_{j=1}^{t-1} (1 - \pi_j) \right] \times [\pi_t]$$

$$= [1 - F(t)] \times \left[ \prod_{j=1}^{t-1} (1 - \beta_j) \right] \times \left[ \prod_{j=1}^{t-1} (1 - \pi_j) \right] \times [\pi_t] \quad (2.39)$$

In the second scenario of natural death at $t$ without tumor present ($\delta=1, d=0$), the following events are experienced simultaneously:

- the event of being free of tumor at or before time $t$, which occurs with probability $\prod_{j=1}^{t} (1 - \lambda_j)$.

- the event of not dying naturally due to a non-tumor related factor before time $t$, which occurs with probability $\prod_{j=1}^{t-1} (1 - \beta_j)$.

- the event of not being terminally sacrificed before time $t$, which occurs with probability $\prod_{j=1}^{t-1} (1 - \pi_j)$.

- the event of natural death without a tumor at time $t$, which occurs with probability $\beta_t$

Hence the likelihood contribution for a subject dying naturally at time $t$ without a tumor is:

$$L_{\{\delta=1,d=0\}} = \left[ \prod_{j=1}^{t} (1 - \lambda_j) \right] \times \left[ \prod_{j=1}^{t-1} (1 - \beta_j) \right] \times \left[ \prod_{j=1}^{t-1} (1 - \pi_j) \right] \times \beta_t$$

$$= [1 - F(t)] \times \left[ \prod_{j=1}^{t-1} (1 - \beta_j) \right] \times \left[ \prod_{j=1}^{t-1} (1 - \pi_j) \right] \times \beta_t \quad (2.40)$$

In the third scenario of being terminally sacrificed at $t$ with tumor present ($\delta=0, d=1$), the following events are experienced simultaneously:
• the event of having developed a tumor, but did not die naturally at or before time $t$, which occurs with the probability
\[
\sum_{j=1}^{t} \left( \prod_{m=1}^{j-1} (1 - \lambda_m)(1 - \beta_m) \right) \times \lambda_j \times \left[ \prod_{m=j}^{t} (1 - \beta_m - \gamma_m) \right]
\]
where $\prod_{m=1}^{j-1} (1 - \lambda_m)(1 - \beta_m)$ is the probability of neither developing tumor nor dying due to factors other than tumor at or before time $j - 1$, $\lambda_j$ is the hazard of developing tumor at time $j$, and $\prod_{m=j}^{t} (1 - \beta_m - \gamma_m)$ is the hazard of not dying naturally (i.e., not dying naturally due to factors either related or unrelated to tumor) from time $j$ to time $t$.

• the event of not being terminally sacrificed before time $t$, which occurs with the probability $\prod_{m=1}^{t-1} (1 - \pi_m)$.

• the event of being terminally sacrificed at time $t$, which occurs with the probability $\pi_t$.

Hence the likelihood contribution for a subject being terminal sacrificed at $t$ with tumor present is:
\[
L_{\{\delta=0,d=1\}} = \left\{ \sum_{j=1}^{t} \left[ \prod_{m=1}^{j-1} (1 - \lambda_m)(1 - \beta_m) \right] \times \lambda_j \times \left[ \prod_{m=j}^{t} (1 - \beta_m - \gamma_m) \right] \right\} \times \left\{ \prod_{m=1}^{t-1} (1 - \pi_m) \right\} \times \pi_t
\]

By defining $\rho_j = \gamma_j/(1 - \beta_j)$, which measures the probability of death at time $j$ due to tumor given that death from non-tumor causes didn’t occur at time $j$, or in other words, the lethality of a tumor at time $j$, the above likelihood contribution can be
written as

\[
L_{(\delta=0,d=1)} = \left[ \prod_{m=1}^{t} (1 - \beta_m) \right] \times \left\{ \prod_{m=1}^{t-1} (1 - \pi_m) \right\} \times \pi_t \\
\times \left\{ \sum_{j=1}^{t} \left[ f(j) \times \left[ \prod_{m=j}^{t} (1 - \rho_m) \right] \right] \right\} 
\]

(2.41)

In the fourth scenario of natural death at \( t \) with tumor present \((\delta=1, d=1)\), the following events are experienced simultaneously:

- the event of having developed a tumor at or before time \( t \) and having not died naturally before time \( t \), which occurs with the probability

\[
\sum_{j=1}^{t} \left[ \prod_{m=1}^{j-1} (1 - \lambda_m)(1 - \beta_m) \right] \times \lambda_j(b) \times \left[ \prod_{m=j}^{t-1} (1 - \beta_m - \gamma_m) \right]
\]

where \( \prod_{m=1}^{j-1} (1 - \lambda_m)(1 - \beta_m) \) is the probability of neither developing a tumor nor dying from non-tumor causes at or before time \( j - 1 \), \( \lambda_j \) is the probability of developing tumor at time \( j \), and \( \prod_{m=j}^{t-1} (1 - \beta_m - \gamma_m) \) is the probability of not dying naturally (i.e., not dying from factors either related or unrelated to tumor) from time \( j \) to time \( t - 1 \).

- the event of not being terminal sacrificed before time \( t \), which occurs with the probability \( \prod_{m=1}^{t-1} (1 - \pi_m) \).

- the event of dying naturally at time \( t \), which occurs with the probability \( \beta_t + \gamma_t \), where \( \beta_t \) is the probability of dying naturally at time \( t \) due to factors unrelated to tumor and \( \gamma_t \) is the probability of dying naturally at time \( t \) due to factors related to tumor.
Hence the likelihood contribution for a subject dying naturally at $t$ with tumor present is:

$$L_{\{\delta=1,d=1\}} = \left\{ \sum_{j=1}^{t} \left\{ \prod_{m=1}^{j-1} (1 - \lambda_m(b))(1 - \beta_m) \right\} \times \lambda_{j1}(b) \times \left[ \prod_{m=j}^{t-1} (1 - \beta_m - \gamma_{m1}) \right] \right\}$$

$$\times \left\{ \prod_{m=1}^{t-1} (1 - \pi_m) \right\} \times (\beta_t + \gamma_{t1})$$

By using $\rho_j = \frac{\gamma_j}{1 - \beta_j}$, the above likelihood contribution can be written as:

$$L_{\{\delta=1,d=1\}} = \left[ \prod_{m=1}^{t-1} (1 - \beta_m) \right] \times \left\{ \prod_{m=1}^{t-1} (1 - \pi_m) \right\} \times \left[ 1 - (1 - \beta_t) \times (1 - \rho_t) \right]$$

$$\times \left\{ \sum_{j=1}^{t} \left\{ f(j) \times \left[ \prod_{m=j}^{t-1} (1 - \rho_m) \right] \right\} \right\}$$

(2.42)

By utilizing the tumor indicator $d$, the two likelihood contribution functions, (2.39) and (2.41), can be combined to one as the likelihood contribution for subject being terminally sacrificed at $t$ with tumor 1 present ($d = 1$) or without ($d = 0$):

$$L_{\{\delta=0\}} = L_{\{\delta=0,d=1\}} \times L_{\{\delta=0,d=0\}}$$

(2.43)

$$= \left[ \sum_{j=1}^{t} f(j) \left( \prod_{m=j}^{t} (1 - \rho_m) \right) \right] \times \frac{1}{d} \left[ 1 - F(t) \right]^{1-d}$$

$$\times \left[ \prod_{j=1}^{t} (1 - \beta_j) \right] \times \prod_{j=1}^{t-1} (1 - \pi_j) \times \pi_t$$

And by combining 2.40 and 2.42, we can get the likelihood contribution for a subject dying naturally at $t$ with tumor present ($d = 1$) or without ($d = 0$):
\[ L_{\delta=1} = L_{\delta=1,d=1} \times L_{\delta=1,d=0} \]  \hspace{1cm} (2.44)

\[ = \left[ \sum_{j=1}^{t} f(j \mid b) \left( \prod_{m=j}^{t-1} (1 - \rho_m) \right) \right]^d \times \left[ 1 - F(t) \right]^{1-d} \times \left[ \prod_{j=1}^{t-1} (1 - \beta_j) \right] \times \left[ \prod_{j=1}^{t-1} (1 - \pi_j) \right] \times \left[ 1 - (1 - \beta_t) \times (1 - \rho_t)^{1-d} \right] \]

And 2.43 and 2.44 can be further combined to:

\[ L = \left( \left[ \sum_{j=1}^{t} f(j) \left( \prod_{m=j}^{t-1} (1 - \rho_m) \right) \right]^d \times \left[ 1 - F(t) \right]^{1-d} \right) \times \left[ \prod_{j=1}^{t-1} (1 - \beta_j) \right] \times \left[ \prod_{j=1}^{t-1} (1 - \pi_j) \right] \times \left[ 1 - (1 - \beta_t) \prod_{k=1}^{K} (1 - \rho_t^k) \right]^{\delta} \pi_t^{1-\delta} \]  \hspace{1cm} (2.45)

Covariates can be included in the model for the distribution of tumor onset time and the distribution of time to natural death due to tumor-independent causes as follows:

\[ F(j \mid x_i) = \Phi(\alpha_j - x_i^T \psi), \; j = 1, \ldots, J \]  \hspace{1cm} (2.46)

\[ G(j \mid x_i) = \Phi(w_j - x_i^T \tau), \; j = 1, \ldots, J \]

where \( F(\cdot \mid x_i) \) is the c.d.f. of the onset time of a tumor and \( G(\cdot \mid x_i) \) is the c.d.f. of the time to natural death due to tumor-independent causes given the covariate vector \( x_i \). Note that \( \alpha_j \)'s and \( w_j \)'s characterize the baseline cumulative distribution functions of tumor onset time and natural death time due to tumor-independent causes, respectively (these parameters are to be estimated); \( \psi \) and \( \tau \) are the regression coefficient vectors.
The authors also extended their model for analyzing data with \( k (k = 1, \ldots, K) \) tumor types, by incorporating a subject-specific random effect, \( b \), which is shared among the onset time distributions of multiple tumors, to account for their correlations. Note that \( b \sim Q(\phi) \), where \( \phi \) is a parameter characterizing \( Q \). For tumor type \( k (k = 1, \ldots, K) \), let \( V_k = j \) if \( I_j \) contains the unknown age at tumor onset and let \( \Delta_k = I(V_k \leq T) \) indicate the current status at death. It is assumed that given \( b \), the onset time distributions of different tumors are independent, i.e.,

\[
F(v_1, \ldots, v_K | b) = F_1(v_1 | b) \cdots F_K(v_K | b).\]

The onset time distribution of tumor \( k \) is hence incorporating covariates as follows:

\[
F_k(j | x_i, b_i) = \Phi(\alpha_{jk} - x_i^T \psi_k - b_i), \quad \text{for } j = 1, \ldots, J; k = 1, \ldots, K, \quad (2.47)
\]

And the likelihood contribution for subject \( i \) in (2.45) is extended to:

\[
L^{(K)} = \left( \prod_{k=1}^{K} \sum_{j=1}^{t} f_k(j | b) \left[ \prod_{m=j}^{t-\delta} (1 - \rho_{mk}) \right]^{\Delta_k} [1 - F_k(t | b)]^{1-\Delta_k} \right) 
\times \left[ \prod_{j=1}^{t-\delta} (1 - \beta_j) \right] \left[ \prod_{j=1}^{t-1} (1 - \pi_j) \right] 
\times \left[ 1 - (1 - \beta_t) \prod_{k=1}^{K} (1 - \rho_{tk})^{\Delta_k} \right]^{\delta} \pi_t^{1-\delta}, \quad (2.48)
\]

To simplify the methods for posterior inference, the authors augmented the observed data \( Y_i = (T_i, \delta_i, d_{i1}, \ldots, d_{ik})^T \) with two latent variables:

1. \( V_{ik} \): the onset time of tumor \( k \) for the \( i \)th subject.

2. \( C_i \): the cause of death for the \( i \)th subject.

Also, to ensure estimability in Bayesian computation, the authors assume the lethality parameter for the \( k \)th tumor \( \rho_k \) is constant across time and subjects.
2.3.4 Accounting for Informative Censoring Through a Copula Function

Zhao et al. (2015) proposed an additive hazards model for current status data with informative censoring, which can be considered as a modification of the model by Zhang et al. (2005). Instead of using shared random effects as in Zhang et al. (2005), Zhao et al. (2015) used a “copula” function to account for the correlation between the event time and the censoring time.

Consider a failure time study that consists of $n$ independent subjects. For subject $i$, let $T_i$ denote the event time and $Z_i$ denote a $p$-dimensional vector of covariates. The authors assumed an additive hazard model for $T$:

$$h_1(t \mid Z) = h_{10}(t) + Z'\beta$$  \hspace{1cm} (2.49)$$

where $h_{10}(t)$ is an unspecified baseline function and $\beta$ is the vector of regression parameters. Let $U_i$ denote the censoring time on subject $i$. The authors assumed a Cox proportional hazards model for $U_i$:

$$h_2(u \mid Z) = h_{20}(u) \exp(Z'\gamma)$$  \hspace{1cm} (2.50)$$

where $h_{20}(c)$ denotes an unspecified baseline hazard function and $\gamma$ a vector of regression parameters.

Let $F_T$ and $F_U$ denote the marginal distribution functions of the $T_i$'s and $U_i$'s given covariates, respectively, and $F$ their joint distribution. It follows from the Theorem 2.3.3 of Nelsen (2006) that there exists a so called copula function $M_\alpha(w, v)$ defined on $I^2 = [0, 1] \times [0, 1]$ with $M_\alpha(w, 0) = M_\alpha(0, v) = 0$, $M_\alpha(w, 1) = w$ and $M_\alpha(1, v) = v$ such that

$$F(t, u) = M_\alpha[F_T(t), F_U(u)]$$
The parameter $\alpha$ stands for the association between $T_i$’s and $U_i$’s. It can also be shown that
\[
\Pr(T \leq t \mid U = u, \mathbf{Z}) = \left. \frac{\partial M_\alpha(w, v)}{\partial v} \right|_{w=F_T(t), v=F_U(u)}
\]
which is denoted by $m_\alpha(F_T(t), F_U(u))$. The copula function allows one to model the correlation and the marginal distribution separately. However, it has been pointed out that given the copula function, the association parameter $\alpha$ is generally not identifiable without prior or extra information. The authors assumed that both the copula function and $\alpha$ are known and focused on estimation of regression parameters $\beta$ and $\gamma$.

Define $H_{10}(t) = \int_0^t h_{10}(s)ds$, $H_{20}(u) = \int_0^u h_{20}(s)ds$, and $\theta = (\beta, \gamma; H_1, H_2)$. Let $f_U$ denote the marginal density function of the $U_i$’s given covariates. Then it can be shown that
\[
F_T(t) = 1 - \exp[-H_{10}(t) - \mathbf{Z}' \beta t]
\]
\[
F_U(u) = 1 - \exp[-H_{20}(u) \exp(\gamma' \mathbf{Z})]
\]
\[
f_U(u) = \exp[-H_{20}(u) \exp(\gamma' \mathbf{Z})]h_{20}(u) \exp(\gamma' \mathbf{Z})
\]

Furthermore, the likelihood function is
\[
L(\theta) = \prod_{i=1}^n \left\{ m_\alpha(F_T(u_i), F_U(u_i))F_U(u_i)^{d_i} \left[ 1 - m_\alpha(F_T(u_i), F_U(u_i)) \right] f_U(u_i) \right\}^{1-d_i}.
\]

By maximizing the likelihood function on a so called sieve space defined on $M(\cdot)$, the estimation of the parameter $\theta$ can be obtained.

Zhao et al. (2015) didn’t measure the correlation between the event time and censoring time by random effects which restricts the type of correlation, but by a copula function $M$ with an association parameter $\alpha$. Since both $M$ and $\alpha$ are not
identifiable in general without prior or extra information, they have to be assumed known. The authors gave two copula models as options. The first copula model is referred to as the FGM model:

\[ M_\alpha(w,v) = wv + \alpha wv(1 - w)(1 - v), \quad -1 \leq \alpha \leq 1. \] (2.54)

The second copula model is referred to as the Frank model:

\[ M_\alpha(w,v) = \log_\alpha [1 + (\alpha^w - 1)(\alpha^v - 1)], \quad \alpha > 0, \alpha \neq 1 \] (2.55)

It is apparent that the assumption of \( M \) and \( \alpha \) is not true in practice. To address this issue, the authors recommended a sensitivity analysis on different copula models and association levels based on AIC or other criteria.

### 2.4 Conditions Under Which Informative Censoring Should Be Considered

Betensky (2000) explored the conditions when the informative censoring mechanism is required for unbiased parameter estimation with current status data, extending the work of Williams and Lagkkos (1977) in the context of right censored data. Let \( n \) denote the number of subjects and let \( i \) index the subjects. Let \( T_i \) and \( U_i \) denote the \( i \)th subject’s event time and examination time, respectively. Let \( d_i = 1 \) if \( T_i \leq U_i \) and \( d_i = 0 \) if \( T_i > U_i \). The observed data are \( \{(U_i,d_i), i = 1, \ldots, n\} \).

It is assumed here that \( T \) and \( U \) are continuous random variables and \( T > 0 \) and \( 0 < U \leq u^* \) with probability one, where \( u^* \leq \infty \). The results extend also to discrete random variables. It is of primary interest to estimate \( F(t) = P(T \leq t) \). In order to write the full likelihood of the data in terms of \( F \), they defined the quantities

\[ a(u) = P(U \in du \mid T \leq u), \quad b(u) = P(U \in du \mid T > u), \]

36
where “$U \in du$” denotes the event that $U$ is contained in an infinitesimal interval about $u$. The full likelihood given the observed data is then

$$L = \prod_{i=1}^{n} F(U_i)^{d_i} \{1 - F(U_i)\}^{1 - d_i} a(U_i)^{d_i} b(U_i)^{1 - d_i}. $$

The constraints imposed on the triplet $\{F(\cdot), a(\cdot), b(\cdot)\}$ defining the model are that $F$ is a proper distribution function, $a$ and $b$ are nonnegative functions, and that every subject is examined by time $u^*$, that is

$$\int_{0}^{u^*} [a(u) F(u) + b(u) \{1 - F(u)\}] du = 1,$$

or equivalently, upon integration by parts,

$$\int_{0}^{u^*} \{A(u^*) - B(u^*) + B(u^*)/F(u^*) - A(u) + B(u)\} dF(u) = 1, \quad (2.56)$$

where $A(u) = \int_{0}^{u} a(x) dx$ and $B(x) = \int_{0}^{x} b(x) dx$.

The examination times are ignorable if and only if (2.56) does not constrain $B(\cdot)$;

$$A(u^*) - B(u^*) + B(u^*)/F(u^*) - A(u) + B(u) = 1/F(u^*) \quad (2.57)$$

for all $u \leq u^*$. This property is equivalent to the constant sum property, $a(u) - b(u) = 0$ for all $u \leq u^*$. The constant sum property allows the simpler likelihood,

$$L = \prod_{i=1}^{n} F(U_i)^{d_i} \{1 - F(U_i)\}^{1 - d_i},$$

to be used for estimation of $F$.

Models with this constant sum property, $S_2$, include independence models, $S_1$, and are included among the models, $S_3$, for which the conditional probability of occurrence of the event prior to the examination time given the examination time is nondecreasing in the examination time. This conditional probability is given by

$$a(u) F(u) / [a(u) F(u) + b(u) 1 - F(u)],$$
which is equal to $F(u)$, a nondecreasing function of $u$, when $a(u) = b(u)$. Thus $S_1 \subseteq S_2 \subseteq S_3$, and it can be shown that the subsets $S_3^C$, $S_3 \cap S_2^C$ and $S_2 \cap S_1^C$ are all nonempty. The property $S_2 \cap S_1^C$ being nonempty means that, independent censoring and some dependent censoring models with the constant sum property can ignore modeling of the censoring process and still obtain unbiased parameter estimates for the event process. As an example given by the authors, consider a model in which the event time $T$ takes the following four values with equal probability: \{1, 2, 5, 6\}; the value of censoring time $U$ depends on the value of the event time with the following rules: if $T \in (2, 5)$, then $U = 3$; if $T \in (1, 6)$, then $U = 4$. Now

- $a(3) = P(U = 3 \mid T \leq 3) = 1/2$
- $a(4) = P(U = 4 \mid T \leq 4) = 1/2$
- $b(3) = P(U = 3 \mid T > 3) = 1/2$
- $b(4) = P(U = 3 \mid T > 4) = 1/2$

Hence $a(u) = b(u)$ and this model belongs to $S_2$, but does not belong to $S_1$ since $U$ depends on $T$.

### 2.5 Summary

In this Chapter 2, we reviewed some typical models for analyzing current status data. Most of the existing models for current status data assume that censoring time is independent of event time. However, when the independent censoring is an unreasonable assumption for some current status data, the use of the independent censoring model on analyzing these data would produce in biased results.
A few authors have developed models for current status data accounting for dependent censoring, mostly motivated by tumorigenicity studies where censoring is caused by natural death which might depend on the tumor onset time. However, these methods require strong assumptions on tumor lethality. To address the limitations of the above methods, we propose Bayesian threshold regression models accounting for dependent censoring for current status data. In the next Chapter, we review some existing threshold regression models that are related to the models we propose.
Chapter 3: THRESHOLD REGRESSION

3.1 Threshold Regression for Right Censored Data

In many lifetime studies, it is natural to interpret and model the duration of time-to-event as following the sample path of a latent stochastic process, with an event occurring once the process reaches a boundary or threshold state. Such models have been appropriately called first hitting time (FHT) models. FHT models have been applied to data from many different fields, including medicine (e.g., Lee et al., 2000), environmental science (e.g., Lee et al., 2009), economics (e.g., Baxter and Rennie, 1997), engineering (e.g., Shaked and Shanthikumar, 1991), and sociology (e.g., Eaton and Whitmore, 1977). For example, FHT models may describe the failure time of an engineering system, the transition time for a stock price change and the length of marriage. FHT models are conceptually appealing in biomedical studies where stochastic processes often underlie different biomedical events, such as the decline of health toward death and the latent development of a tumor. The Wiener diffusion process, which was originally used as a suitable model for many physical processes that exhibit random variation over time, has also been used extensively in FHT models for biomedical data (e.g., Lee et al., 2004; Whitmore and Su, 2007; Balka et al., 2009; Yu
et al., 2009). A Wiener diffusion process \( Y(t) \) starting at \( y_0 \) with drift \( \mu \) and variance \( \sigma^2 \) has the following properties:

1. \( Y(t) \) has independent increments; for any non-overlapping time intervals \((t_1, t_2), (t_3, t_4)\), \( Y(t_2) - Y(t_1) \) and \( Y(t_4) - Y(t_3) \) are independent.

2. \( Y(t_2) - Y(t_1) \) is normally distributed with mean \( \mu(t_2 - t_1) \) and variance \( \sigma^2(t_2 - t_1) \) with \( t_1 < t_2 \).

The Wiener diffusion process is realistic in describing the mechanism of many latent processes encountered in biomedical studies, such as disease progression and health deterioration, in that the bidirectional movements of a Wiener process may well describe the fluctuation in health. Let \( Y(t) \) denote a latent health process with initial state \( Y(0) = y_0 > 0 \) and \( S \) (the event time) be the first time a sample path of the health status process reaches 0, i.e., \( S = \inf\{t : Y(t) = 0\} \). As seen in Cox and Miller (1965), \( S \) follows an inverse Gaussian distribution with probability density function (p.d.f.)

\[
f_T(t|\mu, \sigma^2, y_0) = \frac{y_0}{\sqrt{2\pi\sigma^2t^3}} \exp\left[-\frac{(y_0 + \mu t)^2}{2\sigma^2t}\right]
\]

\((t \geq 0, \sigma^2 > 0, y_0 > 0 \text{ and } -\infty < \mu < \infty)\) and cumulative distribution function (c.d.f.)

\[
F(t|\mu, \sigma^2, y_0) = \Phi\left[-\frac{(y_0 + \mu t)}{\sqrt{\sigma^2 t}}\right] + \exp\left(-\frac{2y_0\mu}{\sigma^2}\right) \Phi\left[\frac{\mu t - y_0}{\sqrt{\sigma^2 t}}\right],
\]

where \( \Phi(\cdot) \) is the c.d.f. of the standard normal distribution. Note that if \( \mu > 0 \), the Wiener process will never hit the boundary at zero with probability \( \lim_{t \to +\infty} [1 - F(t | \mu, \sigma^2, y_0)] = 1 - \exp(-2y_0\mu/\sigma^2) \) (Cox and Miller, 1965). Thus, the Wiener process
accommodates a cure fraction, which is present in some applications. For example, some people in the population are immune to infection since they are vaccinated, or some patients are cured by medical treatments and will not develop the disease again.

Through careful examination of Equations (3.1) and (3.2), it can be seen that both \( f(t|\mu, \sigma^2, y_0) \) and \( F(t|\mu, \sigma^2, y_0) \) depend on \( y_0/\sigma \) and \( \mu/\sigma \) only. Hence we need to fix one of three parameters (\( \mu, y_0, \sigma \)) to avoid over-parameterization. The initial health status (\( y_0 \)) and the rate with which the latent health status degrades (\( \mu \)) are more meaningful, clinically, than the variance parameter \( \sigma^2 \), thus we fix \( \sigma^2 \) to define the measurement scale of the latent process.

In an approach known as threshold regression (TR), covariates may be related to the latent health process through generalized link functions for the two process parameters \( \mu \) and \( \ln(y_0) \). Suppose that there are two covariate vectors \( X = (X_1, X_2, \ldots, X_k) \) and \( Z = (Z_1, Z_2, \ldots, Z_j) \) that are not necessarily identical and are linked to \( \mu \) and \( \ln(y_0) \) respectively with the following forms (Lee and Whitmore, 2006).

\[
\mu = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_k X_k = X'\alpha
\]  
(3.3)

\[
\ln(y_0) = \beta_0 + \beta_1 Z_1 + \cdots + \beta_j Z_j = Z'\beta
\]  
(3.4)

where \( \alpha = (\alpha_0, \ldots, \alpha_k)' \) and \( \beta = (\beta_0, \ldots, \beta_j)' \) are the regression coefficient vectors. Covariates in \( X \) have causal effects on degradation rate \( \mu \), while covariates in \( Z \) have effects on the level that the latent health process has advanced prior to the study (i.e., \( y_0 \)). Depending on the case of application, some covariates may only have causal effects and should only occur in \( X \), while others may have both types of effects and hence should occur in both \( X \) and \( Z \). In a randomized clinical trial, for example, treatment is unrelated to the initial health status \( y_0 \) at the time of randomization,
but could affect health degradation rate $\mu$ afterwards; while age could affect both $y_0$ and $\mu$ (Pennell et al., 2010). A negative coefficient for a covariate linked to $\mu$ means that subjects with larger values of this covariate have a more negative drift and thus their health deteriorates faster, while a negative coefficient for a covariate linked to $\ln(y_0)$ means that subjects with larger values of this covariate have a smaller value of $y_0$ and thus their initial health is worse. Note that $e^{\beta_j}$ can be interpreted as a multiplicative factor of the baseline proximity to the threshold which triggers the event corresponding to a unit increase of covariate $Z_j$: if $\beta_j$ is positive, a unit increase of covariate $Z_j$ lifts $y_0$ to be further from the threshold by the factor of $e^{\beta_j}$, and if $\beta_j$ is negative, a unit increase of covariate $Z_j$ decreases $y_0$ to be closer to the threshold by the factor of $e^{\beta_j}$.

Threshold regression model does not require the assumption of proportional hazards which is sometimes violated. However, the widely used Cox proportional hazards regression requires this assumption. When this assumption is violated, inferences by the Cox proportional hazards assumption should be made cautiously and threshold regression can be used alternatively. We use a leukemia remission study dataset (Garrett, 1997) as an example to illustrate the violation of proportional hazards assumption as follows, and compare the threshold regression with the Cox proportional hazards regression when the proportional hazards assumption is violated. This dataset consists of 42 patients who were monitored to see if they relapsed (relapse: $1 = \text{yes}$, $0 = \text{no}$) and how long (in weeks) they remained in remission (weeks). These 42 patients received two different treatments. For the first treatment, 21 patients received a new experimental drug (drug A), and the other 21 received a standard
drug (treatment1: 1 = drug A, 0 = standard). For the second treatment, 20 patients received a different drug (drug B), and the other 22 received a standard drug (treatment2: 1 = drug B, 0 = standard). White blood cell count, a strong indicator of the presence of leukemia, was recorded in three categories (1 = normal, 2 = moderate, 3 = high); we let \( \text{wbc2} \) indicate if white blood cell count is in category 2 and let \( \text{wbc3} \) indicate if white blood cell count is in category 3. A Stata post-estimation command `estat phtest` which is on the basis of Schoenfeld residuals, can be used to test the proportional hazards assumption after fitting a Cox model. By using this command, it turns out that treatment1 doesn’t violate the proportional hazards assumption after adjusting for \( \text{wbc2} \) and \( \text{wbc3} \), with p-value of this test for treatment1 equal to 0.6948, on the basis of a null hypothesis that the proportional hazards assumption is not violated. However, the p-value of this test for treatment2 is 0.0014, concluding that treatment2 violates the proportional hazards assumption after adjusting for \( \text{wbc2} \) and \( \text{wbc3} \) based on the data. This violation for treatment2 can also be demonstrated by the “log-log” plot and the nonparametric Kaplan-Meier survival estimate which are often used for assessment of the proportional hazards assumption. In Figure 3.1, the curves of -\ln{-\ln(survival)} versus \ln(analysis time) for both the drug B group and the standard group are plotted in a “log-log” plot. If the plotted lines are reasonably parallel, the proportional hazards assumption has not been violated. Since the two curves corresponding to the two groups cross each other in Figure 3.1, the proportional hazards assumption is clearly violated for treatment2. The proportional hazards violation is also suggested by Figure 3.2 where the Kaplan-Meier survival curves for the two groups also cross each other. In Figure 3.3, we overlay the Kaplan-Meier survival curves
and the Cox model predicted curves for treatment2. The Kaplan-Meier curves and Cox model predicted curves are not close to each other, suggesting a poor fit of the Cox model. Obviously this poor fit results from a violation of the proportional hazards assumption for treatment2. Note that in Figure 3.4, we also overlay the Kaplan-Meier survival curves and the predicted curves for treatment2 by threshold regression model, which is based on a Wiener process and does not assume proportional hazards. We can see that the Kaplan-Meier curves and the threshold regression predicted curves match very well. From this example, we can see the advantage of threshold regression model over Cox model when the proportional hazards assumption is violated.

Several early papers considering regression structures for FHT models include Whitmore (1983), Whitmore et al. (1998), Lee et al. (2000) and Lee et al. (2004). For example, Lee et al. (2000) used a threshold regression model based on a bivariate Wiener diffusion process to study progression to death in AIDS, with the drift parameters linked to treatment variables in the form of (3.3) and initial health status parameters of the parent processes linked to baseline covariates in the form of (3.4). Note that, although the variance parameter \( \sigma^2 \) can serve as one of the two free parameters to which covariates may be related, such models are uncommon since \( \sigma^2 \) is not as clinically meaningful as \( \mu \) and \( y_0 \). An example of such a model was proposed in Whitmore (1983), where the author fixed \( y_0 \), estimated \( \sigma^2 \) and linked covariates to \( \mu \). The author applied this model to lifetime data of aluminum-reduction cells installed at different dates, and lifetime data of motorettes operating under different temperatures.
Several extensions have been proposed to increase flexibility in modeling covariate effects. Yu et al. (2009) extended Whitmore (1983)’s model to incorporate an unspecified smooth function of a covariate into the linear link function for $\mu$. Yu et al.’s model is able to accommodate potentially nonlinear effects of covariates without specification of a pre-determined functional form and hence adds flexibility to the modeling structure. Li and Lee (2011) extended the work by Yu et al. (2009) and proposed a more generalized semi-parametric threshold regression model which is able to accommodate a vector of covariates, say, $X$, with their regression coefficients varying with another set of covariates. Li and Lee (2011)’s model reduces to Whitmore (1983)’s model when all regression coefficients are constant, and reduces to Yu et al.’s model when $X$ only contains an intercept. Lee et al. (2010) employed threshold regression under a Markov decomposition to study the effects of time-varying factors (e.g., cumulative pack-years of smoking to date) on the risk of lung cancer onset for female nurses.

3.2 Threshold Regression for Current Status Data

Whitmore et al. (1998) and Lee et al. (2000) proposed a frequentist threshold regression model for right censored and marker data. Tong et al. (2008) extended this approach to current status data and used their method to model time to tumor onset following exposure to a carcinogen. The authors considered a medical study which involves two stochastic processes, with the first process $Y(t)$ representing the latent health process of study subjects, and the second process $X(t)$ representing an observable marker process that is related to $Y(t)$. The authors used a bivariate Wiener diffusion process to model $\{Y(t), X(t)\}$; this means that at any time $t$, $\{Y(t), X(t)\}$
follows a bivariate normal distribution, with mean \( \mu = (Y(0), X(0))' + t(\mu_y, \mu_x)' \) and covariance matrix \( t\Sigma \), where

\[
\Sigma = \begin{pmatrix}
\sigma_{yy} & \sigma_{yx} \\
\sigma_{xy} & \sigma_{xx}
\end{pmatrix}
\]

Let \( \sigma_y = \sigma_{yy}^{1/2} \) and \( \sigma_x = \sigma_{xx}^{1/2} \), then the correlation between process \( Y(t) \) and process \( X(t) \) is

\[
\rho = \frac{\sigma_{yx}}{\sigma_y \sigma_x}.
\]

The closer that \(|\rho|\) is to 1, the closer the marker process \( X(t) \) emulates the latent health process \( Y(t) \).

For the latent health process \( Y(t) \), the authors assumed that \( Y(0) = \delta \) which is an unknown positive number, and the time of event \( S \) is the time when \( Y(t) \) degrades to level 0 for the first time. For the marker process \( X(t) \), they assumed \( X(0) = 0 \). The distribution of the time of event corresponding to the latent health process \( Y(t) \) is in our interest.

For current status data with the marker information obtained at the examination time \( t_i \) for subject \( i \), the observed data are \( \{t_i, d_i, x_i\} \), where \( d_i = I(S_i \leq t_i) \) and \( x_i = X_i(t_i) \) which is the value of the marker process for subject \( i \) at time \( t_i \). Let \( \mu_y \) and \( \mu_x \) denote the drifts of \( Y(t) \) and \( X(t) \) respectively, and let \( \phi(\cdot) \) and \( \Phi(\cdot) \) denote the probability density and distribution functions of the standard normal random variable, respectively. Let \( p_s(x) \) be the probability density function of \( X(t) \) given survival beyond time \( t \), and \( p_f(x) \) be its probability density function given failure before time \( t \). We define \( c_1 \) and \( c_2 \) as

\[
c_1 = c_1(t) = \frac{\delta + \mu_{y,x}(t)}{\sigma_y \sqrt{(1 - \rho^2) t}}
\]

\[
c_2 = c_2(t) = \frac{\delta + \mu_{y,*}(t)}{\sigma_y \sqrt{(1 - \rho^2) t}}
\]
with
\[ \mu_{y,x}(t) = \mu_y t + \rho \sigma_y \sigma_x^{-1} (x - \mu_x t) \text{ and } \mu_{y,x}(t) = \mu_{y,x}(t) - 2\delta (1 - \rho^2). \]

Following Lee et al. (2000), it can be shown that
\[
p_s(x) = \lim_{h \to 0} \frac{1}{h} \Pr\{X(t) \in (x, x+h), S > t\} \\
= \Phi(c_1)q_1(x) - \exp \left[-\frac{2\delta \mu_y}{\sigma_y^2}\right] \Phi(c_2)q_2(x)
\]
and
\[
p_f(x) = \lim_{h \to 0} \frac{1}{h} \Pr\{X(t) \in (x, x+h), S \leq t\} = \lim_{h \to 0} \frac{1}{h} \Pr\{X(t) \in (x, x+h)\} - \lim_{h \to 0} \frac{1}{h} \Pr\{X(t) \in (x, x+h), S > t\} \\
= q_1(x) - p_s(x)
\]

where \( q_1(x) \) and \( q_2(x) \) are probability density functions of normal distributions with means \( \mu_x t \) and \( \mu_x t - 2\delta \rho \sigma_x / \sigma_y \) respectively, with the same variance \( \sigma_x^2(t) \).

Thus the likelihood function of this bivariate model with parameter vector \( \theta = (\delta, \mu_y, \rho, \mu_x, \sigma_x, \sigma_y) \) for a set of data \( \{(t_i, d_i, x_i), i = 1, \ldots, n\} \) is
\[
L(\theta) = \prod_{i=1}^{n} [p_s(x_i)]^{1-\Delta_i} [p_f(x_i)]^{\Delta_i}
\]
The authors obtained estimates of regression coefficients by maximizing this likelihood function.

### 3.3 Bayesian Approaches

Most of the papers listed above formulated their inference procedure using frequentist approaches. Bayesian approaches have some advantages over frequentist approaches, such as the ability to incorporate historical data and the ability to perform exact inferences through the use of Markov chain Monte Carlo (MCMC) methods.
Pettit and Young (1999) proposed a Bayesian FHT model based on a Wiener diffusion process. The method allowed inferences to be made for both the unknown threshold level and the parameters of the degradation process. They used a Gibbs sampler to sample from the posterior distributions of model parameters and predictive distributions of failure times. The authors assumed the final state of the process was observable. This assumption does not hold for most biomedical data for which the underlying stochastic processes are usually latent and hence so are their final states. However, there still exist some applications with observable degradation. For example, Lu and Meeker (1993) modeled the failure time of metals using the size of a fatigue crack as an observable measure of degradation for which a crack length of 1.6 inches was defined as the “failure” level. Tseng and Yu (1997) also considered a lifetime study of light emitting diodes (LED) whose measurable standardized light intensities degrade. The lifetime of an LED is the time when its standardized light intensity hits a fixed constant, say, 50%. Since they assumed an observable final state, Pettit and Young (1999) used the p.d.f. of the degradation level at the right censoring time given the subject survived beyond this time point as the likelihood contribution for a surviving subject. In simulation studies, the authors found that the inferences were more accurate if they took into account the degradation information of the surviving subjects rather than simply treating them as right censored observations.

More recently, Shubina et al. (2005) extended the Bayesian model by Pettit and Young (1999) to incorporate observable marker data and accommodated censored data by augmenting the process degradation level for the right censored subjects through a data augmentation technique. Saebo and Almoy (2004) proposed a
Bayesian FHT model for competing risks data also based on the Wiener diffusion process and Saebo et al. (2005) extended the model to account for genetic relationships by incorporating random effects in the drift parameters of the underlying Wiener diffusion processes.

Pennell et al. (2010) proposed a random-effects Bayesian threshold regression model which incorporated random effects not only on the drift parameter, but also on the initial state parameter of the underlying Wiener diffusion process. In this model, the drift was modeled as a linear function of a random subject-specific intercept and covariates

$$\mu_i = b_i + z_i' \beta, \quad (3.5)$$

where $b_i \sim N(0, \lambda^{-1})$ and $z_i$ and $\beta$ are, respectively, the covariate vector for the $i$th subject and the corresponding regression coefficient vector. The initial state parameter of the $i$th subject $y_{0i}$ was modeled as

$$y_{0i} \sim N_+(u_i' \alpha, \tau^{-1}),$$

where $u_i$ and $\alpha$ are respectively the covariate vector for the $i$th subject and the regression coefficient vector that is linked to the location parameter of $N_+(\cdot)$, which is a normal distribution truncated to the left at 0.

In their competing risks models, Saebo and Almoy (2004) and Saebo et al. (2005) used the survival function involving the cumulative distribution function of a normal distribution as the likelihood contribution for right censored subjects. This likelihood resulted in non-closed forms of the full conditional posterior distributions for the regression parameters and hence Metropolis-Hastings (M-H) algorithm was required for posterior computation. However, since the M-H algorithm does not generate draws
from the full conditional posterior at each iteration of the MCMC, convergence may be slow when imbedded in a Gibbs sampling algorithm. Pennell et al. (2010) avoided these issues by using a modification of the data augmentation technique by Shubina et al. (2005). Pennell et al.'s method involved sampling the value of the $i$th latent process level at right censoring time $u_i$ conditioning on the fact that this subject survived beyond $u_i$ (i.e., the death time $s_i$ is greater than $u_i$), which has p.d.f.

$$f_0(y_i \mid \mu, \sigma^2, y_0, s_i > u_i) = \frac{1}{\sqrt{2\pi \sigma^2 u_i}} \exp\left[-\frac{(y_i - y_0 - \mu u_i)^2}{2\sigma^2 u_i}\right] \times \left[1 - \exp\left(-\frac{-2y_i y_0}{\sigma^2 u_i}\right)\right] I(y_i > 0)$$

(3.6)

The derivation of (3.6) can be found in Cox and Miller (1965) and Lu (1995). The authors used rejection sampling to sample from this distribution and derived an efficient covering function. Inference for regression parameters were obtained using a hybrid MCMC methodology consisting of Gibbs and rejection sampling steps. Hence at each iteration of MCMC, draws were generated from the full conditional posteriors of parameters and the convergence of the hybrid MCMC is fast.

### 3.4 Threshold Regression with Operational Time Scale

In applications of threshold regression, the natural time scale of the underlying process does not need to be calendar time. For example, it may be better to track the lifetime of a car by its usage time, not by the calendar time. In a biomedical context, individuals who are exposed to different conditions over varying calendar time intervals will have different rates of disease progression, depending on the disease risk and health stress to which the subject is exposed during that interval. For example, a worker’s lung health may deteriorate at different rates over different time periods.
because of different levels of occupational exposure to a carcinogen such as diesel exhaust.

A practical transformation from calendar time to operational time is through the following linear transformation (Lee et al., 2004):

\[ r = R(t) = \sum_{k=1}^{K} \alpha_k T_k(t), \quad \text{where} \quad \sum_{k=1}^{K} T_k(t) = t. \quad (3.7) \]

Above we use notation \( r \) and \( t \) to denote operational time (or running time) and calendar time (or clock time) respectively, and \( R(t) \) to denote the transformation of calendar time \( t \) to operational time \( r \). The transformation 3.7 assumes that a subject is exposed to up to \( K \) different conditions over a calendar time interval \([0, t]\), and the conditions are respectively indexed by \( k = 1, \ldots, K \). Parameter \( \alpha_k \) stands for the rate at which the operational time is to advance, \( T_k(t) \) stands for the calendar time during \([0, t]\) that the subject is exposed to condition \( k \), and \( \sum_{k=1}^{K} T_k(t) = t \).

There have been at least two applications of operational time scale threshold regression models to biomedical data. Lee et al. (2004) and Lee et al. (2009) utilized operational time to assess the lung cancer risk and mortality of railroad workers. Lengths of each railroad worker’s stay in different job categories (such as engineer-brakemen, shop worker, and other worker) were used as the \( T_k(t) \) in (3.7) with each \( \alpha_k \) interpreted as the ratio of the rate at which the operational time progresses while in job category \( k \) relative to that while in the reference job category which was retirement. Lee et al. (2009) compared the Kaplan-Meier survival estimates calculated using calendar time scale and operational time scale and found that the estimates over the operational time scale had a clearer differentiation of survival curves between levels of important predictors than the estimates over calendar time scale. The authors concluded that this was due to the ability of the operational time scale
to account for differences in time spent in job categories with varying levels of exposure to diesel exhaust. Lee et al. (2008) also utilized an operational time scale to differentiate between the rate of disease progression before and after switching to the alternative therapy from the primary therapy in a myeloma randomized clinical trial.

Whitmore (2013) described the maximum likelihood estimation for a general threshold regression model with operational time scale defined in (3.7). Let variables $I_k(k = 1, \ldots, K)$ indicate if a subject’s FHT occurs at exposure condition $k$: if yes, $I_k = 1$, otherwise, $I_k = 0$. Note that $\sum_{k=1}^{K} I_k = 1$ since FHT can only occur at one exposure condition. Maximum likelihood estimation for right censored survival data under this context requires the construction of the likelihood function with contributions from both surviving and failing subjects. For a surviving subject, its likelihood contribution is given by its survival probability

$$P(T > t) = P[R(T) > R(t)] = P(R > r) = S(r)$$ (3.8)

which is the survival function on operational time scale. The second equality in (3.8) simply follows the notation of $R = R(t)$ and $r = R(t)$, where $R(\cdot)$ is the linear transformation defined in (3.7). For an observed failure at $r$ on the operational time scale, the likelihood contribution is given by the probability that failure will occur in the infinitesimal interval $(r, r + dr)$. This probability differential equals $f(r) dr$, where $f(r)$ is the p.d.f. of the FHT on operational time scale. By using standard calculus for change of variables,

$$dr = dR(t) = \sum_{k=1}^{K} \frac{\partial R_k(t)}{\partial t} dt = \sum_{k=1}^{K} \alpha_k \frac{\partial T_k(t)}{\partial t} dt = \left[ \sum_{k=1}^{K} \alpha_k I_k \right] dt,$$ (3.9)

we obtain the transformation from $r$ to $t$ in the probability element:

$$f(r) dr = f(r) \left[ \sum_{k=1}^{K} \alpha_k I_k \right] dt = f(r) J dt$$ (3.10)

53
where $J = \sum_{k=1}^{K} \alpha_k I_k$ is the corresponding Jacobian in the transformation of the p.d.f. from $f(r)$ on the operational time scale $r$ to $f(t)$ on calendar time scale $t$. From (3.8) and (3.10), by replacing $r$ with $R(t)$ in the form of (3.7), we obtain the likelihood contribution corresponding to a surviving subject, $S[R(t)]$, and that corresponding to a failing subject, $f[R(t)]J$, which constitutes the likelihood function of a general threshold regression model with $f(\cdot)$ as p.d.f. and $S(\cdot)$ as survival function on the operational time scale.

Whitmore (2013) also constructed simulation studies for different exposure scenarios on the basis of the operational time scale transformation in (3.7). For example, he tested a scenario imitating a three-arm clinical trial with unchanged exposure for each subject, and also tested a scenario where the exposure condition can change during a subject’s lifetime with a given probability. He used a threshold regression model based on the Wiener diffusion process and used the likelihood contributions for the surviving and failing subjects derived above in the simulation studies and found that the maximum likelihood estimates were unbiased.
Figure 3.1: Log-Log Plot by the “treatment2” Variable for the Leukemia Data

Figure 3.2: Kaplan-Meier Plot by the “treatment2” Variable for the Leukemia Data

Figure 3.3: Cox Predicted Plot v.s. Kaplan-Meier Plot by the “treatment2” Variable for the Leukemia Data

Figure 3.4: TR Predicted Plot v.s. Kaplan-Meier Plot by the “treatment2” Variable for the Leukemia Data
Chapter 4: BAYESIAN THRESHOLD REGRESSION
WITH INFORMATIVE CENSORING

4.1 Introduction

In some biomedical studies, investigators need to analyze current status data, where the observation of time to event $S$ is restricted to knowledge of whether or not $S$ exceeds an examination time $U$ (Jewell and van der Laan, 2004). For example, in an animal experiment, tumor onset time usually cannot be observed exactly but only known to be greater or less than an examination time. Many authors have proposed different methods to analyze current status data. Some of these methods assume examination time $U$ is independent of event time $S$. These methods include Cox (Huang, 1996), additive hazards (Lin et al., 1998; Martinussen and Scheike, 2002), accelerated failure time (Rabinowitz et al., 1995) and proportional odds models (Rossini and Tsiatis, 1996), which are reviewed in Section 2.2. However, since independent censoring may not be a reasonable assumption in some current status data applications (e.g., tumorigenicity studies), some authors proposed methods accounting for dependence of examination time $U$ on event time $S$ (e.g., Lagakos and Louis, 1988; Zhang et al., 2005; Dunson and Dinse, 2002). Among these methods, Dunson and Dinse (2002) proposed a Bayesian regression model, also motivated by tumorigenicity
studies. The authors accounted for tumor lethality using a latent binary variable, indicating whether or not death was caused by tumor. This latent binary variable could be augmented during MCMC. Their model is flexible and suitable for studies of tumors whose lethality is either mild (i.e., tumors that cannot cause death) or extremely lethal (i.e., tumors that cause death rapidly). However, one of the limitations of their model is that it assumes constant tumor lethality across subjects; this assumption might be unreasonable since tumor lethality could vary across subjects due to genetic, treatment or environmental factors.

To address the limitations in previous methods for current status data, in this chapter we propose a Bayesian threshold regression model for current status data analysis accounting for informative censoring. As reviewed in Chapter 3, threshold regression is a conceptually appealing method in biomedical studies since it provides inferences on latent health processes. In addition, threshold regression doesn’t assume proportional hazards which might be unreasonable in some applications. We propose to jointly model time to event and time to censoring using latent Wiener processes which fail once they hit a boundary value. Using data augmentation, we sample the unobserved event time and values of the latent processes for those subjects who do not experience an event. Similar to the model proposed by Dunson and Dinse (2002), we augment a latent binary variable which is used to indicate whether or not censoring is related to the event of interest during MCMC. Informative censoring is accounted for based on an assumption that, at the time of event occurrence, the censoring process that is previously independent of the event of interest could be upgraded to be dependent of the event of interest, according to the value of the latent binary variable.
Our approach, along with the approach in Dunson and Dinse (2002), fall in the “shifting mechanism” class of dependent censoring models mentioned in Section 2.3. However, unlike the model proposed by Dunson and Dinse (2002), our model doesn’t assume constant lethality across subjects since we incorporate covariates to model the drift of the censoring process that is dependent of the event of interest. Also, unlike Dunson and Dinse (2002), our model provides insights into the latent processes which characterize tumor development and overall health.

The rest of this chapter is organized as follows. In Section 4.2, we introduce our Bayesian threshold regression model for the current status data, both under independent (Section 4.2.1) and informative censoring (Section 4.2.2). In Section 4.3, we present numerical simulation results comparing the models under independent censoring and informative censoring, demonstrating the importance of accounting for the informative censoring. In Section 4.4, we apply the proposed model to data from a National Toxicology Program (NTP) 2-year rodent carcinogenicity study of chloroprene (National Toxicology Program, 1998).

4.2 Threshold Regression for Current Status Data

4.2.1 Model Under Independent Censoring

Likelihood Formulation

Let $S$ be an event time whose exact value is unknown. Instead, all we know about $S$ is it does or does not exceed censoring time $U$. For instance, in a carcinogenicity study, $S$ is time to tumor and $U$ is examination time, which could be due to natural death or planned sacrifice. The exact tumor onset time is unknown. We define a binary random variable $D$ for such information: $D$ is 1 if $S \leq U$ and 0 if $S > U$. 
Thus \( P(D = 1) = P(S \leq u \mid U = u) = F(u) \). Suppose a random sample of the population is obtained with observed data given by \( \{(d_i, u_i) : i = 1, \ldots, n\} \). The likelihood of these data is

\[
L_{\text{ind}} = \prod_{i=1}^{n} P(D_i = 1)^{d_i} P(D_i = 0)^{1-d_i}
\] (4.1)

Assuming \( S \) follows an inverse Gaussian distribution defined in (3.1) and (3.2). The likelihood becomes

\[
L_{\text{ind}} = \prod_{i=1}^{n} F(u_i \mid \mu_i, \sigma^2, y_0)^{d_i} [1 - F(u_i \mid \mu_i, \sigma^2, y_0)]^{1-d_i}
\]

\[
= \prod_{i=1}^{n} \left\{ \Phi \left[ -\frac{(y_0 + \mu_i u_i)}{\sqrt{\sigma^2 u_i}} \right] + \exp \left( -\frac{2y_0 \mu_i}{\sigma^2} \right) \Phi \left[ \frac{\mu_i u_i - y_0}{\sqrt{\sigma^2 u_i}} \right] \right\}^{d_i}
\times \left\{ \Phi \left[ \frac{(y_0 + \mu_i u_i)}{\sqrt{\sigma^2 u_i}} \right] - \exp \left( -\frac{2y_0 \mu_i}{\sigma^2} \right) \Phi \left[ \frac{\mu_i u_i - y_0}{\sqrt{\sigma^2 u_i}} \right] \right\}^{1-d_i}
\] (4.2)

where the drift of subject \( i \) (\( \mu_i \)) is modeled as a linear function of covariates \( x_i \), i.e., \( \mu_i = x_i' \alpha \)

Tong et al. (2008) used a similar likelihood function to jointly model current status data with an observable marker. While this likelihood works well for frequentist approaches, in the Bayesian framework, it results in complicated posterior distributions involving the standard normal c.d.f. As an alternative, we propose augmenting the unobserved event time \( S \). Data augmentation approaches for unobserved event times for interval censored data have been discussed extensively in the survival literature. We use an approach used by several others (e.g., Sinha et al. (1999), Calle (2003), Gajewski et al. (2004), Komarek et al. (2005), Komarek and Lesaffre (2008)). We sample event times from the distribution of \( S \) given \( U = u \), which is the inverse Gaussian p.d.f. truncated to the right at \( u \), using the inverse c.d.f. method (Devroye and Devroye, 1986). Thus the likelihood contribution for subjects with \( d_i = 1 \) is
\[ f_S(s_i | \mu_i, \sigma^2, y_0, u_i) = \frac{y_0}{\sqrt{2\pi \sigma^2 s_i^3}} \exp \left[ -\frac{(y_0 + \mu_i s_i)^2}{2\sigma^2 s_i} \right] I(0 < s_i \leq u_i), \quad (4.3) \]

For right censored observations (i.e., \(d_i = 0\)), we use a method proposed by Pennell \textit{et al.} (2010) to augment the latent process level at \(u_i\). This algorithm implements a rejection sampling procedure based on a truncated normal covering distribution to sample from the kernel of the p.d.f. for the process value at censoring time \(u_i\), conditioning on the fact that the \(i\)th subject survived beyond \(u_i\). This p.d.f. is proportional to the following:

\[ f_Y(y_i | \mu_i, \sigma^2, y_0, u_i) = \frac{1}{\sqrt{2\pi \sigma^2 u_i}} \exp \left[ -\frac{(y_i - y_0 - \mu_i u_i)^2}{2\sigma^2 u_i} \right] \times \left[ 1 - \exp \left( \frac{-2y_i y_0}{\sigma^2 u_i} \right) \right] I(y_i > 0) \quad (4.4) \]

The derivation of (4.4) can be found in Cox and Miller (1965) and Lu (1995).

Conditional on the augmented data, we obtain the following likelihood:

\[
I_{\text{ind}}^{(2)} = \prod_{i=1}^{n} [f_S(s_i | \mu_i, \sigma^2, y_0, u_i)]^{d_i} [f_Y(y_i | \mu_i, \sigma^2, y_0, u_i)]^{1-d_i} \\
= \prod_{i=1}^{n} \left\{ \frac{y_0}{\sqrt{2\pi \sigma^2 s_i^3}} \exp \left[ -\frac{(y_0 + \mu_i s_i)^2}{2\sigma^2 s_i} \right] I(0 < s_i \leq u_i) \right\}^{d_i} \\
\times \left\{ \frac{1}{\sqrt{2\pi \sigma^2 u_i}} \exp \left[ -\frac{(y_i - y_0 - \mu_i u_i)^2}{2\sigma^2 u_i} \right] \times \left[ 1 - \exp \left( \frac{-2y_i y_0}{\sigma^2 u_i} \right) \right] I(y_i > 0) \right\}^{1-d_i} \quad (4.5) \]

Note that \(\int_{0}^{\mu} f_S(s | \mu, \sigma^2, y_0, u) \, dy = F(u)\) and \(\int_{0}^{\infty} f_Y(y | \mu, \sigma^2, y_0, u) \, dy = 1 - F(u)\), and hence by sampling \(S\) and \(Y\) at each iteration of the MCMC, we perform numerical integration giving us a model equivalent to (4.2).
Prior Specification and Posterior Computation

A Bayesian specification of the model is completed by assigning diffuse conjugate multivariate normal prior, $\mathbf{N}(\mathbf{0}, 100 \times \mathbf{I})$, to $\alpha$ and a diffuse log-normal prior, $\log \mathbf{N}(0, 100)$, to $y_0$.

Posterior inference proceeds using an MCMC algorithm consisting of Gibbs, Metropolis and rejection sampling steps to sample the parameters ($S_i$, $y_i$, $\alpha$, $y_0$) based on their respective full conditional posterior distributions. The necessary conditional distributions are described in Appendix A.1.1, as well as a summary of the sampling algorithm.

4.2.2 Model Under Informative Censoring

Stochastic Model

We presented a threshold regression model for current status data in Section 4.2.1, assuming that the censoring time $U$ is independent of the event time $S$. However, this assumption is not always reasonable. For example, in tumorigenicity studies, $U$ may be time of natural death, which could have been affected by the presence of a tumor. In this case, if the censoring time (i.e., natural death time) is treated as independent of the event time (i.e., tumor onset time), bias would occur when estimating the event time distribution. In this section, we propose a joint model for the event and censoring time distributions to take into account the dependent censoring. Starting in this section, we use $E$ to denote the interval censored event of interest, and $C$ the censoring event.

We denote the stochastic process corresponding to $E$ as $Y_E(t)$ which we model using a Wiener process with initial status $y_{E0}$ and drift $\mu_E$ and we define $S$ as the
first time a sample path of \( Y_E(t) \) reaches the 0 level. The process by which \( C \) occurs may be stochastic (e.g., death prior to the event of interest) or non-stochastic (e.g., end of study, planned sacrifice in an animal study); the latter process is completely independent of \( E \) (and thus will receive no further attention) but the former may depend on \( E \). For example, in an animal tumorigenicity experiment, an animal is examined for a tumor (i.e., \( E \)) at the time of its natural death or terminal sacrifice (i.e., \( C \)). Event \( C \) is independent of \( E \) when \( C \) is by terminal sacrifice. However, \( C \) might depend on \( E \) when \( C \) is by natural death since tumor occurrence may affect survival. Let \( Y_C(t) \) denote the latent censoring process which we model as follows:

\[
Y_C(t) = \begin{cases} 
Y_{C1}(t) & \text{for } 0 < t < s \\
Y_{C1}(t) \text{ with probability } 1 - \pi & \text{for } t \geq s \\
Y_{C2}(t) \text{ with probability } \pi & \text{for } t \geq s
\end{cases}
\]

That is, prior to \( E \) occurring at \( s \), we assume \( Y_C(t) \) follows a Wiener process \( Y_{C1}(t) \) with drift \( \mu_{C1} \) and initial status \( y_{021} \) independent of \( Y_E(t) \). However, after the event we assume that with probability \( \pi \), \( Y_2 \) follows a different process \( Y_{C2}(t) \) with a different drift \( \mu_{C2} \) and initial status \( y_{022} \). For example, in the animal tumorigenicity experiment, the tumor process can be modelled with \( Y_E(t) \) and the death process can be modelled with \( Y_C(t) \). Prior to the tumor onset of an animal, its death process \( Y_{C1}(t) \) is independent of the tumor process. After an animal develops a tumor, it is possible (with probability \( \pi \)) that the tumor onset causes the death process to shift to a new process \( Y_{C2}(t) \) which degrades faster than \( Y_{C1}(t) \). If a tumor is mild it may not affect animal survival. Thus with probability \( 1 - \pi \), the death process \( Y_C(t) \) remains unchanged following tumor onset and still follows \( Y_{C1}(t) \).

We indicate the observable censoring mechanism by \( \delta \), which equals 0 if \( C \) is by a non-random event (e.g., planned end of study time) and 1 if \( C \) is by a random event.
driven by a stochastic process (e.g., death). To simplify modelling of the censoring
process, \( Y_C(t) \), we introduce an indicator \( R \), which equals 1 if \( Y_{C1}(t) \) is upgraded to
\( Y_{C2}(t) \) when \( E \) occurs and which equals 0 otherwise. The indicator \( R \) is unobservable
though values can be augmented as we will later explain. The complete current
status data therefore consists of \( Y = (U, D, \delta) \) for subjects without \( E \) occurring and
\( Y = (U, D, \delta, R) \) for those with \( E \) occurring. There are six different combinations of
d, \( \delta \) and \( R \), and details of likelihood contributions corresponding to these six scenarios
as well as the joint likelihood function for time to \( E \) and time to \( C \) are provided in
Appendix A.2.1.

In our application, we focus on modeling the drift parameters as linear functions
of covariates as in (3.3)

\[
\begin{align*}
\mu_{Ei} &= x'_{Ei} \alpha_E \\
\mu_{C1i} &= x'_{C1i} \alpha_{C1} \\
\mu_{C2i} &= x'_{C2i} \alpha_{C2}
\end{align*}
\]

The covariates linked to \( \mu_{Ei}, \mu_{C1} \) and \( \mu_{C2} \) do not have to be the same if it is
believed that the three processes are affected by different covariates. Some care
is needed when choosing the covariates in each regression equation. When sample
size and/or the event rate is small, only a small number of subjects would have \( R \)
augmented as 1 during the MCMC and thus there would be very little information
available to estimate the parameters of the \( Y_{C2}(t) \) process. Thus when the number of
events is small, we recommend a small number of predictors for \( \mu_{C2} \) and discourage
using categorical predictors. For example, if one covariate is dose, one should treat
it as a continuous variable rather than categorical. The model could be extended to include predictors in the initial status; this extension is fairly straightforward, but unnecessary for our motivating example since animals are usually very similar at the start of the experiment (same age, similar body size, similar genetics due to inbreeding).

**Prior Specification and Posterior Computation**

A Bayesian specification of the model is completed by assigning priors to the parameters $\theta = (\pi, \alpha_E, y_{E0}, \alpha_{C1}, y_{C10}, \alpha_{C2}, y_{C20})$. As pointed out in Section 4.2.2, it is possible to have identifiability problems for the $Y_{C2}(t)$ process in our Bayesian model when the sample size is small. Thus, the priors for some of the parameters in this process should, in general, be less diffuse than the priors on the parameters of the other processes.

We recommend a beta(10, 10) centered at 0.5 for $\pi$; $\pi = 0.5$ is a reasonable default prior when we do not have any prior information on whether the true value of $\pi$ is greater or less than 0.5. If there is strong prior knowledge of the likelihood that a tumor is lethal, this prior could be modified to favor a larger value of $\pi$ than 0.5. We found in the simulation study that this prior also works well for true values of $\pi$ other than 0.5 in reducing the estimation bias compared to a model assuming independent censoring.

We recommend diffuse conjugate multivariate normal priors (e.g., $N(0, 100 \times I)$) for $\alpha_E$ and $\alpha_{C1}$, and a more informative prior for $\alpha_{C2}$ (e.g., $N(0, 10 \times I)$) to accelerate mixing of MCMC. In our simulation studies and application, when a $N(0, 100 \times I)$ was used for $\alpha_{C2}$, the posterior means and the credible intervals of the parameters in the event process and the censoring process unrelated to death were similar to the means.
obtained under a $N(0, 10 \times I)$ prior but the mixing of the MCMC was considerably slower. Finally, we also recommend diffuse log-normal priors (e.g., $\log N(0, 100)$) for $y_{E0}$, $y_{C10}$ and $y_{C20}$. A more informative prior, say $\log N(0, 10)$, could also be assigned to $y_{C20}$ like what we did for $\alpha_{C2}$, however, we found that it didn’t improve the mixing of the MCMC and didn’t affect the posterior means and credible intervals of parameters much in our simulation studies and application.

Posterior inference proceeds using an MCMC algorithm consisting of Gibbs, Metropolis and rejection sampling steps to sample the parameters and latent variables ($R_i, S_i, y_{Ei}, \alpha_E, y_{E0}, y_{C1i}, \alpha_{C1}, y_{C10}, y_{C2i}, \alpha_{C2}, y_{C20}, \pi$) based on their respective full conditional posterior distributions. The necessary conditional distributions are described in Appendix A.2.2, along with a summary of the sampling algorithm.

4.3 Simulation Studies

4.3.1 Description of Data and Methods

In this section, we report results obtained from simulation studies conducted to assess the performance of our methodology. We consider the situation of a continuous covariate $V$ simulated from a normal distribution with mean 0 and standard deviation .5, and an ordinal covariate $W$ with four values (0, 1, 2 and 3) with an equal number of subjects assigned to each level. The variable $W$ can be thought of as the administered dose of a carcinogen in an animal study. For each subject, we assume that the drifts of the three latent processes in our informative censoring model were linked to $V$ and $W$ as in (4.6). The true values for the regression coefficients were: $\alpha_E = (-.16, -.32, -.63)$, $\alpha_{C1} = (-.25, -.63, -.32)$ and $\alpha_{C2} = (-.47, -.63, -.32)$, and the three processes had true initial values of 4.26, 4.71 and 4.71. We also set $\sigma^2_E$, $\sigma^2_{C1}$ and
$\sigma^2_{C_2}$ equal to 1 in the simulation and MCMC. Data were simulated under 4 different scenarios corresponding to different values of $\pi$: 0.9, 0.75, 0.5 and 0.25. A total of 50 datasets were generated for each scenario.

To simulate data, we first generated the value of $R$ for each subject from a Bernoulli distribution with $\text{Pr}(R = 1) = \pi$. We also set the end of study time to be 10 for each subject. We then generated the latent processes $Y_E(t)$ and $Y_C(t) = Y_{C_1}(t)$ simultaneously, by accumulating normally distributed increments of the processes over time increments of length $r = 0.01$ starting from their corresponding initial status values $y_{01}$ and $y_{021}$. The increments were $\Delta_E \sim N(\mu_E \times 0.01, \sigma^2_E \times 0.01)$ for $Y_E(t)$ and $\Delta_{C_1} \sim N(\mu_{C_1} \times 0.01, \sigma^2_{C_1} \times 0.01)$ for $Y_{C_1}(t)$.

If $Y_E(t)$ reached 0 first, we took $d_i = 1$, and then depending on the value of $R_i$, $Y_C(t)$ either continued to follow $Y_{C_1}(t)$ with increments $\Delta_{C_1}$ (if $R_i = 0$) or was updated to $Y_{C_2}(t)$ starting with a new initial status $y_{C_20}$ and accumulating increments $\Delta_{C_2} \sim N(\mu_{C_2} \times 0.01, \sigma^2_{C_2} \times 0.01)$ (if $R_i = 1$). If $Y_C(t)$ reached 0 before $t = 10, \delta_i = 1$; if not $\delta_i = 0$. The sample size, or the number of subjects, of each simulated dataset was 200, as in the tumorigenicity data in Section 4.4. In the simulated datasets, the percentage of right censored observations for the event of interest (i.e., $d_i = 0$) was around 60%, and the percentage of right censored observations (i.e., $\delta_i = 0$) for the censoring event for each dataset range from around 35% ($\pi = 0.25$) to around 30% ($\pi = 0.9$). These censoring rates are also similar to the rates in the tumorigenicity study presented in Section 4.4.

Each dataset was analyzed using the independent and informative censoring models. We assigned the same priors as recommend in Sections 4.2.1 and 4.2.2 to the model parameters. A total of 30,000 MCMC iterations following a 20,000 iteration
burn-in were run for each dataset under both independent and informative censoring models, as described in Sections 4.2.1 and 4.2.2. The MCMC algorithms for the independent and informative censoring models.

4.3.2 Results

Table 4.1 contains the bias of the posterior means of $\alpha_{E0}, \alpha_{E1}, \alpha_{E2}$ and $y_{E0}$ parameters in the event process for different values of $\pi$. The results were based on 50 replications. We found that when censoring was more informative, the bias of the posterior means of $y_{E0}$ and $\alpha_{E0}$ increased for both models. However, the estimates provided by the informative censoring model were less biased, especially when censoring was highly informative (i.e., when value of $\pi$ was large, such as .9, .75 or .5). The biases of the posterior means of $\alpha_{E1}$ and $\alpha_{E2}$ were approximately 10% or less for both models, although bias was smaller for the informative censoring model when the value of $\pi$ was .9 and .75.

Simulation results for the parameters of the censoring processes are in Appendix A.2.3. The biases of the parameters in $Y_{C1}(t)$ (i.e., the censoring process independent of the event process) were close to zero. The biases and sample standard deviations of the posterior means of the parameters in $Y_{C2}(t)$ (i.e., the censoring process dependent of the event process) and the parameter $\pi$ were pretty large; the former is due to the informative priors for $\pi$ and $\alpha_{C2}$ that are not centered on truth and the latter is due to the small number of observations (usually) following $Y_{C2}(t)$; for example, when the true $\pi$ is 0.5 and there are 120 events in a sample of 200 subjects (the average event rate in the simulated datasets), then there are only $120 \times 0.5 = 60$ subjects that would be expected follow $Y_{C2}(t)$ as opposed to the 140 subjects who
wouldn’t be expected to follow $Y_{C2}(t)$. Although the parameter estimates for $Y_{C2}(t)$ and $\pi$ are biased, the biases didn’t affect our estimates of the time to $E$ distribution much. In Figure 4.1, we compare the survival estimates for the event of interest $E$ provided by the informative and independent censoring models to the true survival curves. The curves for the informative censoring model were close to the true curve, indicating that, for the most part, the biases of the underestimated $\alpha_{E0}$ and the overestimated $y_{E0}$ offset each other. In contrast, the curves for the non-informative censoring model were further away from the truth, especially for large $\pi$. In other words, the more informative the censoring, the better the informative censoring model performed relative to the independent censoring model.

### 4.4 Application to Tumorigenicity Study

#### 4.4.1 Description of Data and Analysis Methods

We applied our model to a National Toxicology Program (NTP) 2-year rodent carcinogenicity study of chloroprene, a chemical used to produce synthetic rubber (National Toxicology Program, 1998). Dunson and Dinse (2002) and Tong et al. (2008) also analyzed data from this study in their current status papers but considered different subsets. The experiment incorporated a control group and three dose groups for both sexes of F344/N rats and B6C3F1 mice. Groups of 50 rodents were exposed to chloroprene at concentrations of 0, 12.8, 32, or 80 ppm by inhalation 6 hours per day, 5 days per week for up to 2 years. The occurrence of tumors in various sites was determined through a pathologic examination at the time of death. Many of the animals died naturally or were sacrificed prior to study conclusion, where the latter was done to alleviate suffering of sick animals according to protocol. All
Table 4.1: Simulation results for the parameters of the event process by the informative and independent censoring models for different values of $\pi$

| $\pi$ | Parameters | True Values | Informative Censoring Model | | | Independent Censoring Model | | |
|---|---|---|---|---|---|---|---|
| .9 | $\alpha_{E0}$ | -.16 | -0.11 | 0.15 | -.31 | .16 | |
| | $\alpha_{E1}$ | -.32 | -0.02 | 0.15 | -.05 | .14 | |
| | $\alpha_{E2}$ | -.63 | -0.06 | 0.09 | -.09 | .09 | |
| | $y_{E0}$ | 4.26 | 1.18 | 1.04 | 2.76 | 1.23 | |
| .75 | $\alpha_{E0}$ | -.16 | -0.07 | 0.15 | -.22 | .15 | |
| | $\alpha_{E1}$ | -.32 | -0.00 | 0.14 | -.03 | .14 | |
| | $\alpha_{E2}$ | -.63 | -0.04 | 0.09 | -.07 | .09 | |
| | $y_{E0}$ | 4.26 | 0.81 | 0.98 | 2.02 | 1.03 | |
| .5 | $\alpha_{E0}$ | -.16 | 0.01 | 0.14 | -.09 | .14 | |
| | $\alpha_{E1}$ | -.32 | 0.02 | 0.14 | .01 | .13 | |
| | $\alpha_{E2}$ | -.63 | -0.02 | 0.10 | -.05 | .09 | |
| | $y_{E0}$ | 4.26 | 0.25 | 0.96 | .96 | .94 | |
| .25 | $\alpha_{E0}$ | -.16 | 0.03 | 0.15 | -.02 | .15 | |
| | $\alpha_{E1}$ | -.32 | 0.02 | 0.15 | .02 | .13 | |
| | $\alpha_{E2}$ | -.63 | -0.01 | 0.09 | -.03 | .09 | |
| | $y_{E0}$ | 4.26 | 0.09 | 1.09 | .42 | .92 | |

Note: Bias is calculated by subtracting the true value from the posterior mean of a parameter; SSD is the sample standard deviation of the estimates.
surviving animals were sacrificed at 2 years, regardless of their health or tumor status. For illustration purposes, we confine our attention to lung tumors in F344/N male rats. The dataset we used had a sample size of 200 and the following five variables: censoring time $U$ (in days), censoring indicator $\delta$ ($\delta = 1$: natural death or moribund sacrifice; $\delta = 0$: terminal sacrifice), lung tumor indicator at death time $d$ ($d = 1$: lung tumor present at death; $d = 0$ otherwise), baseline weight and dose of chloroprene (0, 12.8, 32, or 80 ppm). The Kaplan-Meier survival curves for the four different dose groups of the F344/N male rat lung tumor data are given in Figure 4.2. Not surprisingly, rats exposed to high doses (32 and 80 ppm) died faster than rats exposed
Figure 4.2: Kaplan-Meier survival estimates for the four different dose groups of the F344/N male rat lung tumor data (in days)

to low doses (0 and 12.8 ppm). The distribution of death versus dose groups is provided in Table 4.2, where we can see that high dose groups (32 and 80 ppm) have more deaths than low dose groups (0 and 12.8 ppm). The log-rank test comparing the high and low dose groups yielded a p-value 0.0032, suggesting significant difference of survival between high and low dose groups. In Table 4.2, the distribution of lung tumor onset versus dose groups is also provided, where we couldn’t see any obvious trend that the impact of the dose effect has on lung tumor onset.

The focus of our analysis was on the effect of dose on time to lung tumor adjusting for base line weight. The drift of the tumor onset process was linked to the effect of
Table 4.2: Distributions of number of death and lung tumor occurrence v.s. dose groups for the F344/N male rat lung tumor data (N=50/group)

<table>
<thead>
<tr>
<th>Death</th>
<th>Dose (ppm)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>12.8</td>
</tr>
<tr>
<td>Death</td>
<td>37 (74%)</td>
<td>41 (82%)</td>
</tr>
<tr>
<td>Lung Tumor</td>
<td>32 (64%)</td>
<td>32 (64%)</td>
</tr>
</tbody>
</table>

dose and base line weight as:

$$\mu_E = \alpha_{E0} + \alpha_{E1}\text{weight} + \alpha_{E2}\text{dose}$$

And the remaining drifts of the death processes related or unrelated to tumor onset were modeled the same way. Note that dose is treated as a continuous variable; this is common practice in toxicological risk assessment as it allows risk estimation at doses not used in the study.

We used the same priors as recommended in Section 4.2.2. The MCMC algorithm described in Appendix A.2.2 was used to generate 120,000 MCMC iterates, with the first 20,000 discarded as a burn-in. The trace plots for our informative censoring model are given in Appendix A.2.4. We found that mixing was slow for the parameters of the tumor dependent death process, it is due to the small number of observations expected to follow this process. It wouldn’t affect the inferences of parameters of the lung tumor development process much based on our observation in the simulation study.

4.4.2 Results

The results for the informative censoring model are provided in the left panel in Table 4.3. We found strong evidence of an negative effect of chloroprene on the drift
of the death process unrelated to lung tumor \( P(\alpha_{C12} < 0 \mid \text{Data}) > 0.95 \); thus the greater the dose of chloroprene, the faster the death process approaches 0. However, dose of chloroprene had negligible effect on tumor onset and baseline weight did not affect time to tumor onset or time to death unrelated to lung tumor (90% credible intervals of \( \alpha_{E2} \), \( \alpha_{E1} \) and \( \alpha_{C11} \) included zero).

Our findings agree with the results in Dunson and Dinse (2002), where the authors also found strong evidence of a negative effect of chloroprene on death unrelated to lung tumor (the estimated posterior probability of a negative effect exceeded 0.99) and a weak evidence of an effect on lung tumor onset (the estimated posterior probability of a negative effect was 0.81). However, they considered a different subset of data and used a different model: they only analyzed the F344/N rat data in the control group and the group with highest dose of chloroprene and they analyzed the data in discrete time scale and did not allow the impact of a tumor on survival to differ across subjects. In addition, they didn’t adjust for the baseline weight variable in their analysis. We cannot compare our results with those in Tong et al. (2008) since the authors analyzed time to liver tumor instead of lung tumor and only used data from a single dose group.

The posterior estimates of the parameters in the death process related to lung tumor (i.e., \( \alpha_{C20} \), \( \alpha_{C21} \), \( \alpha_{C22} \) and \( y_{C20} \)) and the parameter \( \pi \) are also provided in Table 4.3. However, these results should be interpreted cautiously given the biases observed in the simulation studies described in Section A.2.3. Note that the estimated \( \pi \) value was .395, implying that the true value of the \( \pi \) variable is probably a little below .25, given the biases toward 0.5 observed in the simulation study. There was strong evidence of a negative effect of chloroprene on death related to lung tumor.
since the 90% credible interval of $\alpha_{C22}$ didn’t include zero. However, this evidence should be interpreted cautiously given the negative biases for $\alpha_{C22}$ observed in the simulation studies.

On the right panel of Table 4.3, we also list the posterior estimates of the parameters of the TR model presented in Section 4.2.1, which assumes independent censoring. The estimates of the parameters for the tumor process differed across models; especially the estimates of the initial health status, $y_{E0}$, and the intercept of the drift, $\alpha_{E0}$. We compared the estimated survival curves for lung tumor onset from the two different models in Figure 4.3. Survival probability estimates from the informative censoring model were larger than those from the independent censoring model starting around 200 days, and the estimated survival curves from these two models crossed at around 600 days. The estimated survival curves for lung tumor onset obtained from the model assuming independent censoring began to drop at around 200 days, while those from the model assuming informative censoring began to drop at around 300 days. In comparison, the Kaplan-Meier survival estimates for death began to drop sharply at around 500 days. Thus, the model accounting for informative censoring suggests a 100 day smaller lag between tumor onset time and death time.

4.5 Concluding Remarks

In this chapter, we proposed a Bayesian threshold regression model for current status data accounting for informative censoring. We jointly model time to event and time to censoring using latent Wiener processes which fail once they hit a boundary value. We sample the unobserved event time and values of the latent processes for
Table 4.3: Analysis result of the F344/N male rat lung tumor data by the informative and independent censoring models

<table>
<thead>
<tr>
<th>Param.</th>
<th>Informative Censoring Model</th>
<th>Independent Censoring Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Post. Mean</td>
<td>90% Cred. I.</td>
</tr>
<tr>
<td>$\alpha_{E0}$</td>
<td>-0.119</td>
<td>(-0.193, -0.046)</td>
</tr>
<tr>
<td>$\alpha_{E1}$</td>
<td>-0.855E-04</td>
<td>(-0.505E-03, 0.3344E-04)</td>
</tr>
<tr>
<td>$\alpha_{E2}$</td>
<td>6.67E-06</td>
<td>(-0.199E-03, 0.211E-03)</td>
</tr>
<tr>
<td>$y_{E0}$</td>
<td>78.821</td>
<td>(45.527, 113.267)</td>
</tr>
<tr>
<td>$\alpha_{C10}$</td>
<td>-0.199</td>
<td>(-0.2539, -0.1429)</td>
</tr>
<tr>
<td>$\alpha_{C11}$</td>
<td>-0.241E-03</td>
<td>(-0.613E-03, 0.129E-03)</td>
</tr>
<tr>
<td>$\alpha_{C12}$</td>
<td>-0.184E-03</td>
<td>(-0.363E-03, -3.7E-06)</td>
</tr>
<tr>
<td>$y_{C10}$</td>
<td>151.524</td>
<td>(133.206, 169.017)</td>
</tr>
<tr>
<td>$\alpha_{C20}$</td>
<td>0.062</td>
<td>(-3.503, 3.134)</td>
</tr>
<tr>
<td>$\alpha_{C21}$</td>
<td>-0.010</td>
<td>(-0.034, 0.013)</td>
</tr>
<tr>
<td>$\alpha_{C22}$</td>
<td>-0.029</td>
<td>(-0.065, -0.66E-02)</td>
</tr>
<tr>
<td>$y_{C20}$</td>
<td>255.079</td>
<td>(117.144, 441.024)</td>
</tr>
<tr>
<td>$\pi$</td>
<td>0.395</td>
<td>(0.181, 0.671)</td>
</tr>
</tbody>
</table>

those subjects who do not experience an event. Similar to the approach used in Dunson and Dinse (2002), we augment a latent variable for each subject which indicates if the event and censoring processes are related during MCMC. Informative censoring is accounted for by modeling time to censoring using two different latent censoring processes according to the value of the latent binary variable: one is independent of the event of interest and the other dependent. The drifts of each process are modeled as linear functions of covariates, and hence covariate effects on the distributions of either event and censoring time can be estimated through drifts. Our model provides insights into the latent processes underlying the event and censoring. Another advantage of this model is that it does not assume proportional hazards which might be an unreasonable assumption in some applications and can therefore serve as an attractive alternative to the popular Cox proportional hazards model. In contrast to
Figure 4.3: Estimated survival probability curves for lung tumor onset at the mean baseline weight of 121.545g by the informative and independent censoring models for the F344/N male rat lung tumor data (in days).

Dunson and Dinse (2002), our model is defined in a natural continuous time scale instead of a discrete time scale and our model does not assume the impact of event on censoring is constant across subjects.

We conducted simulation studies on data with different levels of dependent censoring to verify the effectiveness of our model. We compared the joint model accounting for informative censoring with a model assuming independent censoring and found that our joint model resulted in more accurate estimates of regression coefficients and survival probabilities when the censoring depended on the time of event.
We applied our methods to data from a National Toxicology Program (NTP) 2-year rodent carcinogenicity study of the impact of chloroprene on lung tumor onset and found that accounting for informative censoring of the data yielded different estimates of regression coefficients and, as a result, the estimated survival curves for lung tumor onset were different.

A limitation of our model accounting for informative censoring is that convergence of the MCMC is slow under small sample sizes. This limitation can be alleviated by assigning less diffuse priors to $\alpha_{C_2}$ (the coefficients of the drift parameter in the censoring process related to tumor) and $\pi$ (the probability that the event of interest affects the censoring time). In this dissertation, a beta(10, 10) prior was assigned to $\pi$ and a $N(0, 10 \times I)$ prior was assigned to $\alpha_{C_2}$. Our approach makes the a priori assumption that $\pi$ is around 0.5; i.e., we assume that the event of interest has a fifty-fifty chance of affecting time to censoring, or in other words, we assume a priori ignorance of whether the event is going to affect the censoring. We also assume, a priori, that the magnitude of the coefficients of $\alpha_{C_2}$ fall into the range of $(-10, 10)$. Although our approach can lead to biased estimates of the censoring time distribution when these a priori assumptions are not correct, we found in our simulation studies and data application that these priors provided posterior means and credible intervals of the parameters of the event process and the censoring process unrelated to the event that were similar to those obtained under more diffuse priors. Another extension of our model which could alleviate this limitation is to incorporate information from historical control data from NTP studies to elicit informative priors on $\pi$ and $\alpha_{C_2}$. A more detailed discussion is provided in Chapter 6.
Chapter 5: BAYESIAN THRESHOLD REGRESSION FOR MULTIVARIATE CURRENT STATUS DATA WITH INFORMATIVE CENSORING

5.1 Introduction

In Chapter 4 we proposed a threshold regression model for univariate current status data accounting for informative censoring. Sometimes, multivariate current status data arise in biomedical studies. For example, in carcinogenicity studies, tumors can develop in multiple organ sites and development of different tumors may intrinsically correlate with each other. The occurrence of these tumors are usually determined through a pathologic examination at the time of death and thus the onset times are either left or right censored. Further complicating matters, many of these examinations occur at time of natural death, which could be affected by the different types of tumors.

For example, let’s say we are interested in modeling time to lung tumor in an animal carcinogenicity study and an animal dies without a lung tumor but with a liver tumor. If the development processes of a lung tumor and liver tumor are completely unrelated, then the death can be treated as a censoring event which is independent of lung tumor onset. However, if the development of the two tumors
are related, the death should be treated as a censoring event which is related to an animal risk of lung tumor onset. Ignoring the liver tumor data would thus result in biased inferences on time to lung tumor. Thus in order to make unbiased inferences about the distribution of the tumor onset times, we need to model the tumor onset times jointly.

Several methods have been developed for the analysis of censored multivariate interval-censored survival data assuming independent censoring. Guo and Lin (1994) proposed a marginal regression model for grouped failure-time data. Ross and Moore (1999) proposed a linear log odds survival model for the marginal hazard functions and accounted for the between event dependency using a Clayton-Oakes gamma frailty model (Clayton, 1978; Oakes, 1982). Kaplan and Meier (1999b) and Kaplan and Meier (1999a) proposed nonparametric maximum likelihood estimators for bivariate and multivariate interval-censored data, respectively. Kaplan and Meier (2000) proposed a proportional hazards analysis for the multivariate interval-censored data.

Motivated by tumorigenicity studies, Dunson and Dinse (2002) proposed a Bayesian regression model for multivariate current status data subject to informative censoring. The authors accounted for possible dependency among onset times of different tumors by a subject specific latent variable and assumed conditional independence of the event times given this latent variable. The authors accounted for tumor lethality by a latent binary variable, indicating whether or not death was caused by tumor. Subjects with tumor affected deaths indicated by this latent binary variable would have different likelihood contributions than those whose deaths had nothing to do with tumor, and hence informative censoring was accounted for. For more details about their methodology, see Section 2.3.3. Unlike previous methods (e.g., Lagakos
and Louis, 1988), Dunson and Dinse’s method did not require the probability that a tumor is lethal to be specified a prior. However, they assumed constant tumor lethality which may not hold due to genetic, environmental, or physiological differences across subjects. In addition, the computational burden of their model is high; Dunson and Dinse use a discretized time scale and as a result a large amount of regression parameters are needed to estimate the event and examination time distributions.

In this chapter we propose a multivariate Bayesian threshold regression model to accommodate multiple left censored events which can be potentially correlated with each other. We use a random effect shared by the drifts of the processes underlying the events of interest to model the correlation of the event times. The censoring process is modeled using a latent Wiener process whose time scale is changed at the occurrence of an event thus accounting for dependent censoring. Our approach, along with the approach in Dunson and Dinse (2002), fall in the “shifting mechanism” class of dependent censoring models mentioned in Section 2.3.

5.2 Multivariate Threshold Regression Model for Current Status Data

5.2.1 Model Under Independent Censoring

Stochastic Model

Let \( S_j (j = 1, \ldots, K) \) be \( K \) event times whose exact values are not known. We are interested in estimating the full joint distribution of the event times: \( F(s_1, s_2, \ldots, s_K) \). Since the \( S_j \)'s are possibly not independent of each other, we address this correlation by incorporating a subject-specific random effect, \( b_i \), for subject \( i \) in the drift parameter of each event process. For the \( j \)th event process \( Y_j(t) \) which drives the \( j \)th event of interest for subject \( i \), we denote its drift parameter as \( \mu_{ji} \). Similar to (3.5), we link
\( \mu_{ji} \) to the random effect and the covariates as follows

\[
\mu_{ji} = b_i + \alpha_{j0} + \alpha_{j1} x_{j1i} + \cdots + \alpha_{jk} x_{jki} = b_i + \mathbf{x}_{ji}' \alpha_j
\]  

(5.1)

where \( b_i \sim N(0, \lambda^{-1}) \) and \( \mathbf{x}_{ji} \) and \( \alpha_j \) are, respectively, the covariate vector linked to the drift parameter \( \mu_{ji} \) and its regression coefficient vector. Note that it is not necessary to use the same covariates for the drift parameter of each event process.

For the \( i \)th subject, the \( j \)th and \( k \)th processes, \( Y_j(t) \) and \( Y_k(t) \), have the following covariance

\[
\text{COV}[Y_j(t), Y_k(t)] = \frac{t^2}{\lambda}
\]

(5.2)

and correlation

\[
\text{Corr}[Y_j(t), Y_k(t)] = \frac{t}{\sqrt{t + \lambda \sigma_1^2} \sqrt{t + \lambda \sigma_2^2}}
\]

(5.3)

where \( \sigma_1^2 \) and \( \sigma_2^2 \) are the variance parameters of the two processes. When \( \sigma_1^2 = \sigma_2^2 = 1 \), this correlation is equal to

\[
\text{Corr}[Y_j(t), Y_k(t)] = \frac{t}{t + \lambda}
\]

(5.4)

Equation (5.4) implies that the correlation approaches 1 over time, which makes sense in tumorigenicity studies; the more that time progresses, the more likely that the development of the tumors are influenced by related factors.

Given \( b_i \), the event processes of \( i \)th subject, i.e., \( Y_{ji}(t) \)'s, are independent of each other. Therefore the event times of the \( i \)th subject driven by these event processes, i.e., \( S_{ji} \)'s, are independent of each other. Thus

\[
f_S(s_1i, s_2i, \ldots, s_Ki \mid b_i) = \prod_{j=1}^K f_{S_{ji}}(s_{ji} \mid b_i)
\]

(5.5)

81
where \( f_{S_j}(s_{ji} \mid b_i) \) is an inverse Gaussian distribution with probability density function (p.d.f.)

\[
f_{S_j}(s_{ji} \mid b_i) = \frac{y_{j0}}{\sqrt{2\pi \sigma_j^2 s_{ji}^3}} \exp \left[ \frac{-(y_{j0} + \mu_j s_{ji})^2}{2\sigma_j^2 s_{ji}} \right]
\]

(5.6)

where \( \mu_j, y_{j0} \) and \( \sigma_j^2 \) are respectively the drift, initial state, and variance parameters of the underlying Wiener process driving the \( j \)th event of interest \( (\sigma_j^2 > 0, y_{j0} > 0 \text{ and } -\infty < \mu_j < \infty) \). The cumulative distribution function (c.d.f.) of \( S_j \) is

\[
F(s_{ji} \mid b_i) = \Phi \left[ \frac{-(y_{j0} + \mu_j s_{ji})}{\sqrt{\sigma_j^2 s_{ji}}} \right] + \exp \left( -2y_{j0}\mu_j \sigma_j^2 \right) \Phi \left[ \frac{\mu_j s_{ji} - y_{j0}}{\sqrt{\sigma_j^2 s_{ji}}} \right], \quad (5.7)
\]

where \( \Phi(\cdot) \) is the c.d.f. of the standard normal distribution. Conditional on \( b_i \), the likelihood of a sample of \( n \) subject is hence:

\[
L_{ind} = \prod_{i=1}^{n} \prod_{j=1}^{K} F_j(u_i|\mu_j, \sigma_j^2, y_{j0})^{d_{ji}} [1 - F_j(u_i|\mu_j, \sigma_j^2, y_{j0})]^{1-d_{ji}} \]

\[
= \prod_{i=1}^{n} \prod_{j=1}^{K} \left\{ \Phi \left[ \frac{(y_{j0} + \mu_j u_i)}{\sqrt{\sigma_j^2 u_i}} \right] + \exp \left( -2y_{j0}\mu_j \sigma_j^2 \right) \Phi \left[ \frac{\mu_j u_i - y_{j0}}{\sqrt{\sigma_j^2 u_i}} \right] \right\}^{d_{ji}}
\]

\[
\times \left\{ \Phi \left[ \frac{(y_{j0} + \mu_j u_i)}{\sqrt{\sigma_j^2 u_i}} \right] - \exp \left( -2y_{j0}\mu_j \sigma_j^2 \right) \Phi \left[ \frac{\mu_j u_i - y_{j0}}{\sqrt{\sigma_j^2 u_i}} \right] \right\}^{1-d_{ji}} \]

(5.8)

where \( u_i \) is the censoring time for the \( i \)th subject; and \( d_{ji} \) is the censoring indicator for the \( j \)th event of interest of subject \( i \): \( d_{ji} = 1 \) if \( s_{ji} \leq u_i \) and \( d_{ji} = 0 \) if \( s_{ji} > u_i \).

Instead of using a likelihood based on interval censored event times, we augment the unobserved event time of the \( j \)th process for subject \( i \), \( S_{ji} \), using the distribution of \( S_{ji} \) given \( U_i = u_i \), which is the inverse Gaussian p.d.f. truncated to the right at \( u_i \). Thus the likelihood contribution for subjects with \( d_{ji} = 1 \) is

\[
f_{S_j}(s_{ji} | \mu_j, \sigma_j^2, y_{j0}, u_i) = \frac{y_{j0}}{\sqrt{2\pi \sigma_j^2 s_{ji}^3}} \exp \left[ \frac{-(y_{j0} + \mu_j s_{ji})^2}{2\sigma_j^2 s_{ji}} \right] I(0 < s_{ji} \leq u_i), \quad (5.9)
\]
For right censored observations of the \( j \)th process (i.e., \( d_{ji} = 0 \)), we use the method in Pennell et al. (2010) (reviewed in Section 3.3) to augment \( y_{ji} \), the process level of the \( j \)th latent process at censoring time \( u_i \), and use the p.d.f. of \( y_{ji} \) conditional on survival beyond \( u_i \) as the likelihood contribution:

\[
f_{Y_j}(y_{ji} \mid \mu_{ji}, \sigma^2_j, y_{j0}, u_i) = \frac{1}{\sqrt{2\pi \sigma^2_j u_i}} \exp \left[ -\frac{(y_{ji} - y_{j0} - \mu_{ji} u_i)^2}{2\sigma^2_j u_i} \right] \times \left[ 1 - \exp \left( \frac{2y_{ji}y_{j0}}{\sigma^2_j u_i} \right) \right] I(y_{ji} > 0) \tag{5.10}
\]

Conditional on the augmented data, we obtain the following likelihood:

\[
L^{(2)}_{\text{ind}} = \prod_{i=1}^{n} \prod_{j=1}^{K} [f_{S_j}(s_{ji} \mid \mu_j, \sigma^2_j, y_{j0}, u_i)]^{d_{ji}} [f_{Y_j}(y_{ji} \mid \mu_j, \sigma^2_j, y_{j0}, u_i)]^{1-d_{ji}}
\]

\[
= \prod_{i=1}^{n} \prod_{j=1}^{K} \left\{ \frac{y_{j0}}{\sqrt{2\pi \sigma^2_j s_{ji}^3}} \exp \left[ -\frac{(y_{j0} + \mu_j s_{ji})^2}{2\sigma^2_j s_{ji}^2} \right] I(0 < s_{ji} \leq u_i) \right\}^{d_{ji}} \times \left\{ \frac{1}{\sqrt{2\pi \sigma^2_j u_i}} \exp \left[ -\frac{(y_{ji} - y_{j0} - \mu_{ji} u_i)^2}{2\sigma^2_j u_i} \right] \times \left[ 1 - \exp \left( \frac{2y_{ji}y_{j0}}{\sigma^2_j u_i} \right) \right] I(y_{ji} > 0) \right\}^{1-d_{ji}} \tag{5.11}
\]

Note that \( \int_0^u f_{S_j}(s_j \mid \mu_j, \sigma^2_j, y_{j0}, u) \, dy = F(u) \) and \( \int_0^{\infty} f_{Y_j}(y_{ji} \mid \mu_j, \sigma^2_j, y_{j0}, u) \, dy_{ji} = 1 - F(u) \), and hence by sampling \( S_j \) and \( Y_j \) at each iteration of an MCMC algorithm, we perform numerical integration giving us a likelihood equivalent to (5.8).

**Prior Specification and Posterior Computation**

A Bayesian specification of the model is completed by assigning priors to the parameters \( \alpha_j, y_{j0}, \) and \( \lambda \). We recommend diffuse conjugate multivariate normal priors for \( \alpha_j \); diffuse log-normal priors for \( y_{j0} \); and diffuse Gamma prior for \( \lambda \).
Posterior inference proceeds using an MCMC algorithm consisting of Gibbs, Metropolis and rejection sampling steps to sample the parameters \((\alpha_j, y_{j0}, \lambda)\) and latent variables \((S_{ji}, y_{ji}, b_i)\) based on their respective full conditional posterior distributions. The necessary conditional distributions are described in Appendix B.1.2, as well as a summary of the sampling algorithm.

### 5.2.2 Model Under Informative Censoring

#### Stochastic Model

We presented a multivariate threshold regression model for current status data in Section 5.2.1, assuming that the censoring time \(U\) is independent of the multiple event times \(S_j (j = 1, \ldots, K)\). However, this assumption is not always reasonable. For example, in tumorigenicity studies, \(U\) may be time of natural death, which could have been affected by the presence of one or more tumors. In this case, if the censoring time (i.e., natural death time) is treated as independent of the event times (i.e., onset times of different tumors), bias may occur when estimating the event time distribution. In this section, we propose a joint model for the event and censoring time distributions to take into account the dependent censoring. Starting in this section, we use \(E_j\) to denote the \(j\)th interval censored event of interest, and \(C\) the censoring event.

We denote the stochastic process corresponding to \(E_j\) as \(Y_{E_j}(t)\) which we model using a Wiener process with initial status \(y_{Ej0}\), drift \(\mu_{E_j}\) and variance \(\sigma^2_{E_j}\). We define \(S_j\) as the first time a sample path of \(Y_{E_j}(t)\) reaches the 0 level. The process by which \(C\) occurs may be stochastic (e.g., death prior to the event of interest) or non-stochastic (e.g., end of study, planned sacrifice in an animal study); the latter process is completely independent of all \(E_j\) (and thus will receive no further attention). For example, in an animal tumorigenicity experiment, an animal is examined for tumors
at $K$ different sites (i.e., $E_1, \ldots, E_K$) at the time of its natural death or terminal sacrifice (i.e., $C$). Event $C$ is independent of $E_1, \ldots, E_K$ when $C$ is by terminal sacrifice. However, $C$ may depend on the $E_j$’s when $C$ is by natural death since single or multiple tumor occurrences could affect survival.

We presented a threshold regression model for univariate current status data with informative censoring in Section 4.2.2, where informative censoring is accounted for by modeling time to censoring using two latent health processes: one is independent of the event of interest and the other dependent. This approach is reasonable for univariate current status data but not for multivariate current status data. In the case of multivariate current status data, the censoring process could be affected by both number and type of events that occur before the censoring time. Accounting for each possible combination through a different stochastic process would be unwieldy. To address this issue, we use an operational time scale, which is reviewed in Section 3.4, to model the censoring process. Let $Y_C(r)$ denote the latent censoring process which follows a Wiener process with drift $\mu_C$, initial status $y_{C0}$ and variance $\sigma^2_C$.

The censoring process has an operational time scale defined by the following linear transformation:

$$r = \sum_{l=1}^{M} \tau_l t_l, \quad \text{where} \quad \sum_{l=1}^{M} t_l = t. \quad (5.12)$$

where $t$ is the study time in usual calendar time scale. $t_1, \ldots, t_M$ are the amount of time spent in each of $M$ possible event states. For example, in a tumorigenicity study of heart and liver hemangiosarcomas, there are four possible event states for a subject: having neither a heart hemangiosarcoma nor a liver hemangiosarcoma, having only a heart hemangiosarcoma but not a liver hemangiosarcoma, having only a liver hemangiosarcoma but not a heart hemangiosarcoma, and having both a liver and
heart hemangiosarcoma. Each of these event states might have different effect on the degradation rate of the censoring process, which is accounted for by $\tau_1, \ldots, \tau_M$. In this way, the censoring process is modeled to depend on the multiple event processes. One of the $\tau_l$ parameters, say $\tau_1$, has to be set to unity to define the unit of measurement of the operational time scale.

The drift parameters in the event processes and censoring process are modeled as follows

$$
\mu_{Eji} = b_i + x'_{Eji} \alpha_{Ej}, \quad \text{where } j = 1, \ldots, K
$$

(5.13)

$$
\mu_{Ci} = x'_{Ci} \alpha_C
$$

(5.14)

where $b_i$ is as defined in (5.1), $x_{Eji}$ is the covariate vector linked to the drift parameter, $\mu_{Eji}$, of the $j$th event process and $\alpha_{Ej}$ is its regression coefficient vector. The term $x_{Ci}$ is the covariate vector linked to the drift parameter, $\mu_{Ci}$, of the censoring process and $\alpha_C$ is its regression coefficient vector. The model could be extended to include predictors in the initial status; this extension is fairly straightforward, but unnecessary for our motivating example from a carcinogenicity experiment since animals are usually very similar at the start of the experiment (same age, similar body size, similar genetics due to inbreeding).

**Prior Specification and Posterior Computation**

A Bayesian specification of the model is completed by assigning priors to the parameters $\alpha_{Ej}$, $y_{E0}$, $\alpha_C$, $y_C$, $\lambda$, and $\tau$. We recommend diffuse conjugate multivariate normal priors for $\alpha_{Ej}$ and $\alpha_C$, diffuse log-normal priors for $y_{Ej}$ and $y_C$, a Gamma prior diffuse on moderate values for $\lambda$, e.g., Ga(2, 1/2), and a log-normal $N(0, 1)$ prior for $\tau$. The prior for $\tau$ has mode on $\tau = 1$ and is not diffuse. As with the
parameter $\pi$ in the model proposed in Chapter 4, less diffuse priors for informative censoring parameter $\tau$ are needed to accelerate mixing of MCMC. Ideally, the mode would reflect knowledge of the effect of a tumor on mortality. In the absence of this knowledge, we found a good default approach is to shrink the estimates of $\tau$ back toward no effect. We found in our simulation study that even with this prior for $\tau$, our method is able to identify situations when there is informative censoring.

Posterior inference proceeds using an MCMC algorithm consisting of Gibbs, Metropolis and rejection sampling steps to sample the parameters $(\alpha_{Ej}, y_{E0j}, \alpha_{C}, y_{CO}, \lambda, \tau)$ and latent variables $(S_{ji}, y_{Eji}, y_{Cj}, b_i)$ based on their respective full conditional posterior distributions. The necessary conditional distributions are described in Appendix B.2.2, as well as a summary of the sampling algorithm.

5.3 Simulation Studies

5.3.1 Description of Data and Methods

In this section, we report results obtained from simulation studies conducted to assess the performance of the informative censoring model. Without loss of generality, we still consider the case of two events of interest. We consider the situation of two covariates: a continuous covariate $V$ simulated from a normal distribution with mean 2.5 and standard deviation 1, an ordinal covariate $W$ with four values (0, 1, 2 and 3) with an equal number of subjects assigned to each level, and a random effect $b$ following $N(0, 1)$ distribution. The variable $W$ can be thought of as the administered dose of a carcinogen in an animal study. We generated informative censored data and analyzed the data using both the independent and informative censoring models.
For each subject, the drifts of the two latent processes corresponding to the events of interest in our informative censoring model were linked to $V$, $W$ and $b$ as follows

$$\mu_{E_1i} = b_i + 3 - 0.8V_i - 2W_i$$
$$\mu_{E_2i} = b_i + 2.5 - 0.75V_i - 1.8W_i$$ (5.15)

The two processes both had initial values $y_{E0_1} = y_{E0_2} = 3$. We also set $\sigma^2_{E_1}$ and $\sigma^2_{E_2}$ to 1 when simulating and analyzing the data to set the scale of the latent processes.

To simulate data, we first set the end of study time to be 15 for each subject. We then generated the latent event processes $Y_{E1}(t)$ and $Y_{E2}(t)$ simultaneously, by accumulating normally distributed increments of the two processes: $\Delta_{Ej} \sim N(\mu_{Ej} \times dt, \sigma^2_{Ej} \times dt)$, where $dt$ is a small increment of time. We simultaneously accumulated $\Delta_{E1}$ and $\Delta_{E2}$ over time increments of length $dt = 0.01$ until the end of study time $t = 15$. Then, for each subject, we generated the path of the latent censoring process $Y_C(r)$ in operational scale using two steps. In Step 1, we changed the calendar time scale in $t$ to operational time scale in $r$ for each subject using the following transformation:

$$r = R(t) = t_1 + \tau_1t_2 + \tau_2t_3 + \tau_3t_4$$ (5.16)

where $t_1$, $t_2$, $t_3$ and $t_4$ are the amount of time spent in each of the following states, respectively: neither $E_1$ nor $E_2$, $E_1$ but no $E_2$, $E_2$ but no $E_1$, $E_1$ and $E_2$. Note that $\sum_{t=1}^{4} t_i = t$.

In Step 2, we generated the latent censoring processes $Y_C(r)$, by accumulating normally distributed increments of the process: $\Delta_C \sim N(\mu_C \times 0.01, \sigma^2_C \times 0.01)$ until the end of the study time in operational time scale, which was calculated for each subject using (5.16). If the censoring process $Y_C(r)$ reached 0 before the end of study time in operational time scale, $\delta_i = 1$ and the censoring time $t$ in calendar time scale...
was the sum of amount of time spent in each of the event states until $Y_C(r)$ reached 0. If the censoring process $Y_C(r)$ didn’t reach 0 before the end of study time in the operational time scale, $\delta_i = 0$ and the censoring time in calendar time scale was $t = 15$. We took $d_{ji} = 1$ if $Y_{Ej}(t)$ reached 0 before the censoring process reached 0, otherwise, $d_{ji} = 0$. For example, suppose that there is a subject $i$ who got $E_1$ at time $t_1$ and then $E_2$ at time $t_2$ before the end of study time $t = 15$. Then for subject $i$, $T_1(t) = t_1$, $T_2(t) = t_2 - t_1$, $T_3(t) = 0$, $T_4(t) = 15 - t_2$, and $r' = T_1(t) + \tau_1 T_2(t) + \tau_2 T_3(t) + \tau_3 T_4(t)$ is the end of study time in operational time scale. If the simulated censoring process $Y_C(r)$ reached 0 at $r$ before $r'$, we take $\delta_i = 1$, otherwise, $\delta_i = 0$. We then transform $r$ back to $t$, the censoring time in calendar time scale, using (5.16). Obviously, when $\delta_i = 0$, $t = 15$, and in this case, both $d_{1i}$ and $d_{2i}$ equal to 1 since both $E_1$ and $E_2$ occurred before the end of study time $t = 15$. However, when $\delta_i = 1$, $t$ would be less than 15. We took $d_{1i} = 1$ if $t_1 < t$, otherwise $d_{1i} = 0$. We took $d_{2i} = 1$ if $t_2 < t$, otherwise $d_{2i} = 0$.

Data were simulated under 4 different scenarios corresponding to different values of $\{\tau_1, \tau_2, \tau_3\}$: $\{2, 2, 3\}$, $\{1.5, 1.5, 2\}$, $\{1, 1, 2\}$ and $\{1, 1, 1\}$. Note that the first three settings correspond to different scenarios of informative censoring while the last setting corresponds to independent censoring since all $\tau$’s are equal to 1, meaning that degradation of the censoring process $Y_C(r)$ is unaffected by $E_1$ and $E_2$. A total of 50 datasets were generated for each scenario of $\tau$’s. The sample size, or the number of subjects, of each simulated dataset was 200.

For our analysis, we assigned conjugate multivariate normal priors $N(0, 10000 \times I)$ to $\alpha_{Ej}$, $\alpha_C$; log-normal priors $\log N(\log(3), 10000)$ to $y_{Ej}$, $y_C$; log-normal $\log N(0, 1)$ prior to $\tau$. For parameter $\lambda$, its prior may reflect values estimated from prior studies;
we used a diffuse prior over a moderate value, Ga(2, 1/2), as recommended by Pennell et al. (2010). A total of 30,000 iterations following a 30,000 burn-in were run in MCMC for each dataset under both independent and informative censoring models, as described in Sections 5.2.1 and 5.2.2. The MCMC algorithms for the independent and informative censoring models are described in Appendix B.1 and B.2 respectively.

5.3.2 Results

Table 5.1 contains the bias of the posterior means of $\alpha_{Ej0}$, $\alpha_{Ej1}$, $\alpha_{Ej2}$, $y_{Ej0}$, and $\lambda$ for different values of $\{\tau_1, \tau_2, \tau_3\}$. The results are based on 50 replications. The table provides the average biases of the estimated parameters and the sample standard deviations of the estimates (SSDs). In general the estimates provided by the informative censoring model were less biased than those from the independent censoring model under informative censoring, i.e., when $\{\tau_1, \tau_2, \tau_3\}$ was $\{2, 2, 3\}$, $\{1.5, 1.5, 2\}$, or $\{1, 1, 2\}$. However, in the independent censoring scenario of $\tau = \{1, 1, 1\}$, both models performed well. Bias was less for the independent censoring model, but still acceptable (around 10% or less) for the Dependent censoring model.

The simulation results for the parameters of the censoring process in operational time scale are provided in Appendix B.2.3. The biases of the posterior means of the parameters were large at high levels of informative censoring (i.e., when all $\tau$’s were considerably bigger than 1), and decreased with the informative censoring level. Although the parameter estimates of the censoring processes were biased for high informative censoring levels, the biases didn’t affect our estimates of the time to event distributions much.
In Figures 5.1, we compare the survival curves for the first event of interest estimated by the informative and independent censoring models to the true survival curves for the four scenarios of \( \{\tau_1, \tau_2, \tau_3\} \). These curves are calculated at the mean of the continuous covariate \( V \) and the four different levels of categorical covariate \( W \) (i.e., \( W=1, 2, 3 \) or \( 4 \)). We provide a similar comparison plot for the second event of interest in Figure 5.2. The survival estimates provided by the informative censoring model were closer to the true values when censoring was informative. At the value 1 of the ordinal covariate \( W \), the differences between the survival estimates by the informative and independent censoring models were greatest. From Figures 5.1, we can see that, although some parameter estimates from the informative censoring model (especially for the initial state parameter and the intercept parameter of drift) were biased under informative censoring scenarios, those biases offset each other and resulted in survival probability estimates close to truth. In the independent censoring scenario, the survival estimates by the informative and independent censoring models were very similar and close to the truth.

5.4 Application to Tumorigenicity Study

5.4.1 Description of Data and Analysis Methods

We applied our model to data from a National Toxicology Program (NTP) 2-year rodent carcinogenicity study of urethane. Urethane is often found as a by-product of fermentation. Urethane is carcinogenic in many species including mice and monkeys and is potentially carcinogenic to humans (see, for example, Mirvish, 1968; Salmon et al., 1991; Thorgeirsson et al., 1994). At the time of the NTP study, ethanol was not classified as a carcinogen in experimental animals. However, there was evidence
Table 5.1: Simulation results for the parameters of the event processes by the informative and independent censoring models for different values of $\tau$

<table>
<thead>
<tr>
<th>Parameters</th>
<th>True Values</th>
<th>Bias</th>
<th>SSD</th>
<th>Bias</th>
<th>SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{E10}$</td>
<td>3</td>
<td>-0.15</td>
<td>0.55</td>
<td>-0.44</td>
<td>0.64</td>
</tr>
<tr>
<td>$\alpha_{E11}$</td>
<td>-0.8</td>
<td>0.12</td>
<td>0.16</td>
<td>0.59</td>
<td>0.26</td>
</tr>
<tr>
<td>$\alpha_{E12}$</td>
<td>-2</td>
<td>0.08</td>
<td>0.29</td>
<td>0.39</td>
<td>0.25</td>
</tr>
<tr>
<td>$y_{E10}$</td>
<td>3</td>
<td>-0.44</td>
<td>0.37</td>
<td>-1.91</td>
<td>0.55</td>
</tr>
<tr>
<td>$\alpha_{E20}$</td>
<td>2.5</td>
<td>-0.06</td>
<td>0.48</td>
<td>-0.42</td>
<td>0.47</td>
</tr>
<tr>
<td>$\alpha_{E21}$</td>
<td>-0.75</td>
<td>0.08</td>
<td>0.15</td>
<td>0.53</td>
<td>0.28</td>
</tr>
<tr>
<td>$\alpha_{E22}$</td>
<td>-1.8</td>
<td>0.05</td>
<td>0.31</td>
<td>0.41</td>
<td>0.28</td>
</tr>
<tr>
<td>$y_{E20}$</td>
<td>3</td>
<td>-0.38</td>
<td>0.42</td>
<td>-1.84</td>
<td>0.59</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>1</td>
<td>0.29</td>
<td>0.56</td>
<td>0.05</td>
<td>0.55</td>
</tr>
</tbody>
</table>

$\{2, 2, 3\}$

$\alpha_{E10}$ | 3 | 0.19 | 0.60 | -0.17 | 0.58 |
| $\alpha_{E11}$ | -0.8 | -0.00 | 0.20 | 0.34 | 0.28 |
| $\alpha_{E12}$ | -2 | -0.11 | 0.34 | 0.11 | 0.30 |
| $y_{E10}$ | 3 | -0.13 | 0.47 | -1.27 | 0.71 |
| $\alpha_{E20}$ | 2.5 | 0.13 | 0.57 | -0.24 | 0.60 |
| $\alpha_{E21}$ | -0.75 | -0.01 | 0.19 | 0.31 | 0.30 |
| $\alpha_{E22}$ | -1.8 | -0.11 | 0.37 | 0.13 | 0.40 |
| $y_{E20}$ | 3 | -0.01 | 0.52 | -1.06 | 0.95 |
| $\lambda$ | 1 | 0.09 | 0.48 | -0.08 | 0.51 |

$\{1.5, 1.5, 2\}$

$\alpha_{E10}$ | 3 | 0.23 | 0.50 | 0.51 | 0.81 |
| $\alpha_{E11}$ | -0.8 | -0.02 | 0.18 | 0.18 | 0.44 |
| $\alpha_{E12}$ | -2 | -0.16 | 0.30 | -0.43 | 0.51 |
| $y_{E10}$ | 3 | -0.03 | 0.50 | -0.75 | 1.28 |
| $\alpha_{E20}$ | 2.5 | 0.16 | 0.61 | 0.30 | 0.78 |
| $\alpha_{E21}$ | -0.75 | -0.01 | 0.22 | 0.21 | 0.44 |
| $\alpha_{E22}$ | -1.8 | -0.18 | 0.30 | -0.35 | 0.43 |
| $y_{E20}$ | 3 | 0.04 | 0.48 | -0.73 | 1.25 |
| $\lambda$ | 1 | 0.03 | 0.48 | -0.49 | 0.25 |

$\{1, 1, 2\}$

$\alpha_{E10}$ | 3 | 0.34 | 0.65 | 0.09 | 0.56 |
| $\alpha_{E11}$ | -0.8 | -0.13 | 0.20 | -0.00 | 0.19 |
| $\alpha_{E12}$ | -2 | -0.19 | 0.32 | -0.17 | 0.39 |
| $y_{E10}$ | 3 | 0.38 | 0.49 | 0.24 | 0.74 |
| $\alpha_{E20}$ | 2.5 | 0.16 | 0.49 | 0.03 | 0.69 |
| $\alpha_{E21}$ | -0.75 | -0.07 | 0.17 | 0.03 | 0.26 |
| $\alpha_{E22}$ | -1.8 | -0.14 | 0.23 | -0.14 | 0.38 |
| $y_{E20}$ | 3 | 0.34 | 0.42 | 0.13 | 0.80 |
| $\lambda$ | 1 | 0.17 | 0.53 | 0.06 | 0.50 |

Note: Bias is calculated by subtracting the true value from the posterior mean of a parameter; SSD is the sample standard deviation of the estimates.
that consumption of alcoholic beverages could be related to occurrence of malignant
tumors of the oral cavity, liver and other organs. Since urethane is classified as being
possibly carcinogenic to humans and humans are mainly exposed to urethane by
consuming fermented foods and drinking alcoholic beverages, which are also related
to cancer, urethane in combination with ethanol was nominated by the U.S. Food
and Drug Administration, for an in-depth toxicological evaluation by the NTP.

The NTP experiment incorporated a control group and three dose groups of ure-
thane for both sexes of B6C3F1/NCTR mice. Groups of 144 rodents were exposed
to urethane at concentrations of 0, 10, 30, or 90 ppm in the presence of 0%, 2.5%,
Figure 5.2: Comparisons of survival estimates provided by the informative and independent censoring models for the second event of interest ($E_2$) in the simulation studies.

or 5% ethanol in drinking water. Many animals died prior to 2 years due to natural causes or moribund sacrifice (to alleviate suffering of sick animals). All remaining survival animals were terminally sacrificed at 2 years. The occurrence of tumors in various sites was determined through a pathologic examination at the time of death: either natural death, moribund sacrifice or terminal sacrifice. We confine our attention to heart and liver hemangiosarcomas among male mice. The dataset we used had a sample size of 576 and the following six variables: censoring time $U$ (in days), censoring indicator $\delta$ ($\delta = 1$: natural death or moribund sacrifice; $\delta = 0$: terminal survival).
sacrifice), heart hemangiosarcoma indicator at death time $d_1$ ($d_1 = 1$: heart hemangiosarcoma present at death; $d_1 = 0$ otherwise), liver hemangiosarcoma indicator at death time $d_2$ ($d_2 = 1$: liver hemangiosarcoma present at death; $d_2 = 0$ otherwise), baseline weight and dose of urethane (0, 10, 30, or 90 ppm). We combined the data from animals exposed and unexposed to ethanol, because in a preliminary analysis, we found no main effect of ethanol and no interaction effect between urethane and ethanol on either heart or liver hemangiosarcoma onset. Using Wald tests in logistic regression we obtained $p=0.974$ for the main effect of ethanol and $p=0.874$ for the interaction effect between urethane and ethanol on heart hemangiosarcoma onset and $p=0.911$ for the main effect of ethanol and $p=0.892$ for the interaction effect between urethane and ethanol on liver hemangiosarcoma onset.

The Kaplan-Meier survival curves for the four different dose groups are given in Figure 5.3. Not surprisingly, mice exposed to high doses (30 and 90 ppm) died faster than rats exposed to low doses (0 and 10 ppm).

The distribution of death versus dose groups is provided in Table 5.2, where we can see that, the higher the dose groups, the more deaths occur. The log-rank test comparing the dose groups yielded a $p$-value $< 0.0001$, suggesting significant difference of survival between different dose groups.

The distributions of heart and liver hemangiosarcoma onset versus dose groups are provided in Table 5.3 and 5.4 respectively, where we can see that, the higher the dose groups, the more heart and liver hemangiosarcomas occur.

The focus of our analysis was on the effect of urethane dose on time to heart and liver hemangiosarcomas adjusting for baseline weight. The drifts of the tumor onset processes for heart and liver hemangiosarcomas were modeled as a linear function of
Figure 5.3: Kaplan-Meier survival estimates for the four different dose groups of the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data (in days)

Table 5.2: Distribution of number of death v.s. dose groups for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data

<table>
<thead>
<tr>
<th>Death</th>
<th>Dose (ppm)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>No</td>
<td>95</td>
<td>85</td>
</tr>
<tr>
<td>Yes</td>
<td>49</td>
<td>59</td>
</tr>
<tr>
<td>Total</td>
<td>144</td>
<td>144</td>
</tr>
</tbody>
</table>

576
Table 5.3: Distribution of number of heart hemangiosarcomas v.s. dose groups for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data

<table>
<thead>
<tr>
<th>Heart hemangiosarcomas</th>
<th>Dose (ppm)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>No</td>
<td>144</td>
<td>144</td>
</tr>
<tr>
<td>Yes</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>144</td>
<td>144</td>
</tr>
</tbody>
</table>

Table 5.4: Distribution of number of liver hemangiosarcomas v.s. dose groups for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data

<table>
<thead>
<tr>
<th>Liver hemangiosarcomas</th>
<th>Dose (ppm)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>No</td>
<td>138</td>
<td>136</td>
</tr>
<tr>
<td>Yes</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Total</td>
<td>144</td>
<td>144</td>
</tr>
</tbody>
</table>

the random effect $b_i$, dose and baseline weight:

\[ \mu_{\text{heart}_i} = b_i + \alpha_{E10} + \alpha_{E11}\text{weight}_i + \alpha_{E12}\text{dose}_i \]

\[ \mu_{\text{liver}_i} = b_i + \alpha_{E20} + \alpha_{E21}\text{weight}_i + \alpha_{E22}\text{dose}_i \]

The death process was defined on the following operational time scale:

\[ r_i = t_{1i} + \tau_1 t_{2i} + \tau_2 t_{3i} + \tau_3 t_{4i}, \quad \text{and } \sum_{t=1}^{4} t_i(t) = t. \]  

(5.17)

where $t$ is the observed study time of the death process in calendar time scale, $t_{1i}$ is the amount of time that the $i$th subject has neither a heart hemangiosarcoma nor a liver hemangiosarcoma, $t_{2i}$ is the amount of time that the $i$th subject has only a heart hemangiosarcoma but not a liver hemangiosarcoma, $t_{3i}$ is the amount of time
that the $i$th subject has only a liver hemangiosarcoma but not a heart hemangiosarcoma, and $t_{4i}$ is the amount of time that the $i$th subject has both a liver and heart hemangiosarcoma. The drift parameter of the death process was modeled as a linear function of dose and baseline weight:

$$\mu_{\text{death}_i} = \alpha_{C_0} + \alpha_{C_1}\text{weight}_i + \alpha_{C_2}\text{dose}_i$$

We assigned the same priors as in the simulation study to the model parameters. The MCMC algorithm described in Appendix B.2.2 was used to generate 150,000 MCMC iterates, with the first 50,000 discarded as a burn-in. The trace plots for our informative censoring model are given in Appendix B.2.4. We found no evidence of lack of convergence, though mixing was slow for the parameters of the liver hemangiosarcoma development process.

5.4.2 Results

The results for the informative censoring model are provided in the left panel of Table 5.5. We found strong evidence of an adverse effect of urethane on the drifts of both the heart and liver hemangiosarcoma development processes, as well as the death process in the operational time scale, since the posterior means of $\alpha_{E_{12}}$, $\alpha_{E_{22}}$ and $\alpha_{C_2}$ were all negative and their 90% credible intervals did not include zero. These results agree with the proportion of deaths and animals with hemangiosarcoma in each dose group provided in Tables 5.2, 5.3 and 5.4. We found little evidence that baseline weight affects the drifts of the two hemangiosarcoma development processes and the death process since the 90% credible intervals for $\alpha_{E_{11}}$, $\alpha_{E_{21}}$ and $\alpha_{C_1}$ all included zero.

Posterior estimates of the operational time scale parameters are also provided in Table 5.5. Although the posterior mean of $\tau_1$ was less than 1, the 90% credible interval
of $\tau_1$ essentially contained 1 which implies that the degradation rate of the death process in periods when an animal only had a heart hemangiosarcoma was similar to the degradation rate in periods when an animal had no hemangiosarcomas. The posterior means of $\tau_2$ and $\tau_3$ were greater than 1, which implies that the degradation rate of the death process in periods when an animal had a liver hemangiosarcoma only or had both a liver and a heart hemangiosarcoma was bigger than the degradation rate of the death process in periods when an animal had no hemangiosarcomas. The posterior mean of $\tau_3$ was greater than the posterior mean of $\tau_2$, which implies that the degradation rate of the death process in periods when an animal had both a heart and a liver hemangiosarcoma was bigger than the degradation rate of the death process in periods when it had a liver hemangiosarcoma only.
Table 5.5: Analysis result of the B6C3F1/NCTR male mouse heart and liver hemangiosarcoma data by the informative and independent censoring models, as well as the model with right censored data.

<table>
<thead>
<tr>
<th>Para.</th>
<th>Informative Censoring Model</th>
<th>Independent Censoring Model</th>
<th>Model with Right Censored Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Post. Mean</td>
<td>90% Cred. I.</td>
<td>Post. Mean</td>
</tr>
<tr>
<td>$\alpha_{E10}$</td>
<td>0.756</td>
<td>(0.666, 0.846)</td>
<td>0.788</td>
</tr>
<tr>
<td>$\alpha_{E11}$</td>
<td>0.921E-04</td>
<td>(-0.308E-03, 0.486E-03)</td>
<td>0.812E-04</td>
</tr>
<tr>
<td>$\alpha_{E12}$</td>
<td>-0.360E-02</td>
<td>(-0.484E-02, -0.235E-02)</td>
<td>-0.364E-02</td>
</tr>
<tr>
<td>$y_{E10}$</td>
<td>4.090</td>
<td>(3.174, 5.223)</td>
<td>3.606</td>
</tr>
<tr>
<td>$\alpha_{E20}$</td>
<td>0.026</td>
<td>(-0.105, 0.172)</td>
<td>0.213</td>
</tr>
<tr>
<td>$\alpha_{E21}$</td>
<td>-0.291E-03</td>
<td>(-0.717E-03, 0.146E-03)</td>
<td>-0.305E-03</td>
</tr>
<tr>
<td>$\alpha_{E22}$</td>
<td>-0.262E-02</td>
<td>(-0.336E-02, -0.192E-02)</td>
<td>-0.240E-02</td>
</tr>
<tr>
<td>$y_{E20}$</td>
<td>224.100</td>
<td>(129.606, 302.203)</td>
<td>89.854</td>
</tr>
<tr>
<td>$\alpha_{C0}$</td>
<td>-0.030</td>
<td>(-0.037, -0.024)</td>
<td>$\alpha_{C1}$</td>
</tr>
<tr>
<td>$y_{C0}$</td>
<td>42.696</td>
<td>(39.821, 45.645)</td>
<td>$\tau_1$</td>
</tr>
</tbody>
</table>
To help better understand the estimates of $\tau_1$, $\tau_2$ and $\tau_3$, we provide the Kaplan-Meier Survival Estimates and Nelson-Aalen Cumulative Hazard Estimates for subjects with only one of the two hemangiosarcomas and with both of the two hemangiosarcomas versus those without any hemangiosarcoma in Figure 5.4. In our dataset there were 10 animals with only heart hemangiosarcoma, 56 animals with only liver hemangiosarcoma, and 7 animals with both heart and liver hemangiosarcomas. The Kaplan-Meier and Nelson-Aalen estimates of animals with only heart hemangiosarcoma and animals without any hemangiosarcoma were similar, which explains why the 90% credible interval for $\tau_1$ essentially contained 1. In contrast, compared to animals without any hemangiosarcomas, animals with only liver hemangiosarcoma or with both heart and liver hemangiosarcomas had Kaplan-Meier estimates that decreased faster and Nelson-Aalen estimates that increased faster late in study after hemangiosarcomas likely occurred, which explains why the 90% credible intervals for $\tau_2$ and $\tau_3$ did not contain 1.

In the middle panel of Table 5.5, we also list parameter estimates for the TR model presented in Section 5.2.1, which assumes independent censoring. The posterior means of the parameters for the heart hemangiosarcoma development process were similar across models. This makes sense because 1 was just outside the 90% credible interval for $\tau_1$, indicating a weak effect of heart hemangiosarcoma (alone) on the death process. However, things were different for the liver hemangiosarcoma development process: the posterior mean of $y_{E20}$ by the informative censoring model was much larger than that of the independent censoring model, and the posterior mean of $\alpha_{E20}$ by the independent censoring model was much larger than that of the informative censoring model. This big difference also makes sense because 1 was not in the 90%
Figure 5.4: Kaplan-Meier survival estimates and Nelson-Aalen cumulative hazard estimates for subjects with only one of the two hemangiosarcomas and with both of the two hemangiosarcomas versus those without any hemangiosarcoma, for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data (in days)
credible interval for $\tau_2$, indicating a strong effect of the liver hemangiosarcoma (alone) on the death process.

We compared the estimated survival curves for heart hemangiosarcoma onset from the two different models in Figure 5.5 and those for liver hemangiosarcoma in Figure 5.6. The curves correspond to the marginal survival function:

$$S(s_j|\cdot) = \Phi\left[\frac{y_0E_j + \mu_{Ej}s_j}{\sqrt{s_j^2/\lambda + s_j}} - \exp\left[-2y_0E_j\mu_{Ej} + \frac{2y_0^2E_j}{\lambda}\right]\Phi\left[\frac{\mu_{Ej}s_j - y_0E_j - \frac{2y_0E_js_j}{\lambda}}{\sqrt{s_j^2/\lambda + s_j}}\right]\right]$$

This is equivalent to the survival function for randomized drift model proposed by Aalen and Gjessing (2001). The estimated survival curves for the heart hemangiosarcoma onset from the informative censoring model were very close to (just slightly above) those from the independent censoring model; the curves were all decreasing over the first 35 days (approximately) and plateaued at probabilities greater than 0.9 for the rest of the observation period. Thus urethane caused a small number of heart hemangiosarcoma which occurred shortly after exposure to urethane. In addition, it appears that the onset of heart hemangiosarcoma didn’t affect survival. In contrast, for liver hemangiosarcoma, the estimated survival probabilities according to both models was around 0.7 at the end of the study for the highest dose dose (90 ppm) of urethane. For the 90 ppm group, the estimated survival curve from the independent censoring model began to drop around 200 days while the curve for the informative censoring model began to drop around 300 days. Thus the model accounting for informative censoring suggests that liver tumors begin to appear 100 days later than what the independent censoring model suggests.

If we compare these estimated survival curves to the Kaplan-Meier survival estimates for death in Figure 5.3, we can see that the Kaplan-Meier survival estimates
for death for the 90 ppm dose group of urethane keep dropping very slowly before around 400 days, also suggesting the early (before 100 days) onset of heart hemangiosarcoma won’t affect the death time much and this agrees with the estimate of \( \tau_1 \) whose 90% credible interval essentially included 1. However, these Kaplan-Meier survival estimates for death began to drop sharply after around 400 days, which was probably due in part to the onset of liver hemangiosarcoma in animals. This sharp dropping trend agrees with the estimate of \( \tau_2 \) which is greater than 1 (2.436) with 90% credible interval not including 1.

In both the informative and independent censoring models, the estimates of the precision parameter \( \lambda \) of the random effect \( b_i \) is around 26. At day 5 around when the estimated survival curves for heart hemangiosarcoma decreased most rapidly, the estimated correlation between the underlying heart and liver hemangiosarcoma development processes was roughly \( 5/(5 + 26) = 0.16 \). Thus during the time period when heart hemangiosarcoma occurred the most, the liver hemangiosarcoma development process was weakly correlated with the heart hemangiosarcoma development process which suggests that these two hemangiosarcomas were likely triggered by two different mechanisms.

Bailer and Piegorsch (1997) pointed out that some tumors are considered instantaneously lethal, like hemangiosarcomas, and thus the time to death and time to tumor are assumed equivalent. In this situation, it is possible to employ a simpler analysis by treating the time of death with a tumor as tumor onset time, while treating an animal dying during the study or sacrificed at the end of the study without a tumor as a right censored observation. Thus, we employed our TR model presented in Section 5.2.1 which assumes independent censoring on the same NTP heart and
liver hemangiosarcoma data but treated the current status data as right censored data. Other than the fact that right censored data were used, the other settings of the MCMC run were the same as that for the current status data models.

Parameters estimates from the model assuming right censored data are provided in the right panel of Table 5.5. Baseline weight had a negligible effect on the onset of both the heart and liver hemangiosarcomas since the 90% credible intervals for $\alpha_{E_{11}}$ and $\alpha_{E_{21}}$ both included zero; this result was consistent with the results from the current status data models. However, the posterior means of $\alpha_{E_{11}}$ and $\alpha_{E_{21}}$ were negative when the right censored data model was applied and they were positive in the current status data models. The right censored data model also provided a smaller negative effect of urethane dose on the onset of both heart and liver hemangiosarcomas than the current status data models, though conclusions about the presence or absence of an effect were the same. Another important difference between the right censored and current status data models was the remarkably different estimates of the initial states and the intercepts of the drifts. The right censored data model provided larger initial states ($y_{E_{10}}$ and $y_{E_{20}}$) and smaller intercepts ($\alpha_{E_{10}}$ and $\alpha_{E_{20}}$) which suggests a later time to tumor onset. The differences between the initial states and drifts make sense because the right censored data model assumes instantaneous tumor lethality and hence a very small gap between the time of tumor onset and time of death which usually happens late, while the current data models did not assume instantaneous lethality.

Figures 5.5 and 5.6 contain the estimated survival curves for heart and liver hemangiosarcoma onset assuming right censored data. For the heart hemangiosarcoma,
the estimated survival curves from the right censored data model demonstrated totally different patterns than the informative and independent censoring models: the curves stayed around 1 from the start of the study until around 400 days and at which point they began to drop sharply; at around 600 days, they dropped below the curves from the informative and independent censoring models. In contrast to the curves for heart hemangiosarcoma, the estimated survival curves for time to liver hemangiosarcoma from the right censored data model demonstrated patterns similar to those from the informative censoring models except for an additional 100 day lag before the survival curves began to drop. In contrast to the Kaplan-Meier survival estimates for death in Figure 5.3, we can see that the Kaplan-Meier survival estimates for death and the estimated survival curves from the right censored data model both began to drop sharply almost at the same time (around 400 days). This is not surprising since by treating the data as right censored data, we assumed liver hemangiosarcoma was instantaneously lethal.

5.5 Concluding Remarks

In this chapter we proposed a Bayesian threshold regression model for multivariate current status data accounting for informative censoring. In this model, we used a random effect shared by the drifts of the processes underlying the events of interest to account for the correlation of event times. We accounted for informative censoring using a novel operational time scale dependent on the event status of a subject. The drifts of the processes underlying the events of interest were modeled as linear functions of covariates and the random effect. The drift of the censoring process in the operational time scale was modeled as a linear function of covariates. Our
Figure 5.5: Estimated survival probability curves at the mean baseline weight of 14.102g by the informative and independent censoring models, as well as the right censored data model for heart hemangiosarcoma, for the B6C3F1/NCTR mouse heart and liver hemangiosarcoma data (in days).

Bayesian threshold regression model for multivariate current status data shares many same advantages as our model for univariate current status data: the model does not assume proportional hazards, does not assume the censoring process is constant across subjects, and is defined on a natural continuous time scale.

We conducted simulation studies on data with different levels of dependent censoring, to verify the effectiveness of our model. We compared the joint model accounting for informative censoring with a model assuming independent censoring and found that the joint model resulted in more accurate estimates of regression coefficients.
and survival probabilities when the censoring depended on the time of event. When
the censoring time didn’t depend on the time of event, both the model accounting
for informative censoring and the model assuming independent censoring performed
equally well.

We applied both the joint model accounting for informative censoring and the
model assuming independent censoring to data from a National Toxicology Program
(NTP) 2-year rodent carcinogenicity study of the impact of urethane on heart and
liver hemangiosarcoma onsets. The estimates of regression coefficients for the latent
heart hemangiosarcoma development process by the two models were similar, while those for the latent liver hemangiosarcoma development process by the two models were quite different (especially for the initial state parameter and the intercept parameter of the drift). Survival curves estimated by the informative censoring model for heart hemangiosarcoma onset were very close to those estimated by the independent censoring model in each of the four dose groups of urethane. While for liver hemangiosarcomas, the estimated survival curves by the informative censoring model dropped late than those by the independent censoring model. We also compared the estimation results of our models with right censored data model. By comparing the estimated survival curves, it suggests that both the heart and liver hemangiosarcomas are not instantaneously lethal tumors, but the liver hemangiosarcoma is more rapidly lethal than the heart hemangiosarcoma. The difference in lethalities between the heart and liver hemangiosarcomas is also evident by the difference in the estimates of the operational time scale parameters corresponding to these two hemangiosarcomas.

A limitation of this model for multivariate current status data accounting for informative censoring intrinsically assumes that each subject is subject to dependent censoring, which is unlike the model for univariate current status data in Chapter 4. This assumption could be relaxed by an extension of our model to incorporate binary latent variables indicating if the censoring process depends on the event process, similar to the approaches that are used in Chapter 4 or Dunson and Dinse (2002). A more detailed discussion is provided in Chapter 6.
Chapter 6: CONCLUDING REMARKS

6.1 Overview

This dissertation has described Bayesian threshold regression models for current status data subject to informative censoring. To account for informative censoring, we propose to jointly model time to event and time to censoring using latent Wiener diffusion processes.

In Chapter 4, we proposed a Bayesian threshold regression model for univariate current status data accounting for informative censoring. Previous regression models for current status data either assumed independent censoring which may not be a reasonable assumption in some applications, or accounted for informative censoring but with strong assumptions. Among the models accounting for informative censoring, Dunson and Dinse (2002) proposed a Bayesian model motivated by tumorigenicity studies. The authors accounted for tumor lethality using a latent binary variable, indicating whether or not death was caused by tumor. Similar to the model proposed by Dunson and Dinse, we augment a latent binary variable for each subject which indicates if the process driving the event of interest and that driving the event of censoring are related, and informative censoring is accounted for by modeling time to censoring using two different latent censoring processes: one is independent of the
event of interest and the other dependent. The drifts of each process are modeled as linear functions of covariates, and hence covariate effects on the distributions of either event and censoring time can be estimated through drifts.

In Chapter 5, we proposed a Bayesian threshold regression model for multivariate current status data accounting for informative censoring. In this model, we used a random effect shared by the drifts of the processes underlying the events of interest to account for the correlation of event times. We accounted for informative censoring using a novel operational time scale dependent on the event status of a subject. The drifts of the processes underlying the events of interest were modeled as linear functions of covariates and the random effect. And the drift of the censoring process in the operational time scale was modeled as a linear function of covariates.

### 6.2 Strengths and Limitations of Proposed Models

The Bayesian threshold regression model for univariate current status data in Chapter 4 and that for multivariate current status data in Chapter 5 share many same advantages: both models do not assume proportional hazards; in contrast to Dunson and Dinse (2002), both models do not assume the censoring process is constant across subjects and are defined in a natural continuous time scale instead of a discrete time scale.

One difference that needs to be noted between the two models we proposed is the method used to account for dependent censoring. In the model for univariate current status data in Chapter 4, we augment a latent binary variable for each subject which indicates if the event affects the censoring process and hence dependent censoring may not exist for each subject. However, in the model for multivariate current status
data in Chapter 5, we used operational time scale parameters to measure the impact of different events on the censoring process. Thus this approach intrinsically assumes that each subject is subject to dependent censoring, which is unlike the model for univariate current status data in Chapter 4. Because of this difference, the model for multivariate current status data in Chapter 5 is best suited for studies with events certain to affect the censoring process. An example of such study is the one we used in Section 5.4: the animal carcinogenicity study of the impact of urethane on onsets of heart and liver hemangiosarcomas, which are regarded as lethal tumors.

Though motivated by tumorigenicity studies, our informative censoring models can be applied to other types of current status data where censoring time depends on event time. In cross-sectional epidemiology studies, such data arise when the examination time corresponds to a visit to the doctor due to symptoms related to the disease of interest. For example, Dunson and Dinse (2002) mentioned that their informative censoring model could be applied to studies of uterine leiomyoma (fibroids) in which the left censoring time for some women corresponds to a doctor’s visit after they began to have some symptoms related to leiomyoma, such as bleeding or back pain. To avoid biased inferences on the distribution of leiomyoma onset time, the dependence of the examination time on the leiomyoma onset time needs to be taken into account and hence our informative censoring models could be applied to these data.

A limitation of the model for univariate current status data in Chapter 4 is that convergence problems can occur at small sample sizes. We explicitly model a separate death process for animals whose tumor affects survival. If the informative censoring rate (measured by the $\pi$ parameter) and the number of events (e.g., animals with
tumor) are both small, this process is weakly identifiable, resulting in unstable estimates. However, in these scenarios, the stability can be improved by simplifying the model for the drift of the tumor-related death process (i.e., including less predictors in the model for the tumor-related death process after tumor onset), or by assigning stronger priors to the $\pi$ parameter or model parameters in the death process after tumor onset.

6.3 Future Research

Some extensions of our methods could be interesting topics of future research. For the TR model for current status data accounting for informative censoring in Section 5.2.2, the dependence of the censoring process on the event process was introduced by the operational time scale parameters. This approach intrinsically assumes that each subject is subject to dependent censoring. To relax this assumption, one could include binary latent variables indicating if the censoring process depends on the event process, similar to the approaches that are used in Section 4.2.2 or Dunson and Dinse (2002). This approach would allow the censoring events of some subjects to be totally unrelated to their events of interest. One potential extension of our model utilizing this approach would be to model $\tau_l$’s as a function of a binary latent variable, $R_l$ ($R_l = 1$: censoring process depends on the event process; $R_l = 0$: otherwise):

$$\tau_l = (1 - R_l) + R_l\gamma_l$$  \hspace{1cm} (6.1)

where $\gamma_l$ is the degradation rate of the censoring process corresponding to the $l$th of the $M$ possible event states if $R_l$ is 1. If $R_l$ is 0, then $\tau_l$ is 1, which means the degradation rate of the censoring process corresponding to the $l$th event state equals
to that of the first event state. The value of $R_t$ for each subject is augmented during MCMC.

Another research direction is to incorporate into our models observable time-varying marker data which are sometimes available. The use of marker data which are correlated with the latent process underlying time to event can increase the precision of parameter estimates. Tong et al. (2008) explored the joint modeling of the latent process with marker data but the authors did not account for informative censoring. It is of interest to develop a model with the flexibility to incorporate marker data while accounting for informative censoring at the same time. It is important to point out that, when incorporating such time-varying marker data, the correlation between the marker process and the latent process underlying time to event could change over time. For instance, in the animal tumorigenicity study analysis conducted by Tong et al. (2008), the authors used the body weight of animals as the observable marker for the liver tumor development process assuming a constant correlation between the two processes. However, the correlation between body weight and liver tumor development should be weak at the beginning of the study because no liver tumor occurs at this time. However, as cancer develops, serious weight loss could occur hence the correlation between body weight and liver tumor development is probably strong. Hence the assumption of a constant correlation between the marker and the event process is not a reasonable assumption in this setting and future models should remove this restriction.

Another extension is to incorporate information from historical control data from NTP studies to elicit informative priors. When the sample size is small, this approach is particularly useful to improve the precision of parameter estimates. Dunson and
Dinse (2001) proposed a mixed effects model to include the individual animal data from several studies into a single analysis. We could extend our models to use a similar approach for elicitation of informative prior distributions; this extension could be especially useful for the univariate current status model for informative censoring described in Chapter 4, to alleviate convergence issues due to small sample size.

Finally, assessment of fit of our threshold regression models for current status data is another research area remain to be explored. Currently for right censored data, the assessment of fit of the threshold regression models is simply addressed by comparing the estimated survival curves with the Kaplan-Meier estimates. A more formal model diagnostic approach (e.g., an approach based on residuals) would be of interest for the threshold regression models, either for the right censored or current status data. Formal methods for validating predictions from our models would also be an interesting area of future research including possible extensions to correct biases as explained in Bayarri et al. (2007).
Appendix A: CHAPTER 4 SUPPLEMENTAL MATERIAL

A.1 Model Under Independent Censoring: Posterior Computation and Data Augmentation

A.1.1 MCMC Algorithm

Our MCMC algorithm uses Gibbs, Metropolis and rejection sampling steps to sample the parameters ($\alpha$ and $y_0$) and latent variables ($\{s_i : d_i = 1\}$ and $\{y_i : d_i = 0\}$) based on their respective full conditional posterior distributions. The necessary conditional distributions as well as a summary of the sampling algorithm at each iteration are described as follows.

**Step 1.** For left censored subjects, sample event time $S_i$ given observation time $U_i$ from the following inverse Gaussian distribution truncated to the right at $u_i$ using the inverse c.d.f. method:

$$f(s_i | \cdot) = \frac{y_0}{\sqrt{2\pi\sigma^2 s_i}} \exp\left[-\frac{(y_0 + (x'_i \alpha) s_i)^2}{2\sigma^2 s_i}\right] I(0 < s_i \leq u_i), \quad (A.1)$$

where \( | \cdot \) stands for conditioning on all other parameters.
**Step 2.** For subjects whose event time is right censored, sample the process value $y_i$ at the right censoring time using the following kernel:

$$K(y_i) = \frac{1}{\sqrt{2\pi} \sigma u_i} \exp \left[-\frac{1}{2\sigma^2 u_i} \left(y_i - y_0 - (x_i' \alpha) u_i \right)^2 \right] \left(1 - \exp \left[\frac{-2y_0 y_i}{\sigma^2 u_i} \right]\right) I(y_i > 0) \tag{A.2}$$

We used a simplified version of the rejection sampling algorithm proposed by Pennell *et al.* (2010), which used one envelope function instead of two different envelope functions. The steps to implement the rejection sampling algorithm are as follows:

1. Find the unique mode $m_i$ of (A.2) using a numerical search algorithm and let the covering function $g(y_i)$ be $N_+(m_i, u_i)$ where $N_+ (\cdot)$ is a standard normal distribution truncated to the left at 0.

2. Generate $y_i \sim g(y_i)$.

3. Generate $R \sim Unif(0, 1)$.

4. If $R g(y_i) < K(y_i)$, accept $y_i$; otherwise reject $y_i$ and return to the 2nd step.

**Step 3.** Sample the regression coefficient vector $\alpha$ for the drift parameter covariates directly from its full conditional distribution which is a multivariate normal with mean

$$\mu_{\alpha \text{post}} = \left(\sigma^{-2} \left(\sum_{i: d_i = 1} s_i x_i' x_i' + \sum_{i: d_i = 0} u_i x_i' x_i' \right) + \Sigma_{\alpha \text{prior}}^{-1}\right)^{-1} \times \left(-\sigma^{-2} \left(\sum_{i: d_i = 1} x_i y_0 + \sum_{i: d_i = 0} x_i (y_0 - y_i) \right) + \Sigma_{\alpha \text{prior}}^{-1} \mu_{\alpha \text{prior}}\right)$$

and covariance matrix

$$\Sigma_{\alpha \text{post}} = \left(\sigma^{-2} \left(\sum_{i: d_i = 1} s_i x_i' x_i' + \sum_{i: d_i = 0} u_i x_i' x_i' \right) + \Sigma_{\alpha \text{prior}}^{-1}\right)^{-1}$$
where $\mu_{\alpha_{\text{prior}}}$ is the mean vector and $\Sigma_{\alpha_{\text{prior}}}$ is the covariance matrix of the multivariate normal prior distribution for $\alpha$.

**Step 4.** Use the Metropolis-Hastings random walk algorithm to sample from the full conditional posterior distribution of $y_0$, which has kernel:

$$K(y_0 | \cdot) = \exp \left[ -\frac{(\log y_0 - \mu_{\log y_0 \text{prior}})^2}{2\sigma^2_{\log y_0 \text{prior}}} \right] \times \prod_{d_i=1} y_0 \frac{1}{\sqrt{2\pi\sigma^2 s_i}} \times \exp \left[ -\frac{(y_0 + (x_i'\alpha)s_i)^2}{2\sigma^2 s_i} \right] \times \prod_{d_i=0} \frac{1}{\sqrt{2\pi\sigma^2 u_i}} \exp \left[ -\frac{(y_i - y_0 - (x_i'\alpha)u_i)^2}{2\sigma^2 u_i} \right] \left( 1 - \exp \left[ \frac{-2y_0y_i}{\sigma^2 u_i} \right] \right)$$

where $\mu_{\log y_0 \text{prior}}$ is the mean and $\sigma^2_{\log y_0 \text{prior}}$ is the variance of normal prior distribution for $\log(y_0)$.

### A.2 Model Under Informative Censoring: Posterior Computation and Data Augmentation

#### A.2.1 Likelihood Function

Using equations 3.1, 4.3 and 4.4 in Chapters 3 and 4, the likelihood contributions for the six different combinations of $d$ (indicates if event of interest occurs), $\delta$ (indicates if the censoring event driven by a stochastic process occurs) and $R$ (indicates if the censoring process is upgraded to a new process at time of event occurrence) are below. Note that the subscript condition indicator involving $d$, $\delta$ and $R$ means that the likelihood contribution corresponds to a subject with this condition satisfied.
\[ L_{1i} = L_{(d_i=0,\delta_i=0)} = f_Y(y_{Ei} \mid \mu_{Ei}, \sigma^2_E, y_{E0}, u_i) \times f_Y(y_{C1i} \mid \mu_{C1i}, \sigma^2_{C1}, y_{C10}, u_i) \]

\[ L_{2i} = L_{(d_i=0,\delta_i=1)} = f_Y(y_{Ei} \mid \mu_{Ei}, \sigma^2_E, y_{E0}, u_i) \times f_T(u_i \mid \mu_{C1i}, \sigma^2_{C1}, y_{C10}) \]

\[ L_{3i} = L_{(d_i=1,\delta_i=0, R_i=0)} = (1 - \pi) \times f_S(s_i \mid \mu_{Ei}, \sigma^2_E, y_{E0}, u_i) \times f_Y(y_{C1i} \mid \mu_{C1i}, \sigma^2_{C1}, y_{C10}, s_i) \times f_Y(y_{C1i} \mid \mu_{C1i}, \sigma^2_{C1}, y_{C10}, u_i) \]

\[ L_{4i} = L_{(d_i=1,\delta_i=0, R_i=1)} = \pi \times f_S(s_i \mid \mu_{Ei}, \sigma^2_E, y_{E0}, u_i) \times f_Y(y_{C1i} \mid \mu_{C1i}, \sigma^2_{C1}, y_{C10}, s_i) \times f_Y(y_{C2i} \mid \mu_{C2i}, \sigma^2_{C2}, y_{C20}, u_i - s_i) \]

\[ L_{5i} = L_{(d_i=1,\delta_i=1, R_i=0)} = (1 - \pi) \times f_S(s_i \mid \mu_{Ei}, \sigma^2_E, y_{E0}, u_i) \times f_T(u_i \mid \mu_{C1i}, \sigma^2_{C1}, y_{C10}) \]

\[ L_{6i} = L_{(d_i=1,\delta_i=1, R_i=1)} = \pi \times f_S(s_i \mid \mu_{Ei}, \sigma^2_E, y_{E0}, u_i) \times f_Y(y_{C1i} \mid \mu_{C1i}, \sigma^2_{C1}, y_{C10}, s_i) \times f_T(u_i - s_i \mid \mu_{C2i}, \sigma^2_{C2}, y_{C20}) \]

Thus the likelihood likelihood function is:

\[
\prod_{i=1}^n L_{1i}^{(1-d_i)(1-\delta_i)} \times L_{2i}^{1-d_i} \times L_{3i}^{d_i(1-\delta_i)(1-R_i)} \times L_{4i}^{d_i(1-\delta_i)R_i} \times L_{5i}^{d_i\delta_i(1-R_i)} \times L_{6i}^{d_i\delta_iR_i}
\]

(A.3)

### A.2.2 MCMC Algorithm

Our MCMC algorithm uses Gibbs, Metropolis and rejection sampling steps to sample the parameters \((\alpha_E, y_{E0}, \alpha_{C1}, y_{C10}, \alpha_{C2}, y_{C20})\) and latent variables \(\{s_i : d_i = 1\}, \{y_{Ei} : d_i = 0\}, \{R_i : d_i = 1\}, y_{C1i}, y_{C2i}\) based on their respective full conditional posterior distributions. The necessary conditional distributions as well as a summary of the sampling algorithm at each iteration are described as follows.

**Step 1.** For subjects with a left censored event, sample \(R_i\) from its full conditional distribution which is Bernoulli\((p)\) with

\[
p = \frac{A}{A + B}
\]

119
where

\[ A = \pi \times \left[ \frac{yc_{20}}{\sqrt{2\pi\sigma^2_{C2}}(u_i - s_i)^2} \exp \left[ -\frac{(yc_{20} + (x'_{C2i}\alpha_{C2})(u_i - s_i))^2}{2\sigma^2_{C2}(u_i - s_i)} \right] \right]^{\delta_i} \]

\[ \times \left[ \frac{1}{\sqrt{2\pi\sigma^2_{C1}u_i}} \exp \left[ -\frac{(yc_{1i} - y_{C10} - (x'_{C1i}\alpha_{C1})u_i)^2}{2\sigma^2_{C1}u_i} \right] \right]^{(1-\delta_i)} \]

\[ \times \left[ 1 - \exp \left( -\frac{2yc_{10}yc_{1i}}{\sigma^2_{C1}u_i} \right) \right]^{(1-\delta_i)} \]

and

\[ B = (1 - \pi) \times \left[ \frac{yc_{10}}{\sqrt{2\pi\sigma^2_{C1}u_i}} \exp \left[ -\frac{(yc_{10} + (x'_{C11}\alpha_{C1})u_i)^2}{2\sigma^2_{C1}u_i} \right] \right]^{\delta_i} \]

\[ \times \left[ \frac{1}{\sqrt{2\pi\sigma^2_{C1}u_i}} \exp \left[ -\frac{(yc_{1i} - y_{C10} - (x'_{C1i}\alpha_{C1})u_i)^2}{2\sigma^2_{C1}u_i} \right] \right]^{(1-\delta_i)} \]

\[ \times \left[ 1 - \exp \left( -\frac{2yc_{10}yc_{1i}}{\sigma^2_{C1}u_i} \right) \right]^{(1-\delta_i)} \]

\[ \times \frac{1}{\Phi \left[ (yc_{10} + (x'_{C11}\alpha_{C1})s_i) / \sqrt{\sigma^2_{C1}s_i} \right] - \exp \left( -\frac{2yc_{10}(x'_{C11}\alpha_{C1})}{\sigma^2_{C1}} \right) \Phi \left[ (x'_{C11}\alpha_{C1})s_i - y_{C10} / \sqrt{\sigma^2_{C1}s_i} \right] \]

**Step 2.** For subjects with a left censored event, sample \( S \) from its full conditional posterior distribution by a Metropolis-Hastings random walk algorithm. For subjects that experienced a censoring event (i.e., \( i : \delta_i = 1 \)), the full conditional posterior distribution of \( S \) is proportional to

\[
K(s_i | \delta_i = 1, \cdot) = s_i^{-3/2} \exp \left[ -\frac{(y_{E0} + (x'_{Ei}\alpha_E)s_i)^2}{2\sigma^2_Es_i} \right] \\
\times \left[ s_i^{-1/2} \exp \left[ -\frac{(yc_{1i} - y_{C10} - (x'_{C1i}\alpha_{C1})s_i)^2}{2\sigma^2_{C1}s_i} \right] \right]^{R_i} \\
\times \left[ 1 - \exp \left( -\frac{2yc_{10}yc_{1i}}{\sigma^2_{C1}s_i} \right) \right]^{R_i} \\
\times \left[ (u_i - s_i)^{-3/2} \exp \left[ -\frac{(yc_{20} + (x'_{C2i}\alpha_{C2})(u_i - s_i))^2}{2\sigma^2_{C2}(u_i - s_i)} \right] \right]^{R_i} \\
\times I(0 < s_i \leq u_i).
\]
For subjects that didn’t experience a censoring event (i.e., \( i : \delta_i = 0 \)), the full conditional posterior distribution of \( S \) is proportional to

\[
K(s_i \mid \delta_i = 0, \cdot) = s_i^{-3/2} \exp \left[ -\frac{(y_{E0}) + (x'_{Ei} \alpha_E) s_i^2}{2 \sigma^2_E s_i} \right] \\
\times \left[ s_i^{-1/2} \exp \left[ -\frac{(y_{C1i} - y_{C10} - (x'_{C1i} \alpha_{C1}) s_i^2}{2 \sigma^2_{C1} s_i} \right] \right]^{R_i} \\
\times \left[ 1 - \exp \left[ -\frac{2y_{C10} y_{C1i}}{\sigma^2_{C1} s_i} \right] \right]^{R_i} \\
\times \left[ (u_i - s_i)^{-1/2} \exp \left[ -\frac{(y_{C2i} - y_{C20} - (x'_{C2i} \alpha_{C2}) (u_i - s_i))^2}{2 \sigma^2_{C2} (u_i - s_i)} \right] \right]^{R_i} \\
\times \left[ 1 - \exp \left[ -\frac{2y_{C20} y_{C2i}}{\sigma^2_{C2} (u_i - s_i)} \right] \right]^{R_i} \\
\times I(0 < s_i \leq u_i).
\]

**Step 3.** For subjects with right censored event, sample the process value \( y_{Ei} \) at right censoring time from

\[
K(y_{Ei} \mid \cdot) = \frac{1}{\sqrt{2 \pi \sigma^2_E u_i}} \exp \left[ -\frac{(y_{Ei} - y_{E0}) - (x'_{Ei} \alpha_E) u_i)^2}{2 \sigma^2_E u_i} \right] \\
\times \left( 1 - \exp \left[ -\frac{2y_{E0} y_{Ei}}{\sigma^2_E u_i} \right] \right) I(y_{Ei} > 0)
\]

Note that \( y_{Ei} \) is sampled using the data augmentation method described in Step 2 in Section A.1.1.

**Step 4.** Sample the regression coefficient vector \( \alpha_E \) for the drift parameter covariates in the event process directly from its full conditional distribution which is a multivariate normal with mean

\[
\mu_{\alpha_{E\text{post}}} = \left( \sigma^{-2}_E \left( \sum_{i:d_i=1} s_i x_{Ei} x'_{Ei} + \sum_{i:d_i=0} u_i x_{Ei} x'_{Ei} \right) + \Sigma^{-1}_{\alpha_{E\text{prior}}} \right)^{-1} \\
\times \left( -\sigma^{-2}_E \left( \sum_{i:d_i=1} x_{Ei} y_{E0} + \sum_{i:d_i=0} x_{Ei} (y_{E0} - y_{Ei}) \right) + \Sigma^{-1}_{\alpha_{E\text{prior}}} \mu_{\alpha_{E\text{prior}}} \right)
\]

\[121\]
and covariance matrix

\[ \Sigma_{\alpha_{E\text{post}}} = \left( \sigma_{E}^{-2} \left( \sum_{i:d_i=1} s_i x_{E_i} x_{E_i}' + \sum_{i:d_i=0} u_i x_{E_i} x_{E_i}' \right) + \Sigma_{\alpha_{E\text{prior}}} \right)^{-1} \]

where \( \mu_{\alpha_{E\text{prior}}} \) is the mean vector and \( \Sigma_{\alpha_{E\text{prior}}} \) is the covariance matrix of the multivariate normal prior distribution for \( \alpha_E \).

**Step 5.** Use the Metropolis-Hastings random walk algorithm to sample initial state \( y_{E0} \) of the event process from its full conditional posterior distribution, which has the kernel:

\[
K(y_{E0} | \cdot) = \exp \left[ -\frac{(\log y_{E0} - \mu_{\log y_{E0\text{prior}}})^2}{2\sigma^2_{\log y_{E0\text{prior}}}} \right] \times \prod_{d_i=1} \frac{y_{E0}}{\sqrt{2\pi}\sigma_{E_i}^2 s_i} \times \exp \left[ -\frac{(y_{E0} + (x_{E_i}'\alpha_E) s_i)^2}{2\sigma_{E_i}^2 s_i} \right] \times \prod_{d_i=0} \frac{1}{\sqrt{2\pi}\sigma_{E_i}^2 u_i} \exp \left[ -\frac{(y_{Ei} - y_{E0} - (x_{E_i}'\alpha_E) u_i)^2}{2\sigma_{E_i}^2 u_i} \right] \times \prod_{d_i=0} \left( 1 - \exp \left[ \frac{-2y_{E0} y_{Ei}}{\sigma_{E_i}^2 u_i} \right] \right)
\]

where \( \mu_{\log y_{E0\text{prior}}} \) is the mean and \( \sigma^2_{\log y_{E0\text{prior}}} \) is the variance of normal prior distribution for \( \log(y_{E0}) \).

**Step 6.** For subjects who experience the event of interest (i.e., \( d_i = 1 \)), if the event affects the censoring process (i.e., \( R_i = 1 \)), the process value \( y_{C1i} \) at event time \( s_i \) is sampled using the following kernel:

\[
\exp \left[ -\frac{(y_{C1i} - y_{C10} - (x_{C1i}'\alpha_{C1}) s_i)^2}{2\sigma_{C1i}^2 s_i} \right] \left( 1 - \exp \left[ \frac{-2y_{C10} y_{C1i}}{\sigma_{C1i}^2 s_i} \right] \right) I(y_{C1i} > 0).
\]

If the event doesn’t affect the event of interest (i.e., \( R_i = 1 \)), but the random censoring event doesn’t occur (i.e., \( \delta_i = 0 \)), \( y_{C1i} \) is sampled from the following kernel:

\[
\exp \left[ -\frac{(y_{C1i} - y_{C10} - (x_{C1i}'\alpha_{C1}) u_i)^2}{2\sigma_{C1i}^2 u_i} \right] \left( 1 - \exp \left[ \frac{-2y_{C10} y_{C1i}}{\sigma_{C1i}^2 u_i} \right] \right) I(y_{C1i} > 0)
\]
Note that $y_{C1i}$ is sampled using the data augmentation method described in Step 2 in Section A.1.1.

**Step 7.** Sample the regression coefficient vector $\alpha_{C1}$ for the drift parameter covariates in the censoring process unrelated to event directly from its full conditional distribution which is a multivariate normal with mean

$$\mu_{\alpha_{C1\text{post}}} = \left(\sigma_{C1}^{-2}(A_{\alpha_{C1}}) + \Sigma_{\alpha_{C1\text{prior}}}^{-1}\right)^{-1} \times \left(-\sigma_{C1}^{-2}(B_{\alpha_{C1}}) + \Sigma_{\alpha_{C1\text{prior}}^{-1}}\mu_{\alpha_{C1\text{prior}}}\right)$$

and covariance matrix

$$\Sigma_{\alpha_{C1\text{post}}} = \left(\sigma_{C1}^{-2}(A_{\alpha_{C1}}) + \Sigma_{\alpha_{C1\text{prior}}}^{-1}\right)^{-1}$$

where

$$A_{\alpha_{C1}} = \sum_{i:R_i=1} s_i x_{C1i} x_{C1i}^T + \sum_{i:d_i=1,R_i=0} u_i x_{C1i} x_{C1i}^T + \sum_{i:d_i=0} u_i x_{C1i} x_{C1i}^T$$

$$B_{\alpha_{C1}} = \sum_{i:R_i=1} x_{C1i}(y_{C10} - y_{C1i}) + \sum_{i:d_i=1,\delta_i=1,R_i=0} x_{C1i} y_{C10} + \sum_{i:d_i=1,\delta_i=0,R_i=0} x_{C1i} y_{C10} + \sum_{i:d_i=0,\delta_i=1} x_{C1i}(y_{C10} - y_{C1i})$$

and $\mu_{\alpha_{C1\text{prior}}}$ is the mean vector and $\Sigma_{\alpha_{C1\text{prior}}}$ is the covariance matrix of the multivariate normal prior distribution for $\alpha_{C1}$.

**Step 8.** Use the Metropolis-Hastings random walk algorithm to sample the initial state $y_{C10}$ of the censoring process unrelated to event from its full conditional posterior.
distribution, which has the kernel:

\[
K(y_{C10} | \cdot) = \exp \left[ - \frac{(\log y_{C10} - \mu_{\log y_{C10 prior}})^2}{2\sigma^2_{\log y_{C10 prior}}} \right] 
\times \prod_{R_i=1} \exp \left[ - \frac{(y_{C1i} - y_{C10} - (x'_{C1i} \alpha_{C1}) s_i)^2}{2\sigma^2_{C1_i} s_i} \right] 
\times \prod_{d_i=1, \delta_i=1, R_i=0} y_{C10} \exp \left[ - \frac{(y_{C1i} + (x'_{C1i} \alpha_{C1}) u_i)^2}{2\sigma^2_{C1_i} u_i} \right] 
\times \prod_{d_i=1, \delta_i=0, R_i=0} \left( 1 - \exp \left[ - \frac{-2 y_{C10} y_{C1i}}{\sigma^2_{C1_i} u_i} \right] \right) 
\times \prod_{d_i=0, \delta_i=1} y_{C10} \exp \left[ - \frac{(y_{C1i} + (x'_{C1i} \alpha_{C1}) u_i)^2}{2\sigma^2_{C1_i} u_i} \right] 
\times \prod_{d_i=0, \delta_i=0} \exp \left[ - \frac{(y_{C1i} - y_{C10} - (x'_{C1i} \alpha_{C1}) u_i)^2}{2\sigma^2_{C1_i} u_i} \right] 
\times \prod_{d_i=0, \delta_i=0} \left( 1 - \exp \left[ - \frac{-2 y_{C10} y_{C1i}}{\sigma^2_{C1_i} u_i} \right] \right) 
\]

where \(\mu_{\log y_{C10 prior}}\) is the mean and \(\sigma^2_{\log y_{C10 prior}}\) is the variance of normal prior distribution for \(\log(y_{C10})\).

**Step 9.** For the subjects whose censoring is affected by the event of interest (i.e., \(R_i = 1\)), but the random censoring event doesn’t occur by time \(u_i\), sample process level \(y_{C2i}\) using the kernel:

\[
K(y_{C2i} | \cdot) = \exp \left[ - \frac{(y_{C2i} - y_{C20} - (x'_{C2i} \alpha_{C2}) (u^*_i)^2}{2\sigma^2_E(u^*_i)} \right] \times \left( 1 - \exp \left[ - \frac{-2 y_{E0} y_{Ei}}{\sigma^2_E(u^*_i)} \right] \right) 
\times I(y_{C2i} > 0)
\]

where \(u^*_i = u_i - s_i\). Note that \(y_{C2i}\) is sampled using the data augmentation method described in Step 2 in Section A.1.1.
Step 10. Sample $\alpha_{C2}$ directly from its full conditional distribution which is a multivariate normal with mean

$$\mu_{\alpha_{C2}\text{post}} = \left(\sigma_{C2}^{-2} \sum_{i: R_i = 1} (u_i - s_i)x_{C2i}x_{C2i}' + \Sigma_{\alpha_{C2}\text{prior}}^{-1}\right)^{-1} \times \left( - \sigma_{C2}^{-2} \sum_{i: d_i = 1, \delta_i = 1} x_{C2i}y_{C20} + \sum_{i: \delta_i = 1, R_i = 1} x_{C2i}(y_{C20} - y_{C2i}) \right) + \Sigma_{\alpha_{C2}\text{prior}}^{-1} \mu_{\alpha_{C2}\text{prior}}$$

and covariance matrix

$$\Sigma_{\alpha_{C2}\text{post}} = \left(\sigma_{C2}^{-2} \sum_{i: R_i = 1} (u_i - s_i)x_{C2i}x_{C2i}' + \Sigma_{\alpha_{C2}\text{prior}}^{-1}\right)^{-1}$$

where $\mu_{\alpha_{C2}\text{prior}}$ is the mean vector and $\Sigma_{\alpha_{C2}\text{prior}}$ is the covariance matrix of the multivariate normal prior distribution for $\alpha_{C2}$.

Step 11. Use the Metropolis-Hastings random walk algorithm to sample the initial state $y_{C20}$ of the censoring process related to event from its full conditional posterior distribution, which has the kernel:

$$K(y_{C20} \mid \cdot) = \exp \left[ - \frac{(\log y_{C20} - \mu_{\log y_{C20}\text{prior}})^2}{2\sigma_{\log y_{C20}\text{prior}}^2} \right] \times \prod_{\delta_i = 1, R_i = 1} y_{C20} \exp \left[ - \frac{(y_{C20} + (x_{C2i}'\alpha_{C2})u_i^*)^2}{2\sigma_{C2}^2u_i^*} \right] \times \prod_{\delta_i = 0, R_i = 1} \exp \left[ - \frac{(y_{C2i} - y_{C20} - (x_{C2i}'\alpha_{C2})u_i^*)^2}{2\sigma_{C2}^2u_i^*} \right] \times \prod_{\delta_i = 0, R_i = 1} \left( 1 - \exp \left[ - \frac{-2y_{C20}y_{C2i}}{\sigma_{C2}^2u_i^*} \right] \right)$$

where $\mu_{\log y_{C20}\text{prior}}$ is the mean and $\sigma_{\log y_{C20}\text{prior}}^2$ is the variance of normal prior distribution for $\log(y_{C20})$. 

125
A.2.3 Simulation Results for the Parameters of Death Processes

The simulation results for the parameters of the censoring processes are given in Table A.1. From the table we can see that the biases of the parameters in $Y_{C1}(t)$ (i.e., the censoring process independent of the event process) are close to zero. However, the biases and sample standard deviations of the posterior means of the parameters in $Y_{C2}(t)$ (i.e., the censoring process dependent of the event process) and the parameter $\pi$ were pretty large; the former is due to the informative priors for $\pi$ and $\alpha_{C2}$ that are not centered on truth and the latter is due to the small number of observations (usually) following $Y_{C2}(t)$; for example, when the true $\pi$ is 0.5 and there are 120 events in a sample of 200 subjects (the average event rate in the simulated datasets), then there are only $120 \times 0.5 = 60$ subjects that would be expected follow $Y_{C2}(t)$ as opposed to the 140 subjects who wouldn’t be expected to follow $Y_{C2}(t)$. Furthermore, the parameter $\pi$ was underestimated for the scenarios with high true $\pi$ values (i.e., $\pi=.9$, .75 and .5) and overestimated for the scenario with low true $\pi$ value (i.e., $\pi=.25$). This is expected since we assigned a moderate prior which is centered on .5 on the $\pi$ parameter during MCMC. Although the parameter estimates for $Y_{C2}(t)$ and $\pi$ were biased, the biases didn’t affect our estimates of the time to event distribution much. This can be seen in Figure 4.1, where the survival estimates by the informative censoring models are close to the true values.
Table A.1: Simulation results for the parameters of the death processes by the informative censoring model for different values of $\pi$

<table>
<thead>
<tr>
<th>$\pi$</th>
<th>Parameters</th>
<th>True Values</th>
<th>Bias</th>
<th>SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>.9</td>
<td>$\alpha_{C10}$</td>
<td>-0.25</td>
<td>0.02</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C11}$</td>
<td>-0.63</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C12}$</td>
<td>-0.32</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>$y_{C10}$</td>
<td>4.71</td>
<td>-0.00</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C20}$</td>
<td>-0.47</td>
<td>-0.42</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C21}$</td>
<td>-0.63</td>
<td>-0.83</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C22}$</td>
<td>-0.32</td>
<td>-0.50</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>$y_{C20}$</td>
<td>4.71</td>
<td>4.45</td>
<td>3.25</td>
</tr>
<tr>
<td></td>
<td>$\pi$</td>
<td>0.9</td>
<td>-0.35</td>
<td>0.07</td>
</tr>
<tr>
<td>.75</td>
<td>$\alpha_{C10}$</td>
<td>-0.25</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C11}$</td>
<td>-0.63</td>
<td>0.01</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C12}$</td>
<td>-0.32</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>$y_{C10}$</td>
<td>4.71</td>
<td>0.01</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C20}$</td>
<td>-0.47</td>
<td>-0.27</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C21}$</td>
<td>-0.63</td>
<td>-0.82</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C22}$</td>
<td>-0.32</td>
<td>-0.81</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>$y_{C20}$</td>
<td>4.71</td>
<td>5.86</td>
<td>3.52</td>
</tr>
<tr>
<td></td>
<td>$\pi$</td>
<td>0.75</td>
<td>-0.24</td>
<td>0.08</td>
</tr>
<tr>
<td>.5</td>
<td>$\alpha_{C10}$</td>
<td>-0.25</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C11}$</td>
<td>-0.63</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C12}$</td>
<td>-0.32</td>
<td>-0.02</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>$y_{C10}$</td>
<td>4.71</td>
<td>0.03</td>
<td>0.29</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C20}$</td>
<td>-0.47</td>
<td>-0.22</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C21}$</td>
<td>-0.63</td>
<td>-1.46</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C22}$</td>
<td>-0.32</td>
<td>-0.80</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>$y_{C20}$</td>
<td>4.71</td>
<td>5.88</td>
<td>5.44</td>
</tr>
<tr>
<td></td>
<td>$\pi$</td>
<td>0.5</td>
<td>-0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>.25</td>
<td>$\alpha_{C10}$</td>
<td>-0.25</td>
<td>0.03</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C11}$</td>
<td>-0.63</td>
<td>-0.04</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C12}$</td>
<td>-0.32</td>
<td>-0.03</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>$y_{C10}$</td>
<td>4.71</td>
<td>0.00</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C20}$</td>
<td>-0.47</td>
<td>0.14</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C21}$</td>
<td>-0.63</td>
<td>-1.25</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{C22}$</td>
<td>-0.32</td>
<td>-0.97</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>$y_{C20}$</td>
<td>4.71</td>
<td>3.64</td>
<td>7.05</td>
</tr>
<tr>
<td></td>
<td>$\pi$</td>
<td>.25</td>
<td>0.19</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Note: Bias is calculated by subtracting the true value from the posterior mean of a parameter; SSD is the sample standard deviation of the estimates.
A.2.4 Trace Plots of Regression Parameters from the Informative Censoring Model in the Analysis of the F344/N Male Rat Lung Tumor Data

Figure A.1: $\alpha_{E0}$

Figure A.2: $\alpha_{E1}$

Figure A.3: $\alpha_{E2}$

Figure A.4: $y_{E0}$
Figure A.5: $\alpha_{C10}$

Figure A.6: $\alpha_{C11}$

Figure A.7: $\alpha_{C12}$

Figure A.8: $y_{C10}$
Appendix B: CHAPTER 5 SUPPLEMENTAL MATERIAL

B.1 Model Under Independent Censoring: Posterior Computation and Data Augmentation

B.1.1 Likelihood Function

Using equations (5.9) and (5.10) in Chapter 5, the likelihood contributions for the four different combinations of \( d_1 \) (indicates if the first event of interest occurs) and \( d_2 \) (indicates if the second event of interest occurs) are as follows. Note that the subscript condition indicator involving \( d_1 \) and \( d_2 \) means that the likelihood contribution corresponds to a subject with this condition satisfied.

\[
\begin{align*}
L_{1i} &= L_{\{d_1=1, d_2=1\}} = f_{S_1}(s_{1i} | \mu_{1i}, \sigma^2_1, y_{10}, u_i) \times f_{S_2}(s_{2i} | \mu_{2i}, \sigma^2_2, y_{20}, u_i) \\
L_{2i} &= L_{\{d_1=1, d_2=0\}} = f_{S_1}(s_{1i} | \mu_{1i}, \sigma^2_1, y_{10}, u_i) \times f_{Y_2}(y_{2i} | \mu_{2i}, \sigma^2_2, y_{20}, u_i) \\
L_{3i} &= L_{\{d_1=0, d_2=1\}} = f_{Y_1}(y_{1i} | \mu_{1i}, \sigma^2_1, y_{10}, u_i) \times f_{S_2}(s_{2i} | \mu_{2i}, \sigma^2_2, y_{20}, u_i) \\
L_{4i} &= L_{\{d_1=0, d_2=0\}} = f_{Y_1}(y_{1i} | \mu_{1i}, \sigma^2_1, y_{10}, u_i) \times f_{Y_2}(y_{2i} | \mu_{2i}, \sigma^2_2, y_{20}, u_i)
\end{align*}
\]

Thus the likelihood likelihood function is:

\[
\prod_{i=1}^{n} L_{1i}^{d_1i} L_{2i}^{d_1i(1-d_2i)} L_{3i}^{(1-d_1i)d_2i} L_{4i}^{(1-d_1i)(1-d_2i)} \tag{B.1}
\]
### B.1.2 MCMC Algorithm

Our MCMC algorithm uses Gibbs, Metropolis and rejection sampling steps to sample the parameters \((\alpha_1, y_{10}, \alpha_2, y_{20}, \lambda)\) and latent variables \((\{s_{1i} : d_{1i} = 1\}, \{s_{2i} : d_{2i} = 1\}, \{y_{1i} : d_{1i} = 0\}, \{y_{2i} : d_{2i} = 0\}, b_i)\) based on their respective full conditional posterior distributions. The necessary conditional distributions as well as a summary of the sampling algorithm at each iteration are described as follows.

**Step 1.** Sample \(\alpha_1\) directly from its full conditional distribution which is a multivariate normal with mean

\[
\mu_{\alpha_{1\text{post}}} = \left( \sigma_1^{-2} \left( \sum_{i: d_{1i} = 1} s_{1i} x_{1i} x_{1i}' + \sum_{i: d_{1i} = 0} u_i x_{1i} x_{1i}' \right) + \Sigma_{\alpha_{1\text{prior}}} \right)^{-1} \\
\times \left( -\sigma_1^{-2} \left( \sum_{i: d_{1i} = 1} x_{1i}(y_{10} + b_i s_{1i}) + \sum_{i: d_{1i} = 0} x_{1i}(y_{10} + b_i u_i - y_{1i}) \right) \right)
\]

and covariance matrix

\[
\Sigma_{\alpha_{1\text{post}}} = \left( \sigma_1^{-2} \left( \sum_{i: d_{1i} = 1} s_{1i} x_{1i} x_{1i}' + \sum_{i: d_{1i} = 0} u_i x_{1i} x_{1i}' \right) + \Sigma_{\alpha_{1\text{prior}}} \right)^{-1}
\]

where \(\mu_{\alpha_{1\text{prior}}}\) is the mean vector and \(\Sigma_{\alpha_{1\text{prior}}}\) is the covariance matrix of the multivariate normal prior distribution for \(\alpha_1\).

**Step 2.** Sample \(y_{10}\) directly from its full conditional distribution which has the kernel:

\[
K(y_{10} | \cdot) = \exp \left[ -\frac{(\log y_{10} - \mu_{\log y_{10}\text{prior}})^2}{2\sigma_{\log y_{10}\text{prior}}^2} \right] \\
\times \prod_{d_{1i} = 1} y_{10} \times \exp \left[ -\frac{(y_{10} + (x_{1i}'\alpha_1 + b_i)s_{1i})^2}{2\sigma_1^2 s_{1i}} \right] \\
\times \prod_{d_{1i} = 0} \exp \left[ -\frac{(y_{1i} - y_{10} - (x_{1i}'\alpha_1 + b_i)u_i)^2}{2\sigma_1^2 u_i} \right] \left( 1 - \exp \left[ -\frac{-2y_{10} y_{1i}}{\sigma_1^2 u_i} \right] \right)
\]
where $\mu_{\log y_{10 \text{prior}}}$ is the mean and $\sigma^2_{\log y_{10 \text{prior}}}$ is the variance of normal prior distribution for $\log(y_{10})$.

**Step 3.** Sample $\alpha_2$ directly from its full conditional distribution which is a multivariate normal with mean

$$
\mu_{\alpha_{2 \text{post}}} = \left( \sigma^2_{\alpha_2} \left( \sum_{i:d_{2i}=1} s_{2i} x_{2i} x_{2i}' + \sum_{i:d_{2i}=0} u_i x_{2i} x_{2i}' \right) + \Sigma_{\alpha_{2 \text{prior}}} \right)^{-1} \times \left( -\sigma^2_{\alpha_2} \left( \sum_{i:d_{2i}=1} x_{2i} (y_{20} + b_i s_{2i}) + \sum_{i:d_{2i}=0} x_{2i} (y_{20} + b_i u_i - y_{2i}) \right) + \Sigma_{\alpha_{2 \text{prior}}} \mu_{\alpha_{2 \text{prior}}} \right)
$$

and covariance matrix

$$
\Sigma_{\alpha_{2 \text{post}}} = \left( \sigma^2_{\alpha_2} \left( \sum_{i:d_{2i}=1} s_{2i} x_{2i} x_{2i}' + \sum_{i:d_{2i}=0} u_i x_{2i} x_{2i}' \right) + \Sigma_{\alpha_{2 \text{prior}}} \right)^{-1}
$$

where $\mu_{\alpha_{2 \text{prior}}}$ is the mean vector and $\Sigma_{\alpha_{2 \text{prior}}}$ is the covariance matrix of the multivariate normal prior distribution for $\alpha_2$.

**Step 4.** Sample $y_{20}$ directly from its full conditional distribution which has the kernel:

$$
K(y_{20} \mid \cdot) = \exp \left[ - \frac{(\log y_{20} - \mu_{\log y_{20 \text{prior}}})^2}{2\sigma^2_{\log y_{20 \text{prior}}}} \right] \times \prod_{d_{2i}=1} y_{20} \times \exp \left[ - \frac{(y_{20} + (x_{1i}' \alpha_2 + b_i) s_{2i})^2}{2\sigma^2_{s2i}} \right] \times \prod_{d_{2i}=0} \exp \left[ - \frac{(y_{2i} - y_{20} - (x_{1i}' \alpha_2 + b_i) u_i)^2}{2\sigma^2_{u_i}} \right] \left( 1 - \exp \left[ - \frac{2y_{20} y_{2i}}{\sigma^2_{u_i}} \right] \right)
$$

where $\mu_{\log y_{20 \text{prior}}}$ is the mean and $\sigma^2_{\log y_{20 \text{prior}}}$ is the variance of normal prior distribution for $\log(y_{20})$. 133
**Step 5.** Sample $b_i$ directly from its full conditional distribution which is a normal with mean

$$\mu_{b_i, \text{post}} = \frac{-A_{b_i} \sigma_{b_i}^2}{1 + \lambda \sigma_{b_i}^2}$$

and variance

$$\sigma_{b_i, \text{post}}^2 = \frac{\sigma_{b_i}^2}{\lambda \sigma_{b_i}^2 + 1}$$

where

$$\sigma_{b_i}^2 = \left[ d_{1i} \left( \frac{s_{1i}}{\sigma^2_1} \right) + (1 - d_{1i}) \left( \frac{u_i}{\sigma^2_1} \right) + d_{2i} \left( \frac{s_{2i}}{\sigma^2_2} \right) + (1 - d_{2i}) \left( \frac{u_i}{\sigma^2_2} \right) \right]^{-1}$$

and

$$A_{b_i} = d_{1i} \left[ \frac{s_{1i} x_i' \alpha_1 + y_{10}}{\sigma^2_1} \right] + (1 - d_{1i}) \left[ \frac{u_i x_i' \alpha_1 + y_{10} - y_{1i}}{\sigma^2_1} \right] + d_{2i} \left[ \frac{s_{2i} x_i' \alpha_2 + y_{20}}{\sigma^2_2} \right] + (1 - d_{2i}) \left[ \frac{u_i x_i' \alpha_2 + y_{20} - y_{2i}}{\sigma^2_2} \right]$$

**Step 6.** Sample $\lambda$ directly from its full conditional distribution which is a gamma:

$$\lambda \mid b \sim \text{Gamma} \left( \frac{n}{2} + \gamma_1, \frac{1}{2} \sum_{i=1}^{n} b_i^2 + \gamma_2 \right)$$

where $\gamma_1$ is the shape parameter and $\gamma_2$ is the rate parameter of the gamma prior for $\lambda$.

**Step 7.** For subjects whose first event is right censored, sample the process value $y_{1i}$ at right censoring time from:

$$K(y_{1i} \mid \cdot) = \exp \left[ -\frac{(y_{1i} - y_{10} - (x_i' \alpha_1 + b_i) u_i)^2}{2 \sigma^2_i u_i} \right] \left( 1 - \exp \left[ -\frac{-2y_{10} y_{1i}}{\sigma^2_i u_i} \right] \right) I(y_{1i} > 0)$$

Note that $y_{1i}$ is sampled using the data augmentation method described in Step 2 in Section A.1.1.
Step 8. For subjects whose first event is left censored, use the inverse c.d.f. method to sample $S_{1i}$ from its full conditional distribution which has a kernel proportional to the inverse Gaussian p.d.f. truncated to the right at $u_i$:

$$K(s_{1i} | \cdot) = \frac{1}{\sqrt{s_{1i}^3}} \exp \left[ -\frac{[y_{10} + (x_{1i}' \alpha_1 + b_i)s_{1i}]^2}{2\sigma_1^2 s_{1i}} \right] I(0 < s_{1i} \leq u_i)$$

Step 9. For subjects whose second event is right censored, sample the process value $y_{2i}$ at right censoring time from:

$$K(y_{2i} | \cdot) = \exp \left[ -\frac{[y_{2i} - y_{20} - (x_{2i}' \alpha_2 + b_i)u_i]^2}{2\sigma_2^2 u_i} \right] \left( 1 - \exp \left[ \frac{2y_{20}y_{2i}}{\sigma_2^2 u_i} \right] \right)^{1-d_{2i}} I(y_{2i} > 0)$$

Note that $y_{2i}$ is sampled using the data augmentation method described in Step 2 in Section A.1.1.

Step 10. For subjects whose second event is left censored, use the inverse c.d.f. method to sample $S_{2i}$ from its full conditional distribution which has a kernel proportional to the inverse Gaussian p.d.f. truncated to the right at $u_i$:

$$K(s_{2i} | \cdot) = \frac{1}{\sqrt{s_{2i}^3}} \exp \left[ -\frac{[y_{20} + (x_{2i}' \alpha_2 + b_i)s_{2i}]^2}{2\sigma_2^2 s_{2i}} \right] I(0 < s_{2i} \leq u_i)$$

B.2 Model Under Informative Censoring: Posterior Computation and Data Augmentation

B.2.1 Likelihood Function

Using equations (3.1), (5.9) and (5.10) in Chapters 3 and 5, the likelihood contributions for the eight different combinations of $d_1$, $d_2$ and $\delta$ are as follows. Note that the subscript condition indicator involving $d_{1i}$, $d_{2i}$ and $\delta_i$ means that the likelihood
contribution corresponds to a subject with this condition satisfied.

\[
\begin{align*}
L_{1i} &= L_{\{d_{1i}=1,d_{2i}=1,\delta_i=1\}} = f_{S_1}(s_{1i} \mid \mu_{E_{1i}}, \sigma_{E_{11}}^2, y_{E_{10}}, u_{i}) \times f_{S_2}(s_{2i} \mid \mu_{E_{2i}}, \sigma_{E_{22}}^2, y_{E_{20}}, u_{i}) \times f_T(r_{i} \mid \mu_{C_i}, \sigma_{C_i}^2, y_{C0}) \times J \\
L_{2i} &= L_{\{d_{1i}=1,d_{2i}=0,\delta_i=1\}} = f_{S_1}(s_{1i} \mid \mu_{E_{1i}}, \sigma_{E_{11}}^2, y_{E_{10}}, u_{i}) \times f_{Y_2}(y_{2i} \mid \mu_{E_{2i}}, \sigma_{E_{22}}^2, y_{E_{20}}, u_{i}) \times f_T(r_{i} \mid \mu_{C_i}, \sigma_{C_i}^2, y_{C0}) \times J \\
L_{3i} &= L_{\{d_{1i}=0,d_{2i}=1,\delta_i=1\}} = f_{Y_1}(y_{1i} \mid \mu_{E_{1i}}, \sigma_{E_{11}}^2, y_{E_{10}}, u_{i}) \times f_{S_2}(s_{2i} \mid \mu_{E_{2i}}, \sigma_{E_{22}}^2, y_{E_{20}}, u_{i}) \times f_T(r_{i} \mid \mu_{C_i}, \sigma_{C_i}^2, y_{C0}) \times J \\
L_{4i} &= L_{\{d_{1i}=0,d_{2i}=0,\delta_i=1\}} = f_{Y_1}(y_{1i} \mid \mu_{E_{1i}}, \sigma_{E_{11}}^2, y_{E_{10}}, u_{i}) \times f_{Y_2}(y_{2i} \mid \mu_{E_{2i}}, \sigma_{E_{22}}^2, y_{E_{20}}, u_{i}) \times f_T(r_{i} \mid \mu_{C_i}, \sigma_{C_i}^2, y_{C0}) \times J \\
L_{5i} &= L_{\{d_{1i}=1,d_{2i}=1,\delta_i=0\}} = f_{S_1}(s_{1i} \mid \mu_{E_{1i}}, \sigma_{E_{11}}^2, y_{E_{10}}, u_{i}) \times f_{S_2}(s_{2i} \mid \mu_{E_{2i}}, \sigma_{E_{22}}^2, y_{E_{20}}, u_{i}) \times f_Y(y_{C_i} \mid \mu_{C_i}, \sigma_{C_i}^2, y_{C0}, r_{i}) \\
L_{6i} &= L_{\{d_{1i}=1,d_{2i}=0,\delta_i=0\}} = f_{S_1}(s_{1i} \mid \mu_{E_{1i}}, \sigma_{E_{11}}^2, y_{E_{10}}, u_{i}) \times f_{Y_2}(y_{2i} \mid \mu_{E_{2i}}, \sigma_{E_{22}}^2, y_{E_{20}}, u_{i}) \times f_Y(y_{C_i} \mid \mu_{C_i}, \sigma_{C_i}^2, y_{C0}, r_{i}) \\
L_{7i} &= L_{\{d_{1i}=0,d_{2i}=1,\delta_i=0\}} = f_{Y_1}(y_{1i} \mid \mu_{E_{1i}}, \sigma_{E_{11}}^2, y_{E_{10}}, u_{i}) \times f_{S_2}(s_{2i} \mid \mu_{E_{2i}}, \sigma_{E_{22}}^2, y_{E_{20}}, u_{i}) \times f_Y(y_{C_i} \mid \mu_{C_i}, \sigma_{C_i}^2, y_{C0}, r_{i}) \\
L_{8i} &= L_{\{d_{1i}=0,d_{2i}=0,\delta_i=0\}} = f_{Y_1}(y_{1i} \mid \mu_{E_{1i}}, \sigma_{E_{11}}^2, y_{E_{10}}, u_{i}) \times f_Y(y_{2i} \mid \mu_{E_{2i}}, \sigma_{E_{22}}^2, y_{E_{20}}, u_{i}) \times f_Y(y_{C_i} \mid \mu_{C_i}, \sigma_{C_i}^2, y_{C0}, r_{i}) \\
\end{align*}
\]
where

\[ r_i = \tau_1 [d_{1i}d_{2i}q_is_{1i} + d_{1i}d_{2i}(1 - q_i)s_{2i} + d_{1i}(1 - d_{2i})s_{1i} \]
\[ + d_{2i}(1 - d_{1i})s_{2i} + (1 - d_{1i})(1 - d_{2i})u_i] \]
\[ + \tau_2 [d_{1i}d_{2i}q_i(s_{2i} - s_{1i}) + d_{1i}(1 - d12i)(u_i - s_i)] \]
\[ + \tau_3 [d_{1i}d_{2i}(1 - q_i)(s_{1i} - s_{2i}) + d_{2i}(1 - d_{1i})(u_i - s_{2i})] \]
\[ + \tau_4 [d_{1i}d_{2i}q_i(u_i - s_{2i}) + d_{1i}d_{2i}(1 - q_i)(u_i - s_{1i})] \]

with \( q_i = I(s_{1i} \leq s_{2i}) \) and \( J = \sum_{k=1}^{4} I_{k_i} \tau_k \) where \( I_{k_i} \) equals 1 if subject \( i \) is in state \( k \) in Equation (5.16) at the time of its censoring event and 0 otherwise. Thus the likelihood function is

\[
L = \prod_{i=1}^{n} L_{1i}^{d_{1i}d_{2i} \delta_i} \times L_{2i}^{d_{1i}(1-d_{2i}) \delta_i} \times L_{3i}^{(1-d_{1i})d_{2i} \delta_i} \times L_{4i}^{(1-d_{1i})(1-d_{2i}) \delta_i} \]
\[ \times \prod_{i=1}^{n} L_{5i}^{d_{1i}d_{2i}(1-\delta_i)} \times L_{6i}^{(1-d_{1i})(1-d_{2i}) (1-\delta_i)} \times L_{7i}^{d_{1i}d_{2i}(1-\delta_i)} \times L_{8i}^{(1-d_{1i})(1-d_{2i})(1-\delta_i)} \]

**B.2.2 MCMC Algorithm**

Our MCMC algorithm uses Gibbs, Metropolis and rejection sampling steps to sample the parameters \((\alpha_{E1}, y_{E10}, \alpha_{E2}, y_{E20}, \lambda, \alpha_C, y_{C0}, \log \tau)\) and latent variables \(\{s_{1i} : d_{1i} = 1\}, \{s_{2i} : d_{2i} = 1\}, \{y_{E1i} : d_{1i} = 0\}, \{y_{E2i} : d_{2i} = 0\}, y_{Ci}, b_i\) based on their respective full conditional posterior distributions. The necessary conditional distributions as well as a summary of the sampling algorithm at each iteration are described as follows.
**Step 1.** Sample the regression coefficient vector $\alpha_{E1}$ for the drift parameter covariates in the first event process directly from its full conditional posterior distribution which is a multivariate normal with mean

$$
\mu_{\alpha_{E1} \text{post}} = \left( \sigma_{E1}^{-2} \left( \sum_{i:d_{1i}=1} s_{1i} x_{E1i} x'_{E1i} + \sum_{i:d_{1i}=0} u_i x_{E1i} x'_{E1i} \right) + \Sigma_{\alpha_{E1} \text{prior}} \right)^{-1} \times A_{\alpha_{E1}}
$$

and covariance matrix

$$
\Sigma_{\alpha_{E1} \text{post}} = \left( \sigma_{E1}^{-2} \left( \sum_{i:d_{1i}=1} s_{1i} x_{E1i} x'_{E1i} + \sum_{i:d_{1i}=0} u_i x_{E1i} x'_{E1i} \right) + \Sigma_{\alpha_{E1} \text{prior}} \right)^{-1}
$$

where

$$A_{\alpha_{E1}} = -\sigma_{E1}^{-2} \left( \sum_{i:d_{1i}=1} x_{E1i} (y_{E10} + b_i s_{1i}) + \sum_{i:d_{1i}=0} x_{E1i} (y_{E10} + b_i u_i - y_{E1i}) \right)
+ \Sigma_{\alpha_{E1} \text{prior}}^{-1} \mu_{\alpha_{E1} \text{prior}}$$

and $\mu_{\alpha_{E1} \text{prior}}$ is the mean vector and $\Sigma_{\alpha_{E1} \text{prior}}$ is the covariance matrix of the multivariate normal prior distribution for $\alpha_{E1}$.

**Step 2.** Use the Metropolis-Hastings random walk algorithm to sample the initial state $y_{E10}$ of the first event process from its full conditional posterior distribution, which has the kernel:

$$
K(y_{E10} \mid \cdot) = \exp \left[ -\frac{\left( \log y_{E10} - \mu_{\log y_{E10} \text{prior}} \right)^2}{2\sigma_{\log y_{E10} \text{prior}}^2} \right] 
\times \prod_{d_{1i}=1} y_{E10} \times \exp \left[ -\frac{(y_{E10}) + (x'_{E1i} \alpha_{E1} + b_i s_{1i})^2}{2\sigma_{E1}^2 s_{1i}} \right] 
\times \prod_{d_{1i}=0} \exp \left[ -\frac{(y_{E1i} - y_{E10} - (x'_{E1i} \alpha_{E1} + b_i u_i)^2}{2\sigma_{E1}^2 u_i} \right] 
\times \prod_{d_{1i}=0} \left( 1 - \exp \left[ -\frac{-2y_{E10} y_{E1i}}{\sigma_{E1}^2 u_i} \right] \right)
$$

where $\mu_{\log y_{E10} \text{prior}}$ is the mean and $\sigma_{\log y_{E10} \text{prior}}^2$ is the variance of normal prior distribution for $\log(y_{E10})$. 

138
Step 3. Sample the regression coefficient vector $\alpha_{E2}$ for the drift parameter covariates in the second event process directly from its full conditional posterior distribution which is a multivariate normal with mean

$$\mu_{\alpha_{E2} post} = \left( \sigma^{-2}_{E2} \left( \sum_{i:d_{2i}=1} s_{2i}x_{E2i}x'_{E2i} + \sum_{i:d_{2i}=0} u_i x_{E2i}x'_{E2i} \right) + \Sigma_{\alpha_{E2} prior}^{-1} \right)^{-1} \times A_{\alpha_{E2}}$$

and covariance matrix

$$\Sigma_{\alpha_{E2} post} = \left( \sigma^{-2}_{E2} \left( \sum_{i:d_{2i}=1} s_{2i}x_{E2i}x'_{E2i} + \sum_{i:d_{2i}=0} u_i x_{E2i}x'_{E2i} \right) + \Sigma_{\alpha_{E2} prior}^{-1} \right)^{-1}$$

where

$$A_{\alpha_{E2}} = -\sigma^{-2}_{E2} \left( \sum_{i:d_{2i}=1} x_{E2i}(y_{E20} + b_is_{2i}) + \sum_{i:d_{2i}=0} x_{E2i}(y_{E20} + b_iu_i - y_{E2i}) \right)$$

$$+ \Sigma_{\alpha_{E2} prior}^{-1} \mu_{\alpha_{E2} prior}$$

and $\mu_{\alpha_{E2} prior}$ is the mean vector and $\Sigma_{\alpha_{E2} prior}$ is the covariance matrix of the multivariate normal prior distribution for $\alpha_{E2}$.

Step 4. Use the Metropolis-Hastings random walk algorithm to sample the initial state $y_{E20}$ of the second event process from its full conditional posterior distribution, which has the kernel:

$$K(y_{E20} | \cdot) = \exp \left[ -\frac{(\log y_{E20} - \mu_{\log y_{E20} prior})^2}{2\sigma^2_{\log y_{E20 prior}}} \right] \times \prod_{d_{2i}=1} \frac{y_{E20}}{\sqrt{2\pi\sigma^2_{E2}s_{2i}}} \times \exp \left[ -\frac{(y_{E20} + (x'_{E2i}\alpha_{E2} + b_is_{2i})^2}{2\sigma^2_{E2}s_{2i}} \right] \times \prod_{d_{2i}=0} \frac{1}{\sqrt{2\pi\sigma^2_{E2}u_i}} \times \exp \left[ -\frac{(y_{E2i} - y_{E20} - (x'_{E2i}\alpha_{E2} + b_iu_i)^2}{2\sigma^2_{E2}u_i} \right] \times \prod_{d_{2i}=0} \left( 1 - \exp \left[ -\frac{2y_{E20}y_{E2i}}{\sigma^2_{E2}u_i} \right] \right)$$

where $\mu_{\log y_{E20} prior}$ is the mean and $\sigma^2_{\log y_{E20 prior}$ is the variance of normal prior distribution for $\log(y_{E20})$. 

139
Step 5. Sample the random effects $b_i$ directly from its full conditional distribution which is a normal with mean

$$\mu_{b_{i,\text{post}}} = \frac{-A_{b_i}\sigma_{b_i}^2}{1 + \lambda\sigma_{b_i}^2}$$

and variance

$$\sigma_{b_{i,\text{post}}}^2 = \frac{\sigma_{b_i}^2}{\lambda\sigma_{b_i}^2 + 1}$$

where

$$\sigma_{b_i}^2 = \left( d_{1i} \left[ \frac{s_{1i}}{\sigma_{E1}^2} \right] + (1 - d_{1i}) \left[ \frac{u_{i}}{\sigma_{E1}^2} \right] + \delta_{i} \left[ \frac{s_{2i}}{\sigma_{E2}^2} \right] + (1 - \delta_{i}) \left[ \frac{u_{i}}{\sigma_{E2}^2} \right] \right)^{-1}$$

and

$$A_{b_i} = d_{1i} \left[ \frac{s_{1i}x'_{E1i}\alpha_{E1} + y_{E10}}{\sigma_{E1}^2} \right] + (1 - d_{1i}) \left[ \frac{u_{i}x'_{E1i}\alpha_{E1} + y_{E10} - y_{E1i}}{\sigma_{E1}^2} \right]$$

$$+ \delta_{i} \left[ \frac{s_{2i}x'_{E2i}\alpha_{E2} + y_{E20}}{\sigma_{E2}^2} \right] + (1 - \delta_{i}) \left[ \frac{u_{i}x'_{E2i}\alpha_{E2} + y_{E20} - y_{E2i}}{\sigma_{E2}^2} \right]$$

Step 6. Sample $\lambda$ directly from its full conditional distribution which is a gamma:

$$\lambda \mid b \sim \text{Gamma} \left( \frac{n}{2} + \gamma_1, \frac{1}{2} \sum_{i=1}^{n} b_i^2 + \gamma_2 \right)$$

where $\gamma_1$ is the shape parameter and $\gamma_2$ is the rate parameter of the gamma prior for $\lambda$.

Step 7. For subjects whose first event is right censored, sample their process value $y_{E1i}$ at the right censoring time from:

$$K(y_{E1i} \mid \cdot) = \exp \left[ -\frac{[y_{E1i} - y_{E10} - (x'_{1i}\alpha_{E1} + b_i)u_i]^2}{2\sigma_{E1}^2 u_i} \right] \left( 1 - \exp \left[ -\frac{2y_{E10}y_{E1i}}{\sigma_{E1}^2 u_i} \right] \right) \times \mathbb{1}(y_{E1i} > 0)$$

Note that both $y_{E1i}$ and $y_{E2i}$ below are sampled using the data augmentation method described in Step 2 in Section A.1.1.
Step 8. For subjects whose second event is right censored, sample their process value $y_{E_{2i}}$ at the right censoring time from:

$$K(y_{E_{2i}} \mid \cdot) = \exp \left[ - \frac{[y_{E_{2i}} - y_{E20} - (x'_{E_{2i}} \alpha_{E2} + b_i)u_i]_2^2}{2\sigma^2_{E2}u_i} \right] \left( 1 - \exp \left[ \frac{-2y_{E20}y_{E_{2i}}}{\sigma^2_{E2}u_i} \right] \right) \times I(y_{E_{2i}} > 0)$$

Step 9. For subjects whose first event is left censored, use the Metropolis-Hastings random walk algorithm to sample the left censored time of the first event $s_{1i}$ from its full conditional posterior distribution which has the kernel:

$$K(s_{1i} \mid \cdot) = \frac{1}{\sqrt{s_{1i}^3}} \exp \left[ - \frac{[y_{E10} + (x'_{E_{1i}} \alpha_{E1} + b_i)s_{1i}]_2^2}{2\sigma^2_{E1}s_{1i}} \right] \times I(0 < s_{1i} \leq u_i) \times \left( \frac{1}{\sqrt{r_{i}^3}} \exp \left[ - \frac{[y_{C0} + (x'_{Ci} \alpha_{C})r_i]_2^2}{2\sigma^2_{C}r_i} \right] \right)^{\delta_i} \times \left( \frac{1}{\sqrt{r_{i}^3}} \exp \left[ - \frac{[y_{Ci} - y_{C0} - (x'_{Ci} \alpha_{C})r_i]_2^2}{2\sigma^2_{C}r_i} \right] \left( 1 - \exp \left[ \frac{-2y_{C0}y_{Ci}}{\sigma^2_{C}r_i} \right] \right) \right)^{1-\delta_i}$$

where $r_i$ is the censoring time of the $i$th subject on the operational time scale. If $\delta_i = 1$, $r_i$ is the time of random censoring event; if $\delta_i = 0$, $r_i$ is the time of planned (non-random) censoring event.

Step 10. For subjects whose second event is left censored, use the Metropolis-Hastings random walk algorithm to sample the left censored time of the second event $s_{2i}$ from its full conditional posterior distribution which has the kernel:

$$K(s_{2i} \mid \cdot) = \left( \frac{1}{\sqrt{s_{2i}^3}} \exp \left[ - \frac{[y_{E20} + (x'_{E_{2i}} \alpha_{E2} + b_i)s_{2i}]_2^2}{2\sigma^2_{E2}s_{2i}} \right] \times I(0 < s_{2i} \leq u_i) \right)^{d_{2i}} \times \left( \frac{1}{\sqrt{r_{i}^3}} \exp \left[ - \frac{[y_{C0} + (x'_{Ci} \alpha_{C})r_i]_2^2}{2\sigma^2_{C}r_i} \right] \right)^{\delta_i} \times \left( \frac{1}{\sqrt{r_{i}^3}} \exp \left[ - \frac{[y_{Ci} - y_{C0} - (x'_{Ci} \alpha_{C})r_i]_2^2}{2\sigma^2_{C}r_i} \right] \left( 1 - \exp \left[ \frac{-2y_{C0}y_{Ci}}{\sigma^2_{C}r_i} \right] \right) \right)^{1-\delta_i}$$

141
Step 11. Sample the regression coefficient vector $\alpha_C$ for the drift parameter covariates in the censoring process directly from its full conditional distribution which is a multivariate normal with mean

$$
\mu_{\alpha_{C\text{post}}} = \left( \sigma_C^{-2} \sum_i r_i x_C^i x'_C i + \Sigma_{\alpha_{C\text{prior}}}^{-1} \right)^{-1}
$$

$$
\times \left( -\sigma_C^{-2} \left( \sum_{i: \delta_i = 1} x_C^i y_{C0} + \sum_{i: \delta_i = 0} x_C^i (y_{C0} - y_{Ci}) \right) + \Sigma_{\alpha_{C\text{prior}}}^{-1} \mu_{\alpha_{C\text{prior}}} \right)
$$

and covariance matrix

$$
\Sigma_{\alpha_{C\text{post}}} = \left( \sigma_C^{-2} \sum_i r_i x_C^i x'_C i + \Sigma_{\alpha_{C\text{prior}}}^{-1} \right)^{-1}
$$

where $\mu_{\alpha_{C\text{prior}}}$ is the mean vector and $\Sigma_{\alpha_{C\text{prior}}}$ is the covariance matrix of the multivariate normal prior distribution for $\alpha_C$.

Step 12. Use the Metropolis-Hastings random walk algorithm to sample the initial state $y_{C0}$ of the censoring process from its full conditional posterior distribution, which has the kernel:

$$
K(y_{C0} | \cdot) = \exp \left[ -\frac{(\log y_{C0} - \mu_{\log y_{C0\text{prior}}})^2}{2\sigma^2_{\log y_{C0\text{prior}}}} \right]
$$

$$
\times \prod_{\delta_i = 1} y_{C0} \times \exp \left[ -\frac{(y_{C0}) + (x'_C \alpha_C r_i)^2}{2\sigma^2_C r_i} \right]
$$

$$
\times \prod_{\delta_i = 0} \frac{1}{\sqrt{r_i}} \times \exp \left[ -\frac{(y_{Ci} - y_{C0} - (x'_C \alpha_C r_i)^2}{2\sigma^2_C r_i} \right] \left( 1 - \exp \left[ -\frac{2y_{C0} y_{Ci}}{\sigma^2_C r_i} \right] \right)
$$

where $\mu_{\log y_{C0\text{prior}}}$ is the mean and $\sigma^2_{\log y_{C0\text{prior}}}$ is the variance of normal prior distribution for $\log(y_{C0})$.

Step 13. Use the Metropolis-Hastings random walk algorithm to sample the logarithm of the operational time scale parameters, $\log(\tau)$, of the censoring process.
from its full conditional posterior distribution, which has the kernel:

\[ K(\log \tau \mid \cdot) = \exp \left\{ \sum_{\delta_i=1} \left[ -\frac{3}{2} \log(r_i) - \frac{[y_{C0} + (x'_{Ci} \alpha_C)r_i]^2}{2\sigma^2 C r_i} + \log \left( \sum_{k=1}^4 I_{k_i} e^{\log r_k} \right) \right] \right. \\
+ \sum_{\delta_i=0} \left[ -\frac{1}{2} \log(r_i) - \frac{[y_{Ci} - y_{C0} - (x'_{Ci} \alpha_C)r_i]^2}{2\sigma^2 C r_i} \\
+ \log \left( 1 - \exp \left[ \frac{-2y_{C0}y_{Ci}}{\sigma^2 C r_i} \right] \right) \right] \right. \\
- \frac{1}{2} (\log \tau - \mu_{\log r_0})' \Sigma_{\log r_0}^{-1} (\log \tau - \mu_{\log r_0}) \right\} \]

where \( \tau = (\tau_1, \tau_2, \tau_3)' \) is the vector of the three operational time scale parameters and we take \( \tau = (e^{\log r_1}, e^{\log r_2}, e^{\log r_3})' \); \( r_i = t_{1i} + e^{\log r_1}t_{2i} + e^{\log r_2}t_{3i} + e^{\log r_3}t_{4i} \).

Step 14. For subjects whose censoring event driven by the stochastic process \( Y_C(r) \) is right censored at the planned (non-random) censoring time \( r_i \) (i.e., \( \delta_i = 0 \)), sample their process value \( y_{Ci} \) at \( r_i \):

\[ K(y_{Ci} \mid \cdot) = \exp \left[ -\frac{[y_{Ci} - y_{C0} - (x'_{Ci} \alpha_C)r_i]^2}{2\sigma^2 C r_i} \right] \left( 1 - \exp \left[ \frac{-2y_{C0}y_{Ci}}{\sigma^2 C r_i} \right] \right) I(y_{Ci} > 0) \]

Note that both \( y_{Ci} \) is sampled using the data augmentation method described in Step 2 in Section A.1.1.

B.2.3 Simulation Results for the Parameters of Death Processes

The simulation results for the parameters of the censoring process in the operational time scale, \( Y_C(r) \), are given in Table B.1. From the table we can see that the biases of the parameters decreased with informative censoring level; for the independent censoring scenario of \( \tau \), \( \{1, 1, 1\} \), the biases of parameters were close to zero. In the other scenarios, the biases of \( \tau \) were negative due to the fact that we assigned a log \( N(0, 1) \) prior to \( \tau \) which has mode \( \{1, 1, 1\} \) in order to stabilize the
MCMC. As a result, the estimates of the \( \tau \)'s whose true values were greater than 1 were shrunk toward the prior mode \( \{1, 1, 1\} \). The moderate prior variance of \( \tau \) also resulted in moderate biases for other parameters of the censoring process. Although the parameter estimates for the censoring process were biased, the biases didn’t affect our estimates of distribution parameters for time to events \( S_1 \) and \( S_2 \) much by the informative censoring model. As seen in Figures 5.1 and 5.2, the survival estimates provided by the informative censoring models were close to the true curves.
Table B.1: Simulation results for the parameters of the censoring process in the operational time scale.

<table>
<thead>
<tr>
<th>True ( {\tau_1, \tau_2, \tau_3} )</th>
<th>Parameters</th>
<th>True Values</th>
<th>Bias</th>
<th>SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>( {2, 2, 3} )</td>
<td>( \alpha_{C0} )</td>
<td>2.8</td>
<td>0.37</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>( \alpha_{C1} )</td>
<td>-1</td>
<td>-0.14</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>( \alpha_{C2} )</td>
<td>-1</td>
<td>-0.19</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>( y_{C0} )</td>
<td>3</td>
<td>0.27</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>( \tau_1 )</td>
<td>2</td>
<td>-0.55</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>( \tau_2 )</td>
<td>2</td>
<td>-0.54</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>( \tau_3 )</td>
<td>3</td>
<td>-1.16</td>
<td>0.26</td>
</tr>
<tr>
<td>( {1.5, 1.5, 2} )</td>
<td>( \alpha_{C0} )</td>
<td>2.8</td>
<td>0.35</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>( \alpha_{C1} )</td>
<td>-1</td>
<td>-0.13</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>( \alpha_{C2} )</td>
<td>-1</td>
<td>-0.19</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>( y_{C0} )</td>
<td>3</td>
<td>0.25</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>( \tau_1 )</td>
<td>1.5</td>
<td>-0.25</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>( \tau_2 )</td>
<td>1.5</td>
<td>-0.29</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>( \tau_3 )</td>
<td>2</td>
<td>-0.49</td>
<td>0.22</td>
</tr>
<tr>
<td>( {1, 1, 2} )</td>
<td>( \alpha_{C0} )</td>
<td>2.8</td>
<td>0.36</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>( \alpha_{C1} )</td>
<td>-1</td>
<td>-0.13</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>( \alpha_{C2} )</td>
<td>-1</td>
<td>-0.17</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>( y_{C0} )</td>
<td>3</td>
<td>0.24</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>( \tau_1 )</td>
<td>1</td>
<td>-0.07</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>( \tau_2 )</td>
<td>1</td>
<td>-0.06</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>( \tau_3 )</td>
<td>2</td>
<td>-0.46</td>
<td>0.20</td>
</tr>
<tr>
<td>( {1, 1, 1} )</td>
<td>( \alpha_{C0} )</td>
<td>2.8</td>
<td>0.08</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>( \alpha_{C1} )</td>
<td>-1</td>
<td>-0.03</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>( \alpha_{C2} )</td>
<td>-1</td>
<td>-0.05</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>( y_{C0} )</td>
<td>3</td>
<td>0.14</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>( \tau_1 )</td>
<td>1</td>
<td>0.02</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>( \tau_2 )</td>
<td>1</td>
<td>0.04</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>( \tau_3 )</td>
<td>1</td>
<td>0.02</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Note: Bias is calculated by subtracting the true value from the posterior mean of a parameter; SSD is the sample standard deviation of the estimates.
B.2.4 Trace Plots of Regression Parameters from the Informative Censoring Model in the Analysis of the B6C3F1/NCTR Male Mouse Heart and Liver Hemangiosarcoma Data

Figure B.1: $\alpha_{E10}$

Figure B.2: $\alpha_{E11}$

Figure B.3: $\alpha_{E12}$

Figure B.4: $y_{E10}$
Appendix C: USE OF THE MATLAB PACKAGES: 

TRINFORM AND MULTITRINFORM

The two program packages below can be requested from Tao Xiao (xiao.51@osu.edu) or Michael Pennell (pennell.28@osu.edu).

C.1 The TRInform Package

C.1.1 Syntax

In this section, we introduce the use of the TRInform MatLab package which applies the methodology described in Section 4.

The TRInform package contains the TRInform function which is used as the main entrance to the Bayesian inference of the threshold regression model for the informative censoring data. The following parameters are used in the TRInform function:

- *dataname*: specifies input dataset. Such dataset must be a survival dataset with current status data, including at least the examination time variable, the censoring indicator for the event time, and the censoring indicator for the examination time. The examination time variable records the time of the examination to obtain the current status of the event occurrence; The censoring indicator for the event time records the current status of the event occurrence, value 1 should
be used to indicate the subject whose time of event of interest is left censored, and 0 should be used to indicate the subject whose time of event of interest is right censored. The censoring indicator for the examination time indicates the observable censoring mechanism, value 0 should be used to indicate the subject whose examination time is by a random event driven by a stochastic process (e.g., death), and 1 should be used to indicate the subject whose examination time is by a non-random event (e.g., planned end of study time). The dataset can also include other independent variables that will be used in the threshold regression model.

- **savepath**: specifies the path to save the output file with the sample values of all parameters at each iteration after the burn-in; this output file can be used to calculate the credible intervals of the parameters.

- **iter**: specifies the number of MCMC iterations after the burn-in.

- **burn**: specifies the number of MCMC iterations as the burn-in.

- **informcens**: a binary indicator, value 1 specifies to use the informative censoring mechanism and value 0 specifies to use the independent censoring mechanism.

- **t**: specifies the examination time variable.

- **d1**: specifies the censoring indicator variable for the event time. For this variable, value 1 should be used to indicate the subject whose time of event of interest is left censored, and 0 should be used to indicate the subject whose time of event of interest is right censored.
• $d2$: specifies the censoring indicator variable for the examination time. This variable indicates the observable censoring mechanism, value 1 should be used to indicate the subject whose examination time is by a non-random event (e.g., planned end of study time), and 0 should be used to indicate the subject whose examination time is by a random event driven by a stochastic process (e.g., death).

• $\mu_1$: specifies the vector of variables that are linked to the drift parameter of the Wiener stochastic process corresponding to the event of interest. Note that this vector can be a null vector meaning no covariate is linked to this drift parameter.

• $\ln y01$: specifies the vector of variables that are linked to the initial status parameter of the Wiener stochastic process corresponding to the event of interest. Note that this vector can be a null vector meaning no covariate is linked to this initial status parameter.

• $\mu_2$: specifies the vector of variables that are linked to the drift parameter of the Wiener stochastic censoring process before the event of interest occurs. Note that this vector can be a null vector meaning no covariate is linked to this drift parameter.

• $\ln y02$: specifies the vector of variables that are linked to the initial status parameter of the Wiener stochastic process corresponding to the censoring event. Note that this vector can be a null vector meaning no covariate is linked to this initial status parameter.
• *mu3*: specifies the vector of variables that are linked to the drift parameter of
the Wiener stochastic censoring process after the event of interest occurs. Note
that this vector can be a null vector meaning no covariate is linked to this drift
parameter.

• *lny03*: specifies the vector of variables that are linked to the initial status
parameter of the Wiener stochastic censoring process after the event of interest
occurs. Note that this vector can be a null vector meaning no covariate is linked
to this initial status parameter.

• *t1_init*: specifies the variable that is used as the MCMC initial values of the
event occurrence time.

• *y1_init*: specifies the variable that provides the MCMC initial values of the
event process (i.e., \(Y_E(t)\) in Section 4.2.2) value at examination time given the
event of interest is right censored at examination time.

• *y2_init*: specifies the variable that provides the MCMC initial values of the value
of the censoring process which is independent of the event process (i.e., \(Y_{C1}(t)\)
in Section 4.2.2) given that the random censoring event driven by a stochastic
process is right censored at the augmented event occurrence time or planned
examination time.

• *y3_init*: specifies the variable that provides the MCMC initial values of the value
of the censoring process which is dependent of the event process (i.e., \(Y_{C2}(t)\)
in Section 4.2.2) given that the random censoring event driven by a stochastic
process is right censored at the planned examination time.
• $C_{\text{init}}$: specifies the variable that provides the MCMC initial values of the indicator variable $R$ defined in Section 4.2.2, which equals 1 if $Y_{C1}(t)$ is upgraded to $Y_{C2}(t)$ when event of interest $E$ occurs and which equals 0 otherwise.

• $alpha1\_init$: specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the Wiener stochastic process corresponding to the event of interest.

• $sigma2\_1\_init$: specifies the preset value of the inherent variability parameter of the Wiener stochastic process corresponding to the event of interest. Usually this value is set to 1.

• $alpha2\_init$: specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the censoring process which is independent of the event process (i.e., $Y_{C1}(t)$).

• $sigma2\_2\_init$: specifies the preset value of the inherent variability parameter of the censoring process which is independent of the event process (i.e., $Y_{C1}(t)$). Usually this value is set to 1.

• $alpha3\_init$: specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the censoring process which is dependent of the event process (i.e., $Y_{C2}(t)$).
- \textit{sigma2} \_\textit{init}: specifies the preset value of the inherent variability parameter of the censoring process which is dependent of the event process (i.e., \( Y_{C2}(t) \)). Usually this value is set to 1.

- \textit{pic} \_\textit{init}: specifies the initial value of the \( \pi \) parameter which is the probability that the censoring process turns out to be dependent on the event process after the occurrence of the event of interest.

- \textit{beta1} \_\textit{init}: specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the event of interest.

- \textit{Cov\_beta1} \_\textit{init}: specifies a matrix of MCMC initial values for the sampling variance covariance matrix of the Metropolis-Hastings random walk algorithm on the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the event of interest.

- \textit{beta2} \_\textit{init}: specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is independent of the event process (i.e., \( Y_{C1}(t) \)).

- \textit{Cov\_beta2} \_\textit{init}: specifies a matrix of MCMC initial values for the sampling variance covariance matrix of the Metropolis-Hastings random walk algorithm on the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is independent of the event process (i.e., \( Y_{C1}(t) \)).
link function for the initial status parameter of the censoring process which is independent of the event process (i.e., $Y_{C1}(t)$).

- \textit{beta3_init}: specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is dependent of the event process (i.e., $Y_{C2}(t)$).

- \textit{Cov_beta3_init}: specifies a matrix of MCMC initial values for the sampling variance covariance matrix of the Metropolis-Hastings random walk algorithm on the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is dependent of the event process (i.e., $Y_{C2}(t)$).

- \textit{alpha1_prior}: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the Wiener stochastic process corresponding to the event of interest.

- \textit{Cov_alpha1_prior}: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the event of interest.

- \textit{alpha2_prior}: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in
the linear link function for the drift parameter of the censoring process which is independent of the event process (i.e., $Y_{C_1}(t)$).

- **Cov\_alpha2\_prior**: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is independent of the event process (i.e., $Y_{C_1}(t)$).

- **alpha3\_prior**: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the censoring process which is dependent of the event process (i.e., $Y_{C_2}(t)$).

- **Cov\_alpha3\_prior**: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is dependent of the event process (i.e., $Y_{C_2}(t)$).

- **beta1\_prior**: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the event of interest.

- **Cov\_beta1\_prior**: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the event of interest.
• **beta2 prior**: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is independent of the event process (i.e., $Y_{C1}(t)$).

• **Cov_beta2_prior**: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is independent of the event process (i.e., $Y_{C1}(t)$).

• **beta3 prior**: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is dependent of the event process (i.e., $Y_{C2}(t)$).

• **Cov_beta3_prior**: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is dependent of the event process (i.e., $Y_{C2}(t)$).

• **pic_alpha_prior**: specifies the $\alpha$ parameter of the beta prior distribution for the parameter $\pi$ which is the probability that the censoring process turns out to be dependent on the event process after the occurrence of the event of interest.

• **pic_beta_prior**: specifies the $\beta$ parameter of the beta prior distribution for the parameter $\pi$ which is the probability that the censoring process turns out to be dependent on the event process after the occurrence of the event of interest.
The outputs of the TRInform function are:

- **alpha1_inform**: the posterior mean of the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the Wiener stochastic process corresponding to the event of interest.

- **beta1_inform**: the posterior mean of the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the event of interest.

- **alpha2_inform**: the posterior mean of the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the censoring process which is independent of the event process (i.e., $Y_{C1}(t)$).

- **beta2_inform**: the posterior mean of the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is independent of the event process (i.e., $Y_{C1}(t)$).

- **alpha3_inform**: the posterior mean of the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the censoring process which is dependent of the event process (i.e., $Y_{C2}(t)$).

- **beta3_inform**: the posterior mean of the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process which is dependent of the event process (i.e., $Y_{C2}(t)$).

- **pic_inform**: the posterior mean of the parameter $\pi$. 159
C.1.2 Example

The following is the program used to analyze the lung tumor data example in Chapter 4.

TRInform–Example.m

```matlab
function TRInform–Example

informcens=1;
iter=100000;
burn=20000;
pmeanvector=zeros(numsimulation,24);
pmeanvector(1,1)=0;
pmeanvector(1,2)=0;
dataname=strcat('F344-rats');
data=csvread(strcat('statadata',dataname,'.csv'),1,0);
data_male=data(data(:,7)==0,:);
data=[data_male];
data(:, 1) = [];
data(:, 2)=data(:, 2)*7;
realdata=1;
savepath==strcat('C:\');
DamienT1=0;
d1=data(:,1);
t=data(:,2);
d2=data(:,3);
z=data(:,4);
x=data(:,5);
w=data(:,6);
samplesize=length(d1);
t1_init=t/2;
y1_init=2.5*ones(samplesize,1);
y2_init=4.8*ones(samplesize,1);
y3_init=2*ones(samplesize,1);
C_init=ones(samplesize,1);
mu1=[x,z];
lny01=[];
mu2=[x,z];
lny02=[];
mu3=[x,z];
lny03=[];

alpha1_init=[-0.027754,0.000797,0.000163]*sqrt(.1);
betal_init=[1.357894];
```

Cov_beta1_init=diag(.001*ones(length(beta1_init),1));
sigma2_1_init=1;
alpha2_init=[-0.093770,0.000044,-0.000116]*sqrt(.1);
beta2_init=[4.125465];
Cov_beta2_init=diag(.001*ones(length(beta2_init),1));
sigma2_2_init=1;
alpha3_init=[-0.257393,0.000171,0.000104]*sqrt(.1);
beta3_init=[4.943687];
Cov_beta3_init=diag(.001*ones(length(beta3_init),1));
sigma2_3_init=1;
pic_init=.5;

alpha1_prior=[0,0,0];
Cov_alpha1_prior=diag(100*ones(length(alpha1_prior),1));
beta1_prior=[0];
Cov_beta1_prior=diag(100*ones(length(beta1_prior),1));
alpha2_prior=[0,0,0];
Cov_alpha2_prior=diag(100*ones(length(alpha2_prior),1));
beta2_prior=[0];
Cov_beta2_prior=diag(100*ones(length(beta2_prior),1));
alpha3_prior=[0,0,0];
Cov_alpha3_prior=diag(10*ones(length(alpha3_prior),1));
beta3_prior=[0];
Cov_beta3_prior=diag(100*ones(length(beta3_prior),1));

pic_n_prior=1;
pic_alpha_prior=pic_n_prior*10;
pic_beta_prior=pic_n_prior*10;

[alpha1_inform,beta1_inform,alpha2_inform,beta2_inform,alpha3_inform,
 beta3_inform,pic_inform]=...
TRInform(dataname,savepath,iter,burn,informcens,...
 t1_init,y1_init,y2_init,y3_init,C_init,...
 alpha1_init,sigma2_1_init,alpha2_init,sigma2_2_init,
 alpha3_init,sigma2_3_init,pic_init,...
 betal_init,Cov_betal_init,beta2_init,Cov_beta2_init,
 beta3_init,Cov_beta3_init,...
 alpha1_prior,Cov_alpha1_prior,alpha2_prior,Cov_alpha2_prior
 ,alpha3_prior,Cov_alpha3_prior,...
 betal_prior,Cov_betal_prior,beta2_prior,Cov_beta2_prior,
 beta3_prior,Cov_beta3_prior,...
 pic_alpha_prior, pic_beta_prior);
estimations_inform=[alpha1_inform,beta1_inform,alpha2_inform,
 beta2_inform,alpha3_inform,beta3_inform,pic_inform];
C.2 The MultiTRInform Package

C.2.1 Syntax

In this section, we introduce the use of the MultiTRInform MatLab package which applies the methodology described in Section 5.

The MultiTRInform package contains the MultiTRInform function which is used as the main entrance to the Bayesian inference of the multivariate threshold regression model for the informative censoring data. The following parameters are used in the MultiTRInform function:

- `dataname`: specifies input dataset. Such dataset must be a bivariate (but can be extended to multivariate) survival dataset with current status data, including at least the examination time variable, two censoring indicators for the two event times, and the censoring indicator for the examination time. The examination time variable records the time of the examination to obtain the current status of the event occurrence; the two censoring indicators for the two event times record the current status of the two event occurrences respectively, value 1 should be used to indicate the subject whose time of the corresponding event of interest is left censored, and 0 should be used to indicate the subject whose time of the corresponding event of interest is right censored. The censoring indicator for the examination time indicates the observable censoring mechanism, value 1 should be used to indicate the subject whose examination time is by a random event driven by a stochastic process (e.g., death), and 0 should be used to indicate the subject whose examination time is by a non-random event (e.g., planned
end of study time). The dataset can also include other independent variables that will be used in the threshold regression model.

- **savepath**: specifies the path to save the output file with the sample values of all parameters at each iteration after the burn-in; this output file can be used to calculate the credible intervals of the parameters.

- **iter**: specifies the number of MCMC iterations after the burn-in.

- **burn**: specifies the number of MCMC iterations as the burn-in.

- **informcens**: a binary indicator, value 1 specifies to use the informative censoring mechanism and value 0 specifies to use the independent censoring mechanism.

- **t**: specifies the examination time variable.

- **d11**: specifies the censoring indicator variable for the event time of the first event stochastic process. For this variable, value 1 should be used to indicate the subject whose time of the first event of interest is left censored, and 0 should be used to indicate the subject whose time of the first event of interest is right censored.

- **d12**: specifies the censoring indicator variable for the event time of the second event stochastic process. For this variable, value 1 should be used to indicate the subject whose time of the second event of interest is left censored, and 0 should be used to indicate the subject whose time of the second event of interest is right censored.

- **d2**: specifies the censoring indicator variable for the examination time. This variable indicates the observable censoring mechanism, value 1 should be used
to indicate the subject whose examination time is by a non-random event (e.g., planned end of study time), and 0 should be used to indicate the subject whose examination time is by a random event driven by a stochastic process (e.g., death).

- \(mu_1\): specifies the vector of variables that are linked to the two drift parameters of the Wiener stochastic processes corresponding to the two events of interest. Note that this vector can be a null vector meaning no covariate is linked to these two drift parameters.

- \(lny_{01}\): specifies the vector of variables that are linked to the initial status parameters of the two Wiener stochastic process corresponding to the two events of interest. Note that this vector can be a null vector meaning no covariate is linked to these initial status parameters.

- \(mu_2\): specifies the vector of variables that are linked to the drift parameter of the Wiener stochastic censoring process in operational time scale. Note that this vector can be a null vector meaning no covariate is linked to this drift parameter.

- \(lny_{02}\): specifies the vector of variables that are linked to the initial status parameter of the Wiener stochastic process corresponding to the censoring event. Note that this vector can be a null vector meaning no covariate is linked to this initial status parameter.

- \(t11_{\text{init}}\): specifies the variable that is used as the MCMC initial values of the first event occurrence time.
• \textit{y11\_init}: specifies the variable that provides the MCMC initial values of the first event process value at examination time given the first event of interest is right censored at examination time.

• \textit{t12\_init}: specifies the variable that is used as the MCMC initial values of the second event occurrence time.

• \textit{y12\_init}: specifies the variable that provides the MCMC initial values of the first event process value at examination time given the second event of interest is right censored at examination time.

• \textit{y2\_init}: specifies the variable that provides the MCMC initial values of the value of the censoring process in operational time scale given that the random censoring event driven by a stochastic process is right censored at the planned examination time.

• \textit{alpha11\_init}: specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the Wiener stochastic process corresponding to the first event of interest.

• \textit{sigma2\_11\_init}: specifies the preset value of the inherent variability parameter of the Wiener stochastic process corresponding to the first event of interest. Usually this value is set to 1.

• \textit{alpha12\_init}: specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the linear link function for the
drift parameter of the Wiener stochastic process corresponding to the second event of interest.

- \( \textit{sigma2}\_1\_\textit{init} \): specifies the preset value of the inherent variability parameter of the Wiener stochastic process corresponding to the second event of interest. Usually this value is set to 1.

- \( \textit{alpha2}\_\textit{init} \): specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the censoring process in operational time scale.

- \( \textit{sigma2}\_2\_\textit{init} \): specifies the preset value of the inherent variability parameter of the censoring process in operational time scale. Usually this value is set to 1.

- \( \textit{b1}\_\textit{init} \): specifies the variable that is used as the MCMC initial values of the random effect.

- \( \textit{beta11}\_\textit{init} \): specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the first event of interest.

- \( \textit{Cov\_beta11}\_\textit{init} \): specifies a matrix of MCMC initial values for the sampling variance covariance matrix of the Metropolis-Hastings random walk algorithm on the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the first event of interest.
• \textit{beta12.init}: specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the second event of interest.

• \textit{Cov.beta12.init}: specifies a matrix of MCMC initial values for the sampling variance covariance matrix of the Metropolis-Hastings random walk algorithm on the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the second event of interest.

• \textit{beta2.init}: specifies a vector of MCMC initial values for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process in operational time scale.

• \textit{Cov.beta2.init}: specifies a matrix of MCMC initial values for the sampling variance covariance matrix of the Metropolis-Hastings random walk algorithm on the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process in operational time scale.

• \textit{tau.init}: specifies a vector of MCMC initial values for the coefficients in the linear transformation of calendar time as in Equation (5.12).

• \textit{Cov.tau.init}: specifies a matrix of MCMC initial values for the sampling variance covariance matrix of the Metropolis-Hastings random walk algorithm on the coefficients in the linear transformation of calendar time as in Equation (5.12).
• \textit{lambda\_init}: specifies the MCMC initial values of $\lambda$ parameter, which is the precision parameter of the normal distribution of the random effect $b$.

• \textit{alpha1\_prior}: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the Wiener stochastic process corresponding to the first event of interest.

• \textit{Cov\_alpha1\_prior}: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the first event of interest.

• \textit{alpha2\_prior}: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the Wiener stochastic process corresponding to the second event of interest.

• \textit{Cov\_alpha2\_prior}: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the second event of interest.

• \textit{alpha2\_prior}: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the censoring process in operational time scale.
• \textit{Cov\_alpha2\_prior}: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process in operational time scale.

• \textit{beta11\_prior}: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the first event of interest.

• \textit{Cov\_beta11\_prior}: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the first event of interest.

• \textit{beta2\_prior}: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the second event of interest.

• \textit{Cov\_beta12\_prior}: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the second event of interest.

• \textit{beta2\_prior}: specifies the vector of mean values of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient)
in the log linear link function for the initial status parameter of the censoring process in operational time scale.

- **Cov\_beta2\_prior**: specifies the variance-covariance matrix of the multivariate normal prior distribution for the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process in operational time scale.

- **tau\_prior**: specifies the vector of mean values of the multivariate normal prior distribution for the coefficients in the linear transformation of calendar time as in Equation (5.12).

- **Cov\_tau\_prior**: specifies the variance-covariance matrix of the multivariate normal prior distribution for the coefficients in the linear transformation of calendar time as in Equation (5.12).

- **lambda\_gamma1\_prior**: specifies the $\gamma_1$ parameter of the Gamma prior distribution for the parameter $\lambda$, which is the precision parameter of the normal distribution of the random effect $b$. Note that this Gamma prior distribution has mean $\gamma_1/\gamma_2$ and variance $\gamma_1/\gamma_2^2$.

- **lambda\_gamma2\_prior**: specifies the $\gamma_2$ parameter of the Gamma prior distribution for the parameter $\lambda$, which is the precision parameter of the normal distribution of the random effect $b$. Note that this Gamma prior distribution has mean $\gamma_1/\gamma_2$ and variance $\gamma_1/\gamma_2^2$.

The outputs of the *MultiTRInform* function are:
• $\alpha_{11_{\text{inform}}}$: the posterior mean of the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the Wiener stochastic process corresponding to the first event of interest.

• $\beta_{11_{\text{inform}}}$: the posterior mean of the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the first event of interest.

• $\alpha_{12_{\text{inform}}}$: the posterior mean of the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the Wiener stochastic process corresponding to the second event of interest.

• $\beta_{12_{\text{inform}}}$: the posterior mean of the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the Wiener stochastic process corresponding to the second event of interest.

• $\lambda_{\text{inform}}$: the posterior mean of the $\lambda$ parameter.

• $\alpha_{2_{\text{inform}}}$: the posterior mean of the regression coefficients (including the intercept coefficient) in the linear link function for the drift parameter of the censoring process in operational time scale.

• $\beta_{2_{\text{inform}}}$: the posterior mean of the regression coefficients (including the intercept coefficient) in the log linear link function for the initial status parameter of the censoring process in operational time scale.
• \textit{tau}\textsubscript{inform}: the posterior mean of the coefficients in the linear transformation of calendar time as in Equation (5.12).

C.2.2 Example

The following is the program used to analyze the heart and liver hemangiosarcoma data example in Chapter 5.

\texttt{MultiTRInform-Example.m}

```
function MultiTRInform-Example

DataIndep=1;
DamienT1=0;
iter=100000;
burn=50000;
datamenamel=strcat('B6C3F1-mice-male-bivariate');
data=csvread(strcat('statadata\',datamenamel,'.csv'),1,0);
realdata=1;
savepath=strcat('C:\');

informcens=1;

data\_male=data(data(:,8)==0,:);
data=[data\_male];
data(:, 1) = [];
data(:, 3)=data(:, 3)*7;
t=data(:,3);
w=data(:,7);
samplesize=length(t);

t11\_init=t/2;
y11\_init=2.5*ones(samplesize,1);
d11=data(:,1);
t12\_init=t/2;
y12\_init=2.5*ones(samplesize,1);
d12=data(:,2);
t=data(:,3);
tau\_init=[1.0072,2.1181,1.6738];
alpha11\_init=[0.6513,0.00011555,-0.0023185];
beta11\_init=[1.5127];
Cov\_beta11\_init=diag(.002*ones(length(beta11\_init),1));
sigma2\_11\_init=1;
alpha12\_init=[0.11387,-0.00014441,-0.0029678];
beta12\_init=[5.1311];
Cov\_beta12\_init=diag(.002*ones(length(beta12\_init),1));
sigma2\_12\_init=1;
```

172
d2=data(:,4);

z=data(:,5);
x=data(:,6);
b1_init=.5*ones(samplesize,1);

lambda_init=31.295;
y2_init=y11_init;

mu1=[x,z];
lny01=[];
mu2=[x,z];
lny02=[];

alpha2_init=[-0.028353,-6.30E-06,-0.00035174];
beta2_init=[3.798];
Cov_beta2_init=diag(.002*ones(length(bet2_init),1));
sigma2_2_init=1;
Cov_tau_init=diag(.002*ones(length(tau_init),1));

alpha11_prior=[0,0,0];
Cov_alpha11_prior=diag(10000*ones(length(alpha11_prior),1));
beta11_prior=[log(3)];
Cov_beta11_prior=diag(10000*ones(length(beta11_prior),1));
alpha12_prior=[0,0,0];
Cov_alpha12_prior=diag(10000*ones(length(alpha12_prior),1));
beta12_prior=[log(3)];
Cov_beta12_prior=diag(10000*ones(length(beta12_prior),1));
lambda_gamma1_prior=2;
lambda_gamma2_prior=.5;

[alpha11_inform,beta11_inform,alpha12_inform,beta12_inform,lambda_inform
 ,alpha2_inform,beta2_inform,tau_inform]=...
MultiTRInform(dataname,realdata,savepath,iter,burn,
informcens,...
t,d11,d12,d2,mul,lny01,mu2,lny02,...
t11_init,y11_init,t12_init,y12_init,y2_init,...
alpha11_init, sigma2_11_init, alpha12_init, sigma2_12_init, alpha2_init, sigma2_2_init, bl_init,...
beta11_init, Cov_beta11_init, beta12_init, Cov_beta12_init, beta2_init, Cov_beta2_init, tau_init, Cov_tau_init, lambda_init,...
alpha11_prior, Cov_alpha11_prior, alpha12_prior, Cov_alpha12_prior, alpha2_prior, Cov_alpha2_prior,...
beta11_prior, Cov_beta11_prior, beta12_prior, Cov_beta12_prior, beta2_prior, Cov_beta2_prior, tau_prior, Cov_tau_prior,...
lambda_gamma1_prior, lambda_gamma2_prior);
estimations_inform=[alpha11_inform, beta11_inform, alpha12_inform, beta12_inform, alpha2_inform, beta2_inform, lambda_inform, tau_inform];
Bibliography


176


*Biometrics*, 50.


Jewell NP, van der Laan MJ (2004). “Current status data: review, recent develop-

Kaplan EL, Meier P (1999a). “Generalized MLE of a joint distribution function with
multivariate interval-censored data.” *Journal of the American Statistical Associa-
tion*, 69, 155–166.

Kaplan EL, Meier P (1999b). “A nonparametric maximum likelihood estimator for
bivariate interval censored data.” *Statistics in Medicine*, 18, 3089–3100.

Kaplan EL, Meier P (2000). “A proportional hazards model for multivariate interval-
censored failure time data.” *Biometrics*, 56, 940–943.


