Modelling Credit Risk using Survival Analysis and Bayesian Techniques

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Sanket Sunil Mohire

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

April, 2024

Supervisor: Prof. T V Ramanathan

Certificate

This is to certify that this dissertation entitled ‘Modelling Credit Risk using Survival Analysis and Bayesian Techniques ’ towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Sanket Sunil Mohire at Savitribai Phule Pune University, Pune under the supervision of Prof. T V Ramanathan, Professor, Department of Statistics, Savitribai Phule Pune University, Pune , during the academic year 2023-2024.

Prof. T V Ramanathan

Committee:

Prof. T V Ramanathan

Dr. Leelavati Narlikar

This thesis is dedicated to my supervisor for his invaluable guidance, and to my grandfather for his profound love and affection towards me

Declaration

I hereby declare that the matter embodied in the report entitled ‘Modelling Credit Risk using Survival Analysis and Bayesian Techniques ’ are the results of the work carried out by me at the Department of Statistics, Savitribai Phule Pune University, Pune, under the supervision of Prof. T V Ramanathan and the same has not been submitted elsewhere for any other degree.Wherever others contribute, every effort is made to indicate this clearly, with due reference to the literature and acknowledgment of collaborative research and discussions.

Sanket

Sanket Sunil Mohire

Roll No : 20191070

Acknowledgments

My experience working on the MS thesis has been both exciting and insightful, marking the conclusion of my academic journey at IISER. This achievement would have been impossible without the guidance, affection, and support of numerous individuals.

First and foremost, I extend my sincere gratitude to my supervisor, Prof. T.V. Ramanathan, who introduced me to this wonderful topic to work on. Prof. Ramanathan worked tirelessly to understand my interests and kept me inspired throughout the journey. He has an excellent craft of teaching and imparting knowledge, and he made sure that he worked as hard as the student, making it truly "our thesis."

I am very thankful to my expert, Dr. Leelavati, for her constructive feedback and comments on the project. She introduced the "Bayesian" way of thinking at IISER, which became an interest of me and many more. I am grateful to Dr. Anindya and Dr. Anisa for all their support and help during my days at IISER.

Surviving at IISER would have been very difficult without the motivation and constant support from my friends. We all have come a long way, from discussing topics like "Why did you drop biology in +2?" to "How did you fall in love with quantum cryptography or collective animal behavior?" I must thank Arindam, Atharva, Ashutosh, Raghav, Rajeet, Samarth and Shubhankar for all their academic support. There are many more friends who made my stay at IISER memorable, and I am very grateful to them.

The last year was a thrilling experience with many high and low moments. To sail through it wouldn't have seemed possible without Aishwarya, Anushri, Komal, Tejas, Shantaram, Shital, Shreya and Dr. Madhuri. For some days, it would feel like a tea break with all these people was enough to say that the day was well spent.

I am very thankful to IISER Pune for all the wonderful teachers and excellent facilities. IISER Pune has one of the most beautiful campuses in India, and it reminds me every time how privileged I am for my stay here. I would also like to thank DST Inspire for giving me financial support in the form of the Inspire scholarship.

I am immensely grateful to my grandfather (Dada) for always believing in me and staying by my side. I am thankful to my Tai for all the emotional support and what-not sister duties. Lastly, thank you, Aai and Baba, for letting me make my own choices, both small and large. I love you!

Abstract

Credit-granting institutions provide loans to customers, who may sometimes fail to repay their debts, leading to default. To manage this risk, firms use quantitative credit risk management techniques. These methods help them to estimate and regulate credit risk, ensuring that the firm's risk exposure aligns with its risk tolerance. This contributes to the overall stability of the firm and the broader economy. One of the key metrics estimated through quantitative credit risk management techniques is the Probability of Default (PD), which serves as an input for calculating the Expected Loss (EL).

In this thesis, we focus on applying survival analysis techniques to assess the risk of credit default, by calculating the Probability of Default (PD). Survival analysis involves studying subjects over time in anticipation of encountering an event of interest, such as default. We use survival analysis models such as Cox's proportional hazards model and its extension to mixture cure models. These models have a baseline hazard component, which we estimate by approximating it using a linear combination of different basis functions. We use Markov Chain Monte Carlo (MCMC) techniques with Hamiltonian Monte Carlo sampling for the Bayesian analysis of these models. We apply these models to both Bondora credit data and German credit data, comparing them with traditional estimation procedures such as partial likelihood maximization and the EM algorithm. To evaluate the predictive performance, we discuss the use of ROC curves and the adjustments required for ROC curves when dealing with censored data.

Contents

Chapter 0

Introduction

Risk refers to "any event or action that may adversely affect an organization's ability to achieve its objectives and execute its strategies". Everyone, from individuals to organizations and countries, face risks. Risks can happen in many areas, like finance, health, or safety. While we can't get rid of all risks, we can manage them with safety guidelines to reduce how often and how bad things can be. Managing risks means figuring out what risks might happen, how often they might happen, and how bad they might be. By doing this, individuals and organizations can make smart decisions, understanding the risks and rewards of different options while minimizing potential losses. This applies to areas like finance, health, and cybersecurity. Sometimes, risks come from things we can't predict, and they can cost a lot of money. Therefore, managing risks is an important part of keeping organizations healthy and successful in today's world. Quantitative risk management techniques help experts make better decisions about managing risk within reasonable limits.

Banks and financial institutions lend money to customers. But sometimes, customers don't pay back the money they owe or don't follow the terms of their agreements with the bank. The potential loss a bank might face if a borrower doesn't meet his/her obligations is called credit risk. When a customer fails to pay back the financial institution what he/she owes, then the customer is considered as a defaulter. Examples of defaults include not paying back a loan, missing three consecutive loan payments, or not paying credit card bills. Different types of defaults have different levels of seriousness.

It's crucial for institutions that lend money to assess the creditworthiness of borrowers.

CIBIL score, credit score, and probability of default are some ways to measure credit risk quantitatively. In this thesis, we concentrate on estimating the probability of default, which is the likelihood that a customer won't fully or timely repay their loan and will default.

Mathematically, various techniques like regression models, survival analysis, discriminant analysis, and random forests are used for quantification. While logistic regression has been around for a while, survival analysis, traditionally used in medical and engineering fields, is becoming more important in credit scoring. Narain (1992) (25) was the first to apply survival analysis techniques in a credit risk context.

Regression models, particularly logistic regression, are commonly used. However, Stepanova and Thomas (2002) (26) note that while logistic regression and survival analysis achieve similar accuracy, Tong et al. (2012) (8) highlight key advantages of survival analysis methods. Two of them are as follows:

1. Survival analysis naturally considers the most recent data, even if it is censored. On the other hand, in logistic regression, if one is interested in predicting the probability of default within 24 months, one can't include customers who joined within the past 24 months when building the model. Thus, survival analysis techniques can include the censored data, while logistic regression removes this partially observed data.

2. In survival analysis for credit risk scoring, the goal is to model the distribution of the time until default or some related event. As a result, it becomes feasible to calculate probability of default over any chosen time period, whereas logistic regression is limited to predicting over a single fixed time period.

In classical survival analysis, it is assumed that all observations will eventually experience the event of interest, even with censoring. But in credit risk data, many borrowers will repay their loans and avoid defaulting. These borrowers are called "cured" individuals. With cured individuals present, it is uncertain whether censored individuals will experience the event in the future or not. To deal with this, classical survival analysis has been extended with the mixture cure models. Boag (1949) (28) and Berkson and Gage (1952) (27) were the first to introduce this concept.

When we classify a loan to be good or bad, it is important to measure how accurate our model is. One common way to do this is by calculating the probability of classification error, which is when a good loan is mistakenly labeled as bad, or vice versa. However, this approach does not consider the severity of these errors. Mistakenly classifying a good loan as bad isn't as serious as labeling a bad loan as good, which can result in real losses. To address this, we use receiver operating characteristic (ROC) curves to assess the predictive performance of the classifier. ROC curves help us find the best threshold for classification, balancing sensitivity (true positives) and specificity (true negatives). ROC curves are used in various fields, such as medicine, meteorology, and in machine learning to evaluate different classification algorithms.

In our study, we're using mixture cure models, particularly emphasizing the Cox proportional hazards model, to analyze credit data. Our aim is to estimate both the survival probability and the probability of default for borrowers. The thesis is structured as follows:

- Chapter 1 covers survival analysis concepts and Bayesian techniques.
- Chapter 2 focuses on the Bayesian estimation of the Cox proportional hazards model with a more flexible baseline hazard and compare it with the traditional partial likelihood approach using simulated and real datasets.
- In Chapter 3, we introduce the Bayesian mixture cure model and discuss implementing ROC curves when dealing with censored data.
- Chapter 4 involves implementing the Bayesian mixture cure model on a real dataset and comparing its predictive performance with other parametric mixture cure models estimated using the EM algorithm.
- Finally, Chapter 5 outlines potential future research directions.

Chapter 1

Preliminaries: Some Basics of Survival Analysis and Bayesian Estimation

1.1 Survival Data and Survival Analysis

Survival analysis is a branch of statistics that is dedicated to the analysis of time-to-event or survival data. Time-to-event means the time from a well defined origin till the occurence of a particular event of interest. A typical example is time until death of an individual due to a particular terminal disease, after its detection. Similarly, in the context of credit risk, we can consider the time until a loan becomes default. One of the important characteristics of such types of data is that the event of interest may not be observed for all the subjects as it is not possible to follow all the subjects for a longer (may be infinite) time period. This core feature of survival data is known as "censoring"

In "classical" survival analysis, it is generally assumed that even if the data has censoring, all the subjects are susceptible to the event of interest and will experience it eventually. However, it might happen that some fraction of subjects might never experience the event. This is the case, when credit default is the event of interest. A large fraction of loans issued will not default as they will be paid back by the borrower. Since the event never occurs, such subjects are considered as long term survivors and called as “cured” or “non-susceptible”. In order to account for the cure fraction present, mixture cure models are used in survival analysis. In this section, we introduce survival data, we further formulate the credit risk problem in terms of survival analysis,.

1.1.1 Survival data

Survival analysis observes **subjects** spanning **time** until an **event** occurs. Cox (1972)(14) states that survival data requires three elements:

1. Time origin; the time when the subject became at risk
   (In the credit risk context, it is the time when the loan contract begins)

2. Time scale; the passage of time
   (In the credit risk context, the amount of time in months or days from the start of the loan contract)

3. Event or set of events; defined clearly according to our interest
   (an event of default or early repayment)

1.1.2 Hazard and Survival Functions

Let T be the duration until a particular event occurs. In credit risk context, this T represents the time until a borrower defaults on a loan. T represents the duration of survival and it is a non negative random variable. We generally treat T as a continuous variable.

The distribution of T can be represented in three equivalent forms, namely the survival function $S(t)$, hazard rate function $h(t)$ and the cumulative hazard function $H(t)$.

The survival function represents the probability of the subject surviving beyond time t. It is the probability that the event of interest has not occurred till time t. That is:

$$S(t) = P(T > t) = 1 - F(t) = 1 - \int_0^t f(u)du \tag{1.1}$$

where $F(t)$ is the cumulative distribution function and $f(u)$ is the density function of T. The function $S(t)$ is right continuous and monotonically decreasing in time t with $S(0) = 1$. As T is assumed to be continuous random variable, the survival function will be a continuous and strictly decreasing function.

The hazard rate corresponds to the instantaneous risk of occurrence of the event given that the subject of interest has not occurred till time t. It is defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{1.2}$$

The hazard function can take various shapes with the only restriction being $h(t) \geq 0$.

It is defined as the area under the hazard function up to time t, that is

$$H(t) = \int_0^t h(u)du \tag{1.3}$$

When T is a continuous random variable, the Survival function and density function can be written in terms of hazard function as follows: From (1.2),

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{P(T \geq t)} = \frac{f(t)}{S(t)} = -\frac{dln(S(t))}{dt} \tag{1.4}$$

Therefore,

$$S(t) = e^{-\int_0^t h(u)du} = e^{-H(t)} \tag{1.5}$$

and

$$f(t) = h(t)S(t) = h(t)e^{-\int_0^t h(u)du} = h(t)e^{-H(t)} \tag{1.6}$$

The hazard rate $h(t)$ can also be referred to as the default rate in the context of credit risk. However, for consistency, we will use the term "hazard rate." The density function $f(t)$ represents the probability density of the event of interest, which in this case is default. Therefore, when discussing the estimation of the probability of default, we are essentially estimating the density function of default. It's important to understand that the relationships 1.4, 1.5, 1.6 are important because estimating any one of $h(t)$,$f(t)$ or $S(t)$ uniquely determines the other two.

For further details about these basic survival quantities, one can refer to Klein and Moeschberger (2005)(16)

1.1.3 Censoring

One important characteristics of survival data is the existence of censored observations. The exact event time is unobserved for these observations and it is only known to have occurred in certain interval, if it did. Thus such observations provide partial or incomplete information about the event time. Various factors lead to censoring, such as:

- Subjects may drop out of a study or become lost to follow-up,
- the study may conclude before (or start after) the occurrence of the event of interest.

Depending on the kind of information a censored observation provides, censoring can be mainly classified into three types:

Left Censoring

Left censoring occurs when it is only known that the event has occurred before a certain time but the exact time is unknown.

Interval Censoring

Interval censored observations are those for which it is only known that the event has occurred between two certain time stamps but when exactly within the interval is not known.

Right Censoring

In right censored data, events are observed only if they occur before a specified time. The simplest form is Type I censoring, where all censored subjects have the same censoring times because the study ends for everyone at a predetermined time. Other version is generalized

Type I censoring, where subjects may enter the study at different times, but the endpoint is fixed. Type II censoring occurs when the study stops after a certain number of subjects experience the event, with the remaining subjects being right-censored. Regardless of the type, the total number of subjects in the study is predetermined. Random censoring is another type, where censoring times are random variables. For instance, accidental deaths lead to random censoring as patients can't be followed up.

In cases of right-censored data, we typically observe the follow-up time, Y, and a censoring indicator, δ, rather than directly observing the event time, T. Here, Y represents the minimum of the event time T and the censoring time C, while δ indicates whether the event has occurred before censoring ($\delta = 1$ if $T \leq C$, and $\delta = 0$ otherwise). It's commonly assumed that event and censoring times are independent.

In our study, unless stated otherwise, we assume random right censoring. This aligns well with the credit risk data, where censoring occurs when the loans mature without default, end without default, or are terminated due to reasons like borrower's death. All these situations represent instances of right censoring, while other forms of censoring are very rare in the context of credit risk analysis.

1.1.4 Likelihood

Censoring has a far reaching consequence on the construction of the likelihood function given the survival data. It's necessary to recognize that each observation contributes varying levels of information based on its censoring status. In this thesis, our focus will be solely on right-censored data. An observation corresponding to the event time provides information for the likelihood of the event happening precisely at that time. Whereas, a right-censored observation offers information solely about the event not occurring until the observed follow-up time.

Let $(Y_i, \delta_i), i = 1, 2, .., n$ be the n independent and identically distributed random realisations of (Y, δ). Under the important assumption of the independence of event time T and the censored time C, the likelihood function for right censored data is given by,

$$\mathcal{L} = \prod_{i=1}^{n} f(Y_i|\theta)^{\delta_i} S(Y_i|\theta)^{1-\delta_i} \tag{1.7}$$

equivalently, given that $f(Y_i) = h(Y_i|\theta)S(Y_i|\theta)$, this likelihood can also be expressed as:

$$\mathcal{L} = \prod_{i=1}^{n} h(Y_i|\theta)^{\delta_i} S(Y_i|\theta) \tag{1.8}$$

1.1.5 Kaplan-Meier Estimator

Censoring poses a significant impact on the estimator of the survival function. When censoring occurs, the complement of the empirical distribution function cannot be used to estimate the survival function as it exclude censored observations. And excluding censored observations from estimation would result in the loss of the potential information that these observations could contribute. To address this issue for the right-censored data, Kaplan and Meier (1958) (17) introduced a widely used estimator for the survival function, known as the Kaplan-Meier estimator or product-limit estimator. This method serves as a non-parametric alternative to the survival function estimator derived from the empirical distribution function in scenarios without censoring. This estimator is the product over the failure times of the conditional probabilities of surviving to the next failure time. Formally, it is given by

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{1.9}$$

where n_i is the number of subjects at risk at time t_i, and d_i is the number of individuals who fail at time t_i. In the credit risk context, n_i is the number of all the accounts which are not defaulted till time t_i and d_i is the number of the accounts defaulted at time t_i.The Kaplan-Meier estimator is a step function that decreases monotonically, with discontinuities occurring at observed event times.The jump size is influenced by both the count of subjects experiencing the event at each t_i and the arrangement of the censoring times before t_i. It is easy to understand that when there is no censoring, Kaplan- Meier estimator turns out to be the empirical survival function (similar to the empirical distribution function)

1.1.6 Survival Analysis Methods in Credit Risk Modeling

The standard survival analysis methods focuses on data from homogeneous populations. However, real-world data often includes individual characteristics that influence the survival times. Predicting an individual's survival distribution based on their covariates is a common problem of interest. In order to incorporate the influence of associated covariates usually the estimators are adjusted. This adjustment involves the use of conditional probabilities (probability given covariates) in the quantitative definitions of survival and likelihood functions. (Equations (1.1),(1.2),(1.3),(1.7)). In this thesis, we will explore the single event Cox's proportional hazards model and its extension to the mixture cure model. Both study the same random variable i.e the time to default

1. Cox proportional hazards Model: This model assumes that all customers will default eventually. The observed defaults are uncensored cases and the customers not observed to default are right censored.(14)

2. Mixture cure Cox Model: It relaxes the stringent condition of Cox's Model that all customers will eventually default. This model assumes that the population has a fraction of customers not susceptible to default. In this case, the observed defaults are uncensored cases and the customers not observed to default are either from non-susceptible population or right censored.

1.1.7 Cox proportional hazards Model

The Cox proportional hazards model is a fundamental component in survival analysis where the logarithm of hazard for a subject is written as the sum of linear combination of covariates and logarithm of the baseline hazard.Let T_i denote the event time for subject i and C_i represent the corresponding non-informative right censoring time, thus yielding the observable survival time $Y_i = min(T_i, C_i)$. Each observed Y_i, denoted as y_i (where i= 1,..., n), serves either as a recorded event time ($\delta_i = 1$) or as a censoring time ($\delta_i = 0$).Let there be k explanatory covariates organized in the vector $\mathbf{x}_i^T = [x_1, ..., x_k]$ available for each subject and let β be the associated vector of parameters. The observed data becomes $(y_i, \mathbf{x}_i, \delta_i)$ and can be used to estimate the Cox model given as follows:

$$h(t|x_i) = h_0(t)e^{\mathbf{x}_i^T \beta} \tag{1.10}$$

Here, $h_0(t)$ represents the baseline hazard, which is essentially the hazard for an individual when all covariates are set to zero, at time t. And $\mathbf{x}_i^T \beta$ is the linear predictor computed for individual i at time t.

The Cox model is often known as the "proportional hazards" (PH) model due to its characteristic of maintaining proportional hazards between subjects. This means that the hazard ratio between two distinct subjects, such as A and B, remains constant over time.

$$\frac{h_A(t|x_A)}{h_B(t|x_B)} = e^{(x_A^T - x_B^T)\beta}$$

This constant ratio is referred to as the relative risk.

To estimate the parameters $\beta^T = [\beta_1, ..., \beta_k]$ in (1.10),the usual maximum likelihood procedure can be used. However, the maximum likelihood estimation requires a strong assumption of the parametric form of the baseline hazard like constant hazard (Exponential model) or monotonically increasing or decreasing (Weibull model). Since $h_0(t)$ is unspecified for most of the real datasets, it cannot be done directly. So, partial likelihood approach, introduced by Cox (14) and employed by Breslow (1974)(22), is being used to do inference about the model parameters, β.

The general idea is to express a model parameter as a function of other parameters, eliminating it from the likelihood function. This technique is commonly employed when dealing with a 'nuisance' parameter.(i.e the baseline hazard in the case of Cox PH model). In Cox PH model, this is done as follows:

1. Estimate the baseline hazard $h_0(t)$ non-parametrically given β
2. Substitute $h_0(t)$ by its estimator in the likelihood

Assuming that the observed times (y_i) are in increasing order, the partial likelihood function, when there are no uncensored ties (i.e., no two uncensored observations have the same observed time), takes the following form:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\boldsymbol{x}_i^T \beta}}{\sum_{k \in \mathscr{R}(y_i)} e^{\boldsymbol{x}_k^T \beta}} \right]^{\delta_i} \tag{1.11}$$

The numerator depends only on information from the subject who experiences the event of interest at the observed time y_i, whereas the denominator considers information from all subjects in the risk set $\mathscr{R}$ (i.e: those who have not yet experienced an event). Estimation of β is carried out by maximizing (1.11).

Using these estimates of β, the Breslow's estimator for cumulative baseline hazard is given by (in case of no uncensored ties):

$$\hat{H}_0(t) = \sum_{y_i \leq t} \hat{h}_0(y_i) = \sum_{y_i \leq t} \frac{1}{\sum_{k \in \mathscr{R}(y_i)} e^{\boldsymbol{x}_k^T \beta}} \tag{1.12}$$

where

$$\hat{h}_0(t) = \begin{cases} \frac{1}{\sum_{k \in \mathscr{R}(t)} e^{\boldsymbol{x}_k^T \beta}} & \text{if } t \text{ is an event} \\ 0 & \text{otherwise} \end{cases} \tag{1.13}$$

1.1.8 Mixture Cure Models

When survival data includes a cure fraction, we classify observations into two types:

1. Those experience the event are considered susceptible or ‘uncured’
2. Those never experience the event are considered non-susceptible or ‘cured’.

In such a case, survival function, hazard rate etc. will undergo a modification, justifying the use of cure models.

When dealing with a cure fraction, it is a common practice to assume that a cured subject has $T = \infty$, indicating that the event never occurs, while $T < \infty$ for a non-cured subject. As a result, as time $t \to \infty$, a portion of the observations may remain free from the event. Therefore,

$$\lim_{t \to \infty} S(t) > 0$$

This limiting value, represented by $1 - p$, is the proportion of cured or non-susceptible fraction, known as the cure rate. Similarly, the cumulative hazard function is bounded from above:

$$\lim_{t \to \infty} H(t) < \infty$$

It means, as t grows, the accumulated instantaneous risk of experiencing the event does not approach infinity but rather reaches a plateau, indicating that some subjects will not experience the event.

Due to censoring, the susceptibility status is not directly observed. If we denote the susceptibility status as $\mathcal{S}$ (where $\mathcal{S} = 1$ if $T < \infty$), it is clear that an uncensored observation, $\delta = 1$, has $\mathcal{S} = 1$ because $Y = T$. On the other hand, for censored observations, $\delta = 0$, and therefore, $Y = C$. However, since censoring affects both susceptible and non-susceptible subjects: non-susceptible individuals because the event never occurs, and susceptible individuals because follow-up is not infinite, we cannot determine $\mathcal{S}$ in those cases. Hence, the susceptibility status is only partially observed through the censoring indicator.

An implication of the partially observed cure status is when building the likelihood function. In cure survival analysis, observations fall into two categories: censored or uncensored, similar to classical survival analysis. Uncensored observations contribute to the likelihood function through the density function, while censored observations contribute through the survival function. Notably, there's no distinction between cured and uncured censored subjects. Although the likelihood function retains the same form as (1.7), the survival and density functions differ to accommodate the presence of a cure fraction.

When the Kaplan-Meier estimator of the survival function shows a plateau in the right tail, it indicates the presence of a non-susceptible or cured fraction in the population.

1.2 The Bayesian Estimation

Bayesian estimation allows us to easily add external knowledge into statistical inference through prior distributions. Incorporating such knowledge can enhance the precision of estimates, reduce errors, improve small sample properties, and refine survival estimates. However, improper integration of external knowledge may lead to biased estimates and increased error rates (18).

In this section, we discuss some of the Bayesian concepts that we have used in this thesis. Now in Bayesian inference, any kind of statistical question one can ask has to come down to manipulation of the posterior, which is given by the Bayes theorem as

$$Posterior \propto Likelihood \times Prior \tag{1.14}$$

When the model is too complex, obtaining a closed form expression for the posterior is challenging due to the computation of the normalization factor, which involves difficult integration. This is where Markov Chain Monte Carlo (MCMC) methods come in. Instead of directly handling the posterior density, MCMC represents the posterior with a set of samples. These samples allow for efficient computation of expectations, simplifying the process.

1.2.1 Markov Chain Monte Carlo

The term “Markov Chain Monte Carlo” (MCMC) refers to a range of techniques that help us to simulate observations from the posterior density and compute the required estimator using this generate samples :

1. We aim to sample from a complex density or probability mass function π. This density often arises from Bayesian computations, known as posterior density.

2. Markov chains usually attains a stationary distribution if it exists under certain conditions. By simulating from such a Markov chain for a long enough time with certain restrictions, we can obtain a sample from the chain’s stationary distribution.

3. We want to create a Markov chain whose stationary distribution matches the functional

form of the posterior density π.

4. We want to sample values from this Markov chain. After sufficient burn-in of initial samples, the sequence of values of θ's will turn out to be independent samples generated from the posterior density π.

A Markov chain is a stochastic process that evolves over time by transitioning into different states. The sequence of states is denoted by the collection X_i and the transition probabilities satisfy

$$P(X_t|X_{t-1}, X_{t-2}, ..., X_0) = P(X_t|X_{t-1})$$

This property is known as Markov property. It means that the probability distribution of the process at time t, given all of the previous values of the chain, is the same as the probability distribution given only the the previous value. This property helps in determining the distribution of our next value given just our current value.The set of all possible states that a Markov chain can visit is called the state space and the quantity that governs the probability that the chain moves from one state to another state is the transition kernel or transition matrix.

For a Markov chain with a discrete state space and transition matrix P, let π_n be the probability distribution of the states after n transitions with initial starting probability distribution π_0 and π_* be such that $\pi_* P = \pi_*$. Then π_* is a stationary distribution of the Markov chain and the chain is said to be stationary if it attains this distribution. The basic limit theorem for Markov chains says that, under a specific set of assumptions given below,

$$||\pi_* - \pi_n|| \to 0$$

, as $n \to \infty$, where $||.||$ is the total variation distance between the two densities.

The assumptions are as follows:

1. The stationary distribution π_* exists..

2. The chain is irreducible. Irreducible chain means every state can be reached from every other state.

3. The chain is aperiodic. A chain is considered aperiodic if the number of steps needed

to transition between two states is not a multiple of any integer. In other words, the chain is not restricted to moving in cycles of fixed lengths between specific states.

Time Reversibility

Definition 1.2.1. *A Markov chain is time reversible if*

$$(X_0, X_1, ..., X_n) \overset{D}{=} (X_n, X_{n-1}, ..., X_0)$$

The sequence of states moving in the "forward" direction (with respect to time) is equal in distribution to the sequence of states moving in the "backward" direction. Further, the definition above implies that

$$(X_0, X_1) \overset{D}{=} (X_1, X_0)$$

,

The time reversibility property tells us that for all the states, x, y,

$$P(X_0 = x, X_1 = y) = P(X_1 = x, X_0 = y)$$

$$P(X_0 = x)P(X_1 = y|X_0 = x) = P(X_0 = y)P(X_1 = x|X_0 = y)$$

$$\pi(x)P(x, y) = \pi(y)P(y, x)$$

The last line is called as local balance equation. If the local balance equations are satisfied for a transition matrix P and distribution π, then π serves as the stationary distribution of a chain governed by the transition matrix P.

Time reversibility gives us a way to construct a Markov chain that converges to a given stationary distribution. As long as we can show that a Markov chain with a given transition kernel/matrix P satisfies the local balance equations with respect to the stationary distribution π, we can know that the transition distribution will converge to the stationary distribution.

1.2.2 Metropolis-Hastings

We use Metropolis Hastings algorithm to simulate sample from the stationary distribution of the Markov chain. Let $q(Y|X = x)$ be a transition density for X and Y from which we can easily simulate and let $\pi(X)$ be our target density (i.e. the stationary distribution that our Markov chain will eventually converge to). The Metropolis-Hastings procedure is an iterative algorithm where at each stage, there are three steps. Suppose we are currently in the state x and we want to know how to move to the next state in the state space.

1. Simulate a candidate value $y \sim q(Y|X)$. Note that the candidate value depends on our current state x.

2. Let
$$\alpha(y|x) = \min\left\{\frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}, 1\right\}$$
$\alpha(y|x)$ is referred to as the acceptance ratio.

3. Simulate $u \sim Unif(0, 1)$. If $u \leq \alpha(y|x)$, then the next state is equal to y. Otherwise, the next state is still x.

This three step process represents the transition kernel for our Markov chain from which we are simulating. If $K(y|x)$ is the transition kernel embodied by the three steps above, it can be shown that the Markov chain generated by this transition kernel is time reversible.Eventually, we can be reasonably sure that the samples that we draw from this process are draws from the stationary distribution, i.e. $\pi(X)$. For more details about MCMC and Metropolis-Hastings algorithm, one can refer to Peng (2022) (19)

1.2.3 Hamiltonian Monte Carlo Sampling

When exploring the space, we use different transition kernels based on how we choose $q(y|x)$. Some common samplers like random walk, independence sampler, slice sampler, and hit and run sampler are inefficient in exploring the space. While the Gibbs sampler is effective, it doesn't explore the posterior space coherently, and in high dimensions, the chains may get stuck or fail to explore. Additionally, the running times of these algorithms are not ideal when we have access to additional information such as the gradient of posterior. They are unable to utilize this higher-order information efficiently. This is particularly problematic in high dimensions because the probability of selecting a random point with a high enough density decreases, making it essential to propose points in a way that respects the target density. Hamiltonian Monte Carlo (HMC) sampling addresses this issue.

HMC sampling has a very solid physics background and exploring the posterior space using the vector field constructed by the Hamiltonian dynamics gives the coherency for good exploration.

Hamiltonian Monte Carlo

Let $\pi(x)$ be a probability distribution (here x are the parameters of the distribution) that is to be explored using Hamiltonian dynamics.The idea of the Hamiltonian Monte Carlo is to introduce an auxiliary variable and sample jointly from this bigger space. Consider the space $(x, v) \in R^{d+d}$, where $v \in R^d$ is called as the momentum. The joint density of this space is given by

$$q(x, v) = \pi(x)N(v|0, \Sigma) \propto f(x)N(v|0, \Sigma) \tag{1.15}$$

where Σ is a parameter which we can choose. In other words, the joint density q is the product of two independent densities, π on the x part and a normal density on the momentum v part. Let

$$U(x) = -log(f(x)) \quad K(v) = -log(N(v|0, \Sigma)), \tag{1.16}$$

so that,

$$q(x, v) = \frac{1}{Z}e^{-U(x)}e^{-K(v)} \tag{1.17}$$

where Z be the normalization constant. Suppose we have a ball at position x with momentum v, then make this ball move on the log density $U(x)$ by using Hamiltonian dynamics. That is,

$$\frac{dx}{dt} = \frac{dK}{dv} \quad , \frac{dv}{dt} = -\frac{dU}{dx} \tag{1.18}$$

Let the solution at time t by following the Hamiltonian flow say, $\psi_t(x, v)$ be (x_t, v_t) from some initial point (x, v), that is,

$$(x_t, v_t) = \psi_t(x, v)$$

From the conservation properties of the Hamiltonian dynamics, the Hamiltonian flow ψ_t is time reversible. And hence after sufficient samples drawn from this process, eventually we will be sampling from the stationary distribution, that is, $\pi(x)$

Each step of the HMC Markov chain $(X_0, X_1, ...)$ is determined first by sampling a new independent momentum $\xi \sim N(0, \Sigma = I_d)$, and then running Hamiltonian dynamics equations for a fixed time T, that is, $X_i = \psi_T(X_{i-1}, \xi)$.This is called as the idealized HMC.

```
**Input:** First-order oracle for $f : \mathbb{R}^d \to \mathbb{R}$, an initial point $X_0 \in \mathbb{R}^d$, $T \in \mathbb{R}_{>0}$, $k \in \mathbb{N}$
**for** $i = 1$ **to** $k$ **do**
    Sample a momentum $\xi \sim \mathcal{N}(0, I_d)$;
    Set $(X_i, \xi) = (\psi_T(X_{i-1}, \xi), \xi)$;
**end**
**Output:** $X_k$
```

Algorithm 1: Idealized Hamiltonian Monte Carlo Algorithm (20)

The following theorem from (20) asserts that the HMC chain using the above construction preserves the target density.

Theorem 1.2.1. *Let $f : R^d \to R$ be a differentiable function. Let $T \geq 0$ be the step size of the HMC. Suppose $(X, V)$is a sample from the density*

$$\pi(x, v) = \frac{e^{-f(x)-\frac{1}{2}||v||^2}}{\int e^{-f(y)-\frac{1}{2}||w||^2} d\mu(y, w)}$$

Then the density of $\psi_T(X, V)$ is π for any $T \geq 0$. Moreover the density of $\psi_T(X, \xi)$, where $\xi \sim N(0, I_d)$ is also π. Thus, the idealized HMC algorithms preserves π.

This is a very preliminary introduction to the Hamiltonian Monte Carlo and one can refer to Vishnoi (2021)(20), Betancourt, M. (2017) (35) and (21) for more details.

Chapter 2

Bayesian Estimation of Cox Proportional Hazards Model with Flexible Baseline Hazards

As we discussed, Survival analysis involves following a subject till an event occurs. If event does not occur, the subject is considered to be right censored. The Cox proportional hazards model is a fundamental component in survival analysis where the logarithm of hazard for a subject is written as the sum of linear combination of covariates and logarithm of the baseline hazard. Various applications of the Cox model, such as modelling probability of default or estimating survival probabilities are of interest. Regression coefficients of the Cox model are estimated by maximizing the partial likelihood Cox (1972)(14). And the baseline hazard is not required while estimating these coefficients. The popularity of the partial likelihood is mainly due to its

1. ability to skip the estimation of baseline hazard and focusing mainly on the regression coefficients.
2. ability to establish the asymptotic properties of the regression coefficients estimates

However the partial likelihood approach has the following shortcomings:

1. When the size of the sample is small and there is high censoring, partial likelihood

estimates turn out to be less accurate.(Heinze and Dunkler, 2008, (36))

2. The baseline hazard needs to be estimated indirectly.

Roysten, (2011) (2) suggests that there should be explicit estimation of the baseline hazard function. He proposes a method to approximate the logarithm of the baseline hazard using fractional polynomials and restricted cubic splines. But, this method requires the input of partial likelihood estimates of the regression coefficients. The method does not jointly estimate the regression coefficients and the baseline hazard.

Ma et al. (2014) (1) has developed a maximum likelihood approach that can avoid the shortcomings discussed. Sole above estimation of regression coefficients will not cause problem if the inference is just related to covariate marginal effects or relative risk. But if we want to study survival probability or instantaneous probability of hazard or hazard rate, then baseline hazard estimation is required. As suggested in Hosmer et al. (2008)(3), the Breslow method does provide estimate of baseline hazard for which partial likelihood estimates are required as inputs, but the resulatant hazard estimates are very volatile.

In this thesis, we estimate the baseline hazard and regression coefficents using a bayesian approach. We closely follow the methodology discussed in Ma et al. (2014)(1)

2.1 The likelihood function

We use the same notations that we have used while discussing the Cox Proportional hazards model in the previous chapter. Let the data be of the form (y_i, δ_i, x_i) ∀ i=1 to n , and $h_i(y) = h_0(y)e^{x_i^T \beta}$ be the hazard of the i^{th} subject with covariates x_i^T. Let $\beta^T = [\beta_1, ..., \beta_k]$ be the regression covariate parameters. Then as discussed in the previous chapter the uncensored observations contribute to the likelihood function through the density function while censored observations through the survival function. Therefore the likelihood function is

$$L(\beta, h_0(y)) = \prod_{i=1}^{n} L_i(\beta, h_0(y_i)), \tag{2.1}$$

where y_i is the observed time and δ_i is the censor indicator for the i^{th} subject and $L_i(\beta, h_0(y_i)) = f(y_i)^{\delta_i} S(y_i)^{1-\delta_i}$. The log-likelihood then takes the form

$$l(\beta, h_0(y)) = \sum_{i=1}^{n} \{\delta_i log(f(y_i)) + (1-\delta_i) log(S(y_i))\} \tag{2.2}$$

We can substitute in the above equation the following from the fact that

$$log(S(y)) = -H(y) = -H_0(y)e^{x^T\beta}$$

and

$$f(y) = h(y)S(y)$$

Then the log-likelihood can be written as follows:

$$\begin{aligned} l(\beta, h_0(y)) &= \sum_{i=1}^{n} \delta_i [log(h(y_i)) + log(S(y_i))] + (1-\delta_i) log(S(y_i)) \qquad (2.3) \\ &= -\sum_{i=1}^{n} H(y_i) + \sum_{i=1}^{n} \delta_i log(h(y_i)) \end{aligned}$$

Now catering for the covariates, we get

$$l(\beta, h_0(y)) = -\sum_{i-1}^{n} H_0(y_i) e^{x_i^T\beta} + \sum_{i=1}^{n} \delta_i (log(h_0(y_i)) + x_i^T\beta) \tag{2.4}$$

Estimating the baseline hazard h_0 can be very difficult since it can take different shapes. Since we do not want the baseline hazard $h_0(y)$ to be restricted to a few parametric forms, we use a more common approach of replacing $h_0(y)$ by a function with finite dimensions. Assume $\psi_1, \psi_2, ...\psi_m$ form a basis of this finite dimensional space, then we set

$$h_0(y) = \sum_{u=1}^{m} \theta_u \psi_u(y), \tag{2.5}$$

where ψ_u are non negative basis functions. While many suitable non-negative basis functions for $\psi_u(y)$ are possible we focus on the following functions:

1. B-splines

   B-splines are piecewise polynomial functions used in numerical analysis and computer graphics. They consist of basis functions stitched together to form spline curves. B-splines are versatile tools for approximating complex curves. These basis functions are typically defined recursively using a process known as the Cox-de Boor recursion formula.

   Let k be the number of knots, ξ_1 and ξ_k be 2 boundary knots such that $\xi_0 < \xi_1$ and $\xi_k < \xi_{k+1}$. Now define the augmented knot sequence τ as

   - $t_1 \leq t_2 \leq ... \leq t_M \leq \xi_0$
   - $t_{j+M} = \xi_j, j = 1, 2, ...k$
   - $\xi_{k+1} \leq t_{k+M+1} \leq t_{k+M+2} ... \leq t_{k+2M}$

   The actual values of the additional knots beyond the boundaries can be set arbitrarily and it is conventional to make them all same and equal to ξ_0 and ξ_{k+1}

   Let $B_{i,m}(x)$ be the ith B-spline basis of degree m. Then $B_{i,m}(x)$ is defined recursively as

   $$B_{i,m}(x) = \left(\frac{x - t_i}{t_{i+m} - t_i}\right) B_{i,m-1}(x) + \left(\frac{t_{i+m+1} - x}{t_{i+m+1} - t_{i+1}}\right) B_{i+1,m-1}(x) \text{ for } i = 1, 2, ..., k+2M-m \tag{2.6}$$

   and

   $$B_{i,1}(x) = \begin{cases} 1 & t_i \leq x < t_{i+1} \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

   for $i = 1, 2, ...K + 2M - 1$

   Although these B-spline basis functions are defined recursively, they are polynomials and hence closed form of their integrals and derivatives can be calculated.

2. M-Splines (4)

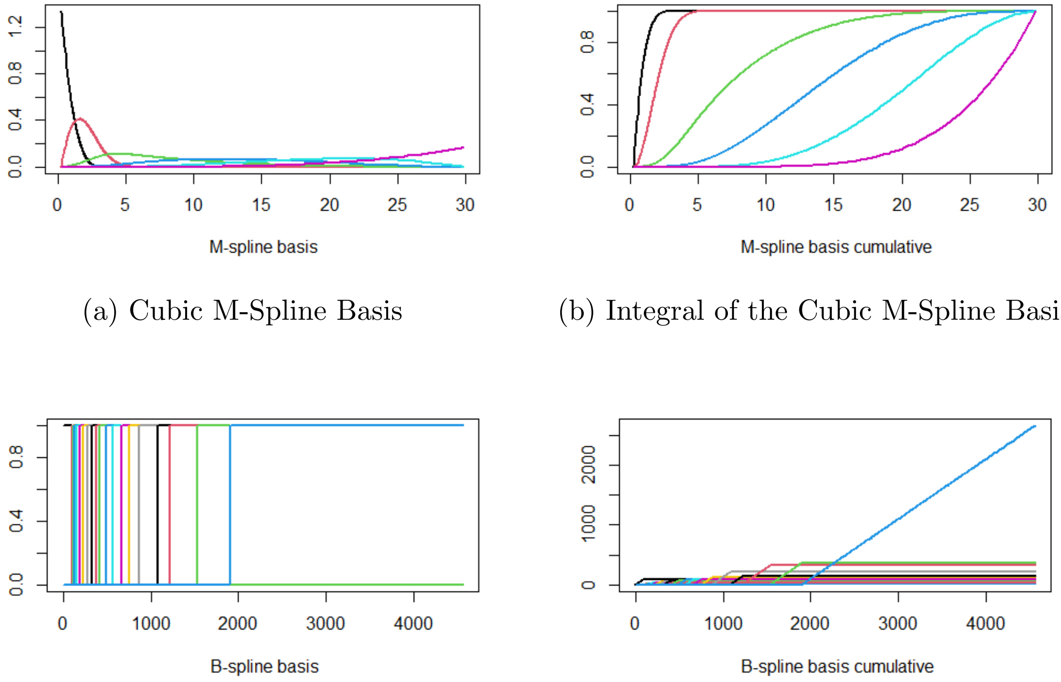
   M-Splines, also refereed as "Curry–Schoenberg B-splines" in De Boor (1978) (33) are considered as normalized B-Splines satisfying

   $$M_{i,m}(x) = \frac{(m+1)B_{i,m}(x)}{t_{i+m+1} - t_i} \tag{2.8}$$

   such that $\int_{t_1}^{t_{K+2M}} M_{i,m}(x) = 1$

3. Indicator functions (0-degree B-Spline basis) which results in a piece-wise constant hazard

Figure 2.1 shows examples of M-splines and B-splines and their integrals



(a) Cubic M-Spline Basis

(b) Integral of the Cubic M-Spline Basis

(c) 0 degree B-Spline basis

(d) Integral of the 0 degree B-Spline basis

Figure 2.1: Examples of M-splines and B-splines and their integrals

The baseline hazard can be made more smooth and flexible by approximating it using Cubic M-Splines and Cubic B-Splines as a function of time. The M-Spline and B-spline basis function matrices can be calculated using the methods from the splines2 R package, Wang (2018)(5) From equation (2.5), the cumulative baseline hazard is given by

$$H_0(y) = \int_0^y h_0(s)ds = \sum_{u=1}^{m} \theta_u \int_0^y \psi_u(s)ds = \sum_{u=1}^{m} \theta_u \Psi_u(y) \tag{2.9}$$

where $\Psi_u(y) = \int_0^y \psi_u(v)dv$. Therefore, the log-likelihood in this case can be written as:

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\sum_{i=1}^{n}\sum_{u=1}^{m} \theta_u e^{x_i^T\beta}\Psi_u(y_i) + \sum_{i=1}^{n} \delta_i(log(\sum_{u=1}^{m} \theta_u \psi_u(y_i)) + x_i^T\beta) \tag{2.10}$$

We wish to jointly estimate $[\beta^T, \boldsymbol{\theta}^T]$ using a Bayesian approach and subject to the constraint $\theta_u \geq 0$ for $u = 1, ..., m$

2.2 The Bayesian Framework

Ma et al. (2014)(1) develop a constrained optimization method that conditionally optimizes one parameter given all the others. But this method is too computationally intensive to implement. Hence we use a Bayesian approach with Hamiltonian Monte Carlo (HMC) which is a form of Markov Chain Monte Carlo (MCMC). HMC calculated the gradients of the posterior distribution and uses it to explore the posterior space more efficiently. We use Stan for computation which implements a specific form of the HMC.

In Bayesian estimation, the regression parameters and spline coefficients are considered as random variables and have to be assigned with prior distributions. However, most of the times these priors are non-informative. But one can choose priors such that they include external judgements and expert insights on the parameters or default rates in the case of credit risk.

2.2.1 Regression Coefficients Priors

Since the regression coefficients can take any value between $(-\infty, \infty)$, one can choose an uninformative prior distribution for the β's such as normal, t, and Cauchy distribution. (15)

2.2.2 Baseline Hazard Priors

The hazard/default rate h_0 is non negative. Since the M-Spline basis functions of the baseline hazard are always non negative, the spline coefficients must also be non negative. However while drawing samples if all θ's tend to zero, then we start getting divergent transitions. Therefore it is advisable to add one more parameter $r \geq 0$ for normalizing constant and estimate θ's such that $\sum_{u=1}^{m} \theta_u = 1$. The hazard can be written as:

$$h_0(y) = r * \sum_{u=1}^{m} \theta_u \psi_u(y)$$

.

A Dirichlet prior can be used for $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)^T$ and a lognormal prior for r as it is non negative.

Posterior Distribution

For Bayesian estimation we use the following independent prior distributions to perform the MCMC with a Hamiltonian dynamics:

1. All regression coefficient $\beta_i \sim N(0, 5)$
2. Spline coefficient
$$\theta \sim Dirichlet([1, ..., 1]_{1*m})$$
3. Normalizing constant $r \sim lognormal(0, 1)$

The log posterior distribution can be obtained using the Bayes theorem as follows:

$$\begin{aligned} log(p(\beta, \boldsymbol{\theta}, \boldsymbol{r}|y, \delta, x) = -\sum_{i=1}^{n}\sum_{u=1}^{m} r * \theta_u e^{x_i^T\beta}\Psi_u(y_i) + \sum_{i=1}^{n} \delta_i(log(\sum_{u=1}^{m} r * \theta_u \psi_u(y_i)) + x_i^T\beta) \\ - \sum_{j=1}^{k}(\frac{\beta_j}{12})^2 - \frac{1}{2}(log(r))^2 - log(r) + \text{constant} \end{aligned} \tag{2.11}$$

where the constant comes from the denominator part of the Bayes theorem.

2.3 Results

In this section, we report the results of a simulation study, where we compare the Bayesian MCMC estimates of the regression coefficients βs with those obtained from the partial likelihood. Additionally, we also compare the baseline hazard estimates obtained through both methods. Then, we fit a Cox model to estimate the probability of default using the Bondora credit data.

2.3.1 Simulation Study

The survival times t_i are simulated from a hazard function given as follows:.

$$h_i(t) = 0.3 * t^2 e^{-0.5x_1 - 2x_2} \tag{2.12}$$

We generate censoring times c_i, independent of t_i, from a uniform distribution $U(0, \nu)$, where ν is chosen to achieve a desired censoring proportion. This process results in independent observations (y_i, δ_i) for $i = 1, ..., n$, where δ_i denotes the censoring indicator.

The model described in (2.12) is a proportional hazards model with baseline hazard $h_0(t)$ as Weibull hazard with scale parameter 0.1 and shape parameter 3 respectively. Regression coefficients are $\beta 1 = -0.5$ and $\beta 2 = -2$ and values for covariates x_1 and x_2 are generated from binomial distributions: $x_1 \sim Bin(1, 0.5)$ and $x_2 \sim Bin(3, 0.4)$.

In order to know how each method is affected by sample size n, and censoring proportion, we choose n=200 and n=2000 with approximate censoring proportions of 18% and 70% for each value of n. Regression coefficients $\beta 1$ and $\beta 2$ are estimated using Cox's partial likelihood approach and spline approximated Bayesian MCMC estimation approach. We approximate $h_0(t)$ as a linear combination of 6 cubic M-spline basis functions, with knots selected to ensure equal event counts in each basis.

The estimated coefficients along with their bias, standard deviation (SD) and mean squared error (MSE) are given in Table 2.1. We observe that for both Partial Likelihood Estimation (PLE) and Bayesian MCMC, the variance increases with censoring proportion but decreases with sample size. Although in these examples the Bayesian MCMC estimates

Table 2.1: Regression coefficients comparison using simulated samples with sizes n=200 and n=2000. For each sample size there are approximately 18%, 70% independent censoring.

		n=200		n=2000	
		19% censoring	69% censoring	17% censoring	70.8% censoring
$\beta 1 = -0.5$	PLE	-0.2042	-0.6248	-0.4632	-0.4483
	Bias	0.2958	-0.1248	0.0368	0.0517
	SD	0.1594	0.2601	0.0501	0.0839
	MSE	0.112906	0.08322705	0.00386425	0.0097121
	Bayesian	-0.0975	-0.5098	-0.4428	-0.4272
	Bias	0.4025	-0.0098	0.0572	0.0728
	SD	0.1408	0.2605	0.0497	0.0845
	MSE	0.1818309	0.06795629	0.00574193	0.01244009
$\beta 2 = -2$	PLE	-1.9871	-2.0807	-2.0973	-1.9208
	Bias	0.0129	-0.0807	-0.0973	0.0792
	SD	0.1619	0.2663	0.0511	0.0780
	MSE	0.02637802	0.07742818	0.0120785	0.01235664
	Bayesian	-1.668	-1.665	-2.036	-1.861
	Bias	0.3320	0.3350	-0.036	0.1390
	SD	0.1408	0.2090	0.0509	0.0718
	MSE	0.1300486	0.155906	0.00388681	0.02447624

are not as efficient as PLE, the plotted Figures (2.2),(2.3),(2.4) and (2.5) of the baseline hazard by both methods show that the Breslow method estimated baseline hazard is having high variability.

2.3.2 Application to Bondora Credit Data

The Bondora dataset (32), obtained from the European P2P lending platform Bondora, consists of 179,236 loan instances spanning from February 2009 to July 2021, and having 112 covariates. These covariates include borrower demographic details, financial information, and loan transaction features. Follow-up time is measured in days from the loan start date to either the contract end date or the default date, with censoring status assigned accordingly. For our analysis, we selected 9 covariates out of the total, which may influence or explain the probability of default. Categorical variables are encoded according to the provided

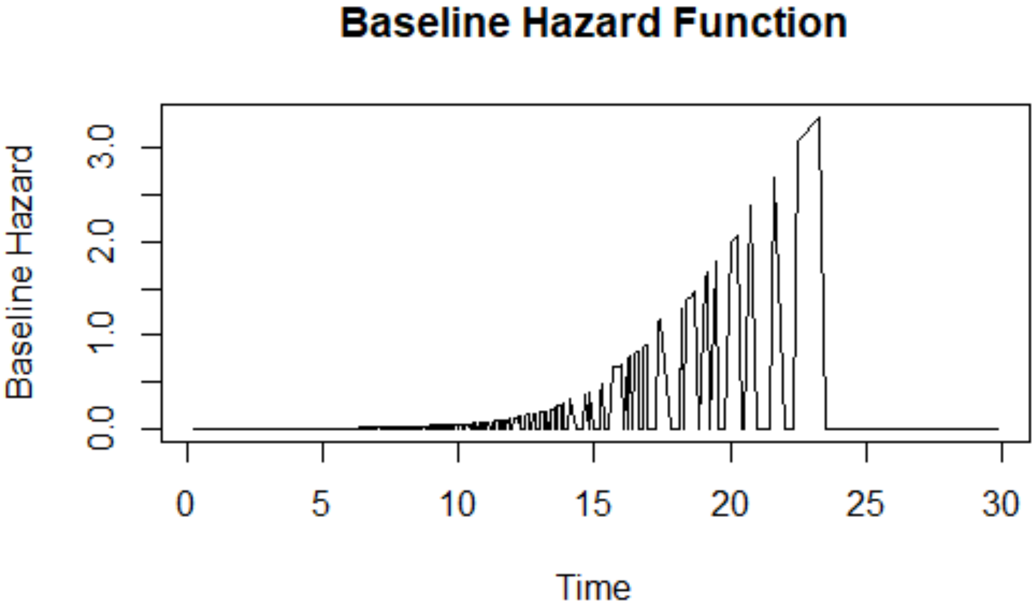


Figure 2.2: Breslow's Baseline hazard Estimate for n=200 and 17% Censoring

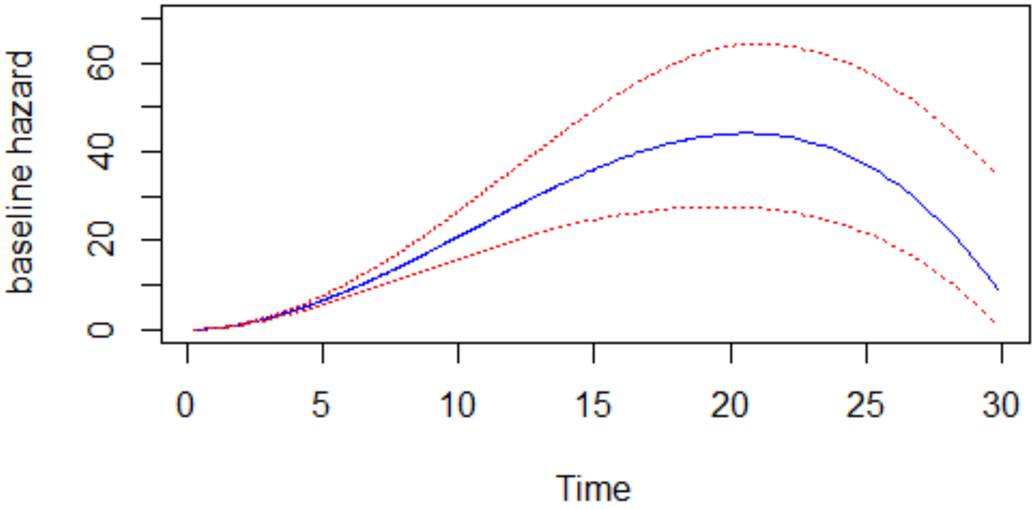


Figure 2.3: Bayesian MCMC Baseline hazard Estimate for n=200 and 17% Censoring

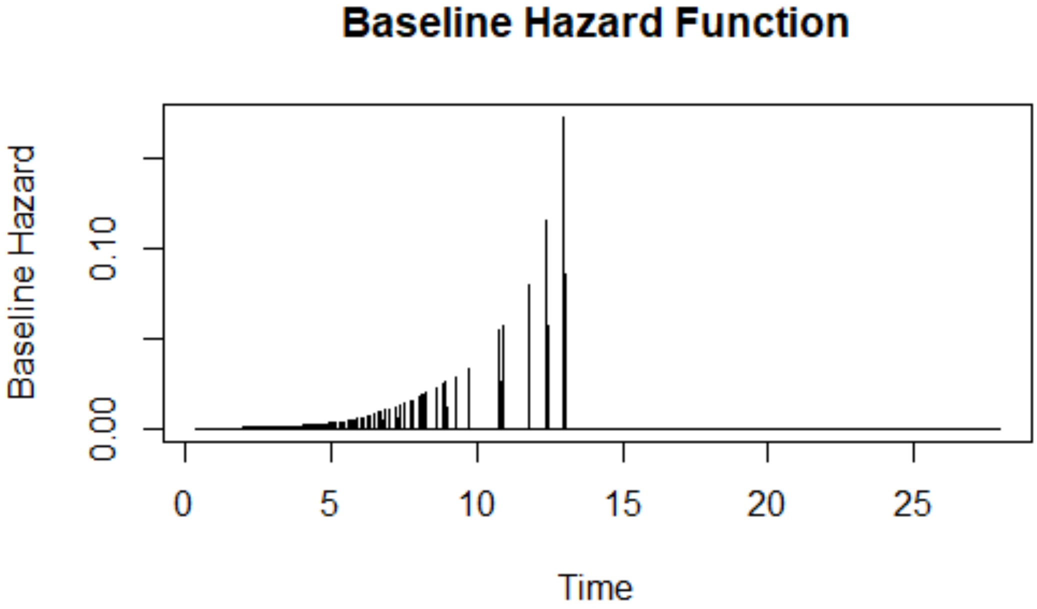


Figure 2.4: Breslow's Baseline hazard Estimate for n=2000 and 70.8% Censoring

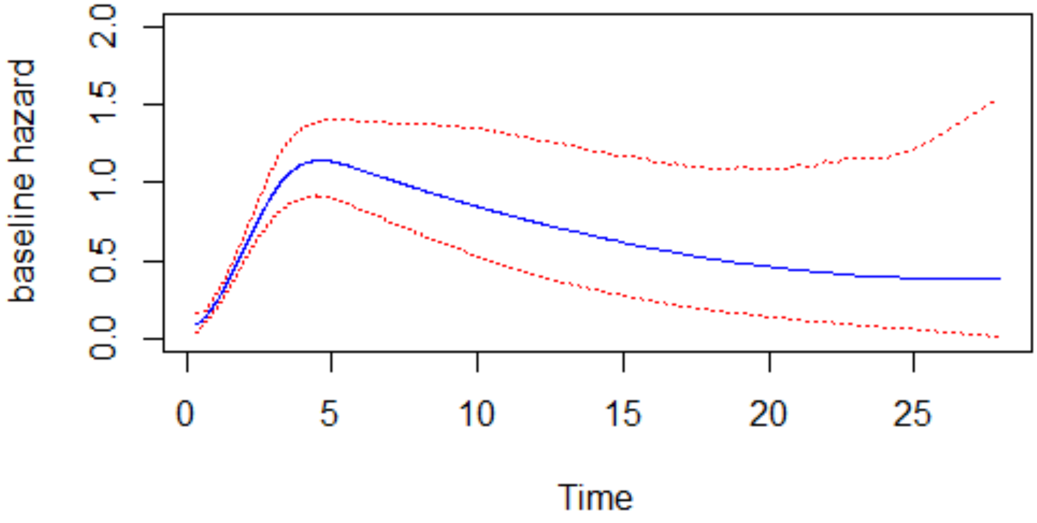


Figure 2.5: Bayesian MCMC Baseline hazard Estimate for n=2000 and 70.8% Censoring

covariate descriptions, while numerical variables are standardized by subtracting the mean and dividing by the standard deviation. A sample of 10,000 observations was chosen from this dataset, with approximately 45.74% of the observations being censored. Further details on the covariates can be found at (31).

We apply the Bayesian MCMC method to fit the Cox's proportional hazards model to the Bondora credit data. The model can be expressed as follows:

$$h(t) = \sum_{u=1}^{20} \theta_u \psi_u(t) e^{\sum_{i=1}^{9} \beta_i x_i} \tag{2.13}$$

We approximate the baseline hazard as a linear combination of 20 indicator functions, with knots chosen to ensure an equal number of observed events within each piecewise constant interval. To compare the regression coefficient estimates obtained from this Bayesian MCMC approach with those from the partial likelihood approach, we select only the significant covariates identified by the PLE method, since we have not developed a significance testing procedure for the Bayesian MCMC method. The selected covariates and the estimated regression coefficients β's are given in Table 2.2. The estimated regression coefficients from both methods are found to be close to each other, indicating that they lead to the same conclusions. The numerical estimates of the baseline hazard, along with 95% credible intervals, are provided in Figure 2.7.

The Breslow baseline hazard estimate given in Figure 2.6 has significantly more variability compared to the smoother estimate obtained from the Bayesian MCMC method. Figure 2.7, shows that the the baseline hazard of default initially peaks and then gradually decreases, suggesting that individuals who survive beyond time 2000 are less likely to default. Our analysis not only yields regression coefficients comparable to those obtained using the partial likelihood approach, but also provides an accurate estimation of the baseline hazard.

While the "stansurv" package (15) offers an option to approximate the baseline hazard using splines and indicator functions in the Cox's Proportional Hazards model, we needed to extend this functionality to the mixture cure model. Therefore, we have developed all the necessary codes independently.

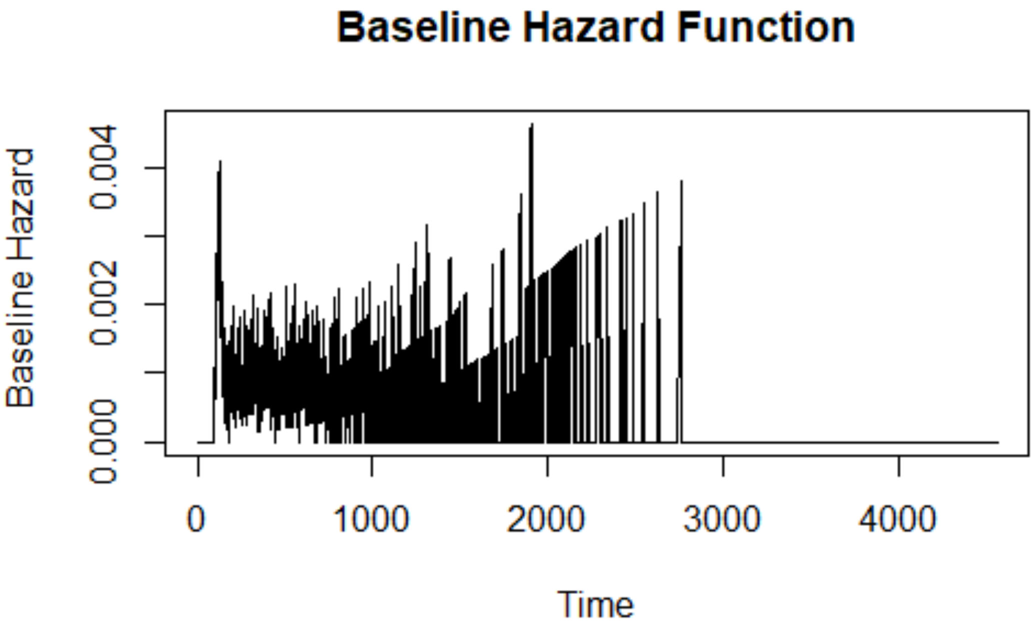


Figure 2.6: Breslow's method estimated baseline hazard for Bondora Credit data

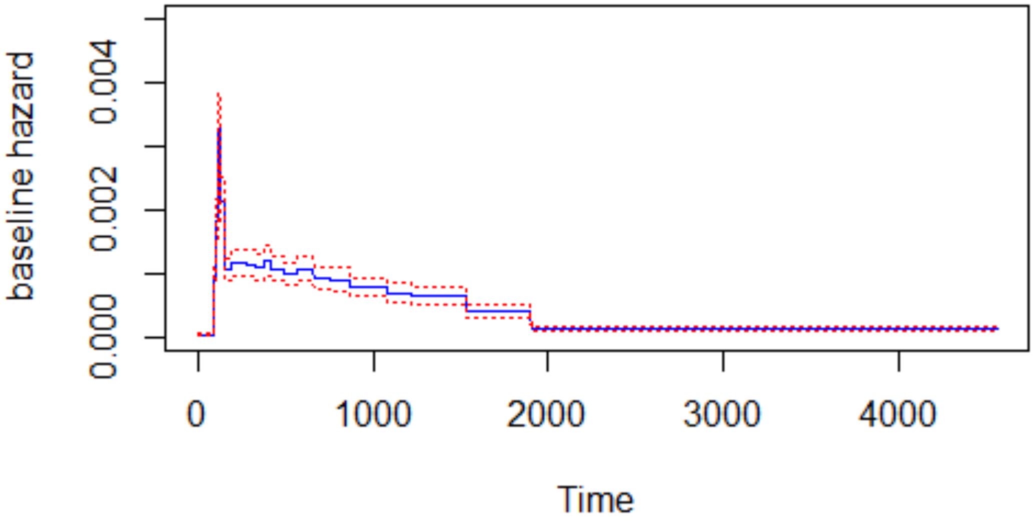


Figure 2.7: Plot of baseline hazard and it's 95% credible interval for Bondora Credit data

Table 2.2: Comparison of the regression coefficient estimates

Covariate Parameter	Partial Likelihood Estimate	Bayesian Estimate
β1 (NewCredit Customer)	0.258750	0.2573
β2 (Verification Type)	-0.102859	-0.1043
β3 (Age)	-0.140354	-0.1410
β4 (Country)	0.409807	0.4089
β5 (Applied Amount)	0.052778	0.05367
β6 (Interest)	0.363773	0.3650
β7 (Use Of Loan)	0.020712	0.02034
β8 (Education)	-0.098560	-0.1013
β9 (Marital Status)	0.077125	0.0762

Chapter 3

Bayesian Estimation of Mixture Cure Cox Model and its Application to Credit Risk Modelling

Mixture cure models originated in medical statistics with the aim of explaining the prolonged survival rates of cancer patients. These models classify patients into two distinct groups: those who achieve a permanent cure and are unlikely to experience a recurrence of cancer, and and those who do not achieve a cure and remain susceptible to the cancer. In this study, we apply mixture cure models to credit risk modelling, where, similar to the medical context, a considerable proportion of subjects may not default, throughout the loan duration.

In standard survival analysis, the survival function is the probability that the subject will not face the event of interest till some stated time t. i.e $S(t) = P(T > t) = 1 - F(t)$ where T is the event time and F is the distribution function of the random variable T. It is assumed that as $t \to \infty$, $S(t) \to 0$ and all the subjects would eventually experience the event of interest. But this need not true in the case of credit data because not every subject would default. A significant portion of accounts might not default throughout the entire loan period. And therefore, the survival function will be plateau at some non-zero levels. These accounts are not susceptible to defaults. Mixture cure models are extensions of classical survival models they are recently becoming popular in the context of credit risk modelling. They are used to model the default rate in terms of two distinct subpopulations Dirick

(2017)(6). In one subpopulation, accounts that are non-susceptible and will not default, while the other subpopulation contains accounts that are susceptible and will default sooner or later.

We have implemented the mixture cure models, which can predict whether borrowers will default or not. These models comprises two components, the incidence, which models the probability of susceptibility of an individual, and the latency, which models the survival distribution for susceptible sub-population. An incidence model is essentially a binary classifier like a logistic regression model or advanced classification methods like tree based methods and clustering algorithms. However we use logistic regression for incidence model as the interpretation becomes somewhat straight. For the latency part, many parametric survival forms have been used in the literature. We use the semi-parametric Cox proportional hazard model that we have introduced in the Chapter 1.

Although the Expectation Maximization (EM) algorithm is frequently used to estimate this model, it does not directly provide estimates of the baseline hazard for susceptible customers or variance estimates for model parameters without further computational steps (7). Therefore, we opt for Bayesian methods of estimation.

3.1 Mixture Cure Model

The mixture cure model differentiates between the two subpopulations of accounts based on the susceptibility to default. The subpopulation that will not default (long term survivors) and another sub-population for those that will eventually default. Hence a binary random variable $\mathcal{S}$ is defined, $\mathcal{S} = 1$ denoting that the account is susceptible to default and will default at some time though it may be censored in the dataset. $\mathcal{S} = 0$ means the account is non-susceptible to default and hence cured. Let δ be the censoring indicator, $\delta = 1$ indicates non censored account and $\delta = 0$ indicates a censored account. Thus, if an account did not default during study period, then it will either not default in the future or is right censored and given sufficient time, it would eventually default. Hence for any account, there are 3 possible states (8):

1. $\delta = 1$ and $\mathcal{S} = 1$; uncensored, susceptible, therefore the account is observed to default;

2. $\delta = 0$ and $\mathcal{S} = 1$; censored, susceptible, hence the account would eventually default;

3. $\delta = 0$ and $\mathcal{S} = 0$; censored, non-susceptible, hence the account will not default.

Note that the susceptibility status of censored subjects remains unobserved.

The mixture cure model is given by

$$S(t|\mathbf{x}, \mathbf{w}) = \pi(\mathbf{w})S(t|\mathcal{S} = 1, \mathbf{x}) + (1 - \pi(\mathbf{w})) \tag{3.1}$$

where $S(t|\mathbf{w}, \mathbf{x})$ is the unconditional survival function for the entire population, $\pi(\mathbf{w})$ is the incidence function representing the proportion of accounts susceptible to default given the covariate vector $\mathbf{w} = (w_1, w_2, ..., w_k)$ and $S(t|\mathcal{S} = 1, \mathbf{x}) = P(T > t|\mathcal{S} = 1, \mathbf{x})$ is the latency function conditional on the account being susceptible to default given the covariate vector $\mathbf{x} = (x_1, x_2, ..., x_s)$. Also note that $\mathbf{x}$ may or may not contain the same covariates as $\mathbf{w}$. Also as $t \to \infty$, $S(t|\mathbf{x}, \mathbf{w}) \to (1 - \pi(\mathbf{w}))$.

The susceptible population proportion given by $\pi(\mathbf{w})$ is modelled using a binary regression model. We use logistic regression because of its convenient parameter interpretation based on the log odds ratio. For the latency part, we use the semi-parametric Cox proportional hazard model. Let α and β be the regression coefficients associated with $\mathbf{w}$ and $\mathbf{x}$ respectively and and $h_0(t)$ be the baseline hazard for the susceptible population. We consider the following specifications:
The incidence part:

$$\pi(\mathbf{w}) = \frac{1}{1 + e^{-\mathbf{w}^T\alpha}} \tag{3.2}$$

The latency part: (given that the account is susceptible to default i.e $\mathcal{S} = 1$) we model the hazard rate /default rate using Cox proportional hazards model given by

$$h(t|\mathcal{S} = 1, \mathbf{x}) = h_0(t)e^{x^T\beta} \tag{3.3}$$

And therefore, the cumulative hazard is

$$H(t|\mathcal{S} = 1, \mathbf{x}_i) = \int_0^t h(t|\mathcal{S} = 1, \mathbf{x}_i) = e^{x^T\beta}\int_0^t h_0(s)ds = H_0(t)e^{x^T\beta} \tag{3.4}$$

where $H_0(t) = \int_0^t h_0(s)ds$ is the cumulative baseline hazard.

Using the relation that $S(t) = e^{-\int_0^t h(s)ds}$, we can write the conditional survival function given that the account is susceptible to default i.e $\mathcal{S} = 1$ as follows:

$$S(t|\mathcal{S} = 1, \mathbf{x}) = e^{-\int_0^t h(s|\mathcal{S}=1,\mathbf{x})ds} = e^{-\int_0^t h_0(s)e^{x^T\beta}ds} = e^{-H_0(t)e^{x^T\beta}} \tag{3.5}$$

Using the relation that $-\frac{dS(t)}{dt} = f(t)$, we can calculate the probability density function associated with the default random variable T as follows:

$$f(t|\mathbf{x}, \mathbf{w}) = -\frac{d}{dt}S(t|\mathbf{x}, \mathbf{w}) = \pi(\mathbf{w})f(t|\mathcal{S} = 1, \mathbf{x}) \tag{3.6}$$

$f(t|\mathcal{S} = 1, \mathbf{x})$ is the conditional probability density function of default at t given that the account i is susceptible.Again using the relation $f(t) = h(t) \cdot S(t)$ the $f(t_i|\mathcal{S} = 1, \mathbf{x})$ can be written in terms of baseline hazard $h_0(t)$ and cumulative baseline hazard $H_0(t)$ as follows:

$$\begin{aligned} f(t|\mathcal{S} = 1, \mathbf{x}) &= h(t|\mathcal{S} = 1, \mathbf{x}) * S(t|\mathcal{S} = 1, \mathbf{x}) \\ &= h_0(t)e^{x^T\beta} * e^{-H_0(t)e^{x^T\beta}} \end{aligned} \tag{3.7}$$

Note that the mixture cure model given in equation (3.1) is completely specified by equations (3.2) and (3.4).

3.2 Methodology and Likelihood

Let the $\mathcal{D} = \{(t_i, \delta_i, \mathbf{w}_i, \mathbf{x}_i); i = 1 \text{ to } n\}$ be the dataset containing all outcomes, and $\delta_i = 1$ if t_i is uncensored and $\delta_i = 0$ otherwise. Then, the uncensored subjects contribute to the likelihood function through the density function $f(t|\mathbf{x}, \mathbf{w})$ while censored observations through the survival function $S(t|\mathbf{x}, \mathbf{w})$.

Hence the likelihood becomes

$$L(\alpha, \beta, h_0(t)) = \prod_{i=1}^{n} f(t_i|\mathbf{x}_i, \mathbf{w}_i))^{\delta_i} S(t_i|\mathbf{x}_i, \mathbf{w}_i)^{1-\delta_i} \tag{3.8}$$

$$L(\alpha,\beta,h_0(t)) = \prod_{i=1}^{n}(\pi(\mathbf{w}_i)f(t_i|\mathcal{S}_i=1,\mathbf{x}_i))^{\delta_i}(1-\pi(\mathbf{w}_i)+\pi(\mathbf{w}_i)S(t_i|\mathcal{S}_i=1,\mathbf{x}_i))^{1-\delta_i} \tag{3.9}$$

The complete likelihood consists of both the logistic and the proportional hazards (PH) component. In cases where there is no non-susceptible fraction, meaning $\pi(\mathbf{w}_i) = 1$, the likelihood function for the mixture cure model simplifies to that of a standard survival model.

Taking logarithm of 3.9,

$$\begin{aligned} l(\alpha,\beta,h_0(t)) = \sum_{i=1}^{n} &\delta_i[log(\pi(\mathbf{w}_i) + log(f(t_i|\mathcal{S}_i=1,\mathbf{x}_i)] \\ &+ (1-\delta_i)log[1-\pi(\mathbf{w}_i)+\pi(\mathbf{w}_i)S(t_i|\mathcal{S}_i=1,\mathbf{x}_i)] \end{aligned} \tag{3.10}$$

$$\begin{aligned} l(\alpha,\beta,h_0(t)) = \sum_{i=1}^{n} &\delta_i[log(\pi(\mathbf{w}_i) + log(h_0(t_i)) + x_i^T\beta - H_0(t_i)e^{x_i^T\beta}] \\ &+ (1-\delta_i)log[1-\pi(\mathbf{w}_i)+\pi(\mathbf{w}_i)e^{-H_0(t_i)e^{x_i^T\beta}}] \end{aligned} \tag{3.11}$$

Estimating the baseline hazard h_0 can be cumbersome since it can take different shapes. Since we do not want the baseline hazard $h_0(t)$ to be restricted to a specific parametric form, we use a more common approach of replacing $h_0(t)$ by a function with finite dimension. Assume $\psi_1, \psi_2, ...\psi_m$ form a basis of this finite dimensional space, then we set

$$h_0(t) = \sum_{u=1}^{m}\theta_u\psi_u(t) \tag{3.12}$$

. where ψ_u are non negative basis functions.

While many suitable non-negative basis functions for $\psi_u(t)$ are possible we focus on the following functions:

1. B-splines
2. M-Splines (4)
3. Indicator functions (0-degree B-Spline basis) which results in a piece-wise constant

hazard

The baseline hazard can be made more smooth and flexible by approximating it using Cubic M-Splines and Cubic B-Splines as a function of time. The M-Spline and B-spline basis function matrices can be calculated using the methods from the splines2 R package (5). From equation (3.12), the cumulative baseline hazard is given by:

$$H_0(t) = \int_0^t h_0(s)ds = \sum_{u=1}^{m} \theta_u \int_0^t \psi_u(s)ds = \sum_{u=1}^{m} \theta_u \Psi_u(t) \tag{3.13}$$

where $\Psi_u(t_i) = \int_0^{t_i} \psi_u(v)dv$.

Therefore the log-likelihood becomes

$$\begin{aligned} l(\alpha, \beta, \boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i (log(\pi(\mathbf{w}_i)) + log(\sum_{u=1}^{m} \theta_u \psi_u(t_i)) + x_i^T \beta - (\sum_{u=1}^{m} \theta_u \Psi_u(t_i)) e^{x_i^T \beta}) \\ +(1-\delta_i) log(1 - \pi(\mathbf{w}_i) + \pi(\mathbf{w}_i) e^{(\sum_{u=1}^{m} \theta_u \Psi_u(t_i)) e^{x_i^T \beta}}) \end{aligned} \tag{3.14}$$

We wish to jointly estimate $[\alpha^T, \beta^T, \boldsymbol{\theta}^T]$ using a Bayesian approach and subject to the constraints $\theta_u \geq 0$ for $u = 1, ..., m$

3.3 Bayesian Framework

Considering the mixture cure model given in (3.1), along with the log-likelihood (3.14), we estimate the parameters using a Bayesian approach with Hamiltonian Monte Carlo Sampling. The choices of the priors for the both, latency β's and incidence α's regression coefficients remain the same as discussed in the previous chapter. For baseline hazard rate, we add a one more parameter $r > 0$ as a normalizing constant and estimate θ's such that

$$\sum_{u=1}^{m} \theta_u = 1 \text{ and } h_0(t) = r \sum_{u=1}^{m} \theta_u \psi_u(t)$$

. A Dirichlet prior is used for $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)^T$ and a lognormal prior for r, as $h_0(t)$ is non negative.

3.3.1 Posterior Distribution

We use the following independent prior distributions to perform the MCMC with Hamiltonian dynamics (15):

1. All latency regression coefficient $\beta_i \sim N(0,6)$

2. All incidence regression coefficient $\alpha_i \sim N(0,6)$

3. Spline coefficient
$$\boldsymbol{\theta} \sim Dirichlet([1, ..., 1]_{1*m})$$

4. Normalizing constant $r \sim lognormal(0,1)$

After determining the likelihood function and fixing the prior distributions for the parameters, the log posterior distribution was obtained using the Bayes theorem as follows:

$$\begin{aligned} log(p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{r}|\mathcal{D}) = \sum_{i=1}^{n}(\delta_i * (log(\pi(\mathbf{w}_i)) + log(r * \sum_{u=1}^{m}\theta_u\psi_u(t_i)) + x_i^T\beta - (r * \sum_{u=1}^{m}\theta_u\Psi_u(t_i))e^{x_i^T\beta}) \\ +(1-\delta_i) * log(1 - \pi(\mathbf{w}_i) + \pi(\mathbf{w}_i)e^{(r*\sum_{u=1}^{m}\theta_u\Psi_u(t_i))e^{x_i^T\beta}})) \\ -\sum_{i=1}^{k}(\frac{\alpha_i}{12})^2 - \sum_{j=1}^{s}(\frac{\beta_j}{12})^2 - \frac{1}{2}(log(r))^2 - log(r) + \text{constant} \end{aligned} \tag{3.15}$$

where the constant comes from the denominator part of the Bayes theorem.

3.3.2 Survival Probability and Probability of Default Prediction

Once the model parameters are estimated, we can predict the survival probability (3.1) and default probability density function (3.6) at time t for a new subject as follows:

Suppose that associated with some account j^* , latency covariate vector x_{j^*} and incidence covariate vector $\mathbf{w}_{j^*}$ is known. Then the predicted survival probability of remaining default free at time $0 < t \leq t_{max}$ where $t_{max} = max(t_i)$ for 1=1,2,... n, denoted by $\hat{S}_{j^*}(t|x_{j^*}, w_{j^*})$ (i.e survival probability at the new data point, from (3.1) can be calculated by taking expectation with respect to the posterior $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{r}|\mathcal{D})$:

$$\hat{S}_{j^*}(t|x_{j^*}, w_{j^*}) = \int S_{j^*}(t|x_{j^*}, w_{j^*}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{r}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{r}|\mathcal{D}) d\boldsymbol{\alpha} d\boldsymbol{\beta} d\boldsymbol{\theta} d\boldsymbol{r} \tag{3.16}$$

Similarly, $\hat{f}_{j^*}(t|x_{j^*}, w_{j^*})$ (i.e probability of default at the new data point, from (3.6) can be calculated by taking expectation with respect to the posterior $p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{r}|\mathcal{D})$:

$$\hat{f}_{j^*}(t|x_{j^*}, w_{j^*}) = \int f_{j^*}(t|x_{j^*}, w_{j^*}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{r}) p(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{r}|\mathcal{D}) d\boldsymbol{\alpha} d\boldsymbol{\beta} d\boldsymbol{\theta} d\boldsymbol{r} \tag{3.17}$$

We approximate the above expectations (3.16) and (3.17) by using samples of MCMC draws from the posterior as follows:

$$\hat{S}_{j^*}(t|x_{j^*}, w_{j^*}) = \frac{1}{K} \sum_{k=1}^{K} S_{j^*}(t|x_{j^*}, w_{j^*}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \boldsymbol{r}^{(k)}) \tag{3.18}$$

$$\hat{f}_{j^*}(t|x_{j^*}, w_{j^*}) = \frac{1}{K} \sum_{k=1}^{K} f_{j^*}(t|x_{j^*}, w_{j^*}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \boldsymbol{r}^{(k)}) \tag{3.19}$$

where $(\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \boldsymbol{r}^{(k)})$ is the k^{th} $(k = 1, ..., K)$ draw from the MCMC independent samples of the posterior distribution

3.4 ROC Curves Estimation

Typically, the Receiver Operating Characteristic, (ROC) curves are one way to assess the predictive performance of a continuous binary classifier. However the classical ROC curve approach (10) is inappropriate since the partially unobserved susceptibility status due to the presence of censoring in survival data. Here we will discuss the adjustments required for

Table 3.1: Possible outcomes of a binary classification.

	$D_i = 0$	$D_i = 1$
$M_i = 0$	true negative	false negative
$M_i = 1$	false positive	true positive

ROC curves when dealing with censored data, as described by Amico, M et al. (2021)(9).

Let D_i be a binary indicator associated with account i (true state of the account) such that,

$$D_i = \begin{cases} 0 & \text{denotes no default} \\ 1 & \text{denotes default} \end{cases}$$

Let M_i be a binary classifier (obtained using estimated incidence) associated with account i (predicted state of the account) such that,

$$M_i = \begin{cases} 0 & \text{if predicted no default} \\ 1 & \text{if predicted default} \end{cases}$$

A classifier is said to have classified the default status of an observation correctly if the values of D and M are the same i.e. when an observation belongs to true positive (TP) or true negative (TN) subjects. Clearly, on the other hand, if an observation belongs to the false positive (FP) or false negative (FN) category, then the classifier has incorrectly classified the default status of an account.

To evaluate if a classifier M, classifies the accounts correctly into default and no default classes, one needs to consider the sensitivity and specificity. Sensitivity tells us the proportion of accounts correctly classified as default when they actually belong to default class, while the specificity indicates the proportion of accounts correctly classified as no default when they actually did not default. When the classifier M is assessed on a continuous scale, it must be divided into two categories in order to conduct binary classification. Therefore, we consider an account i is classified as default when its classifier M_i meets a certain threshold k. That is $M_i \geq k$, for $k \in \mathbb{R}$. Since k can vary, we get multiple sensitivities and specificities.

Since k can take multiple values, there are several possible sensitivities and specificities.To summarize all the information, we need to consider ROC curve, which graphically represents all possible combinations of the sensitivity, and one minus the specificity:

$$Se(k) = Pr(M \geq k|D = 1) \tag{3.20}$$

$$Sp(k) = Pr(M < k|D = 0) \tag{3.21}$$

Equations (3.20) and (3.21) can be obtained from all possible dichotomized versions of the classifier M, based on the value of the threshold k. The ROC curve is sensitivity plotted against one minus specificity for all possible values of $k \in \mathbb{R}$ and is given by:

$$ROC(u) = Se\left\{(1 - Sp)\right\}^{-1}(u), \quad 0 < u < 1 \tag{3.22}$$

where u is an index. A perfect classifier is such that it achieves the probability $Pr(M \geq k|D = 1) = 1$ and $Pr(M \geq k|D = 0) = 0$ for some threshold k. In that case all observations are perfectly classified. Graphically, this corresponds to a point of coordinates $(0, 1)$. On the other hand, an uninformative classifier is such that $Pr(M \geq k|D = 1) = Pr(M \geq k|D = 0)$ for all k. In this situation, the distribution of M is the same in the two classes and the ROC curve is equal to the bisector. Alongside the ROC curve, the area under curve (AUC) is usually defined $AUC = \int_0^1 ROC(u)du$, which summarizes the performance of the classifier M into a single value. An area under the curve (AUC) equal to 1 corresponds to a perfect classifier, while an area under the curve equal to 0.5 corresponds to an uninformative classifier.

3.4.1 Infeasible Estimators

Let T be a non negative random variable which represents the survival time (time to default). We assume that T is subject to random right censoring and instead of observing T we observe the follow up time $Y = min(T, C)$ and censoring indicator $\delta = I(T \leq C)$. C is the censoring time that is supposed to be independent of T given X and W, and $I()$ is the indicator function. Let $(y_i, \delta_i, \mathbf{w}_i, \mathbf{x}_i)$, i=1,2,...n be the iid samples of $(Y, \delta, \mathbf{W}, \mathbf{X})$

A simple and common nonparametric method to estimate a ROC curve consists in estimating the sensitivity and the specificity by their empirical distribution functions given by:

$$\hat{Se(k)} = 1 - \frac{1}{\hat{N1}} \sum_{i=1}^{n} \hat{W_{i1}} \cdot I(M_i \leq k) \tag{3.23}$$

$$\hat{Sp(k)} = \frac{1}{\hat{N0}} \sum_{i=1}^{n} \hat{W_{i0}} \cdot I(M_i \leq k) \tag{3.24}$$

where $\hat{W_{i1}} = I(D_i = 1)$, $\hat{W_{i0}} = I(D_i = 0)$ =1-$\hat{W_{i1}}$, $\hat{N1} = \sum_{i=1}^{n} \hat{W_{i1}}$ and $\hat{N0} = n$ - $\hat{N1}$

When working with survival data having non susceptible population, these estimators cannot be used as the susceptibility status is unobserved. An alternative approach to address this difficulty is to categorize subjects into three types based on the susceptibility threshold proposed by Taylor (1995) (13). This proposal consists in considering an account as non-susceptible if its censored follow-up time is greater than the last uncensored follow-up time, denoted by $\tau = max\{t_i | t_i \text{ is uncensored}\}$. This rule makes sense when there is a clear evidence indicating the existence of a non-susceptible fraction.It is assumed that, when the follow-up period extends beyond the last uncensored event time τ, observations with censored follow-up times greater than most event times can be categorized as non susceptible. Based on this rule, it is therefore possible to distinguish three types of accounts.

1. An uncensored account experiences the event. It then belongs to the susceptible population with certainty, that is, $D = 1$. (It has already defaulted).

2. A censored account with follow-up time $Y > \tau$, we predict it as no default $D = 0$.

3. A censored account with follow-up time $Y \leq \tau$, a probability $Pr(D = 1 | \mathbf{W}, \mathbf{X}, C, T > C)$ replaces the unobserved susceptibility status. When the actual D value is unknown, one can use its probability as an alternative.

As a result, the estimators for sensitivity and specificity functions for survival data having non-susceptible fraction are given by,:

$$\tilde{Se(k)} = 1 - \frac{1}{\tilde{N1}} \sum_{i=1}^{n} \tilde{W_{i1}} \cdot I(M_i \leq k) \tag{3.25}$$

$$\tilde{Sp(k)} = \frac{1}{\tilde{N0}} \sum_{i=1}^{n} \tilde{W_{i0}} \cdot I(M_i \leq k) \tag{3.26}$$

where $\tilde{W_{i1}} = Pr(D_i = 1|\mathbf{W} = \mathbf{w}_i, \mathbf{X} = \mathbf{x}_i, C = C_i, T > C_i)$,
$\tilde{W_{i0}} = Pr(D_i = 0|\mathbf{W} = \mathbf{w}_i, \mathbf{X} = \mathbf{x}_i, C = C_i, T > C_i)$ =1-$\tilde{W_{i1}}$,
$\tilde{N1} = \sum_{i=1}^{n} \hat{W_{i1}}$ and $\tilde{N0} = n$ - $\tilde{N1}$

Using the above sensitivity and specificity estimators, the corresponding ROC curve estimator can be obtained as,

$$\tilde{ROC}(u) = \tilde{Se}\left\{(1 - \tilde{Sp})\right\}^{-1}(u), \quad 0 < u < 1 \tag{3.27}$$

This estimator increases monotonically with u and remains unaffected by strictly increasing transformations of the classifier M, both of which are required characteristics of ROC curves, as outlined by (11). The corresponding estimator for the area under the curve is

$$\tilde{AUC} = \frac{1}{\tilde{N1}\tilde{N0}} \sum_{i=1}^{n} \sum_{j=1}^{n} I(M_j > M_i)\tilde{W_{j1}}\tilde{W_{i0}} \tag{3.28}$$

The development of these estimators (3.25),(3.26),(3.27),(3.28) depends on the decomposition of the sensitivity, the specificity and the area under the curve based on the definition of the conditional probability. Detailed theoretical elements can be found in Section 1 of the supplementary material of y Amico, M et al. (2021)(9).

3.4.2 Feasible Estimators

The $Pr(D = 1|\mathbf{W} = \mathbf{w}_i, \mathbf{X} = \mathbf{x}_i, C = C_i, T > C_i)$ is involved in infeasible estimators (3.25) and (3.26) of the sensitivity and the specificity as well as in the infeasible estimator (3.28) of the area under the curve.Therefore, it is necessary to estimate this quantity in order to obtain estimators that can be used in practice. Based on the definition of the conditional probability, this probability can be written as:

$$Pr(D = 1|\mathbf{W}, \mathbf{X}, C, T > C) = \frac{Pr(T < \infty|\mathbf{W}, \mathbf{X}, C)}{Pr(T > C|\mathbf{W}, \mathbf{X}, C)} = \frac{Pr(T < \infty|\mathbf{W}, \mathbf{X})}{Pr(T > C|\mathbf{W}, \mathbf{X}, C)} \tag{3.29}$$

since T and C are independent given **W** and **X**. Also, as we suppose that the data is coming from the mixture cure model (3.1), this quantity can be further written as:

$$\frac{Pr(T < \infty|\mathbf{W}, \mathbf{X})}{Pr(T > C|\mathbf{W}, \mathbf{X}, C)} = \frac{\pi(\mathbf{w})S(t|\mathcal{S} = 1, \mathbf{x})}{\pi(\mathbf{w})S(t|\mathcal{S} = 1, \mathbf{x}) + 1 - \pi(\mathbf{w})} \tag{3.30}$$

3.5 Results

In this section, we present the results of a simulation study comparing Bayesian MCMC estimates of the latency part coefficients β's and the incidence part coefficients α's with their EM (Expectation Maximization) algorithm estimates. We utilized the EM algorithm via the smcure R package Cao et al.(2012) (12), which employs the Breslow type estimator for the baseline hazard (1.12). Additionally, we compare the Bayesian MCMC estimated baseline hazard under varying sample sizes and censoring proportions. Furthermore, we include simulation results for different parameter dimensions.

For implementation, users can parameterize the baseline hazard by defining the number of observed events within each piece-wise constant interval. Typically, the number of knots is chosen as the cubic root of the total number of events.

3.5.1 Simulation Setting 1

For the simulation study, we first generate survival times using the mixture cure model as specified in (3.1),(3.2) and (3.4) with $n = 200$ and $n = 2000$ observations. We use a Weibull baseline hazard with shape parameter $\lambda = 1.2$ and scale parameter $\nu = 0.8$. Covariates x_1 and x_2 for the incidence part are generated from normal and binomial distributions, respectively, with $x_1 \sim N(0, 1)$ and $x_2 \sim Bin(2, 0.4)$. For the latency part, we generate covariate $w_1 \sim N(1, 1)$. Independent right censoring times for all observations are generated from a uniform distribution to achieve the desired censoring proportion of approximately

25% and 60%. The susceptibility status is simulated using a Bernoulli distribution with probability π_i calculated from the logistic model. The model parameters used are $\beta_1 = 1$, $\beta_2 = -1$, and $\alpha_1 = 4$. Thus, the model used for simulating the data is specified as follows:

$$S(t|x_1, x_2, w_1) = \frac{1}{1 + e^{-4w_1}} * e^{-1.2*t^{0.8}e^{x_1 - x_2}} + (1 - \frac{1}{1 + e^{-4w_1}}) \tag{3.31}$$

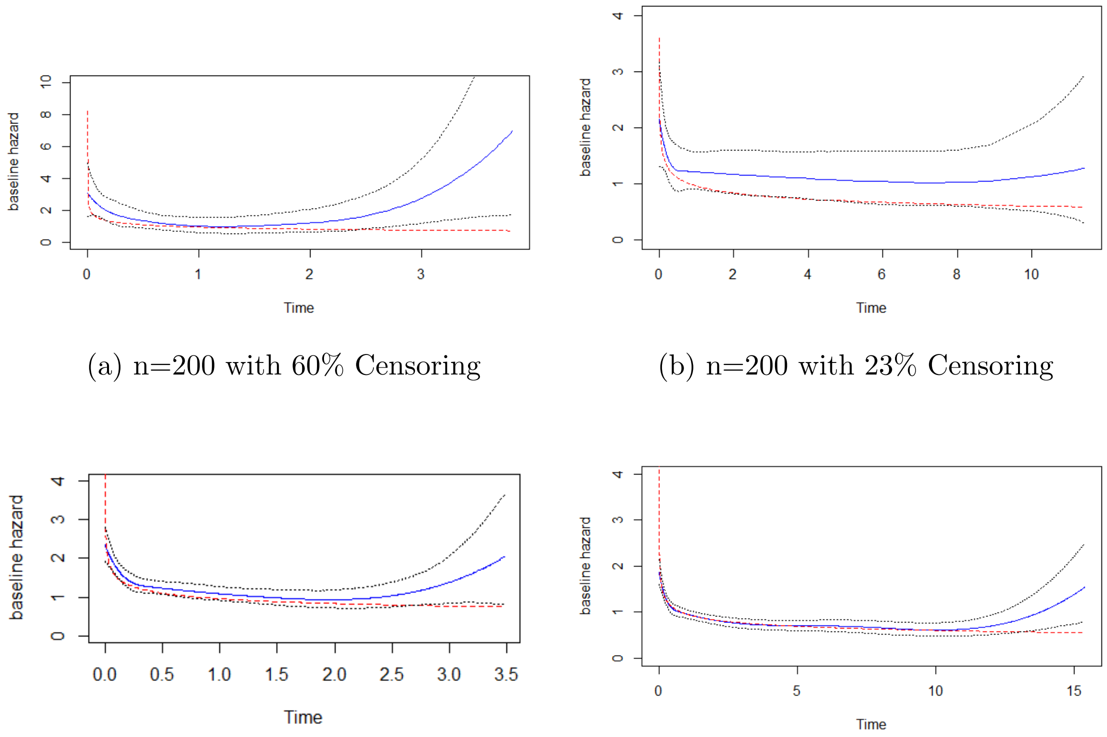
The results from the simulation study are presented in Table 3.2. The Bayesian MCMC approach demonstrates comparable accuracy for coefficients in the survival component (β_1 and β_2) and lower bias for coefficients in the logistic component α_1. The standard deviation (SD) of the estimates follows the trend: $SD_{200,60} > SD_{200,23} > SD_{2000,57} > SD_{2000,27.5}$, indicating that larger sample sizes and lower censoring proportions lead to more precise parameter estimates. Additionally, besides returning estimates for regression parameters, the Bayesian MCMC method also provides estimates for the baseline hazard of the susceptible population. The baseline hazard is parameterized using 6 cubic Mspline basis, with knots selected to ensure a roughly equal number of observed events within each interval. Figure 3.1 illustrates the baseline hazard estimates for different sample sizes and censoring proportions with 95% credible interval. Credible intervals behave similarly to the standard deviation (SD) of covariate parameter estimates. In simpler terms, when there's less data and more censoring, the credible interval tends to be wider. As the sample size increases and censoring decreases, the interval becomes narrower, indicating more precise estimation of the baseline hazard.

3.5.2 Simulation Setting 2

In this simulation study, we check the effect of increasing the number of parameters in both the EM and Bayesian MCMC methods. We use a constant baseline hazard i.e, $\lambda = 1$. We generate data from two models: one with 5 incidence and 4 latency parameters (Table 3.3), and another with 12 incidence and 11 latency parameters (Table 3.4). The covariates are generated as follows: $x_1 \sim N(0, 1)$, $x_2 \sim Bin(2, 0.4)$, $x_3 \sim N(1, 2)$, $x_4 \sim Bin(3, 0.4)$, $x_5 \sim N(3, 4)$, $x_6 \sim Bin(4, 0.4)$, $x_7 \sim N(1, 1)$, $x_8 \sim Bin(2, 0.2)$, $x_9 \sim Bin(3, 0.6)$, $x_{10} \sim N(-3, 2)$, $x_{11} \sim N(-2, 1)$. Both simulations involve generating 2000 observations. Figures 3.5 and 3.6 show the estimated baseline hazards for both models.

Table 3.2: Regression coefficients comparison using simulated samples with sizes N=200 and N=2000. For each sample size there are approximately 25%, 58% independent censoring.

		N=200		N=2000	
		23% censoring	60% censoring	27.5% censoring	57% censoring
$\beta_1 = 1$	EM	0.7726	1.0395	1.047	1.0079
	Bias	-0.2274	0.0395	0.047	0.0079
	SD	0.1172	0.1688	0.0350	0.0495
	MSE	0.0654466	0.03005369	0.003434	0.00251266
	Bayesian	1.0355	1.0344	1.0630	1.0278
	Bias	0.0355	0.0344	0.063	0.0278
	SD	0.0944	0.1397	0.0347	0.0425
	MSE	0.01017161	0.02069945	0.00517309	0.00257909
$\beta_2 = -1$	EM	-0.8003	-0.9617	-0.9472	-1.0715
	Bias	0.1997	0.0383	0.0528	-0.0715
	SD	0.1208	0.2174	0.0429	0.0615
	MSE	0.05447273	0.04872965	0.00462825	0.0088945
	Bayesian	-1.0775	-1.3098	-0.9681	-1.1248
	Bias	-0.0775	-0.3098	0.0319	-0.1248
	SD	0.1254	0.1824	0.0438	0.0601
	MSE	0.02173141	0.1292458	0.00293605	0.01918705
$\alpha_1 = 4$	EM	1.3444	2.1474	3.6444	1.8735
	Bias	-2.6556	-1.8526	-0.3556	-2.1265
	SD	0.3225	1.6428	0.5027	0.3217
	MSE	7.156218	6.130919	0.3791586	4.625493
	Bayesian	4.7148	3.6444	3.9337	3.8572
	Bias	0.7148	-0.3556	-0.0663	-0.1428
	SD	0.9417	1.6342	0.2558	0.4282
	MSE	1.397738	2.797061	0.06982933	0.2037471



(a) n=200 with 60% Censoring

(b) n=200 with 23% Censoring

(c) n=2000 with 57% Censoring

(d) n=2000 with 27.5% Censoring

Figure 3.1: Bayesian MCMC Baseline hazard Estimate for different sample sizes and censoring proportions, Red curve indicates the true baseline hazard

Table 3.3: Estimated parameters from simulation with 5 incidence and 4 latency parameters and constant baseline hazard, N=2000

Incidence parameters	EM Algo	Bayesian MCMC	Latency parameters	EM Algo	Bayesian MCMC
$\alpha_0 = 0$	0.0760	-0.1573			
$\alpha_1 = 1$	0.9545	1.321	$\beta_1 = 1$	1.0210	1.054
$\alpha_2 = -1$	-1.0425	-0.7529	$\beta_2 = -1$	-1.0864	-1.046
$\alpha_3 = 2$	2.0269	2.083	$\beta_3 = 2$	2.0495	2.025
$\alpha_4 = -2$	-2.0759	-2.106	$\beta_4 = -2$	-2.055	-1.982

Table 3.4: Estimated parameters from simulation with 12 incidence and 11 latency parameters and constant baseline hazard, N=2000

Incidence parameters	EM Algo	Bayesian MCMC	Latency parameters	EM Algo	Bayesian MCMC
$\alpha_0 = 0$	0.4427	0.4334			
$\alpha_1 = 1$	1.3925	1.429	$\beta_1 = 1$	0.0986	1.024
$\alpha_2 = -1$	-1.1869	-1.217	$\beta_2 = -1$	-0.0076	-0.9690
$\alpha_3 - 2$	-2.0614	-2.118	$\beta_3 = 2$	0.2128	1.982
$\alpha_4 - 2$	-2.4081	-2.473	$\beta_4 = -2$	-0.0916	-1.881
$\alpha_5 = 3$	3.1634	3.249	$\beta_5 = 3$	0.2163	2.996
$\alpha_6 - 1$	-0.9464	-0.9660	$\beta_6 = -3$	-0.2485	-3.014
$\alpha_7 = 0$	0.1258	1.296	$\beta_7 = 1$	0.0991	1.014
$\alpha_8 = -1$	-1.2773	-1.299	$\beta_8 = -1$	-0.0337	-1.033
$\alpha_9 = 2$	1.8571	1.903	$\beta_9 = 2$	0.1273	2.018
$\alpha_{10} = 2$	2.1510	2.206	$\beta_{10} = 2$	0.1414	2.033
$\alpha_{11} = 3$	3.0232	3.104	$\beta_{11} = 1$	0.0367	0.9224

In this comparison, we observe that the EM algorithm with the Breslow hazard method struggles to accurately estimate the latency parameters when faced with a higher number of parameters. It is also possible that our simulated data, based on a constant hazard, allowed for better estimates using the piecewise constant indicator function for the baseline. However, it is important to note that in scenarios of mixture cure models with numerous parameters, the Breslow method may fail to capture even a constant baseline hazard and the latency parameters.

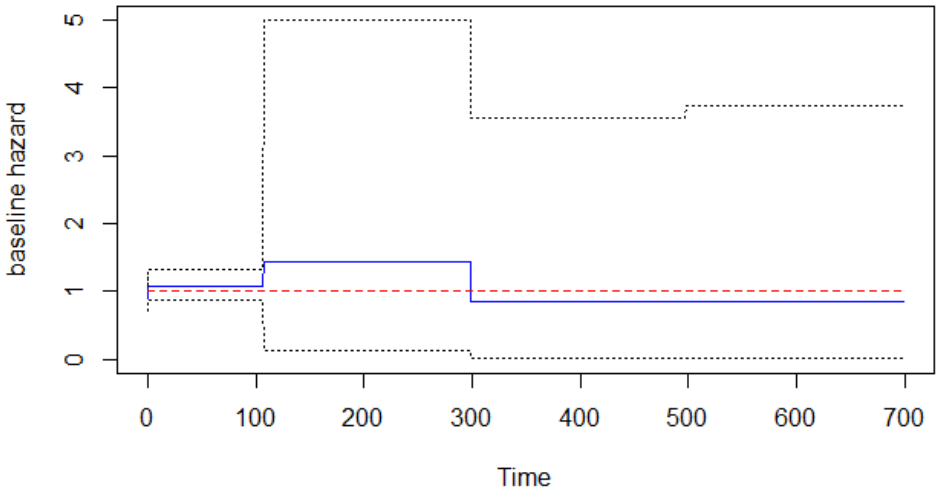


Figure 3.2: Estimated baseline from simulation with 5 incidence and 4 latency parameters with 95% credible interval, Red curve indicates the true baseline hazard

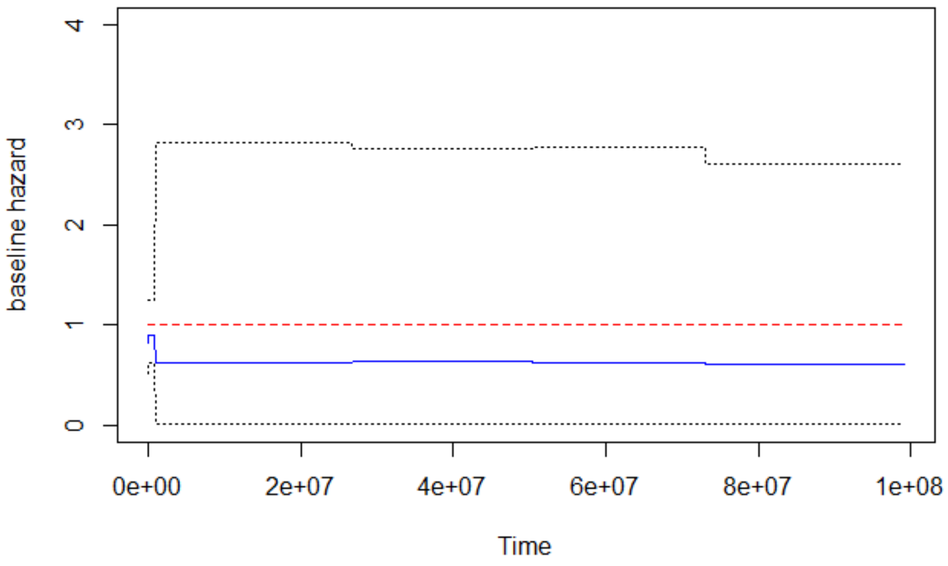


Figure 3.3: Estimated baseline from simulation with 12 incidence and 11 latency parameters with 95% credible interval, Red curve indicates the true baseline hazard

Chapter 4

Data, Implementation and Results

In the previous chapter, we have proposed a Bayesian mixture cure model which has a non parametric baseline hazard for its susceptible population. The model can be used to assess the probability of default of a new loan and its borrower given information about different characteristics. We also discussed how the performance of a mixture cure model can be evaluated using ROC curves and the AUC. In this chapter, we compare our model with other mixture cure models where the the latency part has different parametric forms. The performance of these different models have been compared using the German Credit Data.

4.1 German Credit data

The German credit data (24) is a common dataset in credit risk analysis, with 1000 instances and 20 attributes. Each loan is labeled as either good or bad. In our analysis, we use the duration in months (attribute 2) as the observed follow-time, and credit history (attribute 3) which is also the censoring indicator. A value of A34 in the credit history indicates an uncensored status, while other values indicate censoring. Categorical variables follow a specific format Aij, where i is the variable index and j depends on the number of categories. These variables are encoded as $Aij = j$ for ease of use in the models. Numerical variables are standardized by subtracting the mean and dividing by the standard deviation.Out of the 1000 observations, 293 (29.7%) are considered defaulted. The Kaplan-Meier curve shows a plateau around 40%, with about 2% of censored observations in the plateau. This suggests

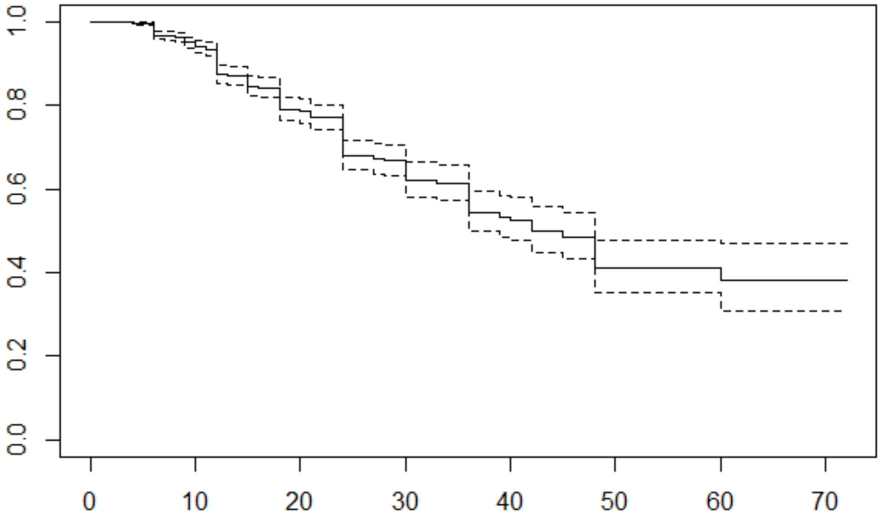


Figure 4.1: Plot of the Kaplan-Meier curve for the German credit data

a cured fraction in the data, making it suitable for our analysis.

4.2 Implementation and Results

We implemented parametric mixture cure models using the “mixcure” package (23). However, for the Bayesian estimation of the parameters associated with the Cox proportional hazards mixture cure model, which is the main focus, we developed our own codes. The codes are given in the Appendix.

There is a challenge in using the complete data likelihood, which includes the unobserved susceptibility status $\mathcal{S}$ for censored observations. In such situations, the unobserved susceptibility status needs to be substituted with its expected value, as described in Tong et al. (2012) (8) and is given by equation (4.1). However, when exploring this posterior using MCMC chains, we encountered numerous divergent transitions, resulting in highly unreliable parameter estimates.

$$E(\mathcal{S}_i) = \begin{cases} 1 & \text{if uncensored observation} \\ Pr(D_i = 1|\mathbf{W} = \mathbf{w}_i, \mathbf{X} = \mathbf{x}_i, C = C_i, T > C_i) & \text{if censored observation} \end{cases} \tag{4.1}$$

where

$$Pr(D_i = 1|\mathbf{W} = \mathbf{w}_i, \mathbf{X} = \mathbf{x}_i, C = C_i, T > C_i) = \frac{\pi(\mathbf{w}_i)S(t|\mathcal{S}_i = 1, \mathbf{x}_i)}{\pi(\mathbf{w}_i)S(t|\mathcal{S}_i = 1, \mathbf{x}_i) + 1 - \pi(\mathbf{w}_i)} \tag{4.2}$$

Another approach of handling missing data problems is to treat the unobserved data as a parameter and assign a suitable prior. However, in cases where the missing data involves discrete parameters, such as binary indicators like susceptibility status, Hamiltonian Monte Carlo (HMC) exploration methods are not applicable. This is because HMC relies on computing gradients, which cannot be done for models with discrete parameters. To overcome this limitation, we use the marginalized likelihood approach, where the complete data likelihood is integrated out with respect to the missing data. This approach has been successfully applied by Basu and Tiwari (2010) (30) in their analysis of cancer data using competing risks mixture cure models.

Choosing the right number of basis functions to approximate the baseline hazard is crucial. Too few functions might not accurately capture the hazard's shape, while too many can lead to overfitting. Therefore, finding an optimal choice is necessary.

We use the German credit data, selecting only significant covariates identified from the Cox proportional hazards model due to limitations with the mixcure() command when handling more than 10 covariates. The selected covariates were used for both the latency and incidence parts. Subsequently, we divide the data into training and test sets in a 7:3 ratio. Next, we apply mixture cure model to the training data, obtaining parameter estimates for both the latency and incidence parts. We then obtain an estimator for the classifier, M, using the incidence estimator. However, we do not estimate the ROC curve using the training data to avoid potential overestimation of M. Instead, we validate our predictions using test data and construct the ROC curve based on these predictions. The ROC curves for various mixture cure models applied to the German credit data are given in Figures 4.2 to 4.7.

Table 4.1 presents the estimated AUC values for different mixture cure models. Notably,

Table 4.1: Estimated AUC values for different mixture cure models

Incidence Model	Latency Model	Estimated AUC
Logistic Model	Cox Proportional hazard with baseline approximated using indicator functions (Bayesian)	0.8669176
	Cox Proportional hazard with baseline estimated using Breslow's Estimator (EM algo)	0.8503274
	Exponential (EM algo)	0.9494817
	Loglogistic (EM algo)	0.9406592
	Log normal (EM algo)	0.9354272
	Weibull (EM algo)	0.9244913

the model with an exponential latency model has superior performance compared to others. Table 4.2 gives the estimated parameters for mixture cure model with exponential latency model.

Among the different approaches using standard parametric forms like exponential, loglogistic, lognormal, and Weibull for the latency model, there is minimal variation in the estimated AUCs. All models perform better than the Cox PH latency model. Bayesian estimation involves generating 4 MCMC chains, each comprising 5000 iterations (3000 warmup, 2000 sampling, thin=1). The Bayesian estimation of the mixture cure model with Cox PH latency model uses a baseline hazard parametrized using 6 piecewise constant indicator functions, with knots selected to ensure an equal number of observed events within each piecewise constant interval. Diagnostic plots of all the estimated parameters are given in Figures 4.11 and 4.12. All chains converge, and the estimates remain stable across different initial hyperparameters. While the mixture cure CPH model with the baseline approximated using indicator functions slightly performed better than the one with Breslow's method estimated baseline, the Bayesian estimated CPH model did not perform well for the German Credit data. This discrepancy may be due to the fact that the data follows an exponential latency model, indicating a constant hazard, wherein the use of more flexible approaches reduces the accuracy.

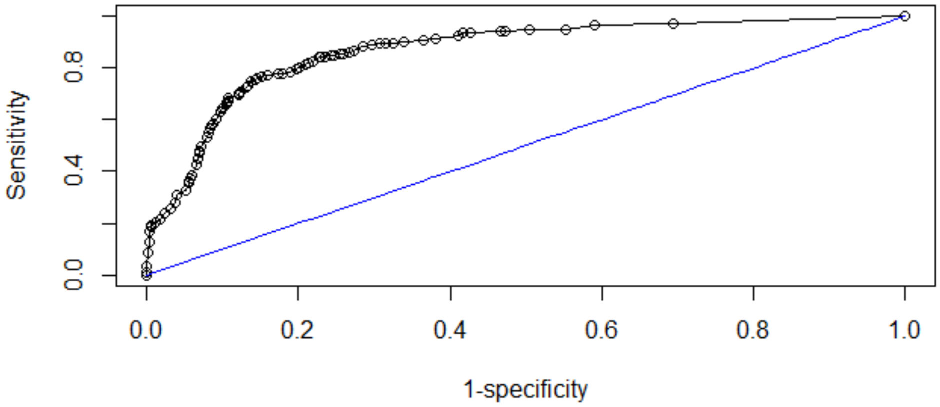


Figure 4.2: ROC curve for mixture cure model with latency model as Cox Proportional hazard with baseline approximated using indicator functions; AUC=0.8669176

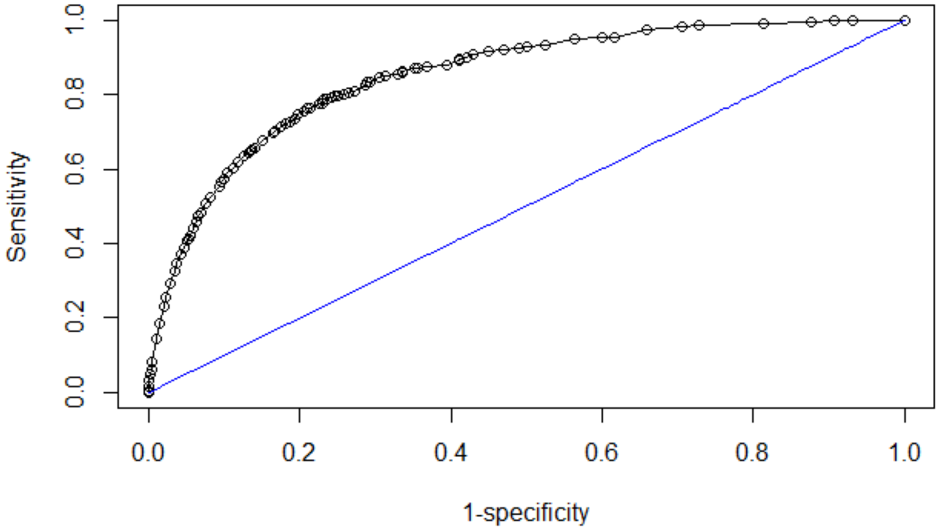


Figure 4.3: ROC curve for mixture cure model with latency model as Cox Proportional hazard with baseline estimated using Breslow's Estimator; AUC=0.8503274

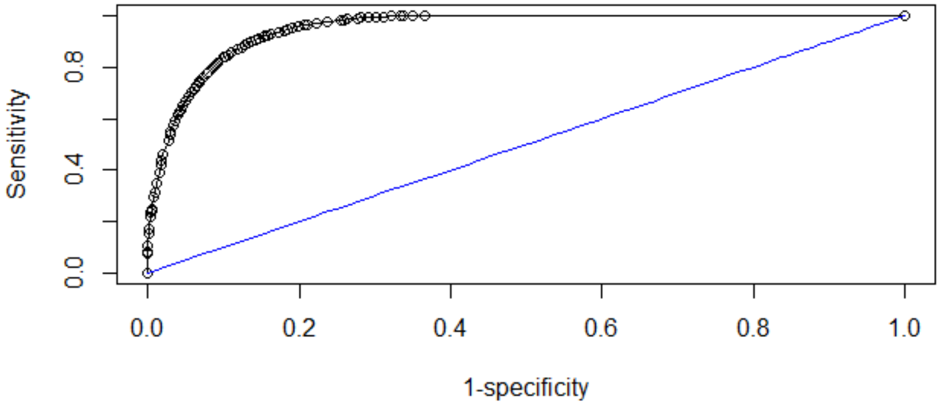


Figure 4.4: ROC curve for mixture cure model with Exponential hazard latency model; AUC=0.9494817

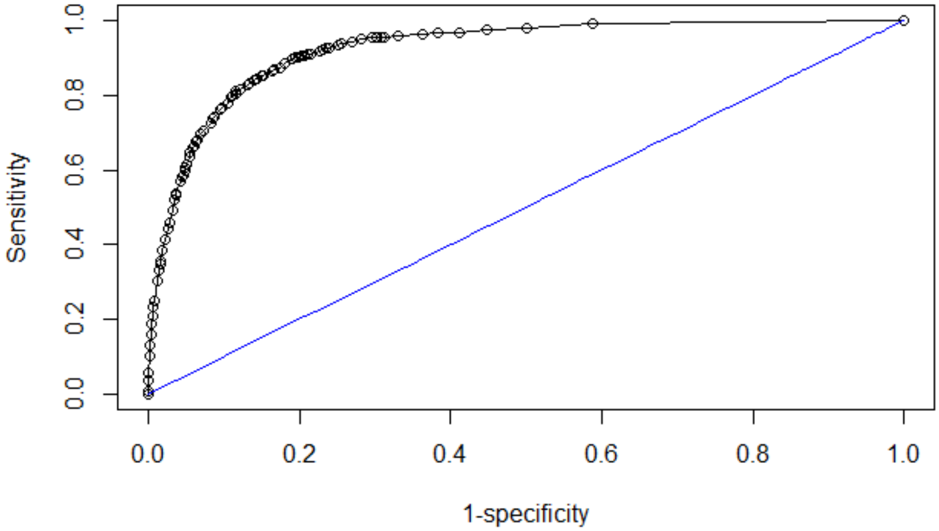


Figure 4.5: ROC curve for mixture cure model with Weibull hazard latency model;AUC=0.9244913

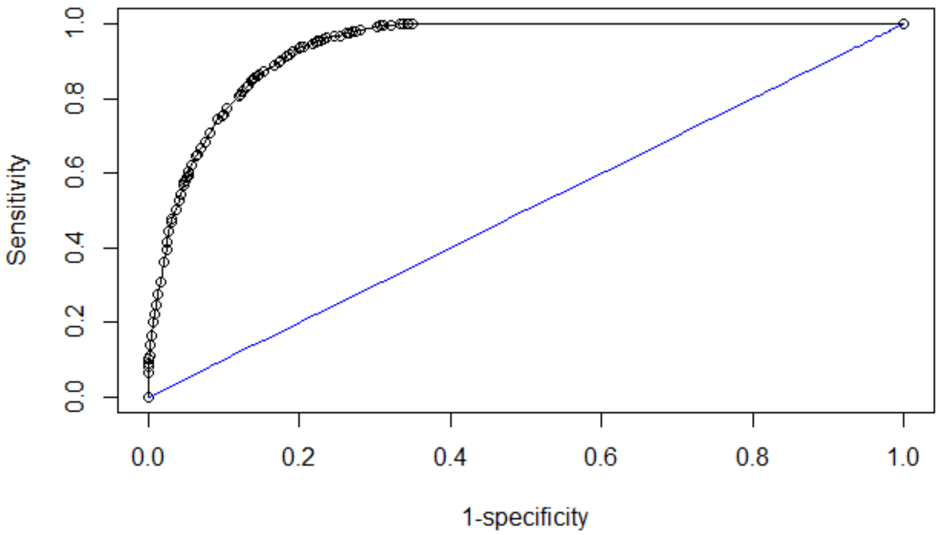


Figure 4.6: ROC curve for mixture cure model with Log-normal hazard latency model; AUC=0.9354272

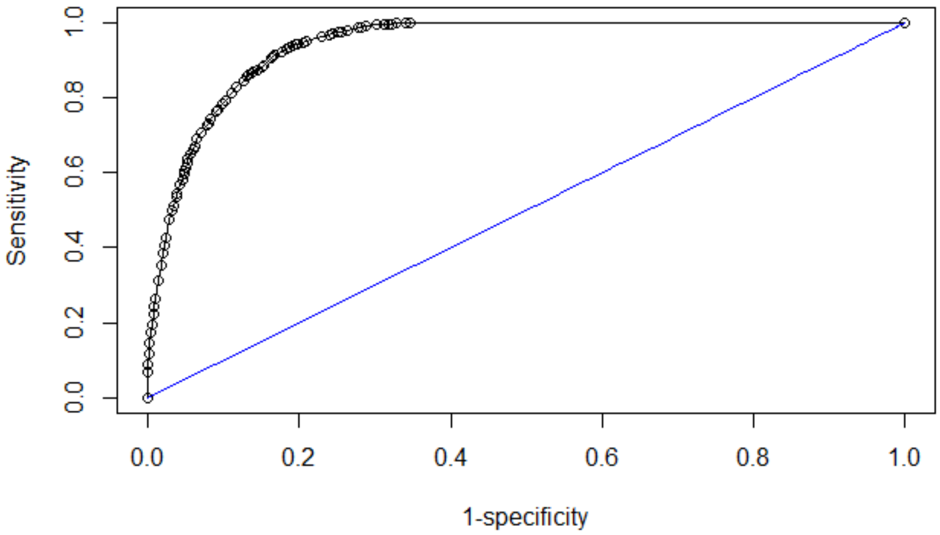


Figure 4.7: ROC curve for mixture cure model with Log-logistic hazard latency model; AUC=0.9406592

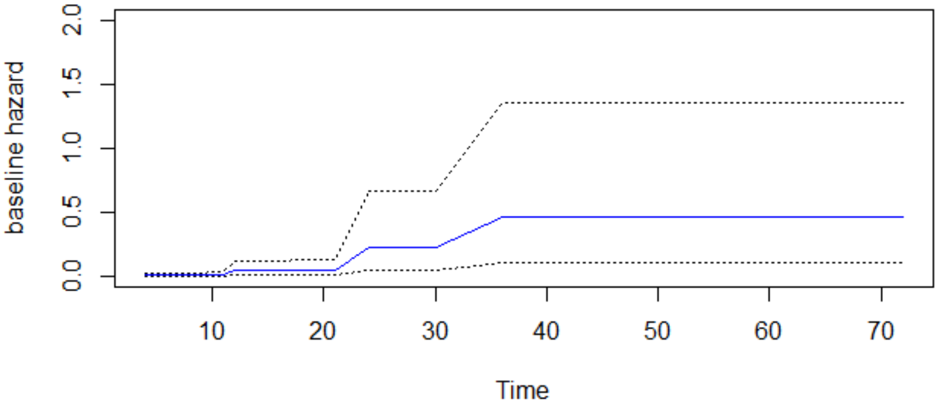


Figure 4.8: Plot of the baseline hazard of the Bayesian Mixture Cure model and its 95% credible interval.

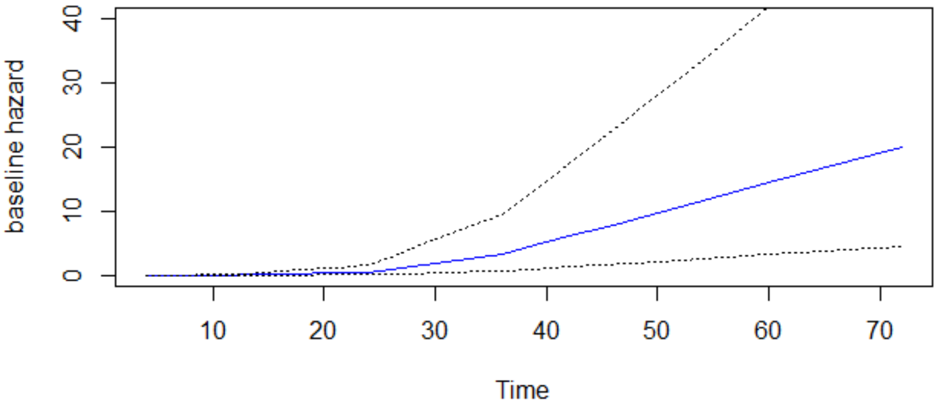


Figure 4.9: Plot of the cumulative baseline hazard of the Bayesian Mixture Cure model and its 95% credible interval.

Table 4.2: Estimated parameters for mixture cure model with Logistic incidence model and exponential latency model

Incidence Parameters	Estimated Value	Latency Parameters	Estimated Value
α_0	-17.74633088	β_0	4.03375525
α_1	0.27205641	β_1	-0.05285746
α_2	-0.03071058	β_2	0.06742929
α_3	1.29987700	β_3	0.64196569
α_4	-0.15250675	β_4	0.12878702
α_5	1.16889452	β_5	0.01894787
α_6	39.10259014	β_6	-0.51814832
α_7	-20.57241671	β_7	-0.58449089
α_8	-0.82047675	β_8	0.61987773

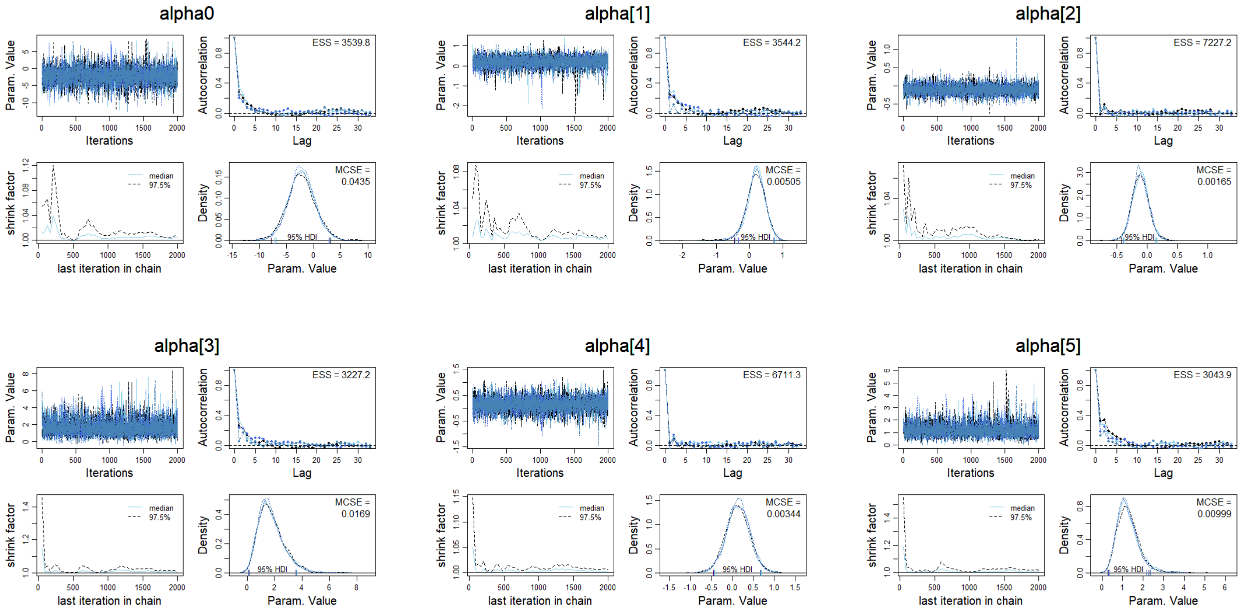
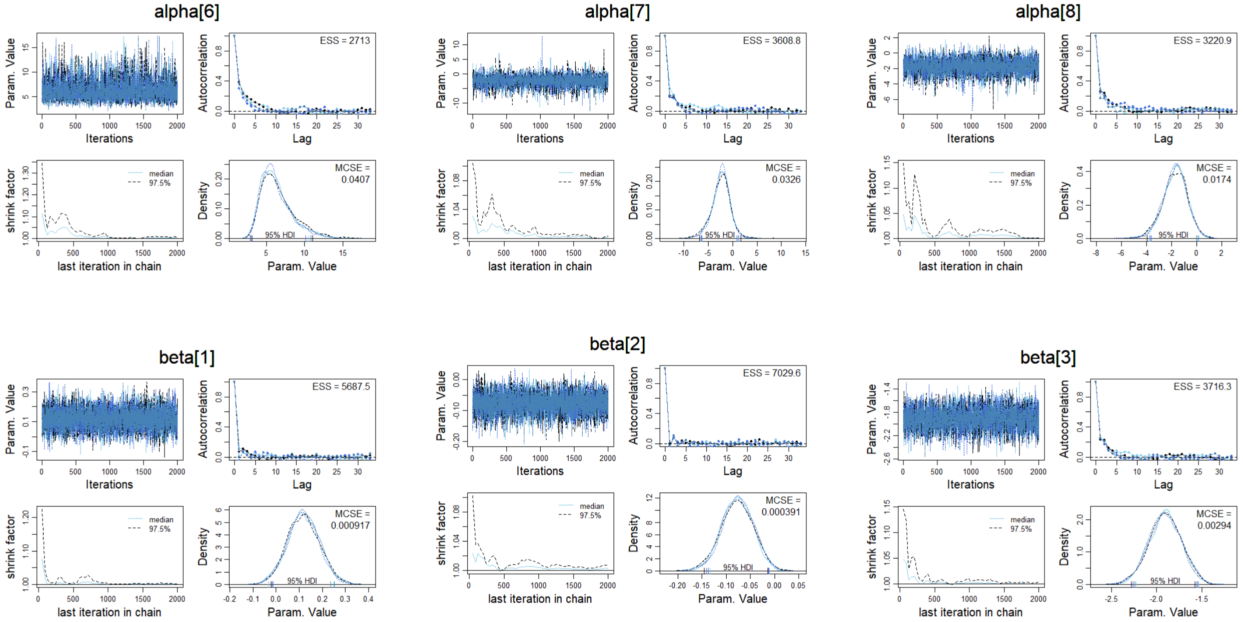


Figure 4.10: MCMC Diagnostics of the Estimated Parameters for the Bayesian Mixture Cure Model



(a) Caption for Figure 5

Figure 4.11: MCMC Diagnostics of the Estimated Parameters for the Bayesian Mixture Cure Model

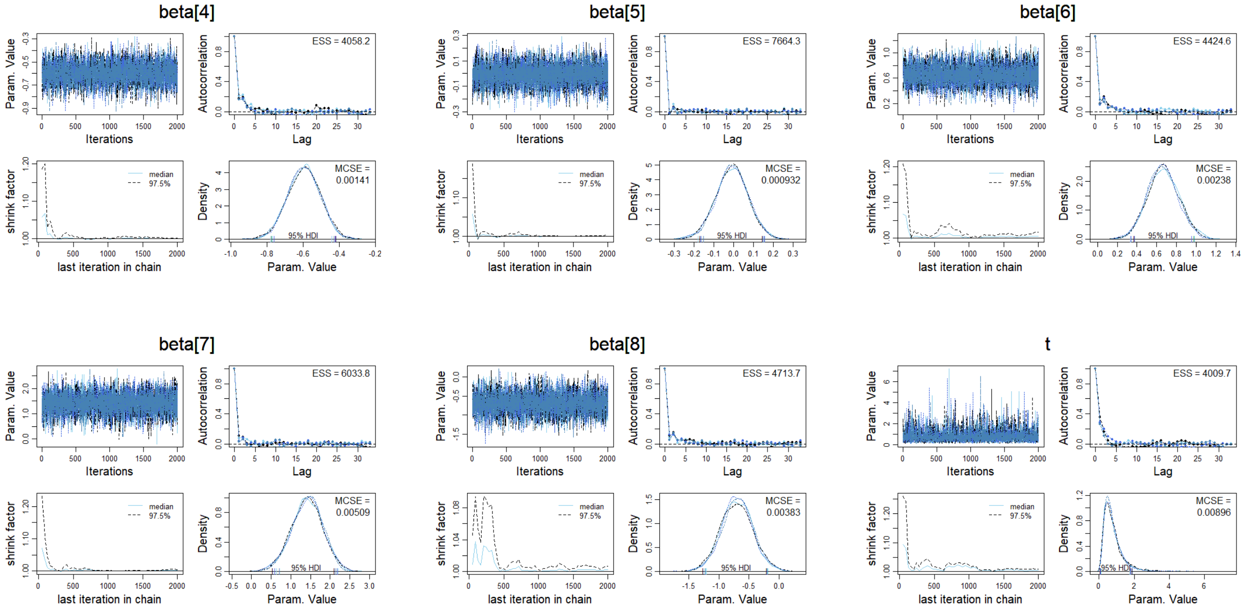


Figure 4.12: MCMC Diagnostics of the Estimated Parameters for the Bayesian Mixture Cure Model

Chapter 5

Conclusions and Future Directions

In Chapter 2, we use a Bayesian approach to estimate the Cox proportional hazards model, where the baseline hazard is approximated using various spline basis functions. We then compare the estimates obtained from this Bayesian approach with the partial likelihood approach using both simulated and real datasets. The analysis of the Bondora dataset reveals that the Bayesian approach not only produces the estimates of the regression coefficients close to those obtained using the partial likelihood approach but also accurately estimates the baseline hazard.

In Chapter 3, we introduce the mixture cure model, where the latency component is modelled using Cox's proportional hazards model. We employ the Bayesian approach to estimate both the regression coefficients and the baseline hazard. We then compare these regression coefficients with those obtained from the EM algorithm, for various sample sizes and levels of censoring using simulated data. The simulation study indicates that larger sample sizes and lower levels of censoring result in more precise estimates for both the regression parameters and the baseline hazard.

In Chapter 4, we apply the Bayesian mixture cure model to the German credit data and evaluate its predictive performance by comparing ROC curves with other parametric mixture cure models estimated using the EM algorithm. We discuss the challenges faced during this process and note that using standard parametric forms for the latency part performs better than using the Cox PH model. However, among mixture cure models with the Cox PH latency model, we find that the Bayesian approach, which utilizes indicator functions to approximate the baseline hazard, performs slightly better than the EM approach with baseline hazard estimated using Breslow's method.

Following are some of the details and future directions that can be looked into:

- For all the mixture cure models discussed, we only focused on the latency part by keeping the incidence part as the logistic function. However, by sacrificing some amount of interpretability, one can substitute the incidence part with more advanced classifiers like random forests and decision trees and improve the predictive performance.

- We can extend the CPH and mixture cure models to include time varying covariates, beacuse credit data is typically in a panel format, where new accounts enter, old accounts leave and each account is observed for a sequential period of time. Therefore, including time varying covariates might improve the default rate prediction.

- Apart from defaults, credit granting institutions also face the competing risk of early repayments. By using a similar Bayesian methodology, CPH model, as well as mixture cure models, can be extended to incorporate this competing risk factor. This results in a model known as multiple events mixture cure model, which assumes that a fraction of customers are susceptible of experiencing either credit default or early repayment (competing risk). Customers not observed to default or early repay fall into either the non-susceptible population (mature cases) or are right-censored. Such a model was introduced in the survival analysis context by Dirick (2015) (6) by extending the parametric competing risk model proposed by Watkins (2014)(34).

- As the mixture cure model involves more parameters, it becomes increasingly complex. It is crucial to address the issue of determining the optimal number of basis functions to prevent overfitting or underfitting. Additionally, identifying which covariates are best suited to explain the incidence part and which are better for explaining the latency part is also essential. Hence, variable selection and hypothesis testing procedures are necessary to assess the significance of the involved covariates.

- When approximating the baseline hazard with spline basis functions, adding a penalty on the spline coefficients can help to ensure smoother results.

- Lastly, one can expand the idea of mixture cure models to fully non-parametric approaches where both incidence and latency are estimated without assumptions about their functional forms. However, this approach may encounter issues related to the curse of dimensionality.

Chapter 6

Appendix

6.1 Simulating Mixture Cure Data

Let $h(t) = h_0(t)e^{\beta' x}$ denote the conventional Cox's Proportional Hazard model with fixed time covariates, where t denotes time, x is the vector of covariates, β is the vector of regression coefficients, and $h_0(t)$ is the baseline hazard function (the hazard function of the outcome occurring for those subjects with $x = 0$).

The survival function for this model is given by $S(t|x) = e^{-H_0(t).e^{\beta' x}}$, where $H_0(t)$ is the cumulative baseline hazard function, which is defined as $H_0(t) = \int_0^t h_0(s)ds$. The distribution function of the event times under the Cox proportional hazards model is $F(t|x) = 1 - e^{-H_0(t).e^{\beta' x}}$. Now using the Inverse CDF Transformation method, we generate the event times T from the Cox PH model as follows (37):

$$T = H_0^{-1}[-log(u)e^{\beta' x}]$$

, where $u \sim Unif(0, 1)$

In our simulation study, we use Weibull hazard function for the baseline hazard. Let $\lambda > 0$ and $\nu > 0$ be the shape and scale parameters of the Weibull distribution respectively. Then the event times T from the Cox PH model with Weibull baseline hazard can be generated as follows:

$$T = \left(-\frac{log(u)}{\lambda.e^{\beta' x}}\right)^{1/\nu}$$

, where $u \sim Unif(0,1)$.

We determine the true susceptibility status by using the logistic function to generate the actual probability of being susceptible. Then, we assign the susceptibility status based on the specified size of the susceptible subpopulation. Next, we independently generate uniform right censoring times for all observations from a distribution with parameters $Unif(0,k)$, where k is adjustable to achieve the desired censoring proportion in the simulated data.

6.2 R Codes

The following codes are implemented for the Bayesian Mixture Cure Model applied on the German Credit data.

```
 # Analysis of the german data
library(Matrix)
library(tidyverse)
library(survival)
library(smcure)
library(rstan)
library(splines2)
library(bridgesampling)
library(mixcure)
options(scipen=930)
library(survival)

gerdata_mod <- read.csv("german_clean.csv",header = TRUE)
gerdata_mod\$X<-NULL
```

```
Cox_model<-coxph(Surv(time,status)~V1+V4+V5+V6+V7+V8+V9+V10+
V11+V12+V13+V14+V15+V16+V17+V18+V19+V20+V21,data=gerdata_mod)
summary(Cox_model)

selected_gerdata <- gerdata_mod[, c(1,2,3,4,5,8,13,16,20,21)
, drop = FALSE]
selected_gerdata <- selected_gerdata[order(selected_gerdata\$time), ]

km_fit <- survfit(Surv(time,status) ~ 1, data = selected_gerdata)
plot(km_fit)

freedom<-6
mixbsmsplc <- bSpline(selected_gerdata\$time,degree = 0
,df=freedom,intercept = TRUE)
B1<-mixbsmsplc

plot(mixbsmsplc[,1]~selected_gerdata\$time,
ylim=c(0,max(mixbsmsplc)), type=’l’, lwd=2, col=1,
     xlab=" M-spline basis", ylab="")
for (j in 2:ncol(mixbsmsplc))
lines(mixbsmsplc[,j]~selected_gerdata\$time, lwd=2, col=j)

mixcbsmsplc <- bSpline(selected_gerdata\$time,degree=0,df=freedom
,integral = TRUE,intercept = TRUE)
B2<-mixcbsmsplc
plot(mixcbsmsplc[,1]~selected_gerdata\$time,
ylim=c(0,max(mixcbsmsplc)), type=’l’, lwd=2, col=1,
     xlab=" M-spline basis cumulative", ylab="")
for (j in 2:ncol(mixcbsmsplc))
lines(mixcbsmsplc[,j]~selected_gerdata\$time, lwd=2, col=j)

extenddf<-cbind(selected_gerdata,mixbsmsplc,mixcbsmsplc)

nrows<-nrow(extenddf)
```

```
random750<-sample(1:nrows,nrows*0.70)
train_data<-extenddf[random750,]
test_data<-extenddf[-random750,]

#Arranging according to time
train_data <- train_data[order(train_data\$time), ]
test_data <- test_data[order(test_data\$time), ]
km_fit2 <- survfit(Surv(time,status) ~ 1, data = train_data)
plot(km_fit2)

X_train <- as.matrix(train_data[, c(1,4,5,6,7,8,9,10)])
B1_train<-as.matrix(train_data[, 11:16])
B2_train<-as.matrix(train_data[, 17:22])
tau<-max(train_data\$time[train_data\$status == 1])

X_test <- as.matrix(test_data[, c(1,4,5,6,7,8,9,10)])
B1_test<-as.matrix(test_data[, 11:16])
B2_test<-as.matrix(test_data[, 17:22])

datalist3<-list(y=train_data\$time,c=train_data\$status
,N=length(train_data\$time),X=X_train, B1=B1_train,B2=B2_train,
num_basis=ncol(B1_train),num_covariates=ncol(X_train),
                y_test=test_data\$time,c_test=test_data\$status,
                N_test=length(test_data\$time)
                ,X_test=X_test, B1_test=B1_test,
                B2_test=B2_test,tau=tau)


#stanmix10ocubic4<-stan(file=’bondora_stan.stan’,
data=datalist3,chains = 1)
stanmix10ocubic4<-stan(file=’genmixcure_new.stan’,data=datalist3,chains = 4,
control=list(max_treedepth=11),iter=5000,warmup =3000)
  print(stanmix10ocubic4)
```

```
source("DBDA2E-utilities.R")
stanmix10o_coda<- mcmc.list(lapply(1:ncol(stanmix10ocubic4), function(x)
{mcmc(as.array(stanmix10ocubic4)[,x,])}))
summary(stanmix10o_coda)
diagMCMC(stanmix10o_coda,parName = c(’t’))
summary(simmix7)


generated_quantities <- as.array(stanmix10ocubic4, "bh0")
dim(generated_quantities)
#view(generated_quantities)
expected_values <- apply(generated_quantities, c(3), mean)
expected_values<-data.frame(expected_values)
expected_values97.5<-apply(generated_quantities, c(3), quantile,probs=0.975)
expected_values2.5<-apply(generated_quantities, c(3), quantile,probs=0.025)
#view(expected_values97.5)
plot(X,Y)
plot(train_data\$time,expected_values\$expected_values,type = "l", col = "blue"
,ylim = c(0,2 ),lty=1,xlab="Time",ylab="baseline hazard")
lines(extenddf\$time,h0_original,type = "l", col = "red",lty=2)
lines(train_data\$time,expected_values97.5,type = "l", col ="black",lty=3)
lines(train_data\$time,expected_values2.5,type = "l", col = "black",lty=3)


generated_quantities2 <- as.array(stanmix10ocubic4, "ch0")
dim(generated_quantities2)
#view(generated_quantities2)
expected_values2 <- apply(generated_quantities2, c(3), mean)
expected_values2<-data.frame(expected_values2)
expected_values97.52<-apply(generated_quantities2, c(3), quantile,probs=0.975)
expected_values2.52<-apply(generated_quantities2, c(3), quantile,probs=0.025)
#view(expected_values97.52)
plot(X,Y)
plot(train_data\$time,expected_values2\$expected_values2,type = "l", col = "blue",
ylim = c(0,40 ),lty=1,xlab="Time",ylab="baseline hazard")
lines(extenddf\$time,h0_original,type = "l", col = "red",lty=2)
lines(train_data\$time,expected_values97.52,type = "l", col ="black",lty=3)
lines(train_data\$time,expected_values2.52,type = "l", col = "black",lty=3)

#test data analysis for bayesian way

wi1_matrix <- as.array(stanmix10ocubic4, "wi1")
dim(wi1_matrix)
view(wi1_matrix)
wi1 <- apply(wi1_matrix, c(3), mean)
wi1<-data.frame(wi1)
wi0<-1-wi1

M_matrix <- as.array(stanmix10ocubic4, "M")
dim(M_matrix)
#view(M_matrix)
M <- apply(M_matrix, c(3), mean)
M<-data.frame(M)

M2_matrix <- as.array(stanmix10ocubic4, "walpha_test")
dim(M2_matrix)
#view(M2_matrix)
M2 <- apply(M2_matrix, c(3), mean)
M2<-data.frame(M2)

test_df<-cbind(test_data\$time,test_data\$status,wi1,wi0,M,M2)
#test_df\$
  #Sensitivity
N1<-sum(test_df\$wi1)
N0<- nrow(test_df)-N1
sensitivity<-function(M,k){
  return(1-(sum(wi1*ifelse(M<=k,1,0))/N1))
}
```

```
sensitivity(-test_df\$M,-1)
#Specificity
Specificity<-function(M,k){
  return((sum(wi0*ifelse(M<=k,1,0)))/N0)
}
Specificity(-test_df\$M,-0.5)

k <- seq(0, 1, length.out = 101)
#view(k)

Roc_Sensitivity<-rep(0, length(k))
Roc_specificity<-rep(0, length(k))
sensitivity(-test_df\$M,-1)
-k[101]
for (i in 1:length(k)){
  Roc_Sensitivity[i]<-sensitivity(-test_df\$M,-k[i])
  Roc_specificity[i]<-Specificity(-test_df\$M[i])

}
Roc<-cbind.data.frame(Roc_Sensitivity,Roc_specificity)

plot(1-Roc\$Roc_specificity,Roc\$Roc_Sensitivity,xlim=0:1,ylim=0:1,
xlab="1-specificity",ylab="Sensitivity")
points(1-Roc\$Roc_specificity,Roc\$Roc_Sensitivity)
lines(1-Roc\$Roc_specificity,Roc\$Roc_Sensitivity)
lines(Roc\$Roc_Sensitivity,Roc\$Roc_Sensitivity,type="l",col="blue")

sum(wi1*wi0\$wi1[10]*(1-ifelse(M<=M\$M[10],1,0)))
wi0\$wi1[1]

#AUC<-function(M){
 # dsum=0
  #for (i in length(M)){
    #  dsum=dsum+sum(wi1*wi0\$wi1[i]*(1-ifelse(M<=M[i],1,0)))
     # print(dsum)
    #}
    #return (length(M)*dsum/((length(M)-1)*(N1*N0)))
  #}

  AUC_matrix <- matrix(0, nrow = length(test_df\$M), ncol = length(test_df\$M))
  for (i in 1:length(test_df\$M)) {
    for (j in 1:length(test_df\$M)) {
      # Apply condition here, for example, product of row and column index
      if (-M\$M[j] > -M\$M[i]){
        AUC_matrix[i, j] <- wi1\$wi1[j]*wi0\$wi1[i]
      }
      else{
        AUC_matrix[i, j] <-0
      }
    }
  }
  AUC<-sum(AUC_matrix)/(N0*N1)
  print(AUC)

distances1 <- sqrt((1-Roc\$Roc_specificity - 0)^2 + (Roc\$Roc_Sensitivity - 1)^2)
distances1
 K<-(which.min(distances1)-1)/100
 closest_point <- Roc[which.min(distances1), ]
 point<-c(1-closest_point\$Roc_specificity,closest_point\$Roc_Sensitivity)


#Fitting mixture cure model using em algorithm

german_em <- smcure(Surv(time,status)˜1
             ,cureform=˜1,
             data=train_data)
```

```
german_em <- smcure(Surv(time,status)~V1+V4+V5
                    ,cureform=~V1+V4+V5,
                    data=train_data,model='ph')

# selected variables : 1,2,3,4,5,8,13,16,20,21 for selected gerdata

german_em_expo <- mixcure(lformula =  Surv(time,status)~V1+V4+V5+V8+V13+V16+V20+V21
                     ,iformula = ~1+V1+V4+V5+V8+V13+V16+V20+V21,
                     lmodel = list(fun = "survreg",dist="exponential"),
                     data=train_data,savedata = TRUE)

german_em_weibull <- mixcure(lformula =  Surv(time,status)~V1+V4+V5+V8+V13+V16+V20+V21
                          ,iformula = ~1+V1+V4+V5+V8+V13+V16+V20+V21,
                          lmodel = list(fun = "survreg",dist="weibull"),
                          data=train_data,savedata = TRUE)

german_em_loglogisticl <- mixcure(lformula =  Surv(time,status)~V1+V4+V5+V8+V13+V16+
V20+V21,iformula = ~1+V1+V4+V5+V8+V13+V16+V20+V21,
                             lmodel = list(fun = "survreg",dist="loglogistic"),
                             data=train_data,savedata = TRUE)



german_em_lognormal <- mixcure(lformula =  Surv(time,status)~V1+V4+V5+V8+V13+V16+
V20+V21,iformula = ~1+V1+V4+V5+V8+V13+V16+V20+V21,
lmodel = list(fun = "survreg",dist="lognormal"),
                          data=train_data,savedata = TRUE)



german_em_gompertz <- mixcure(lformula =  Surv(time,status)~V1+V4+V5+V8+V13+V16+V20+V21
                          ,iformula = ~1+V1+V4+V5+V8+V13+V16+V20+V21,
                          lmodel = list(fun = "survreg",dist="gompertz"),
                          data=train_data,savedata = TRUE)

german_em_ph <- mixcure(lformula =  Surv(time,status)~V1+V4+V5+V8+V13+V16+V20+V21
                          ,iformula = ~1+V1+V4+V5+V8+V13+V16+V20+V21,
                          lmodel = list(fun = "coxph"),
                          data=train_data,savedata = TRUE)




#The test procedure
############################

pd10<-german_em_ph
#summary(pd10)
coef(pd10)
X_test2<-data.frame(X_test)
pred_2<-predict(pd10,
                newdata = X_test2, times = test_data\$time)
#summary(pred_2)
uncure_prob <- pred_2\$cure[, 1]
#pred_2\$times
#mix_surv<-pred_2\$surv
mic_uncure_surv<-pred_2\$uncuresurv
uncure_surv <- sapply(seq_along(mic_uncure_surv), function(i)
mic_uncure_surv[[i]][[i]])

wwi1<-(1-uncure_prob)/((1-uncure_prob)+uncure_surv*uncure_prob)
#view(wwi1)
wwi0<-1-wwi1
MM<-uncure_prob
test_df_em<-cbind(test_data\$time,test_data\$status,wwi1,wwi0,MM)
test_df_em<-data.frame(test_df_em)

NN1<-sum(test_df_em\$wwi1)
NN0<- nrow(test_df_em)-NN1
sensitivityem<-function(MM,k){
  return(1-(sum(wwi1*ifelse(MM<=k,1,0))/NN1))
}

Specificityem<-function(MM,k){
  return((sum(wwi0*ifelse(MM<=k,1,0)))/NN0)
}

k2 <- seq(0, 1, length.out = 101)

Roc_Sensitivityem<-rep(0, length(k2))
Roc_specificityem<-rep(0, length(k2))

for (i in 1:length(k2)){
  Roc_Sensitivityem[i]<-sensitivityem(-test_df_em\$MM,-k2[i])
  Roc_specificityem[i]<-Specificityem(-test_df_em\$MM,-k2[i])

}
Roc_em<-cbind.data.frame(Roc_Sensitivityem,Roc_specificityem)

plot(1-Roc_em\$Roc_specificityem,
Roc_em\$Roc_Sensitivityem,xlim=0:1,ylim=0:1,xlab="1-specificity",ylab="Sensitivity")
points(1-Roc_em\$Roc_specificityem,Roc_em\$Roc_Sensitivityem)
lines(1-Roc_em\$Roc_specificityem,Roc_em\$Roc_Sensitivityem)
lines(Roc_em\$Roc_Sensitivityem,Roc_em\$Roc_Sensitivityem,type="l",col="blue")

wwi1<-data.frame(wwi1)
wwi0<-data.frame(wwi0)
AUC_matrixem <- matrix(0, nrow = length(test_df_em\$MM), ncol = length(test_df_em\$MM))
for (i in 1:length(test_df_em\$MM)) {
  for (j in 1:length(test_df_em\$MM)) {
    # Apply condition here, for example, product of row and column index
    if (-test_df_em\$MM[j] > -test_df_em\$MM[i]){
      AUC_matrixem[i, j] <- wwi1\$wwi1[j]*wwi0\$wwi0[i]
    }
    else{
      AUC_matrixem[i, j] <-0
    }
  }
}
AUCem<-sum(AUC_matrixem)/(NN0*NN1)
print(AUCem)

distances2 <- sqrt((1-Roc_em\$Roc_specificityem - 0)^2 +
(Roc_em\$Roc_Sensitivityem - 1)^2)
#distances2
K2<-(which.min(distances2)-1)/100
closest_point2 <- Roc[which.min(distances2), ]
point<-c(1-closest_point2\$Roc_specificity,closest_point2\$Roc_Sensitivity)
point
K2
#Save all the AUC's here
AUC_expo<-AUCem

AUC_weibull<-AUCem

AUC_cox<-AUCem

AUC_loglogisticl<-AUCem

AUC_lognormal<-AUCe
```

6.3 Stan Codes

Following is the Stan code for the Hamiltonian Monte Carlo Sampling of the posterior obtained from the Mixture Cure Model (Chapter 3):

```
data {
  //Train Data
  int N;
  int num_basis;
  int num_covariates;
  real y[N];
  real c[N];
  matrix[N,num_covariates] X;
  matrix[N,num_basis] B1;
  matrix[N,num_basis] B2;
  real tau;

  // Test Data
  int N_test;
  real y_test[N_test];
  real c_test[N_test];
  matrix[N_test,num_covariates] X_test;
  matrix[N_test,num_basis] B1_test;
  matrix[N_test,num_basis] B2_test;

}

  simplex[num_basis] theta;
  real alpha0;
  row_vector[num_covariates] alpha;
  row_vector[num_covariates] beta;
  //real<lower=0> theta0;
  real<lower=0> t;
```

```
}
transformed parameters{
   vector[N] h0;
   vector[N] H0;
   vector[N] walpha;
   vector[N] xbeta;
   vector[N_test] h0_test;
   vector[N_test] H0_test;
   vector[N_test] walpha_test;
   vector[N_test] xbeta_test;
   walpha=alpha0+to_vector(alpha*X’);
   xbeta=to_vector(beta*X’);
   h0= t*to_vector(theta’*B1’);
   H0= t*to_vector(theta’*B2’);
   walpha_test=alpha0+to_vector(alpha*X_test’);
   xbeta_test=to_vector(beta*X_test’);
   h0_test= t*to_vector(theta’*B1_test’);
   H0_test= t*to_vector(theta’*B2_test’);
}

model {
  //target+= normal_lpdf(beta1 |  1,6);
  target+= normal_lpdf(alpha0 | 0,5);
  target+= normal_lpdf(beta | 0,5);
  target+= normal_lpdf(alpha | 0,5);
  //target+= chi_square_lpdf(theta | 6);
  target+=dirichlet_lpdf( theta | rep_vector(1, num_basis));
  target+= lognormal_lpdf( t | 0, 1);
  //target+= chi_square_lpdf(theta0 | 6);
 //  target+= chi_square_lpdf(theta1 | 6);

  for (i in 1:N){
    if (c[i]==1){
       //Observed data i.e non censored and susceptible
```

```
      target+= -log(1+exp(-walpha[i]))
      -(H0[i])*exp(xbeta[i])
      + log(h0[i])
      + (xbeta[i]);
   } else{
      // unobserved
      target+= -log(1+exp(-walpha[i]))
      + (-walpha[i])
      + log(1+exp(-(-walpha[i]+(H0[i])*exp(xbeta[i]))));

    }
  }


}
generated quantities{
// Baseline Hazard
  vector[N] bh0;
    bh0 = h0;
//Cumulative Baseline Hazard
  vector[N] ch0;
    ch0 = H0;

// Survival Probability and Probability of Default prediction on the test data
  vector[N_test] M;
  vector[N_test] wi1;
  for (j in 1:N_test){
    wi1[j]= if_else(y_test[j]>tau, 1, 0)
    + if_else(y_test[j]<=tau, 1, 0)*if_else(c_test[j]==0, 1, 0)*
         ((exp(-walpha_test[j]))/(exp(-H0_test[j]*exp(xbeta_test[j]))
         + exp(-walpha_test[j])) );
    M[j]=(1/(1+exp(-walpha_test[j]))) ;
  }
}
```

Bibliography

[1] Ma, Jun, Stephane Heritier, and Serigne N. Lô. "On the maximum penalized likelihood approach for proportional hazard models with right censored survival data." Computational Statistics & Data Analysis 74 (2014): 142-156.

[2] Royston, Patrick. "Estimating a smooth baseline hazard function for the Cox model." London: Department of Statistical Science, University College London (2011).

[3] Hosmer, D., Lemeshow, S., and May, S. (2008). Applied Survival Analysis: Regression Modeling of Time to Event Data, (2nd ed.). Wiley-Interscience, Hoboken, New Jersey.

[4] Ramsay, James O. "Monotone regression splines in action." Statistical science (1988): 425-441.

[5] Wang W, Yan J (2018). splines2: Regression Spline Functions and Classes. R package version 0.2.8, URL https://CRAN.R-project.org/package=splines2.

[6] Dirick, Lore, Gerda Claeskens, and Bart Baesens. "Time to default in credit scoring using survival analysis: a benchmark study." Journal of the Operational Research Society 68 (2017): 652-665.

[7] Thackham, Mark. "Survival analysis: applications to credit risk default modelling." PhD diss., Macquarie University, 2022.

[8] Tong, Edward NC, Christophe Mues, and Lyn C. Thomas. "Mixture cure models in credit scoring: If and when borrowers default." European Journal of Operational Research 218, no. 1 (2012): 132-139.

[9] Amico, M., I. Van Keilegom, and B. Han. "Assessing cure status prediction from survival data using receiver operating characteristic curves." Biometrika 108, no. 3 (2021): 727-740.

[10] Heagerty, Patrick J., Thomas Lumley, and Margaret S. Pepe. ”Time-dependent ROC curves for censored survival data and a diagnostic marker.” Biometrics 56, no. 2 (2000): 337-344.

[11] Pepe, Margaret Sullivan. The statistical evaluation of medical tests for classification and prediction. Oxford university press, 2003.

[12] Cai, Chao, Yubo Zou, Yingwei Peng, and Jiajia Zhang. ”smcure: An R-Package for estimating semiparametric mixture cure models.” Computer methods and programs in biomedicine 108, no. 3 (2012): 1255-1260.

[13] Taylor, Jeremy MG. ”Semi-parametric estimation in failure time mixture models.” Biometrics (1995): 899-907.

[14] Cox, David R. ”Regression models and life-tables.” Journal of the Royal Statistical Society: Series B (Methodological) 34, no. 2 (1972): 187-202.

[15] Brilleman, Samuel L., Eren M. Elci, Jacqueline Buros Novik, and Rory Wolfe. ”Bayesian survival analysis using the rstanarm R package.” arXiv preprint arXiv:2002.09633 (2020).

[16] Klein, John P., and P. John. ”Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data.” (1997).

[17] Kaplan, Edward L., and Paul Meier. ”Nonparametric estimation from incomplete observations.” Journal of the American statistical association 53, no. 282 (1958): 457-481.

[18] Bartoš, František, Frederik Aust, and Julia M. Haaf. ”Informed Bayesian survival analysis.” BMC Medical Research Methodology 22, no. 1 (2022): 238.

[19] Peng, Roger D. ”Advanced statistical computing.” Work in progress (2018): 121.

[20] Vishnoi, Nisheeth K. ”An introduction to Hamiltonian Monte Carlo method for sampling.” arXiv preprint arXiv:2108.12107 (2021).

[21] https://www.batisengul.co.uk/post/2021-07-02-intro-to-hmc/

[22] Breslow, Norman. ”Covariance analysis of censored survival data.” Biometrics (1974): 89-99.

[23] Li, Peizhi, Yingwei Peng, Ping Jiang, and Qingli Dong. "A support vector machine based semiparametric mixture cure model." Computational Statistics 35 (2020): 931-945.

[24] Hofmann,Hans. (1994). Statlog (German Credit Data). UCI Machine Learning Repository. https://doi.org/10.24432/C5NC77.

[25] Narain, B. "16. Survival analysis and the credit-granting decision." Readings in Credit Scoring: Foundations, Developments, and Aims (2004): 235.

[26] Stepanova, Maria, and Lyn Thomas. "Survival analysis methods for personal loan data." Operations Research 50, no. 2 (2002): 277-289.

[27] Berkson, Joseph, and Robert P. Gage. "Survival curve for cancer patients following treatment." Journal of the American Statistical Association 47, no. 259 (1952): 501-515.

[28] Boag, John W. "Maximum likelihood estimates of the proportion of patients cured by cancer therapy." Journal of the Royal Statistical Society. Series B (Methodological) 11, no. 1 (1949): 15-53.

[29] BHARADWAJ, SHRUTHI RAVINDRA. "Default Status Prediction in Credit Data using Mixture Cure Models." PhD diss., 2022.

[30] Basu, Sanjib, and Ram C. Tiwari. "Breast cancer survival, competing risks and mixture cure model: a Bayesian analysis." Journal of the Royal Statistical Society Series A: Statistics in Society 173, no. 2 (2010): 307-329.

[31] https://www.bondora.com/en/public-reports

[32] M. Siddhartha, Bondora peer-to-peer lending data, 2020.

[33] De Boor, Carl, and Carl De Boor. A practical guide to splines. Vol. 27. New York: springer-verlag, 1978.

[34] Watkins, John GT, Andrey L. Vasnev, and Richard Gerlach. "Multiple Event Incidence And Duration Analysis For Credit Data Incorporating Non-Stochastic Loan Maturity." Journal of Applied Econometrics 29, no. 4 (2014): 627-648.

[35] Betancourt, Michael. "A conceptual introduction to Hamiltonian Monte Carlo." arXiv preprint arXiv:1701.02434 (2017).

[36] Heinze, Georg, and Daniela Dunkler. ”Avoiding infinite estimates of time-dependent effects in small-sample survival studies.” Statistics in medicine 27, no. 30 (2008): 6455-6469.

[37] Austin, Peter C. ”Generating survival times to simulate Cox proportional hazards models with time-varying covariates.” Statistics in medicine 31, no. 29 (2012): 3946-3958.