UNIVERSITÀ DEGLI STUDI DI CAMERINO

Internation School of Advanced Studies

Computer Sciences and Mathematics

Survival Tree Models and Survival Ensemble Methods in Machine Learning

Doctor of Philosophy in
Mathematics
XXXV cycle

Advisor
Prof. Renato De Leone

Co-Advisor:
Prof. Chunjie Wang

Ph.D. Candidate:
Jia Chen

Academic Year 2020-2023

Acknowledgements

I'd like to extend my deep gratitude to all those who have offered cordial and selfless support in writing this thesis.

Firstly, I am extremely grateful to my supervisor, Prof. Renato De Leone. He guided me, influenced me and helpd me in the process of writing this thesis. It is with his patience, generosity and encouragement that I finally write, revise and perfect my thesis. An appreciation also goes to my co-supervisor, Prof. Chunjie Wang, her significant support in my research and great assistance in my life and career.

Secondly, I am much obliged to all professors who taught or helped me in Camerino, especially to Dr. Cristina Soave and Dr. Zuo, their patience and kind impressed me. I am also very grateful to Dr. Jianjun Wang, Josefi and Elisa Marcelli who assisted me in lots of aspects.

In addition, I would like to thank the members from survival analysis research team in school of mathematics and statistics, CCUT.

Finally, I am also grateful to my family who support me and share with me my worries and frustration and happiness.

Abstract

The aim of this thesis is to construct novel Machine Learning methods to deal with survival data in Survival Analysis. Survival Analysis is considered as a general approach to analyze survival data, and survival data often arises in a variety of applied fields, such as medicine, engineering, economics and so on. Censoring is the main feature of survival data and it implies an information loss. Thus, it is difficult to analyze and model survival data. With the rapid development of Machine Learning, new techniques in Machine Learning have been proposed to tackle survival data that often show better performance compared to traditional statistical methods and tools.

In this thesis, we focus on tree and ensemble methods in Machine Learning for Interval-censored data. Interval-censored data is a general type of survival data. Survival tree is a flexible predictive method for survival data because no specific assumptions are required.

For interval-censored data, the Generalized Log-Rank Test has good discriminative power when appropriate parameters are chosen. We construct a specialized test statistic for Generalized Log-Rank Tests, and propose a new survival tree with hyper-parameters by combining the test statistic with the Conditional Inference Framework for interval-censored data. The effects of tuning hyper-parameters are discussed. Tuning hyper-parameters allows the tree method to become more general and flexible. The new tree method either demonstrates superior performance or remains competitive with the existing tree method for interval-censored data, referred to as ICtree, which is a special case of it. Then, we construct a novel survival ensemble method where the new trees are utilized as base learners, and average prediction weights are applied to estimate the survival function in this ensemble algorithm.

An extensive simulation is carried out to assess the predictive performance of newly proposed tree methods and the new ensemble method. Applications of those novel survival trees in the fields of engineering and medicine are also discussed. The tree methods are applied to a heat exchanger life data and a tooth emergence data. Furthermore, Machine Learning techniques and statistical methods have been applied to establish a credit risk model for evaluating the credit risk of small and medium-sized enterprises.

Contents

List of Figures

List of Tables

List of Symbols

The next list describes the symbols that will be used within the thesis.

$\mathbb{R}$	real numbers
$\mathbb{N}$	natural numbers
Y	response variable
$\mathcal{Y}$	sample space of response variable
$\mathcal{C}$	set of class label for the response variable
K	size of $\mathcal{C}$, i.e., the number of class labels
$\mathbf{X}$	an m-dimensional covariate or m predictor variables
$\mathbf{x}$	an observed value of the covariate $\mathbf{X}$
$\mathcal{X}$	the sample space of $\mathbf{X}$
$\mathcal{L}_n$	sample space of n iid observations
A	Training set
N	The size of training set
$\arg\max_{c \in \mathcal{C}} f$	maximizer of f on the set C
T	a non-negative random variable for the survival time which represents the time of occurrence of the event of interest
C	the censoring time which means the occurrence time of censoring in right censored data
O	the observation time

δ an indicator variable which represents whether the event of interest occurs or not, $\delta \in \{0,\ 1\}$

L left endpoint of an observation interval

R right endpoint of an observation interval

$S(t)$ survival function

$F(t)$ cumulative distribution function

$f(t)$ probability density function

$h(t)$ hazard rate function

$H(t)$ cumulative hazard function

$\{s_j\}_{j=1}^{l}$ l distinct ordered elements of the values of survival time $\{T_i;\ i \in \{1, \ldots, n\}\}$ for n subjects for right censored data; the distinct ordered elements of the values of observed interval endpoints $\{L_i, R_i\}_{i=1}^{n}$

$I(A)$ indicator function. i.e., $I(A)$ is 1 if A is true, otherwise $I(A)$ is 0

$\boldsymbol{\beta}$ regression coefficients

$exp(x)$ exponential function

$Q_\tau(\mathbf{x})$ the τth conditional quantile function

$\sharp\{A\}$ the number of the elements in the set A

Introduction

Survival Analysis

Survival Analysis is a general technique to analyze and model the time to event data. In this case usually the response variable is the time until an event of interest occurs.

Time to event data is common in real world. The event of interest may be a negative individual experience such as disease incidence, tumor recurrence, virus infection, equipment failure; or a positive one, for instance, symptomatic remissions, return to work after treatment, loan repayment, emergence of tooth.

The time of the occurrence of a specific event in medical applications is usually the time until death, and the term "survival time" is utilized to describe the variable which shows that an individual has survived over the period. Therefore, the use of the term of Survival Analysis is natural [50, 83, 233, 236].

Time to event data, also known as survival data, typically consists of two key components:

1. survival time as response variable;
2. other information as covariates, which are also referred to as predictor variables or risk factors.

As an example consider the response of the leukemia patients who are given different treatments. The survival data contain

- the remission time (which is the survival time);
- characteristic information of patient such as age, weight, even the expression characteristics of miRNAs, which are the covariates.

Survival data arise in various disciplines such as medicine, epidemiology, engineering, economics, actuarial science, criminology, and the researches in survival data are meaningful and important. Clearly, the meaning of survival time is different for different fields of

application and events of interest vary. They can be time in weeks until a person goes out of remission, time in months until an employee changes his job, time to death or time to default, or even time from an individual's release until his or her recidivism [50, 211].

Accordingly, the terminology may be different for the same concepts in different fields. For example, in some articles survival time is replaced by the term "failure time" and "failure time data" is used instead of survival data because the events of interest are represented by failure borrowing this concept from industrial field. However, we use "survival time" uniformly throughout this thesis.

It's worth noting that the survival time is not always observed exactly due to cost limitations or losing contact during the observation period. Specifically, the event of interest does not necessarily occur at the end of the study, since the object of observation may withdraw or move away from the study area, or the observation is not continuous due to the cost of the trial. The loss of information in survival data is known as "censoring". Several types of survival data are determined by the way and quantity of information loss. In general, basic types include right censoring, left censoring and interval censoring. In addition, there are other more complex censored data, such as double censored survival data, and Case K interval-censored data. It is a great challenge to handle censored data in Survival Analysis. Well known models and approaches that are available for complete data can not be applied to censored data directly. Consequently, a variety of statistical approaches have been widely developed to deal with this additional difficulty [129].

With the rapid development of Machine Learning techniques, the advantages of this novel technology are emerging in this context, such as non-limitation in assumptions condition, faster computing and good adaptive ability to high dimensional data. Therefore, Machine Learning methods gradually have been applied to censored data to enrich the available techniques for Survival Analysis.

Machine Learning techniques in Survival Analysis

The essence of human intelligence is to learn through practical and theoretical experience. Machine Learning is part of Artificial Intelligence and its core is an intelligent mechanism that learns from data to improve the performance in a remarkable short period. The key to research in Machine Learning is to explore models with underlying structure to construct learning functions based on available data.

The surge in data sizes has had a strong impact on the underlying generation mechanism of models in Machine Learning and greatly improved the learning power. In turn, an intelligent mechanism in Machine Learning is helpful for analyzing the hidden structure of data and understanding the nature of data. Hence, one of development direction of

Machine Learning is to design models with efficient computer-implemented algorithms for analyzing and predicting data. In researches on Machine Learning, computational efficiency and predictive accuracy are the common concerns.

Models in Machine Learning employ a variety of methods involved in various domains and "Machine Learning draws on results from probability and statistic, computational complexity theory, control theory, information theory, philosophy, psychology, neurobiology, and other fields" [168]. Meanwhile, Machine Learning has achieved great success in specific fields, such as medical prediction and diagnosis [58, 140, 207], credit assessment system predicting applicant reliability [34], recognition and analysis of ecological environment [92, 222], speech recognition system [61].

Moreover, Machine Learning algorithms can be successfully utilized to tackle censored data for predicting survival time, estimating the survival probability, and analyzing the relationship between the survival time and covariates based on the practical and theoretical experience. Therefore, a class of novel methods has been proposed in Survival Analysis that combine traditional Survival Analysis methods with advanced machine learning techniques.

Various Machine Learning methods such as Support Vector Machine, Artificial Neural Network and others have been applied to obtain new algorithms for survival data. Among them, tree-based methods have been proved to be extremely effective.

Tree-based methods are very popular Machine Learning methods. Trees are usually constructed using a top-down induction approach. The feature space is partitioned by splitting criteria, forming an upside-down tree structure, the size of which is determined by the stopping and pruning rule.

The splitting rule is crucial for the performance of a tree. The core idea of the splitting rule is to recursively partition the attribute space based on specific criteria. Stopping criteria are also used to determine the size of a tree. Additionally, some tree methods include a pruning procedure to avoid overfitting [42]. A wide variety of tree-based models have been proposed for different splitting rules and prediction tasks [26, 36, 37, 38, 103, 151, 172, 183, 189, 190, 196, 199]. Among them, Classification and Regression Tree (CART) is one f the most classical and commonly used tree methods, and its attractive features are simplicity and interpretability. In addition, it can flexibly and effectively handle high-dimensional problems. Variations of CART for censoring data are useful tools in Survival Analysis. A tree-based method built for survival data is known as survival tree. Most survival trees are variations of tree methods for full data and have good prediction performance in Survival Analysis.

Although tree-based methods have considerable advantages over other methods according to the literature (interpretability, fast calculation) [82, 205], a single tree is unstable due to the high variance of the cut points. Great fluctuations in prediction are induced by

small perturbations in the training data. Since the optimal cut-point heavily depends on training set, this feature badly influences the prediction accuracy of tree-based methods and often leads to overfitting [84, 91].

Therefore, tree-based ensemble methods have been proposed to reduce the variance of single tree while preventing bias from rising sharply [23, 25, 86, 90, 101, 166]. Breiman proposed the concept of "Bagging" in 1994, which is considered as a special case of a wide range of ensemble methods [23]. Random Forests (RF) is a classic and well-known ensemble method, which extend the idea of "Bagging" by growing each tree based on bootstraped data and bootstraped attribute variables [25].

The optimal results obtained by ensemble methods have boosted the development of survival ensemble methods as a new branch of research in Survival Analysis [108, 114, 115, 134, 209]. Most of the methods are constructed by combining one or some statistical tools in Survival Analysis with the core procedure of Random Forests (i.e., bootstrap) and demonstrate good prediction performance.

However, the aforementioned survival tree methods and survival ensembles are just capable to handle right censored data rather than interval-censored data. Interval-censored data is a general data type that can be considered as a natural generalization of right and left censored data. Since the amount of information is more limited, it is hard to deal with the issue comprising this kind of data. However, research on techniques for interval-censored data is particularly important in Survival Analysis. In this thesis, novel survival trees and survival ensembles for interval-censored data are proposed, the performance of those approaches is explored and compared to other existing interval censored survival trees and ensembles methods. Furthermore, the applications of the new approaches to real data in different contexts are discussed.

The thesis is organized as follows.

In Chapter 1, we first introduce the state of the art in Survival Analysis, mainly showing the technical developments and applications in this area based on statistical ideas and principles. The techniques we introduce in Chapter 1 are utilized as helpful tools for the development of novel Machine Learning approaches in Survival Analysis. Secondly, we describe the development of Machine Learning techniques driven by big data. In this part, we also introduce the concept of learning tasks and the characteristics of various popular Machine Learning methods, with a focus on tree models and ensemble methods. Next, we introduce novel approaches of Machine Leaning for Survival Analysis, and a variety of survival trees and survival ensemble methods are outlined in detail.

In Chapter 2, the basic statistical concepts, theories, and classic statistical tools necessary for constructing survival trees and survival ensemble methods are outlined, considering the main tasks that Survival Analysis has to address. We introduce the basic types of data

occurring in Survival Analysis, and corresponding examples in real applications are reported. Then basic and commonly used concepts are described and those concepts depict the possibility of survival over a given time or survival risk. Estimate of survival function, for example, the Kaplan-Meier estimator, is presented. After that, classical parametric methods and flexible semi-parametric models are described. Various statistical distributions and properties for parametric models are discussed. Then the well-known Cox-proportional hazard model is introduced. In addition, the imputation approaches for interval-censored data are explained.

In Chapter 3, a tree-based method and several ensemble algorithms for full data are outlined in detail. These techniques are then applied to construct some new survival trees and survival ensemble methods.

In Chapter 4, we propose new survival trees for interval-censored data based on the frame of conditional inference tree. What is more, ensemble methods that consider the new trees as base learners are implemented, which utilize "weight average" introduced in Quantile Regression Forests. Extensive simulation is performed to explore their predictive performance.

In Chapter 5, we discuss the application of the proposed survival tree methods in a variety of fields. We apply the new approaches to real data from medical and industrial fields. A risk credit model based on Machine Learning techniques and statistical methods in Survival Analysis is constructed to evaluate risk credit of the SMEs. In the last part, we summarize the research results and draw the conclusions of the thesis. Finally, future researches are prospected.

Chapter 1

State of the Art

1.1 Traditional statistical methods in Survival Analysis

The history of research in Survival Analysis can be traced back to the use of the life tables, which collect the surviving information of patients from a certain population and are one of the oldest statistical tools used for demographic analysis [98, 121, 167, 223]. In the early 1950s, many advanced theories of Survival Analysis were inspired by medical and biological applications. The 1950s and 1960s experienced a surge in demands for equipment reliability, and research in the field of reliability engineering was booming. As a consequence, applications of Survival Analysis extended to industry. Furthermore, Survival Analysis was applied to company failure prediction [141, 161], to prognostic personal credit risk [177], and even to criminological research [50]. The theoretical research results of Survival Analysis were increasingly fruitful in this period.

It must be noted that it is unsuitable to apply the classical statistical methods and traditional algorithms directly to predict the survival time and analyze survival data because of censoring. Hence, it is indispensable to construct original techniques and methods to analyze and model survival data, which is the focus of Survival Analysis.

As mentioned earlier, the oldest technique in Survival Analysis is the construction of life table, a nonparametric method measuring mortality and describing the survival of a population. A series of life table estimates have been proposed already in the 1950s [19, 48, 89]. Various techniques based on statistical theories have been developed to estimate survival function which describe the probability that survival time exceeds a time value.

A large literature is focused on nonparametric estimate of survival function. The Kaplan-Meier estimator (also called product limit estimator) has been proposed to estimate the

survival function as a limit of the life table estimator for right censored data [121]. This is a popular nonparametric maximum likelihood estimator (NPMLE) and still widely used until now.

The Nelson-Aalen estimate is another nonparametric approach to estimate the survival function based on right censored data [175, 176] which provides graphical checks of hazard shape for applications of industrial life testing. In fact, the Kaplan-Meier estimator is the product-integral of the Nelson-Aalen estimator of the hazard function which is also related to the probability of failure [2].

Subsequently, extensive studies focused on nonparametric maximum likelihood estimator (NPMLE) of survival function for various types of survival data. A self-consistency estimate characterized by a self-consistency equation was proposed to calculate the nonparametric maximum likelihood estimator which coincides with the Kaplan-Meier estimate [70]. Moreover, some improved iterative procedures based on [70] were proposed for doubly censored data and interval-censored data [227, 229]. Additionally, the iterative Convex Minorant (ICM) algorithm [9, 119] and EM iterative Convex Minorant algorithms [243] were proposed to determine the NPMLE of survival function for interval-censored data.

In this context, it is a basic and important issue to specify the distribution for the underlying survival time in Survival Analysis. Parametric models have been widely used to describe survival time assuming a certain distribution.

The Exponential distribution is often preferred over the normal distribution to describe the survival pattern due to its characteristics. Parametric estimation methods for the case of Exponential distribution are discussed in [73, 74]. Weibull distribution, a generalization of Exponential distribution, has been proposed and applied to the study of reliability and mortality in Survival Analysis [186, 241, 242]. Lognormal distribution, defined as the logarithm of a response variable following Normal distribution, and the Lognormal distribution has been utilized in cancer research. In fact, the investigation showed that the time to onset of some disease can be well approximated by a Lognormal distribution [21, 165]. Gamma distribution well describes the life length of electronic units and provides a useful model of Survival Analysis for industrial reliability problems [162]. Log-logistic distribution is a parametric model where the logarithm of survival time follows Logistic distribution. The rate of spread of virus can be described by this model [178]. Finally, Gompertz distribution is an extension of Exponential distribution and provides a reasonable model in Survival Analysis [95].

The estimator of unknown parameters in those parametric survival models can be obtained by maximum likelihood estimation procedure [147]. Due to the bias of the estimators, correction factors for maximum likelihood estimation and moment estimation have been proposed [251].

In medical practice, the prediction of survival probability for a patient according to his characteristics is indispensable. That aspect, also known as prognosis prediction, is crucial. In fact, it is critical to assess the relationship between survival time and covariates (also named predictor variables or risk factors) in Survival Analysis [128]. Thus, various parametric regression model have been studied in the literature analyzing the significant covariates related to survival time [18, 120, 143, 180]. A parametric regression model is the one that the survival time is assumed to follow a known parametric distribution, such as Exponential distribution. The regression coefficients are estimated by maximum likelihood estimation [128]. Moreover, AIC and BIC are proposed as powerful tools to select the appropriate parametric model with covariates [32, 202].

However, in real world the exact form of the distribution of survival time is hardly known, and that limits the use of the parametric survival models to assess the relationship between survival time and covariates. Moreover, it is difficult to identify the significant covariates. Hence, various semi-parametric models have been proposed for analyzing the underlying survival time. The Proportional hazard (PH) model is one of the most commonly used regression models because of its simplicity and flexibility. In the PH model, it is assumed that the hazard for one individual is proportional to the hazard for another individual. The model initially proposed and applied for right censored data is known as Cox Proportional Hazard model [55, 56]. Some generalizations of PH models to other type of censoring data are explored in the literature [112, 120]. For example, the semi-parametric Cox proportional hazards model was generalized for interval-censored data with time-dependent covariates by utilizing the nonparametric maximum likelihood estimation approach that considered the unknown cumulative hazard function as a step function [39].

There also exists an extensive literature about other continuous semi-parametric regression models used in Survival Analysis, including the proportional odd model, the additive hazards model, the accelerated failure time model, the linear transformation model. These models all specify the effect of covariates on the survival time through survival functions or hazard functions [1, 12, 17, 41, 44, 46, 47, 117, 127, 138, 171, 182].

In real applications, it is common and important to evaluate whether or not the survival functions of two (or more) samples from two (or more) continuous population are statistically equivalent, i.e., whether two (or more) samples follow the same distribution. For example, a comparison between treatment group and control group is used to assess the effectiveness of treatment in a clinical trial. Various sample tests are widely used to attempt to determine the difference between the estimators of two survival functions for the two groups.

It is extensively reported in literature the use of tests for two samples in the comparison of two survival distributions or two estimators obtained from survival data [120, 163].

Log-rank test [184] has been proposed by generalizing the Savage test [200] to survival data and has been widely used in applications of Survival Analysis due to its excellent ability to detect alternatives. More general a class of log-rank test with various weights has been extensively studied [102, 219].

Several linear rank test for right censored data are available in [187]. The generalized Wilcoxon test [87] is a distribution-free two sample test which extends the Wilcoxon test to samples with right censoring. The generalization of the Kruskal-Wallis test [27] extends the test of Kruskal-Wallis [136] and Gehan's generalization of Wilcoxon's test [87] to test the equality of K continuous survival functions. The power of those tests for right censored data has been discussed in [148].

We refer the interested reader to [39, 40, 65, 214] for comprehensive surveys about historical development and recent advances on statistical methods in Survival Analysis for interval-censored data.

1.2 Overview of Machine Learning based techniques

Due to breakthrough in data storage technology, automatic data acquisition and transmission, massive amounts of data have been made available in all sorts of fields since the 1980s, which has boosted the development of Machine Learning technology. One of the major objectives of research in Machine Learning is to construct models and practical algorithms that are capable of dealing with large amounts of data through self-learning and self-improvement. Hence, the research in Machine Learning is important in the era of big data [137, 220, 224].

In the last 20 years, many algorithms and models from Machine Learning have been proposed for specific learning tasks. These algorithms are applied to many fields, and are "used routinely to discover valuable knowledge from large commercial databases containing equipment maintenance records, loan applications, financial transactions, medical records and the like " [168]. Moreover, these algorithms often outperform other classical approaches for certain problems in some fields.

1.2.1 Learning tasks

Both Supervised learning and Unsupervised learning are subcategory of learning models, and aim to determine a model or an algorithm based on training data. However, the dataset for supervised learning consist of labeled data, comprising a large number of training instances, where each instance has a label (i.e., response variable) and feature vectors (i.e., covariate or predictor variable). In the supervised learning process, a predictive model is

trained using the training datasets and the objective is to predict the response variable value for new data with unknown label value. Differently, unsupervised learning techniques construct learning models from unlabeled training examples, where the objective is to uncover patterns, structures, and relationships within data without explicit supervision in the form of labeled outputs.

Semi-supervised learning is a learning model that utilizes the training datasets in which unlabeled and labeled training examples are mixed together [220]. Another not so common learning model is weakly supervised learning [255].

The classification and the regression tasks are two problems at the core of supervised learning. Roughly stated, when the labeled values of the datasets are discrete, then the task is classification. On the contrary, we talk of regression when the label values of the training examples are continuous. Both classification and regression are predictive tasks.

A more detailed explanation of classification is given in [220]: "The goal in classification is to assign an unknown pattern to one out of a number of classes that are considered to be known."

An example of frog sound classification is presented next. The goal of the automatic frog sound identification system is to predict which species of frog makes a specific frog sound. In this sample, the covariates include spectral centroid, signal bandwidth and threshold-crossing rate, while frog species are the labels of training datasets [111]. Another example is the prediction of the credit rates of an enterprise, given an enterprise for which the credit risk is unknown. The goal of a credit risk assessment system is to predict the credit rate based on the information on the enterprise that may include tax amount, number of upstream and downstream enterprises, length of business operation. A classification model based on the credit rates datasets will be presented in Section 5.3.

An example of regression is to predict the international rice prices given the volume of rice export of the last three months. In this case, international rice prices are the labels of the training datasets which are continuous values. The information about volume of rice export are covariates. The goal is to predict the price of rice given the volume of rice export in the previous three months [53]. Another example of regression concerns the stock excess returns. Here the stock excess returns are considered as the continuous labels in the training data, and the aim is to predict them given the current related market condition and information [72].

A range of Machine Learning approaches have emerged for the tasks of classification and regression [220]. It is worth noting that the approaches for regression in Machine Learning tend to be more data-driven and no assumptions are required for most of them, while traditional statistical regression model are based on strict mathematical assumptions [245].

1.2.2 Basic approaches in Machine Learning

Common Machine Learning approaches include Bayesian learning method, logistic regression, Support Vector Machine (SVM), Decision trees and ensemble methods, Artificial Neural Network (ANN) and more.

A brief introduction about those basic approaches is presented next. Tree-based methods will be described in detail in Subsection 1.2.3.

Naive Bayes is a learning method commonly utilized for classification. The method has solid theoretical foundation, which is the Bayes theorem. The Bayes theorem describes the relationship between the posterior probability and the prior probability. The core idea of Naive Bayes is to makes optimal decision based on probability. Specifically, assume that the event of interest has K possible outcomes, i.e., the set of class label for the response variable is $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$. Moreover, let $\mathbf{X} = (X_1, \ldots, X_m)$ be a vector of m features and assume that each feature is independent from the other features. For a given new instance with $\mathbf{X} = \mathbf{x}$, the prediction outcome $B(\mathbf{x})$ by Naive Bayes learning method is

$$B(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} P(c|\mathbf{x})$$

where $P(c|\mathbf{x})$ is the posterior probability that the new instance belongs to class c under the condition $\mathbf{X} = \mathbf{x}$, and it is calculated based on a formula about the posterior probability and the prior probability. The prior probability is obtained from the training data.

Naive Bayes is highly utilized in classification of natural language text documents and other practical classification problems. However, the assumption of conditional independence among features is required and that limits the application of the method [164, 192].

Although the term contains the word "regression", the task of Logistic regression is classification. Assume that the response variable y is a binary variable where the values are 0 or 1. Let $\mathbf{X} = (X_1, \ldots, X_m)$ be the explainable variable. For a new instance with explainable variable $\mathbf{X} = \mathbf{x}$, the following formula is utilized for discriminant,

$$P(y = 1|\mathbf{x}) = g(\theta^T \mathbf{x}) = \frac{1}{1 + exp(-\theta^T \mathbf{x})}$$

where $g(t) = \frac{1}{1+exp(-t)}$ is the sigmoid function, and θ is the vector of parameters [69]. If the probability $P(y = 1|\mathbf{x}) > 0.5$, then the instance is predicted to be positive class. Here 0.5 is the default threshold value. In some cases, a value different from 0.5 can be chosen for the threshold. This technique has been extended to multinomial logistic regression when the event has more than two outcomes [57, 75, 110].

Support Vector Machine (SVM) was proposed originally as a binary classifier, and later extended to multi-classification. Moreover, SVM can be also used for regression. The major techniques of SVM have continued to develop for 20 years from the 1990s to the 2010s.

The basic model is a linear classifier which can classify linearly separable dataset [232]. Moreover, kernel functions can be utilized to map linearly non-separable dataset to a high-dimensional feature space in which linear separation is implemented. Slack variables are also introduced to relax the constraints for better handing the case of non-separable data [54].

Support Vector Machines (SVM) are known for their ability to handle small-sized datasets and high-dimensional feature spaces. However, it is important to note that the performance of SVM can still be influenced by factors such as the choice of the kernel, tuning of hyperparameters, and the presence of class imbalance. Moreover, training SVMs on large-scale datasets can pose challenges in terms of storage space and computational requirements.

Artificial Neural Network (ANN) learning methods were inspired by neurobiology. They were originally proposed around 1958 [195]. ANN are techniques used for training real-valued or vector-valued functions over samples [104, 146], and they are versatile models that can be used for both classification and regression tasks.

A complex network structure is formed by the interconnection of a large number of neurons in ANN. In Multi-Layer Perceptrons (MLPs), neurons are organized in layers with arcs connecting nodes in one layer to nodes in the next layer. A single neuron is the fundamental unit of Multi-Layer Perceptrons (MLPs). It receives a set of input values and each input value which represents a feature, is multiplied by a corresponding weight. Then an output resulting from a weighted sum of the inputs is produced and a specific activation function is utilized to introduce non-linearity for the output of the neuron.

MLPs consists of multiple layers of interconnected neurons, including an input layer, one or more hidden layers, and an output layer. The neurons in each layer are interconnected with the neurons in the subsequent layer through weighted connections. The input layer receives the input data, passes it through the hidden layers, and finally produces the output in the output layer.

Back propagation (BP) [141] is one of the most common method used in Neural Networks, particularly for training MLPs. Gradient descent method is utilized as learning rule, the adaptive weight of neurons and threshold of the network are gradually adjusted via back propagation so as to minimize the sum of squares of error of the network.

The advantage of ANNs is that they are robust and, therefore, can be used with training data with noise or missing data, and are capable of approximating any nonlinear functions to arbitrary accuracy, given sufficient network capacity and appropriate training. The major drawback is that the final decision-making procedure by the ANN model is inexplicable, which decreases its credibility and acceptability. Besides, a large number of parameters are involved in the learning process, such as the initial value of weights, thresholds and

so on. Finally, the great computational requirements of Artificial Neural Networks pose a significant challenge.

In the next subsection, the development of tree-based methods in Machine Learning is introduced in detail.

1.2.3 Tree-based methods

Tree-based methods, particularly Decision Trees, are attractive learning methods due to their flexibility and interpretability, and have been applied to a wide range of classification and regression problems. In addition to Decision Trees, tree-based methods include Random Forest and other ensemble methods that utilize an ensemble of Decision Trees.

Decision Trees share some common features. They partition the training data into subsets based on some specific splitting rules based on the attribute values. However, they may differ in their specific criteria used for splitting and constructing the tree.

A tree is usually constructed by top-down induction. The feature space is partitioned by using a splitting criterion and forms an upside-down tree structure, the size of which is determined by the stopping and pruning rule. For the classification task, a tree-based classifier divides the training data into several mutually exclusive groups, each of which is labelled by the categories more represented within it. For the regression task, the regression tree constructs a piecewise linear regression function.

Moreover, Decision Trees have the ability to utilize as much information from the training datasets as possible during the learning process. This can lead to relatively large decision trees, especially when the datasets are complex or contain a large number of features.

Next, we provide a historical overview of the research on Decision Trees. A variety of tree models based on different splitting rule have been proposed in literature since the 1950s. Automatic Interaction Detection (AID) was the original regression tree proposed in 1963 by Morgan and Sonquist [170]. Sethi and Chatterjee constructed an efficient classification tree by utilizing a new criterion minimizing the costs [204] in 1977. In the same year, invariant imbedding principle of dynamic programming was applied for a novel recursive approach [183]. Then Rounds [196] designed a novel tree using an optimal classification strategies via using the most informative features in 1980. Later Diday considered nearest neighbours with constraints to induce a tree-structured classifier [62]. Li and Dubes in 1986 utilized a permutation statistic to design a new criterion and proposed a tree with automatic selection of the topology [151].

During the same period, the tree-building method for classification known as ID3 [190] was proposed. It uses information gain as a criterion to choose a splitting attribute. The attribute with the highest information gain is chosen as the splitting attribute at each node

of the tree. The proposed method was also capable to deal with the case that the data were noisy or missing.

C4.5 [189] is a modified version of ID3 algorithm where maximization of information gain ratio is utilized as a splitting criterion. While ID3 is prone to select attributes with more values based solely on their information gain, C4.5 introduces information gain ratio to overcome this bias towards attributes with a large number of distinct values. In addition, C4.5 algorithm adopts pruning technique to overcome over-fitting, and it can be applied to continuous or discrete attribute. However, both ID3 and C4.5 are only able to deal with the case of discrete labels, i.e. the design of them is for classification task rather than regression task.

Later CART (Classification And Regression Tree) [26] was proposed as an extension of C4.5. CART is a binary tree method which can handle classification task as well as regression task. For the classification task, the Gini impurity index is employed to choose the splitting attribute. For the regression task, least squares deviation is utilized instead.

The CART algorithm is known for its ability to decrease the size of the decision tree and reduce the computing complexity compared to algorithms like ID3 and C4.5. It can also be utilized in the case of a large of predictor variables, i.e. it can handle high-dimensional data effectively. What's more, CART algorithm is also useful in identifying important predictor variables [216]. It adopts a greedy search approach, but it has an undesirable weakness: biased variable selection. This bias occurs because it evaluates all possible splitting points for each candidate variable and selects the one that optimizes the chosen criterion (Gini impurity or least squares deviation) [122]. The predictor variable with more distinct values has the higher chance to be selected as a splitting variable [64, 157]. This process does not consider interactions between variables.

In 1994, Smoothed and Unsmoothed Piecewise-Polynomial Regression Trees (SUPPORT) [38] has been proposed, where the splitting rule is based on the sign of residuals and cross-validation estimates of mean square error. SUPPORT was later extended to a general piecewise linear model and was also suitable for censored data [37].

In 2000, Li proposed Principal Hessian Directions Regression Trees (PHDRT) [150], where Principal Hessian Directions (PHD), a dimensionality reduction technique, was utilized for partition. The prediction of PHDRT is more accurate than SUPPORT due to flexible partition rules. In addition, PHDRT was designed and explained from a geometric viewpoint which is quite different from the trees mentioned above.

GUIDE method [157] was propsed in 2002. Here the sign of residuals was utilized in a way similar to SUPPORT and two-sample t tests for unequal means and variances were used to split elements within a node. The split selection procedure was implemented in two steps to avoid selection bias from greedy search. What's more, it overcomes the drawbacks

of SUPPORT that fails to deal with categorical predictor variables and to detect pairwise interactions.

Quantile regression trees [36] extended the GUIDE method. Data were recursively partitioned based on the residuals from the quantile regression model. The advantage of quantile regression trees is that they model the relationship between the response and the covariates not only just at the centre but also at the tails of the responses distribution.

ID3, C4.5 and CART are all considered as "axis-parallel decision tree" [103, 246], because the attribute space is partitioned by axis-parallel hyperplane. However, this way of partition probably leads to a more complicated tree structure if the true decision boundaries that are not aligned with the attribute axes. A series of variation of the axis-parallel trees were proposed referred as "oblique decision tree" [13, 103, 172, 173, 244, 246]. A multivariate linear combination of attributes and binary quantization are obtained by an oblique partition. In general, for these methods, the structure is simplified and performance is better compared to traditional tree based on axis-parallel split. However, an initialization with best axis-parallel splits is still required, with high running-time cost which is the problem of oblique decision trees [246].

We refer the interested reader to [199] where a survey of different tree-structured classifier methodology is discussed.

Although tree methods have many advantages, the primary disadvantage is their sensitivity to small variations in the training data. i.e., even a limited noise in the training data can greatly change the selection of the splitting variable and splitting values, and then overfitting occurs easily. Bagging [23] has been proposed to overcome the instability of trees, and enhance the predictive accuracy. Bagging is a technique for classification and regression that involves utilizing bootstrap samples drawn from the original data to train base learners. If the base learners are sensitive to small variations, such as the case of CART, the average prediction error can be reduced significantly by this technique [24].

Random Forest (RF) [25], a modified version of bagging, is a very popular Machine Learning method with powerful predictive ability even for high dimensional data. It is an ensemble of tree predictors inspired by [6]. The simple predictors are generated based on bootstrap subsets of the training data and randomly sampled subsets of features. That will reduce variance compared to classical tree methods [90].

Quantile Regression Forests (QRF) [166] have been proposed in 2006 as a generalization of RF. QRF is a regression ensemble method that utilizes the CART algorithm as its base learners. It is worth noting that QRF calculate the weighted mean of the observed response variable within each leaf node rather than averaging the prediction of base learners. In fact, averaging prediction is equivalent to the weighted mean based on conditional mean estimation of the response variable considering full data [166]. The weight based scheme

proposed in Quantile Regression Forests show better performance when more complex data such as censoring data are considered [10].

Generalized Random Forests [10] have been constructed by Athey in 2019. They were developed for the quantile regression, conditional average partial effect estimation and heterogeneous treatment effect estimation. Several prognostic model-based Random Forests for ordinal regression have been proposed in 2020 [31]. The approaches are powerful in case of proportional and non-proportional odds. In addition, Han proposed Double Random Forest (DRF) [101] in 2020. This is a novel classification ensemble method inspired by the principle and basic properties of Random Forest. The novelty is that sufficiently large size of trees are generated. DRF improves the predictive accuracy of ensemble method of trees compared to Random Forest.

1.3 Machine Learning techniques in Survival Analysis

1.3.1 Overview of survival trees and survival ensembles

Compared to traditional models and methods in Survival Analysis, Machine Learning methods can more flexibly and effectively handle high-dimensional problems. In addition, no specific assumptions are required [236]. Due to the attractiveness of tree methods in Machine Learning, researchers in Survival Analysis have concentrated their attention on the generalization of them for censored data.

Survival tree is a type of tree-based method specifically designed for analyzing censored data in Survival Analysis. A variety of survival trees are proposed for right censored data arising in real world. A comprehensive review of tree-structured methods for survival data with right censoring is presented in [22], where more than 20 survival tree methods developed from 1985 up to 2008 were discussed. In general, there are two aspects to be considered in all common splitting criterion:

1. one is to maximize within-node homogeneity and in this case the distance between estimates of survival functions or the residuals are usually considered as the measure of within-node homogeneity [96, 124, 158];

2. the other is to maximize between-node heterogeneity. In this case, various test statistics measuring between-node separation are applied for splitting rules, giving rise to different version of survival trees. For example, logrank statistic [51, 145, 203], a parametric likelihood ratio statistic [60], Wilcoxon-Gehan statistics and Kolmogorov-Smirnov statistics [52, 203] are all utilized to measure the heterogeneity between nodes.

However, the majority of tree-structured methods presented in [22] suffer from biased variable selection, the cause of which was mentioned earlier in Subsection 1.2.3. Conditional inference trees [107] is also a tree method where a permutation test framework is utilized. It overcomes this weakness and is not biased towards covariates with many values due to the special framework. Thus unbiased variable selection is implemented by conditional inference trees for full data or right censored data flexibly.

As mentioned before, ensemble methods can resolve the problem of high variance, and thus the performance of prediction is better than tree methods. A series of novel survival ensemble methods have been proposed where the existing or original survival trees are the base learners.

Bagging method was extended to bagging survival trees [108] for right censored data where the survival trees were the base learners and the estimated survival functions were given as predictive outcomes [108]. Relative risk forests were proposed in [114], where the base learners (survival trees) were constructed from bootstrapped data and randomly selected subset of covariates, with the splitting rules involving the Poisson likelihood.

Kretowska proposed a survival ensemble method [134], and introduced the survival dipolar trees [133] as base learners. Hothorn [109] assigned inverse probability of censoring weights to the loss function and combined this loss function with the basic procedures of Random Forest to construct a survival ensemble. The estimation of log-survival time were obtained based on the prediction weights in the ensemble method.

Random Survival Forests (RSF) [115] have been proposed by Ishwaran in 2008. RSF considered four different splitting rules: log-rank splitting rule, conservation-of-events splitting rule, log-rank score rule and random log-rank splitting rule. Ensemble mortality was defined as a novel predicted outcome for right censored data in RSF, and the predictive outcome was obtained using the Nelson-Aalen estimator of the cumulative hazard function.

Wang proposed a novel survival ensemble in 2017 [234]. The method combined random subspace, bagging and extended space techniques to build an extended variable space and improved the prediction capability of classifier. In [254], a survival ensemble algorithm has been proposed by following the basic frame of Random Forest where partial least squares regression and the Buckley-James estimator were utilized. New survival unbiased trees and ensembles were developed in 2019 based on a new version of loss function for right-censored data where an extension of the censoring unbiased transformation was applied to construct the loss function [209].

While research of survival trees and ensemble methods for right censored data is quite extensive, relatively few studies are devoted on other types of survival data especially on interval-censored data due to complexity of data.

A survival tree was developed for left-truncated and right-censored data by Wei and

Simonoff [238]. Moreover, a survival tree method for interval-censored data was proposed by Yin in 2002 [250], but the method suffered from bias of splitting variable selection. In 2017, $ICtree$ was proposed as an unbiased tree method for interval-censored data [237], which embedded the log-rank score into the Conditional Inference Frame. An ensemble method ($ICcforest$) [249] based on $ICtree$ was suggested in 2019 for interval-censored data. In 2022, interval censored recursive forests (ICRF) [49] were proposed by employing Wilcoxon rank sum test for splitting step at each iteration. ICRF are suitable for Case I interval-censored data which is a special type for interval-censored data.

In this thesis, we propose an unbiased interval-censored tree by combining Generalized Log-Rank Tests (GLRT) with the Conditional Inference Frame which extends the works from [237]. We also construct a survival ensemble considering the novel interval-censored trees as base learners.

1.3.2 Other Machine Learning techniques for survival data

Artificial Neural Networks (ANN) has been also successfully extended to handle censored data in Survival Analysis and they have demonstrated good performance in modeling and predicting survival outcomes. A proportional hazards Neural Network [77] was proposed for survival data in 1995. The basic idea was that the linear part of the Cox proportional risk model was replaced by the relationship between input variable and output in Artificial Neural Networks. A Partial Logistic ANN (PLANN) method was proposed in [193] to determine the relationship between predictor variables and response showing better predictive performance.

Several Deep Learning Survival Analysis approaches are utilized in assessing the survival risk of patients, that can help physicians make appropriate clinical decisions [118, 123, 191]. Survival Neural Additive Model (SurvNAM) [3, 230] was designed as an extension of a linear combination of Neural Networks. SurvNAM combined the flexibility and predictive power of ANNs with the ability to provide explanations for survival predictions. A specified loss function tailored for survival analysis was utilized in SurvNAM.

Support Vector Machine methods have been extended to handle right-censored data and various modified loss functions are utilized [11, 16, 125, 206]. Recently, Least Squares Support Vector Regression model (LS-SVR) has been proposed for more complex censored data, including interval-censored data, left-truncated and right-censored data [156].

We refer the interested reader to [236] for a comprehensive survey about a variety of Machine Learning techniques and their application to right censored data.

Chapter 2

Basic concepts and traditional statistical models in Survival Analysis

In this chapter we first introduce some basic data types and common terminology in Survival Analysis, which will be used in our thesis. Then some basic statistical methods and traditional models in Survival Analysis will be discussed. The techniques covered here will be applied in later chapters.

2.1 Categories of survival data

Censoring is considered as a unique feature in survival data. In fact, data in survival study are often incomplete due to time and cost of the study. In many cases, only partial information of survival time rather than exact survival time can be obtained, and in this case, we say that censoring occurs or the data is censored. Censored data is easy confused with missing data. However, they are quite different, although for both of them there is a loss of information. Missing data can not provide any information about the time of occurrence of event of interest, whereas censoring shows a range that the survival time falls in or provides some information about occurrence of event or not at some time. There are several types of censoring depending on the contexts of study and the observation methods: right censoring, left censoring and interval censoring.

2.1.1 Right censoring

Right censoring occurs when:

- for an individual the event of interest does not occur during the study period. Considering the study cost and practical operation, the study period can not be infinite, so it is possible that the event occurs after the study ends;
- the contact of an individual during the process of study is lost and the observation of the person can not be completed;
- an individual withdraws or has to be withdrawn from the study due to some reasons, for example, died in a car accident which is not the event of interest.

In those cases, the time at which the study ends or he (or she) fails to be contacted or withdraws from the study is referred as censoring time. Censoring occurs when the event of interest (e.g., death, failure, relapse) has not occurred for a participant up until the censoring time. In fact, the person is still surviving at the censoring time, and we don't know the exact time of occurrence of event of interest. The data is naturally called right censored, since the survival time which is incomplete is on the right side of the observation time.

Let T denote a non-negative random variable for the survival time which represents the time of occurrence of the event of interest. Let C denote the censoring time that is the occurrence time of censoring. Censoring time may be one of following cases: the ending time of study, the time when the person is lost during the study or the time when the person withdraws from the study.

Let δ be an indicator variable which represents whether the event of interest occurs or not. Here, δ=1 means that the survival time is observed exactly, in another words, the individual experiences the event before the end of study. Instead, δ= 0 means that survival time is censored and we only know that the survival time is subsequent to the censoring time.

The observed data for the individual can be represented by a pair of random variables (O, δ) where O is equal to T if the time of event of interest is known exactly, or equal to C if it is censored. Therefore, $O = \min\{T, C\}$ represents the known observation time.

An example of right censored data is reported in [211]. A clinical trial which is designed to explore the effects of 6-mercaptopurine (6-MP) on remissions in acute leukemia is considered. The event of interest is going out of remission for the patients with acute leukemia.

To better explain right censored data, consider the following example. Suppose that the length of the study period is 11 months, data have been collected since the beginning of

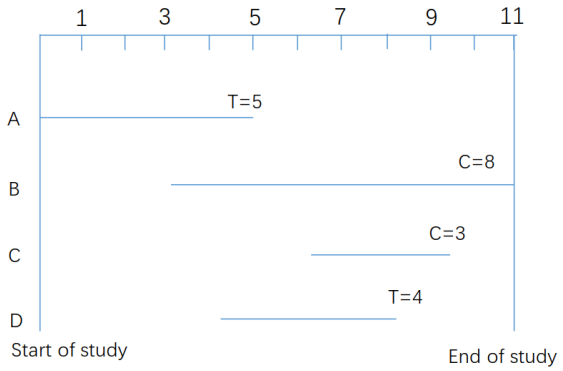


Figure 2.1: Right censoring.

the study. Patient A is followed from the beginning of the study until the leukemia return at months 5, so for patient A, the survival time T is 5 and the time of event of interest is observed exactly. Patient B enters the study at month 3 and is observed for the remaining time of study. He is still in remission (without experimenting the event) until the end of study, which means that the survival time for patient B is censored and it is not observed exactly. In this case, the censored time is 8 months. Patient C enters the study at month 6 and he is followed up until month 9. which means his censored time is 3 months. Finally, patient D is observed starting at month 4 until going out of remission at month 8. Then the survival time is 4 months and it is not censored. These cases are illustrated in Figure 2.1.

Table 2.1 reports the survival time data for the 4 patients we considered. The information on the indicator variable is outlined.

Table 2.1: Survival data for 4 patients.

Individual	Observed data
	(O, δ)
A	(5,1)
B	(8,0)
C	(3,0)
D	(4,1)

$\delta = 0$ indicates that the survival data is censored, while when $\delta = 1$ the time of event of interest is observed exactly.

2.1.2 Left censoring

If the event of interest has occurred before the start of a study or the observation for the specific individual, the exact time of occurrence of event is unknown. Therefore, we only know that the survival time is antecedent to the time when the study starts or the individual enters the study, and the exact time of occurrence of event is unknown. When the survival time T is less than the censoring time C, i.e., $T \leq C$, left censoring occurs.

An interesting example is the study of the time until first marijuana use among high school boys in California [100]. In this survey, the following question is asked to 191 California high school boys:"When did you first use marijuana?" Three kinds of answers are usually given:

- the boy can remember the exact age when he first used marijuana.
- the boy has used marijuana but he fails to remember the exact age of first use.
- the boy never used marijuana.

For the first answer, the survival data is not censored. In the second case, the exact age experiencing the event of interest is not known but only knows that the age of first use marijuana is prior to the current age, and therefore, survival time is left censored. For the third answer, the exact age is still not known, and the survival time that the event occurs is larger than the boy's current age, so in this case it is right censored.

If both left censoring and right censoring may occur in a study, then the survival data is defined as doubly censored [227].

2.1.3 Interval censoring

If the study subject or the event of interest can not be observed continuously, the survival time is not always observed exactly. In many cases, it is only known to fall in an interval $(L, R]$ (the endpoints of an interval usually represent two observation time). This type of censoring is called interval censoring.

Interval censoring is a more general type of censoring in Survival Analysis, and both left-censoring and right-censoring are special cases of it. When $R = \infty$ is the right endpoint of an observed interval or $L = 0$ represents left endpoint of an observed interval, then the interval censoring becomes a natural generalization of the common right censoring case or the less common left censoring case.

An exact observation corresponds to the case $L = R$. If we can only observe each study subject once and either $L = 0$ or $R = \infty$, we referred to it as Case I interval-censored data or current status data. Otherwise, if the interval-censored data include some finite intervals

and L is not always equal to 0, then it is called Case II interval-censored data [40, 214]. In this thesis, when dealing with interval-censored data we refer to Case II interval-censored data, unless specifically we point out that we are considering Case I interval-censored data.

Interval-censored data is common in real world. For instance, in clinical trials or longitudinal study, patients have periodic follow-ups which are possibly discontinuous. Hence the patient's exact time of event of interest is not observed, but it is only known to occur within an interval. In these cases, interval censoring data occurs. In [14, 15], a study about cosmetic effects in women with early breast cancer is presented, where a control trail is carried out to compare radiotherapy alone to radiotherapy and adjuvant chemotherapy. Patients on different treatment regimes visit the clinician every 4-6 months, and the measure of breast retraction are labeled with three levels: none, moderate, severe. The event of interest in this case is the first occurrence of a moderate or severe breast retraction. Since the patients are not observed continuously, the time of occurrence of the event of interest is only known to be in an interval between two visits. Therefore, the data in this study is interval-censored data. If the occurrence of moderate or severe breast retraction for a patient is not observed by the end of study, then right censored data occurs [128].

2.2 The censoring mechanism

Consider the case of right censored data, if the survival time T and the censoring time C are independent, then the censoring mechanism is referred as non-informative censoring [120, 128]. As further explained in [128], "non-informative censoring means that knowledge of a censoring time for an individual provides no further information about this person's likelihood of survival at a future time had the individual continued on the study."

For interval-censored data, censoring mechanism is non-informative if the survival time is independent of the censoring intervals.

Consider the following example. Assume that a longitudinal survey is carried out on the emergence time of a permanent tooth. We will discuss in detail this application in Section 5.1. In this case, the events of interest is the emergence time of a specified permanent tooth and survival time T is the age of emergence of the particular permanent tooth. In fact, the observed time is prespecified in the study and the schedule of observed time does not consider the dental status of the subjects. Obviously, the time of emergence of a permanent tooth can not be observed exactly, and we only know that it occurs within a time interval between observations. What's more, the time of emergence of a permanent tooth does not dependent on the observation time which is just the end point of the interval. Therefore, the censoring mechanism is non-informative.

Throughout the whole thesis we assume that the censoring mechanism is non-informative.

2.3 Basic functions in Survival Analysis and the estimation method

2.3.1 Basic functions and survival curves

Let the nonnegative random variable T represent the survival time from a homogeneous population. The characterization of the probability distribution of T is important and useful in Survival Analysis. This is accomplished through several functions: survival function, hazard rate function and probability density function. These functions are closely related to each other and illustrate the distribution of T from different perspectives.

The survival function $S(t)$ describes the probability that an individual survives beyond time t, i.e.,

$$S(t) = P(T > t), \quad 0 < t < \infty. \tag{2.1}$$

The survival function satisfies

$$S(t) = 1 - F(t),$$

where $F(t)$ is the cumulative distribution function. Thus the survival function is the integral of the probability density function $f(t)$, that is

$$S(t) = \int_t^\infty f(\tau)d\tau \tag{2.2}$$

where $f(t)$ is the probability density function.

Survival functions satisfy the following basic properties:

- they are monotone, nonincreasing. In fact, the probability of surviving exceeding t decreases with the increase of t;
- $S(0) = 1$ and $S(\infty) = 0$, since the event of interest can not be experienced at the start of study, and no one survives if the the study period is infinite.

The survival function is fundamental in Survival Analysis and commonly used to compare two or more survival patterns. In Figure 2.2, survival function for black males and white females are compared (data is taken from the report published by the United States Department of Health and Human Services in 1990 [83]. The report yearly publishes the survival probabilities for all causes of mortality by race and sex.)

The survival curve for white females lies above the one for black males, and the graph indicates that white females have a better survival probability than black males.

Another basic function is the hazard rate function which describes the instantaneous increment of probability per unit of time for the event to occur, given that the individual

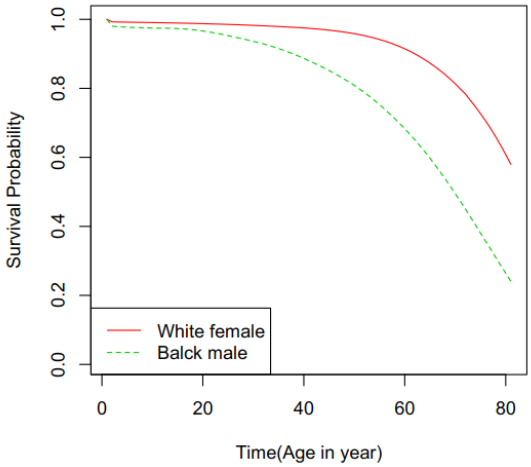


Figure 2.2: Comparison of survival curves.

is surviving at time t. More precisely, the hazard rate function $h(t)$ is defined as

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t}. \tag{2.3}$$

According to the definition, $h(t)$ is nonnegative and has no upper bound.

Let the survival time T be a continuous random variable. The relationship among $S(t)$, $h(t)$ and $f(t)$ can be inferred via the definition of $h(t)$:

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0^+} \frac{P(t \leqslant T < t + \Delta t)}{P(T \geqslant t)\Delta t} \\ &= \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{S(t)\Delta t} \\ &= \frac{f(t)}{S(t)}. \end{aligned} \tag{2.4}$$

Moreover,

$$h(t) = -\frac{d}{dt}(\ln(S(t))) \tag{2.5}$$

follows from Formula (2.2) and (2.4). The integral of the hazard rate function is called cumulative hazard function:

$$H(t) = \int_0^t h(\tau) d\tau. \tag{2.6}$$

Then, from Formula (2.5) and (2.6), it follows that

$$H(t) = -\ln(S(t)). \tag{2.7}$$

From Formula (2.7), it is easy to infer that

$$S(t) = \exp(-H(t)). \tag{2.8}$$

Therefore, once one of these basic functions is known, then all the others can be determined. Later in Subsection 2.4.1 an example will be presented on the calculation of these basic functions.

Considering the data of the survival probabilities by sex and race published by the United States Department of Health and Human Services in 1990 [83]. The graph of the hazard rate function is drawn in Figure 2.3, and we can infer that the hazard rate for black males is much higher than for white females. The conclusion is consistent with the one from Figure 2.2.

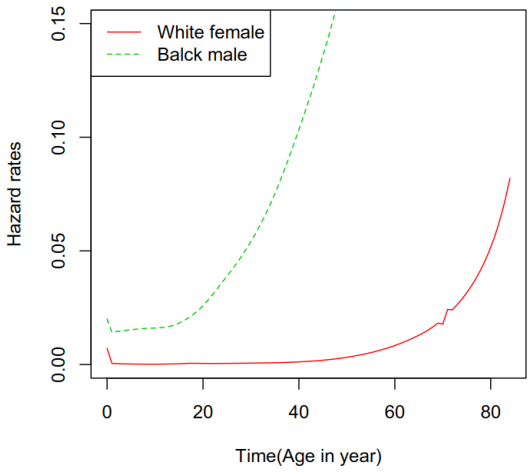


Figure 2.3: Comparison of hazard rate function.

2.3.2 Estimation of survival function

Right censoring

The survival function is a fundamental concept in Survival Analysis, and is often utilized for the prediction of survival probabilities, comparison of two or more samples, or graphical comparison of several clinical treatments. Thus the estimation of the survival function is essential and it is a primary task in Survival Analysis.

The Kaplan-Meier (KM) estimator is a well-known nonparametric estimator of the survival function for right censored data [121].

Let n denote the number of the independent subjects from a homogeneous population with survival function $S(t)$, and assume that right censored data are observed for these

subjects. Let T_i be the survival time and C_i be the censoring time respectively for subject i, $i = 1, \ldots, n$. The observed data is $\{(O_i, \delta_i)\}_{i=1}^n$, where O_i is the observed time of subject i, and δ_i is the indicator variable of subject i.

Let $\{s_j\}_{j=1}^l$ denote the distinct ordered values for the survival time $\{T_i, i = 1, \ldots, n\}$, i.e., assume that the events of interest occur at l distinct time $\{s_j\}_{j=1}^l$ and $s_j < s_{j+1}$ for $j = 1, \ldots, l-1$. Let d_j be the number of events occurring at time s_j. Moreover, let r_j be the number of individuals at risk up to time s_j, which is the number of individuals who survive over s_j or just experience the event at s_j. All quantities $\{s_j, d_j, r_j\}_{j=1}^l$, can inferred from the observed data.

The Kaplan-Meier (KM) estimator is defined as follows

$$\hat{S}(t) = \begin{cases} 1, & \text{if } \ t < s_1; \\ \prod\limits_{s_j \leq t} \left(1 - \dfrac{d_j}{r_j}\right), & \text{if } \ s_1 \leq t. \end{cases} \tag{2.9}$$

Due to the form of the estimator, it is also called Product-Limit estimator, and is a step function.

To better understand Formula (2.9) we present an example again on the effects of 6-mercaptopurine (6-MP) on remissions in acute leukemia mentioned in Subsection 2.1.1. The data is from a clinical trial [211] and the event of interest is relapse of acute leukemia. Table 2.2 shows the information for 10 patients under 6-MP treatment.

Table 2.2: Survival data for 10 patients under 6-MP treatment.

Patient i	Observed data (O_i, δ_i)	Patient i	Observed data (O_i, δ_i)
1	(6,1)	6	(9,0)
2	(6,1)	7	(10,0)
3	(6,0)	8	(10,1)
4	(6,1)	9	(11,0)
5	(7,1)	10	(13,1)

$\delta_i = 0$ indicates that the survival data for patient i is censored while when $\delta_i = 1$ the time of event of interest is observed exactly.

In this example $n = 10$. For patients 1, 2, 4, 5, 8, 10, the relapse of acute leukemia are exactly observed at time 6, 6, 6, 7, 10, 13 respectively, so the number of distinct survival time l is equal to 4, and $s_1 = 6,\ s_2 = 7,\ s_3 = 10,\ s_4 = 13$. At time $s_1 = 6$, three patients relapsed, so $d_1 = 3$. Since before $s_1 = 6$, no patient experienced the event, i.e., all patients are at risk, and we have $r_1 = 10$. At time $s_2 = 7$, only one patient relapse, so $d_2 = 1$. Besides, one patient is censored at time $t = 6$, so he is not considered any longer at time

$t \geq 7$, so at time $s_2 = 7$, six patients are still at risk and $r_2 = 6$. Thus, when $s_2 \leq t < s_3$, the KM estimator is

$$\hat{S}(t) = \left(1 - \frac{d_1}{r_1}\right)\left(1 - \frac{d_2}{r_2}\right).$$

More detail of the calculation of KM estimation are exhibited in Table 2.3.

Table 2.3: Kaplan-Meier estimator for 10 patients under 6-MP treatment.

Time t	s_j	No. of event	No. of at risk	KM estimator
		d_j	r_j	$\hat{S}(t)$
$t < s_1$	-	0	10	1
$s_1 \leq t < s_2$	6	3	10	$1 - \frac{3}{10}$
$s_2 \leq t < s_3$	7	1	6	$\left(1 - \frac{3}{10}\right)\left(1 - \frac{1}{6}\right)$
$s_3 \leq t < s_4$	10	1	4	$\left(1 - \frac{3}{10}\right)\left(1 - \frac{1}{6}\right)\left(1 - \frac{1}{4}\right)$
$s_4 \leq t$	13	1	3	$\left(1 - \frac{3}{10}\right)\left(1 - \frac{1}{6}\right)\left(1 - \frac{1}{4}\right)\left(1 - \frac{1}{3}\right)$

The Kaplan-Meier estimator for 10 patients under 6-MP treatment are calculated at time t. The estimator is step function.

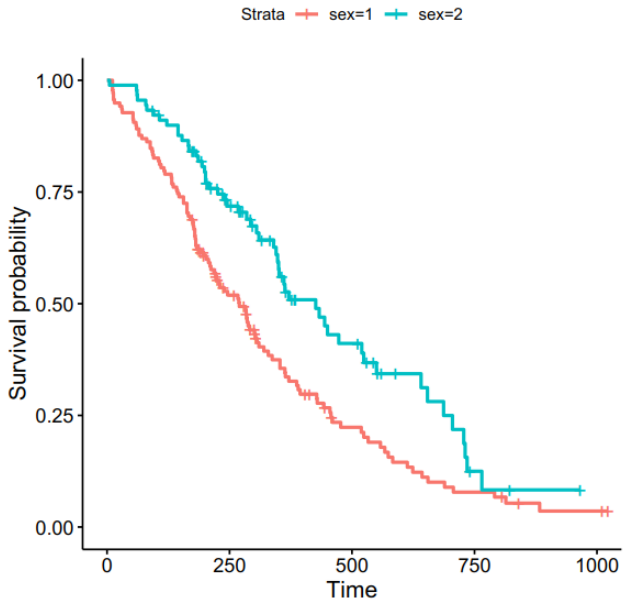


Figure 2.4: Comparison of Kaplan-Meier estimator of survival function for 228 male (sex=1) and female (sex=2) patients with advanced lung cancer.

In Figure 2.4, the Kaplan-Meier estimators functions are shown visually based on the data from the report of 228 patients with advanced lung cancer from the North Central

Cancer Treatment Group [159]. When we just consider 90 patients, the Kaplan-Meier estimators functions are shown in Figure 2.5, and it is easy to see that is a step function. The plots display the KM curves for female and male patients respectively, the unit of time in these plots is days. The survival curve is drawn only up to days 1022 which is the largest time of observation. The analysis of the results of estimators shows that the survival probability of female patients with advanced lung cancer is higher than the one of the male patients over a period of two years.

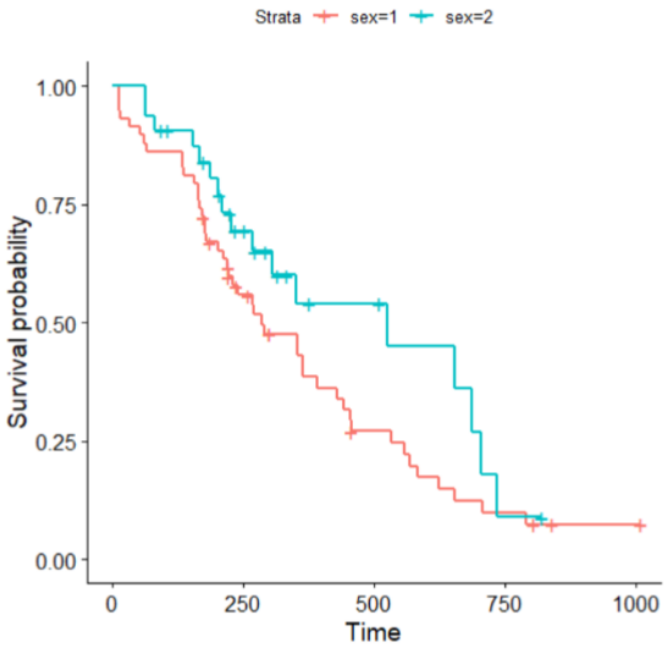


Figure 2.5: Comparison of Kaplan-Meier estimator of survival function for 90 male (sex=1) and female (sex=2) patients with advanced lung cancer.

Interval censoring

Turnbull [229] proposed in 1976 a nonparametric maximum likelihood estimator (NPMLE) as a generalization of the Kaplan-Meier estimator for interval-censored data. The NPMLE is specifically designed to estimate the survival function when the event times are observed within intervals rather than an exact value. First, we introduce the likelihood function for Case II interval-censored data.

Assume that interval-censored data consists of n independent subjects from a homogeneous population with survival function $S(t)$. Let T_i be the survival time for subject i, and suppose that $(L_i, R_i]$ is the censored interval to which T_i belongs, $i = 1, \ldots, n$. Let

$\{s_j\}_{j=0}^{l}$ denote the distinct ordered elements of the values of observed interval endpoints $\{L_i, R_i\}_{i=1}^{n}$ and 0. Moreover, define that $p_j = S(s_{j-1}) - S(s_j)$, $j = 1, \ldots, l$. The likelihood function can be expressed as

$$L_S(\mathbf{p}) = \prod_{i=1}^{n} (S(L_i) - S(R_i)) = \prod_{i=1}^{n} \sum_{j=1}^{l} \alpha_{ij} p_j \tag{2.10}$$

where $\alpha_{ij} = I(s_j \in (L_i, R_i])$, $i = 1, \ldots, n$, $j = 1, \ldots, l$, $\mathbf{p} = (p_1, \ldots, p_l)^T$.

Next, a detailed illustration of those notation is provided through an example. Consider an interval-censored dataset {(0,2], (1,4], (2,7], (3,4]} consisting of four intervals, so $n = 4$. In this case $\{s_j\}_{j=1}^{5} = \{0,\ 1,\ 2,\ 3,\ 4,\ 7\}$ and $l = 5$. It is obvious that $s_j = 7$ when $j = 5$. What's more, $7 \in (L_i, R_i]$, only when $i = 3$; otherwise, $7 \notin (L_i, R_i]$. Thus for $j = 5$

$$\alpha_{i5} = \begin{cases} 1, & \text{if } \ i = 3; \\ 0, & \text{if } \ i \neq 3. \end{cases}$$

With similar calculation for the cases that $j = 1, \ldots, 4$, the following 4 by 5 matrix is obtained with all the values of α_{ij}:

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

It is clear that the likelihood function is only determined by the values of the survival function S at the points s_j, $j = 1, \ldots, l$. According to [159, 229], the estimate of S is calculated by maximizing $L_S(\mathbf{p})$ with respect to $\mathbf{p}$ under the conditions

$$\begin{aligned} &\sum_{j=1}^{l} p_j = 1 \\ &p_j \geq 0, \quad j = 1, \ldots, l. \end{aligned} \tag{2.11}$$

It is usually assumed that the NPMLE of S takes the form of a step function, i.e., it is constant between s_{j-1} and s_j, $j = 1, \ldots, l$.

Unlike the Kaplan-Meier estimator, a closed form is not available for the NPMLE for the survival function for Case II interval-censored data and it has to be obtained by iterative algorithms. Self-consistency algorithm [229] originally proposed by Turnbull in 1976 is an iterative algorithm to determine NPMLE of survival function. This algorithm is in essence an EM algorithms [159], where a conditional expectation is calculated in the E-step and then the conditional expectation is maximized in the M-step.

Convergence speed of the self-consistency algorithm is very slow. Therefore, the Iterative Convex Minorant algorithm (ICM) [9] was introduced in 1992 to speed up the convergence. The method was improved in 1998 [119] via the maximization of a quadratic function substituting the maximization of the likelihood function. Later, a hybrid algorithm EM-ICM was developed to improve the performance by combining the self-consistency and ICM algorithms.

Anderson-Bergman proposed a novel implementation of the EM-ICM algorithm in 2017 [7]. It is very efficient and linear time is allowable at each iteration. In this novel algorithm, the likelihood function is approximated by a second order Taylor expansion. Then the Weighted Pool Adjacent Violators Algorithm (WPAVA) is applied for maximizing the likelihood function with the constraint condition. This improved implementation of the EM-ICM algorithm will be utilized in our new survival tree and ensemble methods in Chapter 4. The algorithm is implemented in the R package "icenReg" [8].

It's worth noting that these are just a few methods for estimating the survival function, and there are other methods and variations available depending on the specific context and data characteristics. The choice of the method should be based on the underlying assumptions, data structure, and research objectives.

2.4 Parametric models

In this section, we assume that the survival time follows specific distributions. Under these assumption, the survival function and hazard function have simpler expressions. In particular, the following parametric models are described: Exponential distribution model, Weibull model, Log-mormal model, Log-logistic model.

2.4.1 Exponential distribution model

The Exponential distribution model is one of the simplest parametric models, and it is commonly applied in Survival Analysis. For example, the survival time of a chronic disease patient is assumed to follow an Exponential distribution in [33, 79, 256]. A nonnegative random variable T representing the survival time is from an Exponential distribution, if the survival function is

$$S(t) = \exp(-\lambda t), \quad t > 0 \tag{2.12}$$

where λ is a positive parameter. It is easy to infer the density function from Formula (2.2) and (2.12):

$$f(t) = \lambda \exp(-\lambda t), \quad t > 0. \tag{2.13}$$

The hazard rate function $h(t)$ can be derived via Formula (2.5) and it is constant

$$h(t) = \lambda, \quad t > 0. \tag{2.14}$$

That implies that the survival hazard is fixed over time. The characteristics of the survival curves for different parametric values λ are shown in Figure 2.6.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$ denote an m-dimensional vector of covariates[1]. In an Exponential regression model it is assumed that the hazard rate function of T given $\mathbf{X} = \mathbf{x}$ is

$$h(t, \mathbf{x}) = \lambda \exp(-\mathbf{x}^T \boldsymbol{\beta}), \quad t > 0 \tag{2.15}$$

where λ is a positive parameter and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^T$ is a vector of regression coefficients. The Exponential regression model can be considered as a special case of Cox proportional hazards models which will introduced in Section 2.5. In fact, it is just the Cox proportional hazards model when the λ is replaced by a function of t.

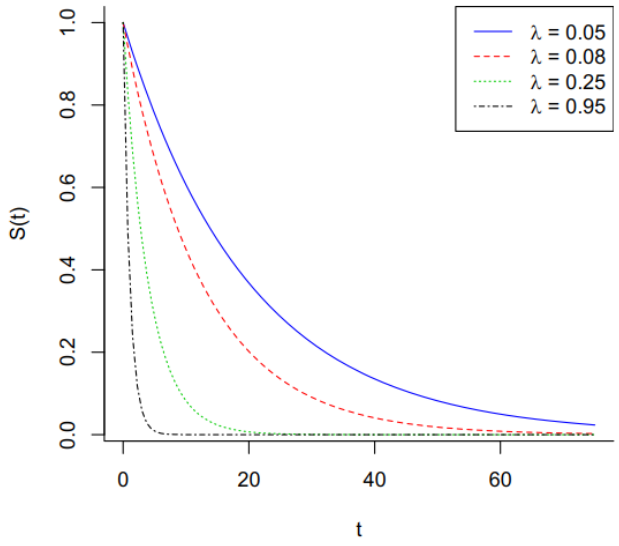


Figure 2.6: Survival curves under the Exponential distribution for different values of λ.

2.4.2 Weibull distribution

The hazard rate function $h(t)$ is just a constant in the Exponential distribution model. Therefore, the Exponential distribution model is too restrictive and can not be suitable

[1]It is worth noting that "covariates" is a term commonly used in statistics, while in Machine Learning the term that has the same meaning is "predictive variables". The term "covariates" will be used consistently through this chapter.

in many real situations. A general distribution model known as the Weibull distribution is utilized more often to reflect the inner properties of real survival data in some context. For instance, the failure life of ST-37 steel which is a type of steel is assumed to follow the Weibull distribution in [241, 242].

The Weibull model is a two-parameter model, the survival function is

$$S(t) = \exp\left(-\lambda t^{\gamma}\right), \quad t > 0, \tag{2.16}$$

where $\lambda > 0$ is a scale parameter, and $\gamma > 0$ is a shape parameter.

It is obvious that the Exponential distribution is a special case of the Weibull distribution when the shape parameter γ is set to 1.

The hazard rate function under the Weibull distribution is

$$h(t) = \lambda\gamma(t^{\gamma-1}), \quad t > 0. \tag{2.17}$$

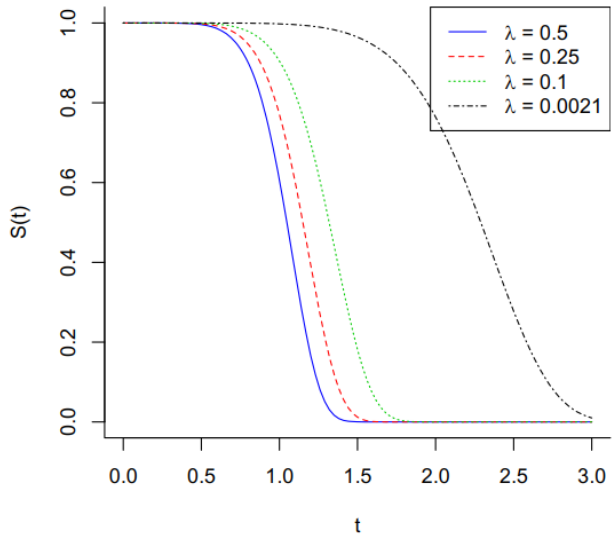


Figure 2.7: Survival curves under the Weibull distribution for $\gamma = 7$ and different values of λ.

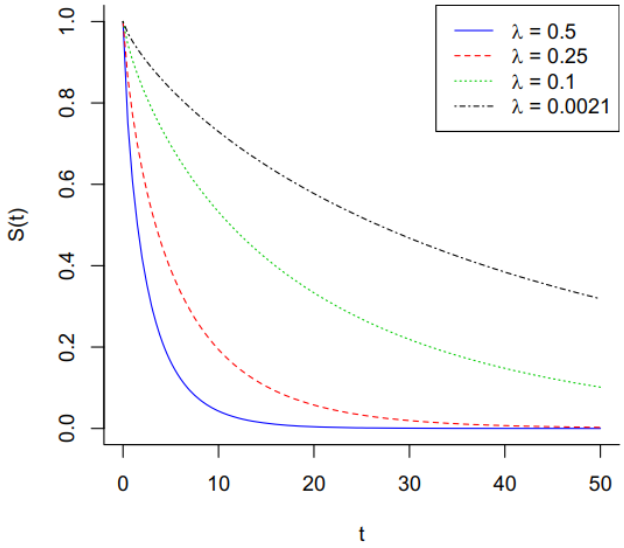


Figure 2.8: Survival curves under the Weibull distribution for $\gamma = 0.8$ and different values of λ.

Figure 2.7 shows the survival curves under the Weibull distribution for various values of the scale parameter with the shape parameter γ fixed to 7. Moreover, Figure 2.8 exhibits the survival curves under the Weibull distribution for different values of the scale parameter with the shape parameter γ fixed to 0.8. The two figures show the influence of the value of γ on the shape of survival curves.

The shape of the hazard rate function is very flexible and includes curves of increasing hazard rate over time as well as curves of decreasing hazard rate over time. If the shape parameter γ is greater than 1, then the Weibull distribution has an increasing failure rate over time. Instead, when the shape parameter γ is less than 1, it has a decreasing failure rate over time.

In Figures 2.9 and 2.10, we show the cases of $\gamma > 1$ and $\gamma < 1$ respectively.

The probability density function under the Weibull distribution is given by

$$f(t) = \gamma\lambda t^{\gamma-1} \exp(-\lambda t^{\gamma}), \quad t > 0 \tag{2.18}$$

that can be easily obtained via (2.2) and (2.16).

Finally, for the Weibull regression model, the hazard function is

$$h(t, \mathbf{x}) = \lambda\gamma(t^{\gamma-1}) \exp(\mathbf{x}^T\boldsymbol{\beta}), \quad t > 0 \tag{2.19}$$

given $\mathbf{X} = \mathbf{x}$. This is the generalization of the Exponential regression model, and when $\gamma = 1$ the Exponential regression model is obtained.

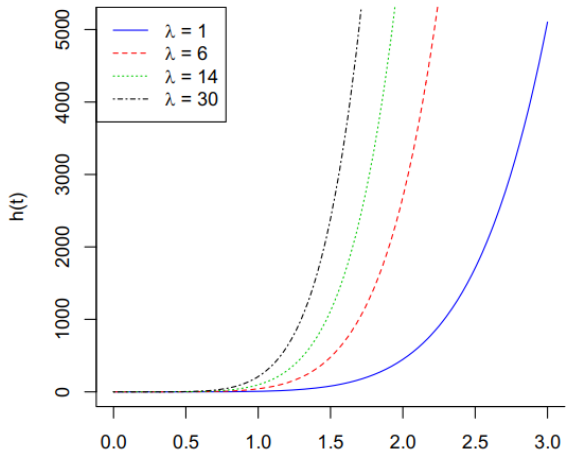


Figure 2.9: Hazard rate curves under the Weibull distribution for $\gamma = 7$ and different values of λ.

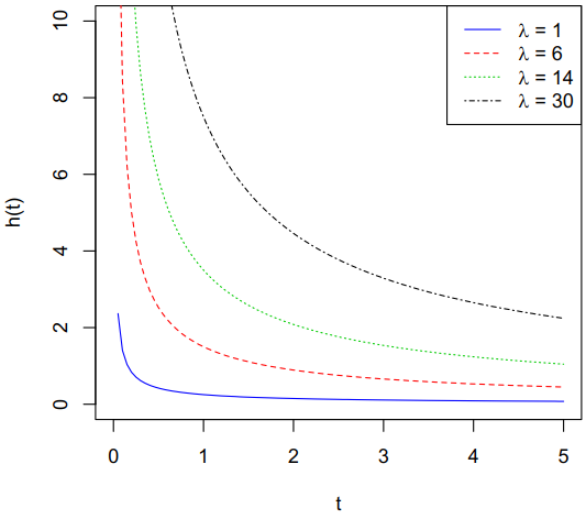


Figure 2.10: Hazard rate curves under the Weibull distribution for $\gamma = 0.8$ and different values of λ.

2.4.3 Log-normal distribution

If $log(T) = \mu + \sigma W$ where W follows a standard Normal distribution with probability density $\phi(w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right)$, then the random variable T follows the Log-normal distribution. The distributions of the survival time in certain diseases are well approximated by a Log-normal distribution as reported in the literature [80, 106, 179]. Under this distribution, the density function of T is

$$f(t) = \frac{1}{\sqrt{2\pi}}(\sigma t)^{-1} \exp\left(-\frac{(logt - \mu)^2}{2\sigma^2}\right), \quad t > 0, \tag{2.20}$$

where $\sigma > 0$ an μ are parameters.

The survival function is given by

$$S(t) = 1 - \Phi\left(\frac{log(t-\mu)}{\sigma}\right), \quad t > 0, \tag{2.21}$$

where $\Phi(w)$ is the normal cumulative distribution function $\Phi(w) = \int_{-\infty}^{w} \phi(s)ds$.

Figure 2.11 shows that the survival curves under the Log-normal distribution for various values of the parameter σ and $\mu = 0$.

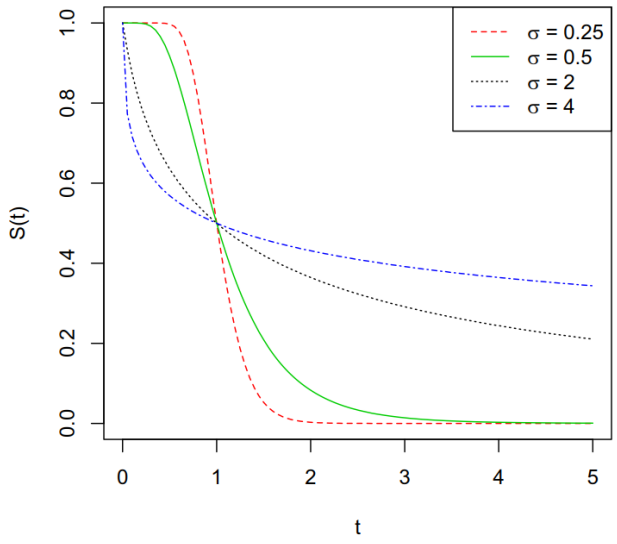


Figure 2.11: Survival curves under the Log-normal distribution for $\mu = 0$ and different values of σ.

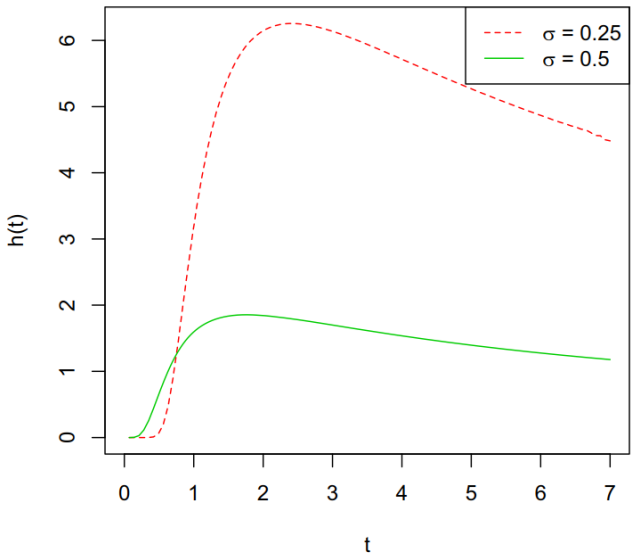


Figure 2.12: Hazard rate curves under log-normal distribution for $\mu = 0$ and different values of σ .

The hazard rate function can be derived based on the relationship between $f(t)$ and $S(t)$. Figure 2.12 shows the hazard rate function under Log-normal models for various values of the parameter σ and μ fixed to 0. The curves of the hazard rate function seem to be hump-shaped in those cases: the curve starts to increase monotonically to reach the top point, then decreases as time t further increases.

2.4.4 Log-logistic distribution

As for the Log-logistic distribution, we suppose that $log(T) = \mu + \sigma W$ but now W follows the Logistic distribution with symmetric density

$$\frac{\exp(w)}{(1+\exp(w))^2}.$$

The survival function under Log-logistic distribution has a relatively simple form:

$$S(t) = \frac{1}{1+\lambda t^{\alpha}}, \quad t > 0, \tag{2.22}$$

while the hazard rate function is given by

$$h(t) = \frac{\alpha\lambda t^{\alpha-1}}{1+\lambda t^{\alpha}}, \quad t > 0, \tag{2.23}$$

where $\alpha = 1/\sigma > 0$ and $\lambda = \exp(-\mu/\sigma) > 0$.

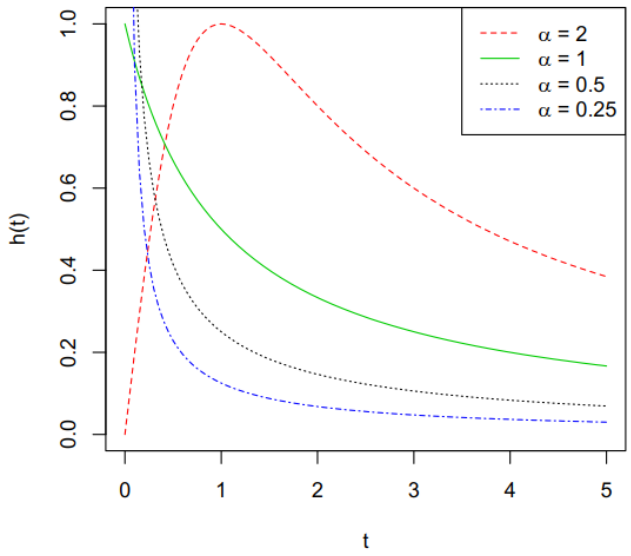


Figure 2.13: Hazard rate curves under the Log-logistic distribution for $\lambda = 1$ and different values of α.

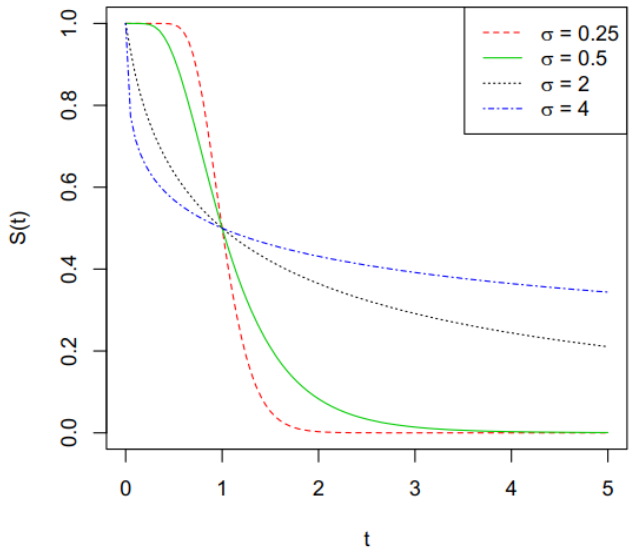


Figure 2.14: Survival curves under the Log-logistic distribution for $\lambda = 1$ and different values of α.

The Figures 2.13 and 2.14 show the hazard rate function and the survival function

under Log-logistic distribution for different values of the shape parameter.

2.5 Semi-parametric models

The proportional hazards (PH) model is the most popular and widely utilized semi-parametric model in Survival Analysis. It was firstly proposed by Cox in 1972 [55]. Therefore, the proportional hazards model is also referred as the Cox model.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$ denote the m-dimensional vector of covariates and let $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)^T$ be the regression coefficients. The hazard rate function in the PH model is

$$h(t, \mathbf{x}) = h_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta}), \quad t > 0, \tag{2.24}$$

given covariates $\mathbf{X} = \mathbf{x}$. Here $h_0(t)$ is called the baseline hazard function which is arbitrary and unspecified. The conditional cumulative function in PH model is

$$H(t, \mathbf{x}) = H_0(t) \exp(\mathbf{x}^T \boldsymbol{\beta}), \quad t > 0, \tag{2.25}$$

where $H_0(t) = \int_0^t h_0(\tau) d\tau$ is the baseline cumulative hazard function.

The survival function of T, given covariates $\mathbf{X} = \mathbf{x}$, has the following expression based on (2.7)

$$\begin{aligned} S(t, \mathbf{x}) &= \exp\left(H(t, \mathbf{x})\right) \\ &= (S_0)^{\exp(\mathbf{x}^T \boldsymbol{\beta})}, \quad t > 0, \end{aligned} \tag{2.26}$$

where $S_0(t) = \exp\left(-\int_0^t h_0(\tau) d\tau\right)$ is the baseline survival function.

The ratio of hazard functions for two individuals with different values of covariates can be determined by Formula (2.24). We assume that the value of covariate $\mathbf{X}$ for one individual is $\mathbf{x_1} = (x_{11}, x_{21}, \ldots, x_{m1})$, and for the other is $\mathbf{x_2} = (x_{12}, x_{22}, \ldots, x_{m2})$. The ratio of hazard rate functions is

$$\begin{aligned} \frac{h(t, \mathbf{x_1})}{h(t, \mathbf{x_2})} &= \frac{h_0(t) \exp({\mathbf{x_1}}^T \boldsymbol{\beta})}{h_0(t) \exp({\mathbf{x_2}}^T \boldsymbol{\beta})} \\ &= \exp((\mathbf{x_1} - \mathbf{x_2})^T \boldsymbol{\beta}) \\ &= \exp\left(\sum_{k=1}^{m} \beta_k (x_{k1} - x_{k2})\right). \end{aligned} \tag{2.27}$$

Suppose that covariates do not involve time t, i.e., $\mathbf{X}$ is a time-independent variable. Thus, the ratio of hazard functions is constant, i.e., the survival hazards for one subject is proportional to the survival hazard for the other subject over time.

As an example of the use of ratio of hazard function, consider a study where the patients receive a certain treatment or a placebo. Assume that the covariate $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$ is univariate that is m=1. In this case, $\mathbf{X}$ is a binary variable denoting the treatment status (placebo=1, treatment=0). Consider two patients where Patient 1 receives a treatment and Patient 2 receives a placebo, the ratio of hazard rate functions is

$$\frac{h(t, \mathbf{X} = \mathbf{1})}{h(t, \mathbf{X} = \mathbf{0})} = \exp(\beta_1).$$

Hence the hazard for the patient with treatment is proportional to the patient with placebo.

As mentioned in Chapter 1, one of the most important tasks in Survival Analysis is to assess the effect of covariates on survival time. Parametric regression models, such as the Exponential regression model or the Weibull regression model, can show a complete functional form where the values of the unknown parameters are derived by appropriate estimation methods. The Cox PH model is also a regression model that can predict the individual hazard and specify the relationship between survival time and covariates. However, the Cox PH model is a semi-parametric model due to the unspecified properties of the baseline function, and it is more flexible and commonly utilized than parametric regression models.

It is worth noting that if the covariates $\mathbf{X}$ are time-dependent, then the assumption of proportional hazards is not satisfied. For example, in a study about effects of surgery on cancer patients, the patients are randomly divided into two groups: the ones in group A undergo surgery, the ones in group B receive radiation therapy without surgery. Assume that surgery or not is the only variable of interest, and introduce a binary variable $\mathbf{X}$ as a covariate where $\mathbf{X} = 1$ denotes surgery and $\mathbf{X} = 0$ denotes radiation therapy without surgery. It is obvious that the operation to remove the tumor is complicated and brings a great risk to patients. But once the patients recover from the critical period following the operation, the patients' survival time will be longer and the risk will be lower than the ones without surgery.

Thus, the curve of hazard rate function for the individuals undergoing surgical operation is above the one for the individual without undergoing surgery at beginning. And the curve of hazard rate function for surgery group decreases over time and gradually goes below the one for the individuals without surgery. The two curves cross after a certain time, which means the ratio of hazard rate is greater than one at beginning and then less than one later. In this case the ratio of hazard rate is not constant but rather varying over time. Thus the Cox model is inappropriate to describe this case. An extension of the Cox PH model for time dependent variables have been proposed to deal with these situations [129].

For right censored data, Cox PH model utilizes the partial likelihood to estimate the unknown parameter, and the partial likelihood estimate is a less complex inference approach.

Consider the observation data $(O_i,\delta_i,\mathbf{x}_i)$, $i = 1,\ldots,n$, where $\mathbf{x}_i$ is the value of covariate $\mathbf{X}_i$ for the ith subject and (O_i,δ_i) are defined as in Subsection 2.3.2. For simplicity, assume no ties in the observation data. Let $\{s_j\}_{j=1}^{l}$ denote the distinct ordered elements of the values of the survival time $T_i; i = 1,\ldots,n$, similarly to what was done in Subsection 2.3.2. Let $R\{s_j\}$ be the risk set at time s_j, i.e., the set of all individuals who survive over s_j or just experience the event at s_j. Clearly, the individuals who do not experience the event of interest prior to s_j are in $R\{s_j\}$. The partial likelihood is expressed as follows for the discrete time:

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{l} \frac{\exp(\mathbf{x}_{\{s_j\}}^T \boldsymbol{\beta})}{\sum_{k\in R\{s_j\}} \exp(\mathbf{x}_k^T \boldsymbol{\beta})}, \quad t > 0, \tag{2.28}$$

where $\mathbf{x}_{\{s_j\}}$ denotes the values of covariates for the individuals who experience the event at s_j, and $\mathbf{x}_k$ denotes the values of covariates for the individuals who belong to $R\{s_j\}$ if $k \in R\{s_j\}$. The parameters $\boldsymbol{\beta}$ can be estimatied by maximizing the logarithm of $L(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. For the sake of convenience, let

$$L_j = \frac{\exp(\mathbf{x}_{\{s_j\}}^T \boldsymbol{\beta})}{\sum_{k\in R\{s_j\}} \exp(\mathbf{x}_k^T \boldsymbol{\beta})}, \quad j = 1,2,\ldots,l.$$

Next we will apply (2.28) to the data shown in Table 2.4 which introduces the information of 8 patients under 6-MP treatment and satisfies the assumption of no ties in the observation data. The background of data has been explained in Subsection 2.3.2, but now the new information on covariates are added. Assume that the covariates $\mathbf{X}$ are univariate, and let $\mathbf{X}$ denote the remission status of patients, where $\mathbf{X} = 1$ means that the status is complete remission and $\mathbf{X} = 2$ means that the status is partial remission.

Table 2.4: Survival data for 8 patients under 6-MP treatment.

Patient i	Observed data $(O_i, \delta_i, \mathbf{x_i})$	Patient i	Observed data $(O_i, \delta_i, \mathbf{x_i})$
1	(6,1,2)	5	(10,0,1)
2	(6,0,1)	6	(10,1,2)
3	(7,1,1)	7	(11,0,1)
4	(9,0,1)	8	(13,1,1)

Assume that no ties in the observation data. $\delta_i = 0$ indicates that the survival data for Patient i is censored, while the time of event of interest for Patient i is observed exactly when $\delta_i = 1$. $\mathbf{X}_i = 1$ indicates that the status is complete remission for Patient i, while the status is partial remission for Patient i when $\mathbf{X}_i = 2$.

As shown in Table 2.4, $n = 8$, and the distinct ordered elements of the values of the

survival time $\{s_j\}_{j=1}^{l}$ are just $\{6,\ 7,\ 10,\ 13\}$, and $l = 4$. The partial likelihood is the product of L_j, $j = 1, \ldots, l$. We explain the calculation of L_1 through an example.

Patient 1 experienced the event of interest at survival time $s_1 = 6$, so $\mathbf{x}_{\{s_1\}}$ is the covariate for Patient 1, and thus the numerator of L_1 is

$$\exp(\mathbf{x}_{\{s_1\}}^T \boldsymbol{\beta}) = \exp(\mathbf{x}_1^T \boldsymbol{\beta}) = \exp(2\beta_1).$$

Since Patient 2 is censored at time $s_1 = 6$ and Patient 1 just experience the event at time $s_1 = 6$, then all the patients belong to the risk set $R\{s_1\}$, and

$$\sum_{k \in R\{s_1\}} \exp(\mathbf{x}_k^T \boldsymbol{\beta}) = \sum_{k=1}^{8} \exp(\mathbf{x}_k^T \boldsymbol{\beta}) = 2\exp(2\beta_1) + 6\exp(\beta_1).$$

Therefore,

$$L_1 = \frac{2\exp(2\beta_1)}{2\exp(2\beta_1) + 6\exp(\beta_1)}.$$

As mentioned earlier, the partial likelihood is the product of the likelihood $\{L_j\}_{j=1}^{l}$ which represents the probability of experience the event at s_j. The formula only focuses on the probabilities for the individuals who experience the event of interest and does not consider the probability for the individuals who are censored. The construction of partial likelihood is more complicated when ties are present. We refer the interested reader to [28, 55, 68] for this case.

For interval censored data, a new methodology was developed in [81] for fitting proportional hazards model. Suppose that interval-censored data are available from n independent subjects and $\mathbf{X_i}$ is the m-dimensional vector of covariates for subject i. Let $(L_i, R_i]$ be the censored interval to which response variable T_i belongs, $i = 1, \ldots, n$. Let $\{s_j\}_{j=0}^{l}$ be defined as in Subsection 2.3.2, where $s_0 = 0$ and $s_l = \infty$. Then, the likelihood function is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \sum_{j=1}^{l} \alpha_{ij} (S(s_{j-1})^{\exp(\mathbf{x_i^T}\boldsymbol{\beta})} - S(s_j)^{\exp(\mathbf{x_i^T}\boldsymbol{\beta})}) \tag{2.29}$$

given covariate $\mathbf{X_i} = \mathbf{x_i}$, where $\alpha_{ij} = I((s_{j-1}, s_j] \subseteq (L_i, R_i])$, $i = 1, \ldots, n$, $j = 1, \ldots, l$. A Newton-Raphson iteration scheme can be applied to maximize the likelihood estimation for computing the value of the parameter $\boldsymbol{\beta}$. In [181] an extended ICM method was applied for maximum likelihood estimation replacing the Newton-Raphson method achieving better computational power. The method was revised and implemented by R package "icenReg" [8], which will be utilized in Subsection 4.1.4 to compare Cox PH model to the survival trees for the interval-censored data.

Besides Cox PH model, there are other semi-parametric models, such as the proportional odds model, the additive hazard model, and so on [41, 152, 153]. Since they are not utilized in this thesis, so we do not discuss them.

2.6 Imputation approaches

Imputation, first proposed by Rubin for sample surveys [197], is also applied to handle missing data. Over time, the application of imputation techniques has expanded to other areas of research, including Survival Analysis. In Survival Analysis, imputation techniques can be used to handle the interval-censored data by imputing survival time based on the interval information. In general, imputation approaches can be employed to transform interval-censored data to right censored data, allowing for the utilization of various techniques based on right censored data in Survival Analysis. Following we discuss the single imputation method and the multiple imputation method.

2.6.1 Single imputation

Single imputation refers to the assignment of a single value to a missing or censored data point. Let $(L_i, R_i]$ be the censored interval within which survival time T_i occurs, $i = 1, \ldots, n$. It is assumed that the survival time for each observation, denoted as T_i, is a value within the interval $(L_i, R_i]$. Specifically, we can select the mid point of interval $(L_i, R_i]$ as the value for T_i which is referred as middle imputation, or the left point L_i can be assigned to T_i which is known as left imputation. If T_i is set to be equal to R_i, then the method is called right imputation. What's more, we can also assign to T_i a value which is randomly selected in the interval $(L_i, R_i]$.

Consider middle imputation as an example. For subject i, the original observed data is $(L_i, R_i]$. If $R_i = \infty$, then the observed data is just right censored, and we keep it. Otherwise, let $T_i = \frac{L_i+R_i}{2}$, and the data $(L_i, R_i]$ is replaced by $(O_i = \frac{L_i+R_i}{2}, \delta_i = 1)$. Then we have an alternative right censored data for original data. As mentioned in the example, $R_i = \infty$ is a special case and it is just a right censored data. In this case, there is no need to assign a value to T_i, we only keep it.

It is worth noting that if the width of finite intervals are enough narrow, then the outcomes of three approaches (left imputation, right imputation, and middle imputation) only differs slightly.

Single imputation is easy to implement, but the inferences from it is not reliable enough. If the intervals are wide or there is significant overlapping among them, single imputation may not provide an accurate approximation and could lead to biased results. So this imputation scheme has been extended to multiple imputation, which can overcome the drawbacks of single imputation.

2.6.2 Multiple imputation

In general, multiple imputation assigns M values ($M > 1$), rather than a single one to a missing or censored value. Correspondingly, M complete datasets are created based on the M imputed values.

Multiple imputation can also utilized for interval-censored data [45, 217, 218, 240]. A procedure for multiple imputation referred as data augmentation algorithm is shown in Algorithm 1. It is a two-step process that involves filling in censored data points multiple times, creating multiple right censored datasets, and then analyzing each dataset separately. The results from the multiple analyses are then combined to produce final estimate of the survival function or other valid statistical inferences.

Algorithm 1 The data augmentation algorithm [197]

Given

- observation intervals $(L_i, R_i]$ for subject i, $i = 1, \ldots, n$.
- $\mathbf{X_i}$, an m-dimensional vector of covariates for subject i, $i = 1, \ldots, n$.
- right censored data $\{O_i^{(0)}, \delta_i^{(0)}, \mathbf{x_i}\}_{i=1}^n$ constructed by single imputation based on observation interval-censored data $\{(L_i, R_i]\}_{i=1}^n$, where $\mathbf{x_i}$ is the value of $\mathbf{X_i}$.
- an initial Kaplan-Meier estimate of survival function $\hat{\mathbf{S}}^{(0)}$ which is obtained based on right censored data $\{O_i^{(0)}, \delta_i^{(0)}, \mathbf{x_i}\}_{i=1}^n$.
- an integer $M \geq 2$.
- a threshold value α.

Procedure

Let $l = 0$;

Repeat steps 1, 2, and 3.

1. Increase l by 1.
2. M right censored datasets $\{O_i^{(k,l)}, \delta_i^{(k,l)}; \mathbf{x_i}; i = 1, \ldots, n\}_{k=1}^M$ are given, and M Kaplan-Meier estimate $\hat{\mathbf{S}}^{(k,l)}(t)$ are calculated based on the right censored datasets, $k = 1, \ldots, M$.
   a) For $k = 1, \ldots, M$, construct a right censored dataset $\{O_i^{(k,l)}, \delta_i^{(k,l)}; \mathbf{x_i}; i = 1, \ldots, n\}$.
      - If $R_i = \infty$, then let $O_i^{(k,l)} = L_i$, and $\delta_i^{(k,l)} = 0$, i.e., when the observed data is just right censored, just keep it.
      - If $R_i < \infty$, then let $O_i^{(k,l)}$ be a random number generated from $\hat{\mathbf{S}}^{(l-1)}(t)$ satisfying the condition $O_i^{(k,l)} \in (L_i, R_i]$, and $\delta_i^{(k,l)} = 1$.

b) For $k = 1, \ldots, M$, Kaplan-Meier estimate $\hat{\mathbf{S}}^{(k,l)}(t)$ is given based on the right censored dataset $\{O_i^{(k,l)}, \delta_i^{(k,l)}; \mathbf{x_i}; i = 1, \ldots, n\}$.

3. $\hat{\mathbf{S}}^{(l)}(t) = \frac{1}{M} \sum_{k=1}^{M} \hat{\mathbf{S}}^{(k,l)}(t)$.

until $\|\hat{\mathbf{S}}^{(l)} - \hat{\mathbf{S}}^{(l-1)}\| < \alpha$.

Output

Let $l^* = l$. The estimator of survival function $\hat{\mathbf{S}}^*(t) = \frac{1}{M} \sum_{k=1}^{M} \hat{\mathbf{S}}^{(k,l^*)}(t)$ and right censored data $\{O_i^*, \delta_i^*, \mathbf{x_i}\}_{i=1}^n$.

Imputation approaches can transform interval-censored data to right censored data and provide the estimate of survival function. Therefore, it will be utilized to construct survival trees for interval-censored data in Chapter 4.

Chapter 3

Conditional inference trees and ensemble methods

Tree methods and ensemble methods are very popular Machine Learning methods and have been extensively applied in different fields. In this chapter, conditional inference trees and ensemble methods are discussed. It is well known that most of the tree methods suffer from splitting variable bias. Conditional inference trees overcome this drawback and, hence, have the advantage of unbiasedness. Another weakness of tree methods is high variance. Ensemble methods based on trees can reduce the variance and have been demonstrated to improve the prediction performance. Random Forest (RF) is a classic and well-known ensemble method for classification and regression. Double Random Forest has remarkable predictive performance for classification and improves RF for classification task. What's more, quantile regression forests, a modification of Random Forest that utilizes a novel weights scheme, can achieve outstanding performance for regression task. Conditional inference forest as an ensemble method borrows the basic frame from Random Forest, but chooses conditional inference trees rather than CART as base learners. These tree algorithms and ensemble methods will be utilized to construct novel survival trees in Chapter 4 and to solve various practical problems in Chapter 5.

3.1 Conditional inference trees

Conditional inference trees are also referred as "ctree" [107]. They are commonly used tree methods and are implemented within the framework of conditional inference. They can handle continuous, censored, ordered, and multivariate response variables. The advantage of them is that they implement unbiased variable selection.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$ denote an m-dimensional covariate vector, and let Y denote the response variable. Let $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \ldots \times \mathcal{X}_m$ be the sample space of $\mathbf{X}$. Let $\mathcal{Y}$ be the sample space of response variable Y, and let $\mathcal{L}_n = \mathcal{Y} \times \mathcal{X}$ be the sample space of n iid observations, specifically, $\mathcal{L}_n = \{(Y_i, X_{1i}, X_{2i}, \ldots, X_{mi})^T; i = 1, \ldots, n\}$.

The construction of conditional inference trees is a two-step procedure. The variable selection and splitting points selection are separate, which guarantees the unbiasedness of the variable selection.

The two steps are :

- Step 1: variable selection. The association between response Y and covariates $\mathbf{X}$ is measured based on p-values according to a pre-specified significant level. Then the component X_{j_0} of $\mathbf{X}$ with strongest association with response Y is selected as splitting variable.

- Step 2: splitting points selection. The optimal splitting point for the selected variable X_{j_0} is searched using a two-sample test.

In Step 1, under the global null hypothesis of independence between the components X_j of the covariates $\mathbf{X}$ and the response Y, a test statistic $T_j(\mathcal{L}_n, \omega)$ is constructed, $j = 1, \ldots, m$:

$$T_j(\mathcal{L}_n, \omega) = \text{vec}\left(\sum_{i=1}^{n} w_i g_j(X_{ji}) h(Y_i, (Y_1, \ldots, Y_n))^T\right) \in \mathbb{R}^{b_j q}, \tag{3.1}$$

where

- $g_j : \mathcal{X}_j \to \mathbb{R}^{b_j}$ is a transformation of the covariate X_j, $j = 1, \ldots, m$;

- $h : \mathcal{Y} \times \mathcal{Y}^n \to \mathbb{R}^q$, called the influence function, depends on all the observation values of the response variable in a permutation symmetric way;

- vec operator converts a $b_j \times q$ matrix into a $b_j q$ column vector by column-wise combination.

Here $\omega = (w_1, \ldots, w_n)^T$ is a weight vector related to the importance of the observation for a given node. More specifically, w_i is 1 when the ith observation falls into a given node, otherwise is 0.

Then, an extended permutation test is applied to calculate p-values based on the test statistics $T_j(\mathcal{L}_n, \omega)$. The computed value is denoted by $p_{(j)}$. Finally, X_{j_0} is selected as splitting variable if $p_{(j_0)}$ is the smallest value among $p_{(j)}$, $j = 1, \ldots, m$.

The algorithm will stop if the global null hypothesis of independence between response variable and covariates can not be rejected according to the prespecified significant level used in step 1.

In step 2, a two-sample test is executed based on the selected splitting variable X_{j_0} and the test statistic

$$T_{j_0}^{\mathcal{A}}(\mathcal{L}_n, \omega) = \text{vec}\left(\sum_{i=1}^{n} w_i I(X_{j_0 i} \in \mathcal{A}) h(Y_i, (Y_1, \ldots, Y_n)^T\right) \in \mathbb{R}^q. \tag{3.2}$$

Here $\mathcal{A}$ is an arbitrary subset taken from the sample space $\mathcal{X}_{j_0}$ and $I(\cdot)$ is the indicator function.

The test statistic $T_{j_0}^{\mathcal{A}}(\mathcal{L}_n, \omega)$ is a special case of the statistic given in (3.1) and is used to measure the discrepancy between the samples

$$\{(Y_i, X_{1i}, X_{2i}, \ldots, X_{mi})^T \in \mathcal{L}_n | X_{j_0 i} \in \mathcal{A}; i = 1, \ldots, n\}$$

and

$$\{(Y_i, X_{1i}, X_{2i}, \ldots, X_{mi})^T \in \mathcal{L}_n | X_{j_0 i} \notin \mathcal{A}; i = 1, \ldots, n\}$$

in a given node.

The optimal splitting $\mathcal{A}^0$ is found by exhaustive search for the subset maximizing the test statistic of two-sample test over all possible subsets $\mathcal{A}$ of the sample space $\mathcal{X}_{j_0}$.

The specific choices of g_j and the influence function h in ctree depend on the requirements and assumptions of the extended permutation test, two-sample test, and the type of data that are analyzed. Additionally, ctree can be efficiently applied to both full data and right-censored data with appropriate choices of g_j and the influence function h.

The framework of conditional inference trees will be utilized later for the construction of our novel interval-censored tree. Possible choices for g_j and h will be discussed in detail in Subsection 4.1.

3.2 Random Forest

Random Forest (RF) is a parallel ensemble method for classification and regression. CART [26] is commonly used for base learners in RF.

The core idea of RF is the application of bootstrap. As improvement of bagging trees, decision trees in RF are built on bootstrapped training data and a subset of the m predictor variables which can influence the selection of the optimal binary splitting variable and splitting value.

It is important that only a subset of predictor variables is considered in building the decision trees in RF. If all the predictor variables are considered, then almost all the trees

will select strong variables for splitting, which leads to an extreme similarity for all the decision trees in RF. Consequently, the high variance with respect to single tree fails to be resolved.

Algorithm 2 presents in detail RF for classification.

Algorithm 2 Random Forest for classification [25]

Given

- A: training set of size N with m predictor variables (i.e., an m dimensional covariates).
- C: set of class label for the response variable.
- K: size of C, i.e., the number of class label.
- q: size of the bootstrapped subset of predictor variables randomly selected from the m predictor variables.
- B: number of classifiers (decision trees).

Procedure

For b=1 to B

1. Choose bootstrapped data from the training data A and randomly select q variables out of the total m predictor variables.
2. Construct a tree classifier T_b as follows:
   i) Initially T_b consists of a single root node that is also an open node (that is, it needs to be further explored) and the instances in the root node are the bootstrapped data A_b.
   ii) Repeat the following steps until all nodes are closed:
   - Choose an open node.
   - Verify if this node can be closed. The closure condition is the stopping rule proposed in CART [26]. If this node can be closed, then it can not be split further and it is a terminal node; otherwise, split it into two children nodes based on the following steps:
     – pick the best variable and cut-point by utilizing the q predictors and bootstrapped training data A_b based on the splitting rule proposed in CART for classification;

– split the instances into two subsets and assign the two subsets to children nodes which represent sub-regions of the data. These new created nodes will be added to the list of open nodes.

Output

For a given new instance Z, produce a prediction outcome $G(Z)$ by aggregating the B classifiers utilizing a majority vote. Let $T_b(Z)$ be the class prediction value of the bth classifier. Then

$$G(Z) = \arg\max_{c_j \in C} \sum_{b=1}^{B} I(T_b(Z) = c_j)$$

The whole procedure of RF for regression is similar to Algorithm 2, the only difference is that the prediction outcome $G(Z)$ is obtained by averaging the prediction from the B individual regression trees. More specifically, let $T_b(Z)$ be the regression predictions of the bth regression tree, then

$$G(Z) = \sum_{b=1}^{B} \frac{1}{B} T_b(Z).$$

3.3 Double Random Forest

It is known that bigger trees have better predictive performance for classification. However, Han et al. [101] found that "the largest tree created with the minimum node size parameter may not be sufficiently large for the best performance of RF", i.e., the size of trees in RF was unsatisfactory for prediction. Hence, Double Random Forest (DRF) was inspired by the idea of generating bigger trees compared to those in RF.

Similar to RF, Double Random Forest (DRF) is a parallel ensemble method for classification, where CART is used for the base classifiers. As mentioned before, a classifier in RF is constructed based on bootstrapped data from the training data and a subset of predictor variables, which avoids high variance and instability. However, bootstrap data leads to fewer unique instances falling into the nodes of a classifier, which is a barrier to grow larger trees. DRF overcomes this difficulty. The full procedure of DRF is given in Algorithm 3.

Algorithm 3 Double Random Forest [101]

Given

- A: training set of size N with m predictor variables.
- C: set of class label for the response variable.

- K: size of C, i.e., the number of class label.
- B: number of classifiers in DRF.
- q: size of the bootstrapped subset of predictor variables randomly selected from all m predictor variables.

Procedure

For b=1 to B,

1. Choose bootstrapped data A_b from the training data A and randomly select q variables out of the total m predictor variable.
2. Construct a tree classifier T_b using the entire training set A as follows:
   i) Initially T_b consists of a single root node that is also a open node (that is, it needs to be further explored) and the instances in the root node are the entire training set A.
   ii) Repeat the following steps until all nodes are closed:
   - Choose an open node. Let s be the chosen node.
   - Verify if this node can be closed. The closure condition is the stopping rule proposed in CART [26]. If this node can be closed, then it can not be split further and it is a terminal node; otherwise, split it into two children nodes based on the following steps:
     – let $A_b(s)$ denote all instances in the open node s, and let N_s be the number of instances in $A_b(s)$. If $N_s > 0.1 \times N$, draw a bootstrap sample $A_b^*(s)$ from $A_b(s)$, otherwise $A_b^*(s) = A_b(s)$.
     – pick the best splitting variable and cut-points based on the q predictors and $A_b^*(s)$ based on the splitting rule proposed in CART for classification.
     – split the instances in the open node s into two subsets and assign the two subsets to children nodes which represent sub-regions of the data. These new created nodes will be added to the list of open nodes.

Output

For a given new instance Z, produce a prediction outcome $G(Z)$ by aggregating the B classifiers using a majority vote. Let $T_b(Z)$ be the class prediction value of the bth classifier. Then

$$G(Z) = \arg\max_{c_j \in C} \sum_{b=1}^{B} I(T_b(Z) = c_j)$$

The main differences between RF and DRF are outlined below by contrast.

- For DRF:

1. all individual classifiers utilize the same data A;
2. DRF searches for the best splitting variable and splitting point based on a random bootstrapped sample and a randomly selected subset of predictors variables at each node of the classifiers. However, once the best splitting variable and splitting point are determined, the original data in the father node are utilized to generate the children nodes.

- For RF:

1. each classifier is constructed by using different bootstrapped data randomly sampled from the training data;
2. the splitting rules are applied to the bootstrapped sample rather than to the entire training data.

As explained before, DRF tends to construct larger trees compared to RF. In fact, in DRF, the entire training data instead of a bootstrapped sample is used to construct the trees.

3.4 Quantile Regression Forests

Let $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}^T$ denote an m-dimensional covariate (that is, m predictor variables). Let Y denote the response variable. The quantity

$$F(y|\mathbf{X} = \mathbf{x}) = P(Y \leq y|\mathbf{X} = \mathbf{x}) \tag{3.3}$$

is the conditional distribution function of Y given covariates $\mathbf{X} = \mathbf{x}$. The τth conditional quantile function is defined as

$$Q_\tau(\mathbf{x}) = \inf\{y : F(y|\mathbf{X} = \mathbf{x}) \geq \tau\}. \tag{3.4}$$

Regression analysis usually provides an estimate of the conditional mean $E(Y|\mathbf{X} = \mathbf{x})$ given $\mathbf{X} = \mathbf{x}$. However, conditional mean only shows one feature of the conditional distribution and other important characteristics are ignored. In contrast, conditional quantiles can provide additional information on a conditional distribution. Thus quantile regression is a useful tool for data analysis. A comprehensive survey on the development of quantile regression is given in [131]. The estimate of the conditional quantile is commonly obtained by minimizing the expected loss function defined by a weighted absolute deviation.

Quantile Regression Forests are a novel way to estimate the conditional quantile based on Random Forest. First, the weights $\{\omega_i(\mathbf{x})\}_{i=1}^n$ given $\mathbf{X} = \mathbf{x}$ for n observations are

obtained by utilizing the prediction of the trees from Random Forest. Then the estimate of the conditional distribution function of the random variable Y is given based on

$$F(y|\mathbf{X}=\mathbf{x}) = P(Y \leq y|\mathbf{X}=\mathbf{x}) = E(I(Y \leq y)|\mathbf{X}=\mathbf{x}).$$

To estimate the conditional distribution function at a particular quantile, the Quantile Regression Forests algorithm calculates the weighted mean of the indicator function for that quantile across the ensemble of trees.

Quantile Regression Forests are explained step by step and calculation of weights is shown in Algorithm 4.

Algorithm 4 Quantile Regression Forests[166]

Given

- $A = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$: training set with size N, where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{mi})^T$ is the observation value of an m dimensional covariate vector for subject i (that is $x_{1i}, x_{2i}, \ldots, x_{mi}$ are the observation values of m predictor variables for subject i), y_i is the observation value of response variable for subject i;
- m: the dimension of a covariate vector or the size of predictor variables;
- B: number of base learners in Quantile Regression Forests;
- τ is a particular quantile value, $0 < \tau < 1$.

Procedure

For b=1 to B

1. Construct a regression tree T_b using a randomly selected subset of predictor variables and bootstrapped data from the training set A similar to RF. The best splitting variable and splitting point can be obtained as for the CART algorithm [26].
2. For each tree T_b and given $\mathbf{X} = \mathbf{x}$, let $\mathcal{R}_b^{l(\mathbf{x})}$ be the terminal node to which $\mathbf{x}$ is assigned, and for each observation $(\mathbf{x}_i, y_i) \in A$, let

$$\omega_i(\mathbf{x}, b) = \frac{I(\mathbf{x}_i \in \mathcal{R}_b^{l(\mathbf{x})})}{\sharp\{j \in \{1, \ldots, N\} | \mathbf{x}_j \in \mathcal{R}_b^{l(\mathbf{x})}\}}$$

   be the corresponding weight, $i = 1, \ldots, N$.
3. Given $\mathbf{X} = \mathbf{x}$, compute the average observation weights

$$\omega_i(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \omega_i(\mathbf{x}, b), \quad i = 1, \ldots, N.$$

Output

The estimation of conditional distribution function is

$$\hat{F}(y|\mathbf{X}=\mathbf{x}) = \sum_{i=1}^{N} \omega_i(\mathbf{x}) I(y_i \leq y)$$

for all $y \in \mathbb{R}$ and the estimation of the τth conditional quantile function is

$$\hat{Q}_\tau(\mathbf{x}) = \inf(y : \hat{F}(y|\mathbf{X}=\mathbf{x}) \geq \tau)$$

for each τ.

Quantile Regression Forests provide an estimate of the distribution function. Then an estimate of conditional quantiles can be calculated from the distribution function. The ways that trees grow in Quantile Regression Forests are the same as in Random Forest, and the key point is the use of the average observation weights $\omega_i(\mathbf{x})$ given $\mathbf{X} = \mathbf{x}$ [166]. Later we will apply this approach to construct a new survival forest for interval-censored data.

3.5 Conditional Inference Forests

As always, let $\mathbf{X} = \{X_1, X_2, \ldots, X_m\}^T$ denote an m-dimensional covariate vector (that is, m predictor variables). Let Y denote the response variable. Conditional Inference Forests [109] usually referred as "cforest" is constructed by utilizing ctree as base learners and growing each ctree using bootstrap sample and a random subset of predictor variables. Conditional Inference Forests can also predict the survival time given covariate $\mathbf{X} = \mathbf{x}$ for right censored data. The prediction of the survival ensemble method is obtained by minimizing a loss function with the prediction weight proposed in [166]. Conditional Inference Forests can address the problem of complex data (right censored data) as well as unbiasness. The procedure of Conditional Inference Forests is outlined in Algorithm 5.

Algorithm 5 Conditional Inference Forests [109]

Given

- $A = \{(y_i, \delta_i, \mathbf{x}_i)\}_{i=1}^N$: observed learning sample, where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{mi})$ is the observation value of an m dimensional covariate vector (that is, $x_{1i}, x_{2i}, \ldots, x_{mi}$ are the observation values of m predictor variables).
- B: the number of ctree in Conditional Inference Forests.

Procedure

For b=1 to B

1. Construct a ctree T_b using a randomly selected subset of predictor variables and bootstrapped data from the training set A. The best splitting variable and splitting point are obtained based on the splitting criterion of ctree.
2. For each ctree T_b and $\mathbf{X} = \mathbf{x}$, let $\mathcal{R}_b^{l(\mathbf{x})}$ be the terminal node to which $\mathbf{x}$ is assigned. For each observation $((y_i, \delta_i, \mathbf{x}_i)) \in A$, let

$$a_i(\mathbf{x}) = \begin{cases} \sum_{b=1}^{B} I(\mathbf{x}_i \in \mathcal{R}_b^{l(\mathbf{x})}), & \text{if } \delta_i = 1; \\ 0, & \text{if } \delta_i = 0. \end{cases}$$

   be a prediction weight, $i = 1, \ldots, N$.

Output

Given covariate $\mathbf{X} = \mathbf{x}$, an estimate of survival time is given by

$$\hat{f}(\mathbf{x}) = \arg\min_{y \in \mathbb{R}} \sum_{i=1}^{N} a_i(\mathbf{x}) L(y_i, y)$$

where $L(y_i, y)$ is the loss function and $y = f(\mathbf{x})$. When $L(y_i, y) = (y_i - f(\mathbf{x}))^2$, then the prediction is simplified as

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{N} \frac{a_i(\mathbf{x}) y_i}{\sum_{j=1}^{N} a_j(\mathbf{x})}$$

In Algorithm 5, the prediction weight is similar to the average observation weight in [166] and "measures the similarity of $\mathbf{x}$ to $\mathbf{x}_i$". In addition, Algorithm 5 can be applied to right censored data, because the base learner used in Conditional Inference Forests is the conditional inference trees (ctree), which can be adapted to handle right-censored data by appropriately choice of g_j and influence function h.

Chapter 4

New survival trees and ensemble methods for interval-censored data

Various survival trees and ensemble methods have been proposed and implemented for right censored data. However, there is only a limited literature on survival tree methods for the interval-censored data due to the complexity of data.

Here, we propose a novel unbiased interval-censored tree that combines Generalized Log-Rank Tests (GLRT) with Conditional Inference Frame (CIF) introduced in Subsection 3.1. GLRT is a test procedure for survival comparison used for the interval-censored data, which is a generalization of Log-Rank Test, and has stronger test ability when appropriate parameters are selected [215]. We adopt several special statistics that depend on parameters based on GLRT, and new tree methods with hyper-parameter denoted by $ICS_1^{\rho,\gamma}tree$ and $ICS_2^{\rho,\gamma}tree$ are derived. We also discuss several different tree models based on other rank tests statistics. A comparison of the prediction performance for $ICtree$, $ICS_1^{\rho,\gamma}tree$, $ICS_2^{\rho,\gamma}tree$ and other interval-censored trees is reported.

The results show that $ICS_1^{\rho,\gamma}tree$ is a competitive interval-censored tree method benefitting from the properties of GLRT and tuning of hyper-parameter ρ, γ. Appropriate hyper-parameter values in $ICS_1^{\rho,\gamma}tree$ improve the performance of prediction and allow $ICS_1^{\rho,\gamma}tree$ to have more extensive adaptability to data.

In Section 4.1, we first review GLRT, and then we introduce some new interval-censored tree methods. The methods combine some new test statistics of GLRT or other rank tests with CIF. Furthermore, we explore a new survival ensemble algorithm referred as $ICS_1^{\rho,\gamma}forests$ utilizing $ICS_1^{\rho,\gamma}tree$ as base learners in Section 4.2. The aggregation scheme

is borrowed from the quantile regression forests and conditional inference forests (cforests) which are introduced in Subsection 3.4 and 3.5.

An extensive simulation is carried out in Section 4.3. Hyper-parameter tuning is implemented by a grid searching cross-validation technique, and the predictive accuracy of the new tree methods are evaluated. Finally, we compare the performance of $ICS_1^{\rho,\gamma} forests$ to $ICcforest$ in Section 4.4.

4.1 A new survival tree for interval-censored data

4.1.1 Overview of Generalized Log-Rank Tests

Consider n independent subjects from k different populations. Let n_l denote the number of subjects from population l with survival function S_l, $l = 1,2,\ldots,k$. Let T_i be the survival time at which the event of interest for subject i happens, $i = 1,2,\ldots,n$, where $\sum_{l=1}^{k} n_l = n$. In addition, let $\mathbf{w}_i$ be the k-dimensional indicator vector related to the subject i whose lth element is 1 if the subject belongs to population l, and 0 otherwise. Moreover let $(L_i, R_i]$ be the censoring interval:

$$(L_i, R_i] = \begin{cases} (0, U_i] & \text{if} \quad T_i \leq U_i \\ (U_i, V_i] & \text{if} \quad U_i < T_i \leq V_i \\ (V_i, \infty) & \text{if} \quad T_i > V_i \end{cases} \tag{4.1}$$

where U_i and V_i are non-negative random variables independent of T_i such that $U_i < V_i$, i.e., we assume non-informative interval censoring. Right-censoring occurs when $R_i = +\infty$. Without loss of generality, we will assume that $k = 2$, but the Generalized Log-Rank Tests (GLRT) are also applicable when $k > 2$. Let $\xi(x)$ be a function over the interval (0,1) that satisfies the condition

$$\lim_{x \to 0} 1 - \xi(1-x) = \lim_{x \to 1} 1 - \xi(1-x) = c_0, \tag{4.2}$$

where c_0 is a constant. Let $\widehat{S}$ be a non-parametric maximum likelihood estimator (NPMLE) of the survival function. The estimator $\widehat{S}$ can be obtained by the improved EMICM algorithm [7, 8], which is an extended form of the early EMICM algorithm [9]. They are introduced and discussed in detail in Subsection 2.3.2. The Generalized Log-Rank Tests are applied to the interval-censored data for the comparison of the survival function. The test statistic is $U_\xi = \sum_{i=1}^{n} \mathbf{w}_i U_{\xi,i}$ where

$$U_{\xi,i} = \frac{\xi(\widehat{S}(L_i)) - \xi(\widehat{S}(R_i))}{\widehat{S}(L_i) - \widehat{S}(R_i)} \tag{4.3}$$

is the score statistic. Note that condition (4.2) is a sufficient condition that guarantees U_ξ to possess an asymptotic normal distribution.

Different score statistic $U_{\xi,i}$ can be obtained by using different functions $\xi(x)$ in (4.3). For example, when

$$\xi(x) = xlogx,$$

$U_{\xi,i}$ is the logrank score statistic. In [215], it was proposed to use

$$\xi(x) = (logx)x^{\rho+1}(1-x)^\gamma$$

with parameters ρ and γ, which inspired us to construct other new functions $\xi(x)$ with parameters ρ and γ and derive a new interval-censored tree with hyper-parameter.

4.1.2 A new interval-censored tree based on Generalized Log-Rank Tests

Our idea for a new interval-censored tree stems from the more extensive test ability of GLRT when appropriate values for the parameters are selected. We combine the Generalized Log-Rank Tests with Conditional Inference Frame to construct new trees for interval-censored data to improve the predictive ability. Specifically, some special score statistics are originally constructed or considered and they are utilized in Generalized Log-Rank Tests. Then those score statistics are assigned to the influence function h in Conditional Inference Frame to construct new trees.

In more detail,

$$\xi_1(x) = (logx)x^{\rho+1}(1-x)^\gamma$$

is considered to obtain the new score statistic $U_{\xi,i}$ denoted by $U^{\rho,\gamma}_{\xi_1,i}$. Moreover, the case of

$$\xi_2(x) = (tan(\frac{\pi}{2}(x-1)))x^{\rho+1}(1-x)^\gamma$$

is also proposed and a new score statistic $U^{\rho,\gamma}_{\xi_2,i}$ is derived. The new score statistics are combined with the Conditional Inference Frame for conditional inference trees to construct new interval-censored tree methods.

With reference to conditional inference trees in Subsection 3.1, let ${Y_i}^T = (L_i, R_i)$ where L_i and R_i are the endpoints of censoring interval specified in (4.1). Here Y_i is a bivariate response variable. Assume the covariates are numeric. In step 1 of CIF, let $g_j(x) = x$ where $x \in \mathcal{X}_j$ and let $h = U^{\rho,\gamma}_{\xi_1,i}$, so $b_j = q = 1$. The test statistic is

$$T_j(\mathcal{L}_n, \omega) = \sum_{i=1}^{n} w_i X_{ji} U^{\rho,\gamma}_{\xi_1,i} \in \mathbb{R}, \tag{4.4}$$

and it is used for extended permutation test.

In step 2 of CIF, let $g_{j_0}(x) = I(x \in \mathcal{A})$ where $x \in \mathcal{X}_{j_0}$ and $\mathcal{A} \subset \mathcal{X}_{j_0}$, so $b_j = 1$. Moreover, let $h = U^{\rho,\gamma}_{\xi_1,i}$, and hence $q = 1$. The test statistic is now

$$T^{\mathcal{A}}_{j_0}(\mathcal{L}_n, \omega) = \sum_{i=1}^{n} w_i I(X_{j_0 i} \in \mathcal{A}) U^{\rho,\gamma}_{\xi_1,i} \in \mathbb{R} \tag{4.5}$$

and it is used for two-sample test.

A new interval-censored tree is obtained, which is denoted by $ICS_1^{\rho,\gamma}tree$. Note that the existing $ICtree$ is a special case of $ICS_1^{\rho,\gamma}tree$ when $\rho = \gamma = 0$.

Similarly, when the $U^{\rho,\gamma}_{\xi_1,i}$ in (4.4) and (4.5) is replaced by $U^{\rho,\gamma}_{\xi_2,i}$, a new tree is constructed with hyper-parameters ρ and γ called $ICS_2^{\rho,\gamma}tree$. Note that since our improved tree methods are based on CIF, the time complexity is the same as CIF.

Algorithm 6 reports the construction of the $ICS_1^{\rho,\gamma}tree$. The choice of the hyper-parameters is discussed in Algorithm 7.

Algorithm 6 $ICS_1^{\rho,\gamma}tree$

Given

- (ρ, γ): selected values for the hyper-parameter;
- $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$: m dimensional covariates;
- Y: response variable;
- $\mathcal{L}_n = \{(Y_i, X_{1i}, X_{2i}, \ldots, X_{mi})^T; i = 1,2, \ldots, n\}$: the sample space of n iid observations.

Procedure

1. Initially $ICS_1^{\rho,\gamma}tree$ consists of a single root node that is also an open node (that is, it needs to be further explored) and the n observations are all assigned to the root node.
2. Repeat the following steps until all nodes are closed:
   - Choose an open node.
   - Verify if this node can be closed. The closure condition is the stopping rule introduced later. If this node can be closed, then it can not be split further and it is a terminal node; otherwise, it can be split into two children nodes, the splitting criterion are specified according to the following steps:

(i) Test the global null hypothesis of independence between any component of the covariates $\mathbf{X}$ and the response Y by permutation test. The test statistic is

$$T_j(\mathcal{L}_n, \omega) = \sum_{i=1}^{n} w_i X_{ji} U^{\rho,\gamma}_{\xi_1,i},$$

(a) If the global null hypothesis can be rejected according to a pre-specified nominal level α, then X_{j_0} satisfying the j_0th component of covariates $\mathbf{X}$ with the strongest association to Y is selected as the splitting variable, i.e., the index j_0 for which the p value is the smallest is selected.

(b) If the global null hypothesis can not be rejected, then the node can be closed.

(ii) If the node is not closed and X_{j_0} is selected as splitting variable in step (i), a two-sample test is utilized to search the optimal binary split $\{X_{j_0} \in \mathcal{A}\}$ and $\{X_{j_0} \notin \mathcal{A}\}$ where the null hypothesis is that the two samples are statistically equivalent. The test statistic is

$$T^{\mathcal{A}}_{j_0}(\mathcal{L}_n, \omega) = \sum_{i=1}^{n} w_i I(X_{j_0 i} \in \mathcal{A}) U^{\rho,\gamma}_{\xi_1,i}.$$

Exhaustive search is carried out for every allowable split, and $\mathcal{A}^0$ is the one such that the test statistic is maximum among all the allowable subset $\mathcal{A} \subset \mathcal{X}_{j_0}$.

(iii) Once the split point is determined, the instances are divided into two subsets, and each subset is assigned to a child node, both of which represent sub-regions of the data. These new created nodes will be added to the list of open nodes.

Output

1. The estimate of survival function $\hat{S}(t|\mathbf{x})$ is calculated for each terminal node based on the novel EM-ICM algorithm given by Anderson-Bergman [7] which is introduced in Subsection 2.3.2.
2. The survival time is given by the median of the Survival function.

For $ICS_2^{\rho,\gamma}tree$, a similar procedure can be utilized, the difference being the score statistic $U^{\rho,\gamma}_{\xi_2,i}$ rather than $U^{\rho,\gamma}_{\xi_1,i}$.

What's more, other rank tests are considered for the construction of other novel survival trees, The performance of novel survival trees based on other rank test is compared with

that of $ICS_1^{\rho,\gamma}tree$ $ICS_2^{\rho,\gamma}tree$, so as to explore the predictive performance of $ICS_1^{\rho,\gamma}tree$ $ICS_2^{\rho,\gamma}tree$.

4.1.3 Interval-censored tree methods based on other rank tests

The $G^{\rho,\gamma}$ test is a class of rank test for interval-censored data which is used to evaluate whether the survival function under each group of data is equivalent or not [94]. The test statistic is $U_G = \sum_{i=1}^{n} \mathbf{w}_i c_{\rho,\gamma,i}$, where

$$c_{\rho,\gamma,i} = \frac{\widehat{S}(L_i)B(1-\widehat{S}(L_i);\gamma+1,\rho) - \widehat{S}(R_i)B(1-\widehat{S}(R_i);\gamma+1,\rho)}{\widehat{S}(L_i) - \widehat{S}(R_i)} \tag{4.6}$$

and the incomplete beta function $B(\cdot)$ is defined as

$$-B(1-t,\gamma+1,\rho) = -\int_0^{1-t} x^{\gamma}(1-x)^{\rho-1}dx = \lambda(t).$$

Here n, L_i, R_i, $\widehat{S}$ and $\mathbf{w}_i$ have the same meaning as in Section 4.1.1. Note that the $G^{\rho,\gamma}$ test coincides with the log-rank test when $\rho = \gamma = 0$ and $\lambda(t) = log(t)t^{\rho}(1-t)^{\gamma}$.

The tree method $ICG^{\rho,\gamma}tree$ is obtained when $c_{\rho,\gamma,i}$ is utilized instead of $U_{\xi_1,i}^{\rho,\gamma}$ in (4.4) and (4.5).

A generalized Wilcoxon test for interval-censored data [239] is also applied to CIF, where

$$U_{i,W} = \widehat{S}(L_i) + \widehat{S}(R_i) - 1, \tag{4.7}$$

is used instead of $U_{\xi_1,i}^{\rho,\gamma}$ in (4.4) and (4.5) to construct a novel survival tree. Here $U_{i,W}$ is the rank score of generalized Wilcoxon test and is utilized for comparison of survival function of two groups. We refer to this tree as $ICWtree$.

Table 4.1 summarizes all novel survival trees proposed in this thesis for interval-censored data.

Table 4.1: Novel trees for interval-censored data

Type of Trees	Rank Tests	Score Statistic / Influence Function	Hyper-parameter
$ICS_1^{\rho,\gamma}tree$	Generalized Log-Rank Tests	$U_{\xi_1,i}^{\rho,\gamma}$	(ρ,γ)
$ICS_2^{\rho,\gamma}tree$	Generalized Log-Rank Tests	$U_{\xi_2,i}^{\rho,\gamma}$	(ρ,γ)
$ICWtree$	Generalized Wilcoxon tests	$U_{i,W}$	-
$ICG^{\rho,\gamma}tree$	$G^{\rho,\gamma}$ test	$c_{\rho,\gamma,i}$	(ρ,γ)

4.1.4 Hyper-parameter tuning

It is important to choose appropriate hyper-parameter values to adapt our tree methods to the dataset. We use grid search cross validation [135] for hyper-parameter optimization. The whole hyper-parameter tuning procedure is reported in Algorithm 7.

Algorithm 7 Hyper-parameter tuning

Given

- A: training set;
- N_A: size of training set A;
- $[a_\rho, b_\rho] \times [a_\gamma, b_\gamma]$: two-dimensional search space of hyper-parameter (ρ, γ);
- B_1: the number of subinterval that $[a_\rho, b_\rho]$ is divided into;
- B_2: the number of subinterval that $[a_\gamma, b_\gamma]$ is divided into;
- $\triangle s_\rho$: step size in direction of ρ;
- $\triangle s_\gamma$: step size in direction of γ;
- K: the fold number of cross validation;
- $\phi(\cdot)$ is a measure for evaluation of the predictive performance.

Procedure

Consider the two-dimensional search space $[a_\rho, b_\rho] \times [a_\gamma, b_\gamma]$.

- Divide $[a_\rho, b_\rho]$ into B_1 subinterval $[\rho_{i-1}, \rho_i]$ of equal width $\triangle s_\rho = (b_\rho - a_\rho)/B_1$, $i = 1, \ldots, B_1$, and $\rho_0 = a_\rho$.
- Divide $[a_\gamma, b_\gamma]$ into B_2 subinterval $[\gamma_{j-1}, \gamma_j]$ of equal width $\triangle s_\gamma = (b_\gamma - a_\gamma)/B_2$, $j = 1, \ldots, B_2$, and $\gamma_0 = a_\gamma$.
- For i=1 to B_1, j=1 to B_2
  (a) randomly split the training set A into K non-overlapping groups of approximately same size. The kth group is considered as test set A_k^{test}, while the remaining parts together compose the training set A_k^{train}, $k = 1, \ldots, K$.
  (b) use $\phi(\cdot)$ for evaluating the predictive accuracy of the tree model with hyper-parameter (ρ_i, γ_j) by K-fold cross validation, where the model is fitted in A_k^{train} and evaluated in A_k^{test}, $k = 1, \ldots, K$. The outcome of K-fold cross validation is denoted by $\phi_{ij}(\cdot)$.

Output

The best hyper-parameter values $(\tilde{\rho}, \tilde{\gamma}) = (\rho_{i^*}, \gamma_{j^*})$, where $\phi_{i^*,j^*}(\cdot)$ is the best value of $\phi_{ij}(\cdot)$ obtained.

4.2 New ensemble method for interval-censored data

We also propose a novel ensemble method referred as $ICS_1^{\rho,\gamma}forests$ for interval-censored data where $ICS_1^{\rho,\gamma}tree$ is chosen as the base learner. The weight average of non-parametric maximum likelihood estimator (NPMLE) is utilized for the estimate of survival function in this method. Here the construction of the weights is similar to the method in the quantile regression forests described in Chapter 3.

Assume that the interval-censored data consists of n independent subjects from a homogeneous population with survival function $S(t)$. Let T_i be the survival time for subject i, and suppose that $(L_i, R_i]$ is the censored interval to which T_i belongs, $i = 1, \ldots, n$. Let $\{s_j\}_{j=0}^{l}$ denote a set of the distinct ordered values of observed interval endpoints $\{(L_i, R_i]\}_{i=1}^{n}$ and 0. Moreover, let $p_j = S(s_{j-1}) - S(s_j)$, $j = 1, \ldots, l$. The non-parametric likelihood function is

$$L_S(\mathbf{p}) = \prod_{i=1}^{n} (S(L_i) - S(R_i))^{a_i(\mathbf{x})} = \prod_{i=1}^{n} (\sum_{j=1}^{m} \alpha_{ij} p_j)^{a_i(\mathbf{x})} \tag{4.8}$$

where $\alpha_{ij} = I(s_j \in (L_i, R_i])$, $j = 1, \ldots, l, i = 1, \ldots, n$; $\mathbf{p} = (p_1, \ldots, p_l)^T$ and $a_i(\mathbf{x})$ are the weights which will be introduced later in Algorithm 8.

Following (2.11) in Subsection 2.3.2, the estimate of $\mathbf{p}$ was calculated by maximizing $L_S(\mathbf{p})$ with respect to $\mathbf{p}$ on the simplex unit, i.e.,

$$\begin{aligned} \hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_l)^T \quad = \quad & \arg\max_{\mathbf{p}} L_S(\mathbf{p}) \\ & s.t. \qquad \sum_{j=1}^{l} p_j = 1, \\ & \qquad\quad p_j \geq 0, \quad j = 1, \ldots, l \end{aligned} \tag{4.9}$$

The optimal solution $\hat{\mathbf{p}}$ in (4.9) is obtained by the EM-ICM algorithm [7] which has been introduced in Subsection 2.3.2.

Based on $\hat{\mathbf{p}}$, the estimate of the survival function, given covariates $\mathbf{X} = \mathbf{x}$, is given by

$$\hat{S}(t|\mathbf{x}) = \begin{cases} 1, & \text{if } \ 0 = s_0 \leq t < s_1; \\ \hat{p_j} + \hat{S}(s_{j-1}|\mathbf{x}), & \text{if } \ s_j \leq t < s_{j+1}, j = 1, \ldots, l-1; \\ 0, & \text{if } \ t \geq s_l. \end{cases} \tag{4.10}$$

The procedure for $ICS_1^{\rho,\gamma}forests$ is described in detail in Algorithm 8.

Algorithm 8 $ICS_1^{\rho,\gamma} forests$

Given

- $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$: training set, where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{mi})^T$ is an m dimensional covariate vector and ${y_i}^T = (L_i, R_i)$ is the censored interval;
- m: the dimension of covariates or the size of predictor variables;
- (ρ, γ): selected values for the hyper-parameter of the base learner $ICS_1^{\rho,\gamma} tree$;
- B: the number of base learners.

Procedure

For b=1 to B

1. construct an $ICS_1^{\rho,\gamma} tree$ T_b by randomly selecting a subset of predictor variables and using bootstrapped data from the training set A.
2. For each observation $(\mathbf{x}_i, y_i) \in A$, let

$$a_i(\mathbf{x}, b) = \frac{I(\mathbf{x}_i \in \mathcal{R}_b^{l(\mathbf{x})})}{\sharp\{j \in \{1, \ldots, N\} | \mathbf{X}_j \in \mathcal{R}_b^{l(\mathbf{x})}\}}.$$

   where $\mathcal{R}_b^{l(\mathbf{x})}$ is the terminal node of T_b to which $\mathbf{x}$ is assigned.

Output

The estimate of survival function $\hat{S}(t|\mathbf{x})$ based on (4.8) and (4.10), where the average observation weights is

$$a_i(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} a_i(\mathbf{x}, b), \quad i = 1, \ldots, N.$$

4.3 Computational experiments for tree methods

Let T_i be the survival time for subject i, and assume that $(L_i, R_i]$ is the censored interval to which T_i belongs. If $T_i \in (\tau_j, \tau_{j+1}]$, then let $L_i = \tau_j$ and $R_i = \tau_{j+1}$ for some j, where $\tau_{j+1} > \tau_j$, $\tau_0 = 0$ and $\tau_l = +\infty$, $j = 1, \ldots, l-1$. The gap between adjacent endpoints of the interval is denoted by $\text{len}_j = \tau_j - \tau_{j-1}$.

We assume that the censoring mechanism is non-informative, i.e., the survival time of interest is independent of the censoring intervals. Let T_i be randomly generated from a specific distribution $F(t)$, and let len_j be a random variable from a specified distribution $G(x)$ for each j. The type of interval-censoring in this experiment is Case II. The censoring mechanism is similar to [237] for the purposes of comparison.

The observations are right censored when the survival time T fall into interval $(\tau_l, +\infty)$. The right censoring rate is controlled by adjusting the number of intervals. Three right censoring rates are considered to explore the impact of right censoring rate on the performance of the tree methods: 0%, 20% and 40%.

4.3.1 Classification

In real world, the distribution of lifetime or time of occurrence of event of interest is uncertain. A good model for classification can distinguish the observations groups with high hazard rates from the ones with low hazard rates. For example, the parameter value λ in Exponential distribution corresponds to the hazard rate, so a good model should be able to distinguish between the groups of observations from an Exponential distributions with different parameters λ_1 and λ_2. More in general, a good model can distinguish the group of observations from distributions with different parameter values, and correctly classify the samples into the corresponding categories from a pool of distribution with different parameters. What's more, a survival tree method can select the important covariates [225].

To evaluate the classification performances of tree models, a tree structured synthetic dataset is constructed.

Tree-structured data

As shown in Figure 4.1, tree-structured datasets are constructed which are easily divided into several categories according to the character of covariates.

Let $X = \{X_1, \ldots, X_6\}^T$ be a 6 dimensional vector of covariates where

- X_1 is from a discrete uniform distribution on $\{1,\ 2,\ 3,\ 4\}$;
- X_2 and X_6 follow the Bernoulli distribution with p=0.5;
- X_3 and X_4 are uniform in the interval $[1,\ 3]$;
- X_5 is from a discrete uniform distribution on $\{1,\ 2,\ 3,\ 4,\ 5\}$.

The survival time of interest T is determined by the values of X_1, X_2, X_3 according to the tree relationship shown in Figure 4.1. More specifically, T is generated from a distribution $F(t)$ with four different parameter value. We consider five data scenarios for the distribution of $F(t)$:

1. Exponential distribution with parameter $\lambda \in \{0.08,\ 0.25,\ 0.6,\ 0.95\}$;
2. Weibull distribution with shape parameter α=7 and scale parameter $\lambda \in \{1,\ 6,\ 14,\ 30\}$, which have an increasing failure rate with time, (we refer to this case as Weibull1);

3. Weibull distribution with shape parameter α=0.8 and scale parameter $\lambda \in \{1,\ 6,\ 14,\ 30\}$, which have a decreasing failure rate with time, (this case is denoted by Weibull2);

4. Log-normal distribution with parameter pairs (μ, σ) which is equal to (0, 0.25),(2, 0.25), (3, 0.4),(4, 0.1) respectively;

5. Log-logistic distribution with shape parameter and scale parameter pairs (α, λ) which is equal to $(0.2,\ 0.8)$, $(0.02,\ 3)$,$(0.03,\ 4)$,$(0.05,\ 2)$ respectively.

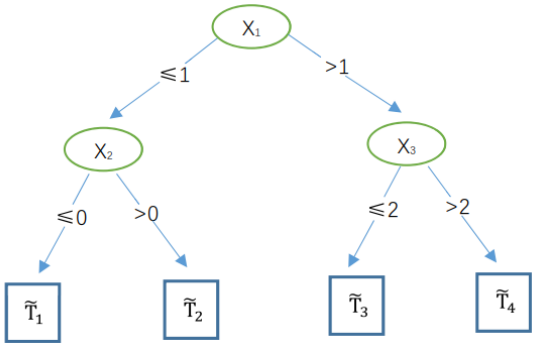


Figure 4.1: Tree-structured data

The distribution $G(t)$ of the interval length len_j follows a uniform distribution on the interval [0.7, 1.5]. Let the sample size be N. The ith observation is denoted by

$$O_i = (L_i, R_i, X_1, X_2, X_3, X_4, X_5, X_6)^T \in \mathcal{L}_n, \quad i = 1, \ldots, N,$$

where $(L_i, R_i)^T = Y_i$ is the response variable.

We use the first data scenario in the simulation as an example to illustrate the simulative data. The quantities $\tilde{T}_1$, $\tilde{T}_2$, $\tilde{T}_3$, $\tilde{T}_4$ are from an Exponential distribution with parameter $\lambda \in \{0.08,\ 0.25,\ 0.6,\ 0.95\}$ respectively and exactly determined by the values of X_1, X_2, X_3 according to the tree relationship shown in Figure 4.1. As we known, λ is the constant value of the hazard function $h(t)$. The simulated data belong to four classes which have different level of hazard rates from Class1 to Class4, The survival function of each class is $S(t) = \exp(-\lambda t)$, $t > 0$. Specifically, the samples belonging to the first class (Class1) has the lowest hazard rate and the survival function is $S(t) = \exp(-0.08t)$; the ones belonging to the fourth class (Class4) has the highest hazard rate, and the survival function is $S(t) = \exp(-0.95t)$. The categories are associated with the values of covariate, and each observation belongs to only one class.

Obviously, a perfect tree method based on this simulated data should grow the recursive structure shown in Figure 4.2. The terminal nodes ID are 3, 4, 6 and 7, corresponding to

the four classes. Each observation will fall in a terminal node according to the values of the covariate of the sample and the recursive structure of a tree. Finally, the category of each observation is labelled by the class corresponding to the terminal node.

For example, if an observation satisfies $X_1 = 1, X_2 = 1, X_3 = 2.3, X_4 = 1, X_5 = 5, X_6 = 0$, then the observation should be assigned to terminal node ID 4 according to Figure 4.1, and the category of it is Class2.

Measure for classification accuracy

Let $N^*(O_i)$ be the terminal node ID into which O_i should fall according to Figure 4.2, $i = 1, \ldots, N$, and let N be the sample size. Moreover, let $N^{**}(O_i)$ be the terminal node ID in which the observation O_i falls according to a specific tree method, $i = 1, \ldots, N$. If $N^*(O_i) = N^{**}(O_i)$, then the observation O_i is correctly classified by the tree methods, $i = 1, \ldots, N$.

```
[[1]]
[1]root
|    [2] X1 <= 1
|    |      [3] X2 <= 0*
|    |      [4] X2 > 0*
|    [5] X1 > 1
|    |      [6] X3 <= 2*
|    |      [7] X3  > 2*
```

Figure 4.2: Recursive structure of the dataset

We define a performance measure cr for the classification task as the ratio between the number of observations falling into the correct terminal node and N, i.e., the ratio of observations labelled correctly:

$$cr = \frac{\sum_{i=1}^{N} I(N^*(O_i) = N^{**}(O_i))}{N},$$

where $I(\cdot)$ is the indicator function.

This quantity is a reasonable measure to evaluate the correctness of a tree method covering the tree-structured dataset, because the correctness of the tree structure generated by tree methods are closely related to its value. If a tree method grows a tree structure with wrong splitting as shown in Figures 4.3 and 4.4, some of the observations will be misclassified. This in turn leads to lower values of cr. For example, an observation with covariates X_1=2, X_3=1.7 should belong to the third category (Class3) and it will be assigned

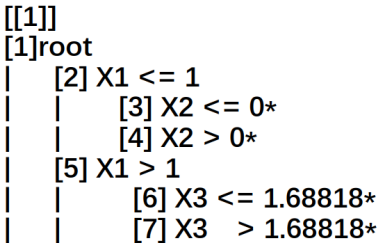


Figure 4.3: Wrong splitting structure grown by a tree method

to terminal node 6 according to data structure shown in Figure 4.2. However, the splitting criterion for the tree method in Figure 4.3 is inconsistent with the character of data, and this observation is incorrectly assigned to terminal node 7. Even worse, in some cases no observation falls in the correct terminal nodes due to the wrong tree structure grown by a tree method. A very incorrect tree structure grown by a tree method is reported in Figure 4.4. In this case, no terminal nodes 6 and 7 are presented, and the observations that should belong to terminal nodes 6 or 7 fail to be correctly classified. The more the splitting by tree methods is wrong, the lower the value of cr is.

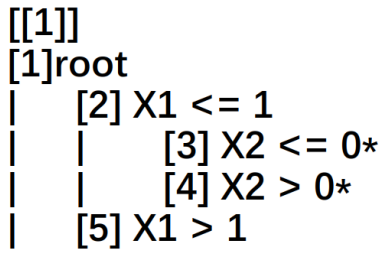


Figure 4.4: Wrong splitting structure grown by a tree method

Hyper-parameter tuning

The hyper-parameter tuning procedure (Algorithm 7) is executed to determine the appropriate hyper-parameter values of tree methods for tree-structured data. In our computational experiment, the sample size N is 1200, and the ratio of training set to test set is 5:1, i.e., the size of training set A is 1000 and the size of test set D is 200. The training set A is used for tuning (ρ, γ) and test set D is for evaluation. Therefore, the process of choice of hyper-parameter and evaluation is separated.

In our experiments the two-dimensional search space of the hyper-parameter (ρ, γ) has

been set to (0.01,0.81) × (0.01,0.81) and $B_1 = B_2 = 40$. Here cr is taken as measure of the predictive performance $\phi(\cdot)$ for the different tree methods such as $ICS_1^{\rho,\gamma}tree$. A 5-fold cross validation is applied. Specifically, $\phi_{\rho_i,\gamma_j}(\cdot)$ is the value of cr which is calculated by utilizing $ICS_1^{\rho_i,\gamma_j}tree$. The optimal hyper-parameter values $(\tilde{\rho},\tilde{\gamma})$ for $ICS_1^{\rho,\gamma}tree$ are

$$(\tilde{\rho},\tilde{\gamma}) = \arg\max_{(\rho_i,\gamma_j)}\{\phi_{\rho_i,\gamma_j}(\cdot)\}.$$

Table 4.2: Optimal hyper-parameter values $(\tilde{\rho},\tilde{\gamma})$ for classification under Exponential distribution.

Type of Trees	0%*	20%*	40%*
	$(\tilde{\rho},\tilde{\gamma})$	$(\tilde{\rho},\tilde{\gamma})$	$(\tilde{\rho},\tilde{\gamma})$
$ICS_1^{\rho,\gamma}tree$	(0.03,0.09)	(0.13,0.05)	(0.05,0.03),(0.05,0.21)
$ICS_2^{\rho,\gamma}tree$	(0.69,0.55)	(0.69,0.49)	(0.03,0.03)

* Right censoring rate of tree-structured data.

Table 4.3: Optimal hyper-parameter values $(\tilde{\rho},\tilde{\gamma})$ for classification under Weibull1 distribution.

Type of Trees	0%*	20%*	40%*
	$(\tilde{\rho},\tilde{\gamma})$	$(\tilde{\rho},\tilde{\gamma})$	$(\tilde{\rho},\tilde{\gamma})$
$ICS_1^{\rho,\gamma}tree$	(0.57,0.33)	(0.53,0.37)	(0.09,0.13)
$ICS_2^{\rho,\gamma}tree$	(0.21,0.01) -	(0.21,0.05) -	(0.25,0.21) (0.29,0.21)

* Right censoring rate of tree-structured data.

Table 4.4: Optimal hyper-parameter values $(\tilde{\rho},\tilde{\gamma})$ for classification under Weibull2 distribution.

Type of Trees	0%*	20%*	40%*
	$(\tilde{\rho},\tilde{\gamma})$	$(\tilde{\rho},\tilde{\gamma})$	$(\tilde{\rho},\tilde{\gamma})$
$ICS_1^{\rho,\gamma}tree$	(0.09,0.09)	(0.09,0.01)	(0.41,0.01)
$ICS_2^{\rho,\gamma}tree$	(0.61,0.45)	(0.41,0.17)	(0.21,0.01)

* Right censoring rate of tree-structured data.

Consider the first scenario where T is from an Exponential distribution with different right censoring rate. The optimal hyper-parameter values chosen by Algorithm 7 are partially shown in Table 4.2 for different right censoring rates. We also obtain the optimal

values of hyper-parameter for the other scenarios where T is from Weibull1 distribution, Weibull2 distribution and Log-normal distribution. The results are listed in Tables 4.3 - 4.5. Since the accuracy is quite low for the case where T is from Loglogistic distribution, the tuning is ignored.

Table 4.5: Optimal hyper-parameter values $(\tilde{\rho}, \tilde{\gamma})$ for classification under Log-Normal distribution.

Type of Trees	0%*	20%*	40%*
	$(\tilde{\rho}, \tilde{\gamma})$	$(\tilde{\rho}, \tilde{\gamma})$	$(\tilde{\rho}, \tilde{\gamma})$
$ICS_1^{\rho,\gamma}tree$	(0.17,0.01)	(0.05,0.03)	(0.09,0.09)
	-	-	(0.07,0.03)
$ICS_2^{\rho,\gamma}tree$	(0.41,0.01)	(0.21,0.01)	(0.29,0.37)

* Right censoring rate of tree-structured data.

From the results obtained, it is clear that the optimal values of the hyper-parameter depend on the data and the tree model. Moreover, right censoring rate also affects the choice of hyper-parameter values. We find that in most cases the optimal hyper-parameter values are around (0.05, 0.05), (0.5, 0.03), (0.05, 0.21) for $ICS_1^{\rho,\gamma}tree$. Instead, for $ICS_2^{\rho,\gamma}tree$, they are around (0.21, 0.01), (0.69, 0.03), (0.69, 0.51). Moreover, the value of cr for $ICG^{\rho,\gamma}tree$ is much lower than the one of $ICS_1^{\rho,\gamma}tree$ and $ICS_2^{\rho,\gamma}tree$, and thus selections of hyper-parameter values are not executed for $ICG^{\rho,\gamma}tree$.

Evaluation of predictive accuracy for classification

The predictive classification performance of various tree models is also evaluated. The choice for the training set A and test set D are same as in Subsection 4.3.1. The mean value of cr from 1000 trials, denoted by $\overline{cr}$, is calculated on the test set D. Different values of the hyper-parameters are considered for our tree models. Here, values around (0.05, 0.05), (0.5, 0.03) and (0.5, 0.21) are adopted for $ICS_1^{\rho,\gamma}tree$ under various distributions. Note that they are not always the optimal hyper-parameter values but just recommended values.

Tables 4.6 - 4.9 show the predictive performance of the implemented methods. In these tables, 0%, 20% and 40% denote the right censoring rate of tree-structured data, and '*' corresponds to the higher values of $\overline{cr}$ of the proposed tree methods than $ICtree$.

As can be seen from the tables, $ICS_1^{\rho,\gamma}tree$ in all cases almost performs better than $ICtree$ under all scenarios. According to the result of Table 4.6 - 4.9, the value $(\rho, \gamma) = (0.05,\ 0.03)$ is recommended for the interval-censored data of which the right censoring rate is lower than 40%, and $(\rho, \gamma) = (0.5,\ 0.21)$ is appropriate for data with over 40% right censoring rate.

Table 4.6: Predictive accuracy under Exponential distribution for different right censoring rates.

Type of Trees	$Hyper-parameter$	0%	20%	40%
	(ρ,γ)	$\overline{cr}$	$\overline{cr}$	$\overline{cr}$
$ICtree$	-	0.5130	0.4416	0.2723
$ICS_1^{\rho,\gamma}tree$	(0.05,0.03)	0.5155*	0.4466*	0.2752*
	(0.05,0.05)	0.5145*	0.4456*	0.2746*
	(0.05,0.21)	0.5137*	0.4465*	0.2761*
	(0.69,0.03)	0.4023	0.4393	0.2457
$ICWtree$	-	0.4974	0.069	0.018
$ICS_2^{\rho,\gamma}tree$	(0.69,0.49)	0.5073	0.4479*	0.2650
	(0.07,0.05)	0.5083	0.3458	0.2744*
	(0.71,0.61)	0.5083	0.4437*	0.2669

Table 4.7: Predictive accuracy under Weibull1 distribution for different right censoring rates.

Type of Trees	$Hyperparameter$	0%	20%	40%
	(ρ,γ)	$\overline{cr}$	$\overline{cr}$	$\overline{cr}$
$ICtree$	-	0.9164	0.9059	0.9212
$ICS_1^{\rho,\gamma}tree$	(0.05,0.03)	0.9165*	0.9076*	0.9192
	(0.05,0.05)	0.9175*	0.9070*	0.9211
	(0.05,0.21)	0.8916	0.8843	0.9248*
	(0.57,0.33)	0.9225*	0.9097*	0.9152
	(0.53,0.37)	0.9160	0.9132*	0.9169
$ICWtree$	-	0.9202*	0.9211*	0.4210
$ICS_2^{\rho,\gamma}tree$	(0.21,0.01)	0.9440*	0.9413*	0.9324*
	(0.29,0.21)	0.9364*	0.9337*	0.9348*

Although $(\rho,\gamma) = (0.05,\ 0.03)$ and $(\rho,\gamma) = (0.05,\ 0.21)$ are not always the best hyper-parameter values for all different data, values of (ρ,γ) around them still result in good performance and adaptability of the $ICS_1^{\rho,\gamma}tree$ method to the data. Moreover, $ICS_2^{\rho,\gamma}tree$ performs better on specific data than $ICS_1^{\rho,\gamma}tree$ and $ICtree$. However, as a whole, $ICS_2^{\rho,\gamma}tree$ shows great instability.

Table 4.8: Predictive accuracy under Weibull2 distribution for different right censoring rates.

Type of Trees	*Hyperparameter*	0%	20%	40%
	(ρ, γ)	$\overline{cr}$	$\overline{cr}$	$\overline{cr}$
$ICtree$	-	0.6358	0.5871	0.4997
$ICS_1^{\rho,\gamma}tree$	(0.05,0.03)	0.6383*	0.5881*	0.4994
	(0.05,0.05)	0.6385*	0.5884*	0.5002*
	(0.05,0.21)	0.6264	0.5827	0.49325
	(0.09,0.09)	0.6397*	0.5874*	0.4993
	(0.41,0.01)	0.5972	0.5739	0.5029*
$ICWtree$	-	0.5810	0.4352	0.3175
$ICS_2^{\rho,\gamma}tree$	(0.61,0.45)	0.6276	0.5812	0.4973
	(0.41,0.17)	0.5963	0.5972*	0.5074*
	(0.21,0.01)	0.1433	0.3366	0.5101*

Table 4.9: Predictive accuracy under Log-Normal distribution for different right censoring rates.

Type of Trees	*Hyperparameter*	0%	20%	40%
	(ρ, γ)	$\overline{cr}$	$\overline{cr}$	$\overline{cr}$
$ICtree$	-	0.9181	0.9215	0.9265
$ICS_1^{\rho,\gamma}tree$	(0.05,0.03)	0.9232*	0.9235*	0.9291*
	(0.05,0.05)	0.9200*	0.9193	0.9267*
	(0.05,0.21)	0.6264	0.8965	0.9275*
	(0.17,0.01)	0.9234*	0.9196	0.9239
$ICWtree$	-	0.9294*	0.9335*	0.4126
$ICS_2^{\rho,\gamma}tree$	(0.21,0.01)	0.5514	0.9510*	0.9448*
	(0.29,0.37)	0.5061	0.8593	0.9395*
	(0.41,0.01)	0.9495*	0.9404*	0.9304*
	(0.71,0.61)	0.9219*	0.9188*	0.9267*

Table 4.10: Predictive accuracy under Loglogistic distribution for different right censoring rates.

Type of Trees	$Hyperparameter$	0%	20%	40%
	(ρ, γ)	$\overline{cr}$	$\overline{cr}$	$\overline{cr}$
$ICtree$	-	0.2797	0.2709	0.2648
$ICS_1^{\rho,\gamma}tree$	(0.05,0.05)	0.2976*	0.2922*	0.2669*
	(0.05,0.21)	0.2849*	0.2874*	0.2779*
$ICWtree$	-	0.3067*	0.2232	0.30956*
$ICS_2^{\rho,\gamma}tree$	(0.71,0.61)	0.2797	0.2815*	0.2805*

Performance of $ICWtree$ is seriously affected by the level of right censoring rate and the type of distribution for T, and predictive accuracy sharply deteriorate in case of high right censoring rate.

All in all, $ICS_1^{\rho,\gamma}tree$ slightly improves the predictive accuracy with appropriate values of hyper-parameter. Although Algorithm 7 can search for the optimal values of hyper-parameter for tree methods for a specific dataset, the value $(\rho, \gamma) = (0.05,\ 0.03)$ and $(\rho, \gamma) = (0.05,\ 0.21)$ are recommended values for hyper-parameter and parameter tuning can be omitted to save time cost. Moreover the right censoring rate should be considered for the choice of hyper-parameter values.

In Section 2.6, the imputation methods have been introduced. In fact, the combination of imputation methods and conditional inference trees can also construct tree models for interval-censored data which are called imputation trees and denoted by $IMtree$ for simplicity. As mentioned before, different imputation can be utilized, such as left imputation, right imputation and middle imputation. Here we show the results of the middle imputation. Only $ICS_1^{0.05,0.05}tree$ is discussed here as a representative of our new tree algorithms. We compare the performance of imputation trees with respect to our trees on datasets of 200, 500 and 1000 observations (i.e., sample sizes $N = 200$, $N = 500$ and $N = 1000$). The results of this comparison are presented in Tables 4.11 - 4.12.

As expected, the performance of survival trees improves as the sample size increases. In addition, the predictive ability for classification of $IMtree$ is lower than that of $ICtree$ in all cases. In various cases $ICS_1^{0.05,0.05}tree$ shows advantages over the other two tree methods. The outcome of the simulation are only presented for the Exponential distribution and Weibull2 distribution. The performance under other data scenarios varies slightly with an increase in the sample size. One possible explanation is that the predictive values are already very high, leaving little room for further improvement.

Table 4.11: Predictive accuracy under Exponential distribution for different right censoring rates.

Type of Trees	*sample size*	0%	20%	40%
	N	$\overline{cr}$	$\overline{cr}$	$\overline{cr}$
$ICtree$	500	0.7151	0.7018	0.6460
	1000	0.8423	0.8550	0.8096
$ICS_1^{0.05,0.05}tree$	500	0.7216	0.7089	0.6480
	1000	0.8430	0.8564	0.8137
$IMtree$	200	0.5103	0.4420	0.2649
	500	0.7023	0.6911	0.6432
	1000	0.8349	0.8404	0.8170

Table 4.12: Predictive accuracy under Weibull2 distribution for different right censoring rates.

Type of Trees	*sample size*	0%	20%	40%
	N	$\overline{cr}$	$\overline{cr}$	$\overline{cr}$
$ICtree$	500	0.8629	0.809	0.6752
	1000	0.8645	0.8667	0.8230
$ICS_1^{0.05,0.05}tree$	500	0.8670	0.8659	0.6761
	1000	0.8671	0.8685	0.8272
$IMtree$	200	0.6270	0.5874	0.4750
	500	0.8566	0.800	0.6738
	1000	0.8638	0.8661	0.8170

4.3.2 Regression

Synthetic tree-structured data is also constructed to explore the performance of classification of tree models in Subsection 4.3.1. Here we discuss the predictive performance of tree models for estimating the survival time, and consider more general data structure for simulation.

Complex structure data

The following setting is used to evaluate the predictive performance of the new survival trees with hyper-parameter. The setting is similar to the one proposed in [237] for the purpose of comparison. The dimension of the covariates are 6, where

- X_1 is from discrete uniform distribution on {-1, 0};

- X_2 is uniform in the interval [-0.5, 0];
- X_3 and X_5 is from Bernoulli distribution with $p = 0.5$;
- X_4 and X_6 are uniform in the interval [0, 1].

Moreover, the covariates are independent. The following distributions for the survival time are utilized:

1) Exponential with parameter $\lambda = e^{\vartheta}$;
2) Weibull with increasing hazard, scale parameter $\lambda = 4e^{\vartheta}$ and shape parameter $\alpha = 7$ (denoted as Weibull1);
3) Weibull with decreasing hazard, scale parameter $\lambda = 4e^{\vartheta}$ and shape parameter $\alpha = 0.7$ (denoted as Weibull2);

The value of the parameter ϑ depends on the covariates X_1 and X_2. Specifically,

$$\vartheta = -\sin\left[(X_1 + 2X_2) \cdot \pi\right] + \frac{3}{2}\sqrt{-X_1 - 2X_2}.$$

The censoring mechanism is non-informative as assumed at the beginning of this section. The gap between censoring intervals is randomly generated from uniformly in [0.5, 1]. Different right-censoring rates are also considered.

A measure of predictive accuracy for regression

Mean Square Error (MSE) is utilized as a measure to evaluate the predictive accuracy of survival time:

$$\frac{1}{N}\sum_{i=1}^{N}(T_i - \hat{T}_i)^2,$$

where N is the sample size, T_i is the true survival time of interest for subject i and $\hat{T}_i$ is the predicted survival time for ith subject, $i = 1, \ldots, N$.

Nevertheless, it is well-known that the results obtained from MSE can often be inaccurate and unsatisfactory [97].

Integrated Brier Score (IBS) is a popular method measuring average discrepancies between true status of event of interest and estimated predicted value, which was first proposed for right censoring data [97]. Later it was extended to interval-censored data [226]:

$$IBS = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\bar{T}}\int_0^{\bar{T}}[\hat{S}_i(t) - \bar{I}\{T_i > t\}]^2 dt, \tag{4.11}$$

where T_i is the true survival time of interest of subject i, $\hat{S}_i(\cdot)$ is an estimator of the survival function for the ith subject, $\bar{T} = \max\{T_j\}_{j=1}^N$. Here the quantities of $\bar{I}\{T_i > t\}$ are given by the formula

$$\bar{I}\{T_i > t\} = \begin{cases} \frac{\hat{S}_i(t)-\hat{S}_i(R_i)}{\hat{S}_i(L_i)-\hat{S}_i(R_i)}, & \text{if } L_i < t \leq R_i; \\ 1, & \text{if } t \leq L_i; \\ 0, & \text{if } t > R_i. \end{cases}$$

Mean integrated squared error denoted by MIE is an another measure for survival data [108], which is calculated by integrating the square difference between the two curves with respect to time and averaging over all observations:

$$MIE = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\max(\bar{T})} \int_0^{\max(\bar{T})} [\hat{S}_i(t) - S_i(t)]^2 dt, \tag{4.12}$$

In (4.12), $S_i(t)$ is the true survival function for the ith subject, which the meanings of T_i and $\hat{S}_i(\cdot)$ are same as defined in (4.11).

Computational results

The dataset is composed of 1000 elements and is divided into training set and test set with ratio 4:1. The training set A is utilized to construct the tree models. Then the predicted performance is evaluated using the test data D.

Boxplots for IBS are drawn from 1000 trials on the test data D. The methods are numbered as shown in Table 4.13. We utilized $(\rho, \gamma) = (0.05,\ 0.05)$ for $ICS_1^{\rho,\gamma}tree$, $(\rho, \gamma) = (0.69,\ 0.5)$ for $ICS_2^{\rho,\gamma}tree$, and $(\rho, \gamma) = (0.01,\ 0.01)$ for $ICG^{\rho,\lambda}tree$. These values are the recommended in Subsection 4.3.1. In addition, for $ICtree$, the Proportional Hazards model for interval-censored data (an extended form of Cox model) denoted by $ICPH$ is also utilized for comparison. In order to evaluate the amount of information loss caused by interval-censoring, we also consider conditional inference trees and the Proportional Hazards model using true survival time T, which are denoted by $Ttree$ and TPH respectively.

Table 4.13: Numbering of the various methods utilized in Figure 4.5.

No.	model	No.	model	No.	model	No.	model
1	$Ttree$	2	$ICtree$	3	$ICS_1^{\rho,\gamma}tree$	4	$ICWtree$
5	$ICG^{\rho,\gamma}tree$	6	$ICS_2^{\rho,\gamma}tree$	7	TPH	8	$ICPH$

The computational results of Boxplots for IBS based on various methods are shown in Figure 4.5. We note that $ICS_1^{\rho,\gamma}tree$ has slightly better performance than $ICtree$ in this setting, especially for the Weibull distribution with increasing hazards. In addition,

$ICS_1^{\rho,\gamma}tree$ and $ICS_2^{\rho,\gamma}tree$ show similar performance that is better than $ICPH$, which demonstrates the flexibility of the methods. Moreover, $ICWtree$ and $ICG^{\rho,\gamma}tree$ seem to be more sensitive to right censoring data. Although there is no noticeable difference with $ICS^{\rho,\gamma}$ when the right-censoring rate is zero, the value of IBS rapidly increases as the right-censoring rate increases. The performance of $ICWtree$ and $ICG^{\rho,\gamma}$ are worse than $ICS^{\rho,\gamma}$'s with over 20% right-censoring rate. Instead, the value of IBS for $Ttree$ is much lower than tree methods because it has no informative loss from interval censoring.

Table 4.14: Numbering of the various methods utilized in Figure 4.6.

No.	model	No.	model	No.	model
1	$ICtree$	2	$ICS_1^{\rho,\gamma}tree$	3	$ICWtree$
4	$ICG^{\rho,\gamma}tree$	5	$ICS_2^{\rho,\gamma}tree$		

MIE is also applied to evaluate the predicted performance of tree methods. The boxplots are drawn from 1000 trials to compare the proposed tree methods with $ICtree$. The outcomes for the obxplots of MIE are shown in Figure 4.6. Also in this case the tree methods are specified in Table 4.14.

The results of the predictive performance evaluated by MIE validate that $ICS_1^{\rho,\gamma}tree$ has advantage in predictive accuracy over other tree methods. Besides, the values of MIE increase with the increase of the right censoring rate of survival data, which means the higher right censoring rate lead to lower predictive accuracy. These results are consistent with those provided by IBS.

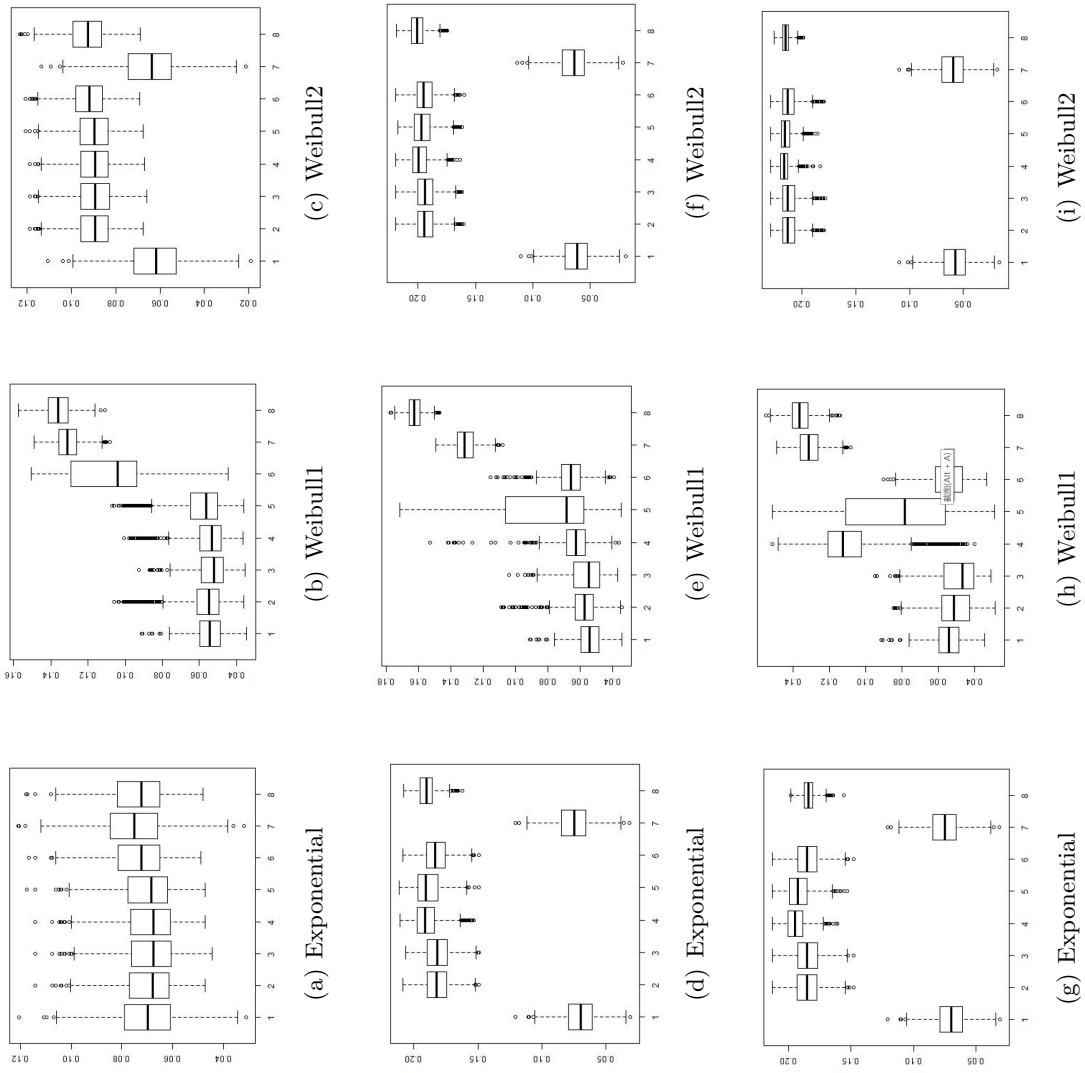
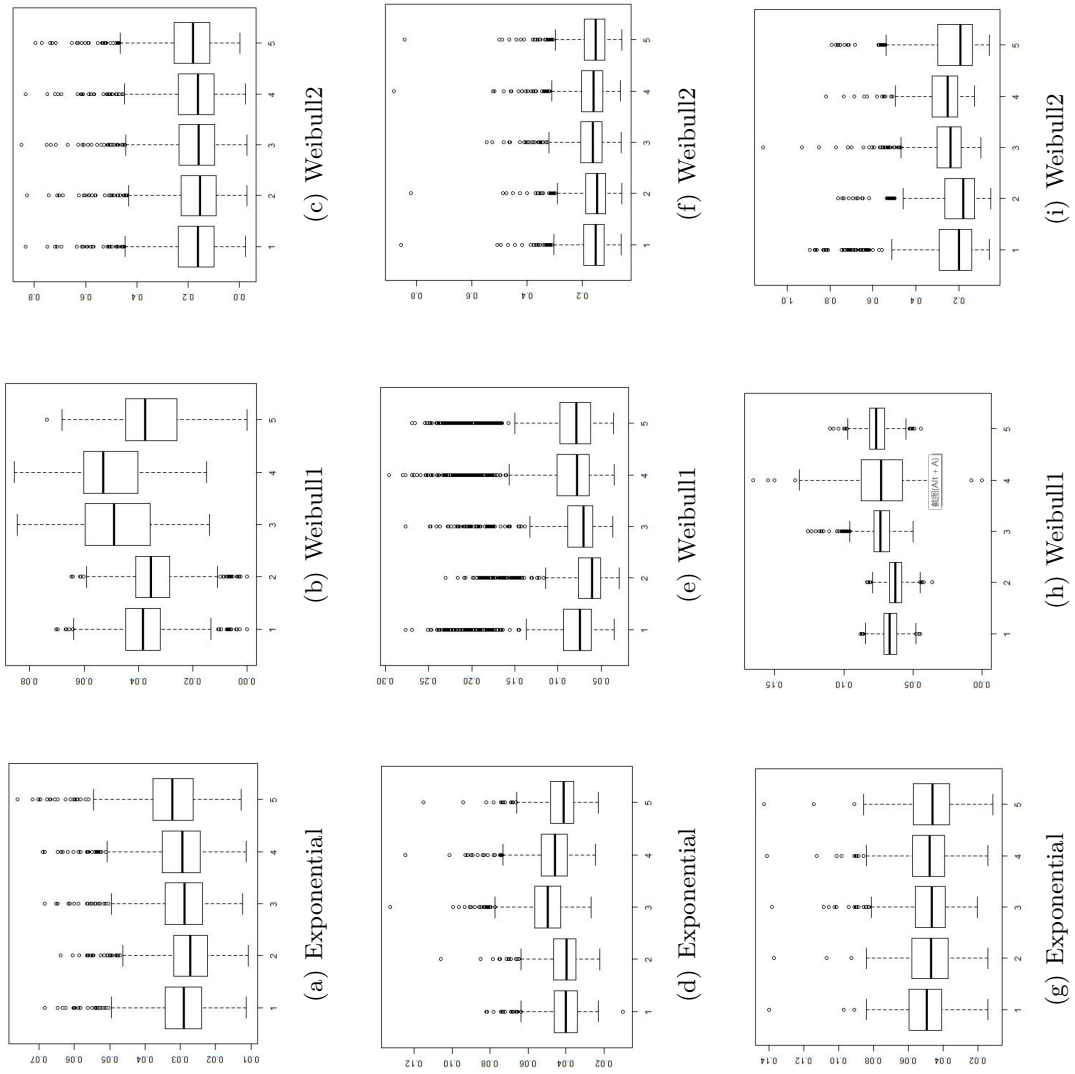


Figure 4.5: IBS boxplots with sample size $N = 200$. First row corresponds to 0% right censoring rate, second row corresponds to 20% right censoring rate and third row corresponds to 40% right censoring rate.



(a) Exponential (b) Weibull1 (c) Weibull2

(d) Exponential (e) Weibull1 (f) Weibull2

(g) Exponential (h) Weibull1 (i) Weibull2

Figure 4.6: *MIE* boxplots with sample size $N = 200$. First row corresponds to 0% right censoring rate, second row corresponds to 20% right censoring rate and third row corresponds to 40% right censoring rate.

4.4 Computational experiments for $ICS_1^{\rho,\gamma}forests$

In this section we discuss the predicted performance of $ICS_1^{\rho,\gamma}forests$ for interval-censored data. The setting of the simulated data is a complex structure data which is the same to the one described in Subsection 4.3.2.

Here, the number B of $ICS_1^{\rho,\gamma}tree$ is 100. The size of the subset of the covariates is $\sqrt{m}$ where m is the number of the covariate. We consider two scenarios: in the first scenario, the size of the dataset is 600; in the second scenario the size of dataset is 1500. The dataset is divided into training set and test set by ratio 2:1, i.e., in the first scenario, the size of test set is 200; and in the second scenario, the size of test is 500. We use the training set to construct $ICS_1^{\rho,\gamma}forests$ and evaluate their predicted performance in the test data. The boxplots of IBS are obtained by executing 1000 trials on the test data. We compare $ICS_1^{\rho,\gamma}forests$ for different values of (ρ,γ) to $ICcforest$ proposed in [249]. The results are reported in Table 4.15. We utilize $(\rho,\gamma) = (0.05,\ 0.05)$, $(\rho,\gamma) = (0.41,\ 0.03)$ and $(\rho,\gamma) = (0.01,\ 0.41)$ for $ICS_1^{\rho,\gamma}forests$. These values are the recommended value as discussed in Subsection 4.3.1. In addition to $ICcforest$, Proportional Hazards model for interval-censored data ($ICPH$) and $ICS_1^{0.05,0.05}tree$ are also utilized for comparison.

Table 4.15: Numbering of the various methods in Figures 4.7 and 4.8

No.	model	No.	model	No.	model
1	$ICcforest$	2	$ICS_1^{0.05,0.05}forests$	3	$ICS_1^{0.41,0.03}forests$
4	$ICS_1^{0.01,0.41}forests$	5	TPH	6	$ICS_1^{0.05,0.05}tree$

The computational results are shown in Figures 4.7 and 4.8. The predictive accuracy of $ICS_1^{\rho,\gamma}forests$ is comparable $ICcforest$, and in some cases better than $ICcforest$ in this setting. In fact, when $(\rho,\gamma) = (0.05,\ 0.05)$, $ICS_1^{\rho,\gamma}forests$ outperforms $ICcforest$ in cases of Exponential distribution and Weibull distribution with decreasing hazards, while when $(\rho,\gamma) = (0.41,\ 0.03)$, $ICS_1^{\rho,\gamma}forests$ has the lower value of IBS than $ICcforest$ in all the cases. The predictive performance of $ICS_1^{\gamma}forests$ and $ICcforest$ is better than $ICS_1^{0.05,\ 0.05}tree$ in cases of Exponential distribution and Weibull distribution with decreasing hazards. From the experiments we note that the tree and ensemble methods are all better than Proportional Hazards model TPH. A possible reason is that the Proportional Hazards model is more suitable for linear data rather than the complex structure data we used in Subsection 4.3.2. Moreover, from Figures 4.5 and 4.6, we note that the value of IBS increases rapidly when the right censoring rate increases. What's more, the predictive performances improves with the increase of sample size. In conclusion, we can assert that the predictive ability has been slightly improved by $ICS_1^{\rho,\gamma}forests$ compared to $ICcforest$

and $ICS_1^{\rho,\gamma}tree$.

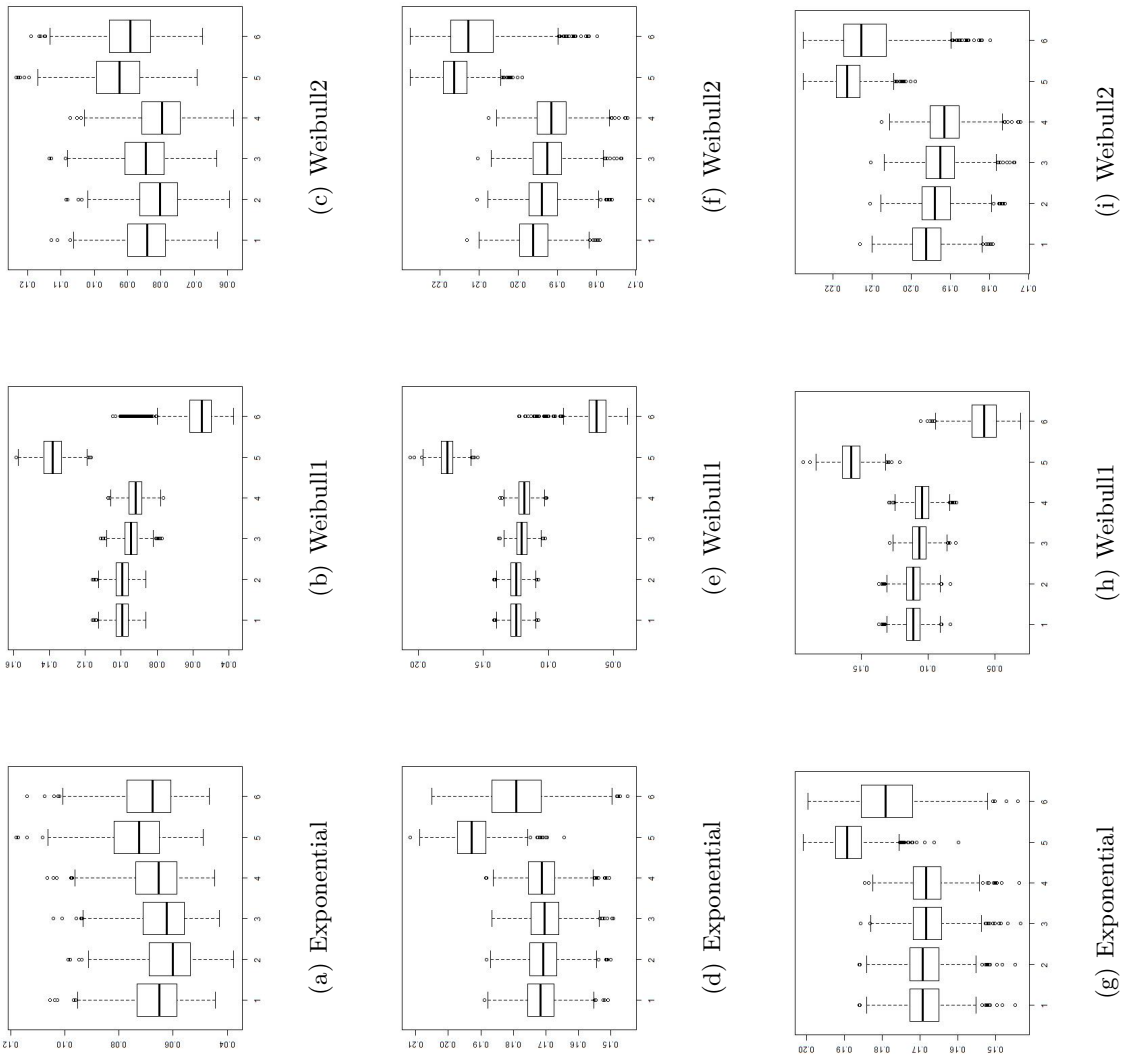
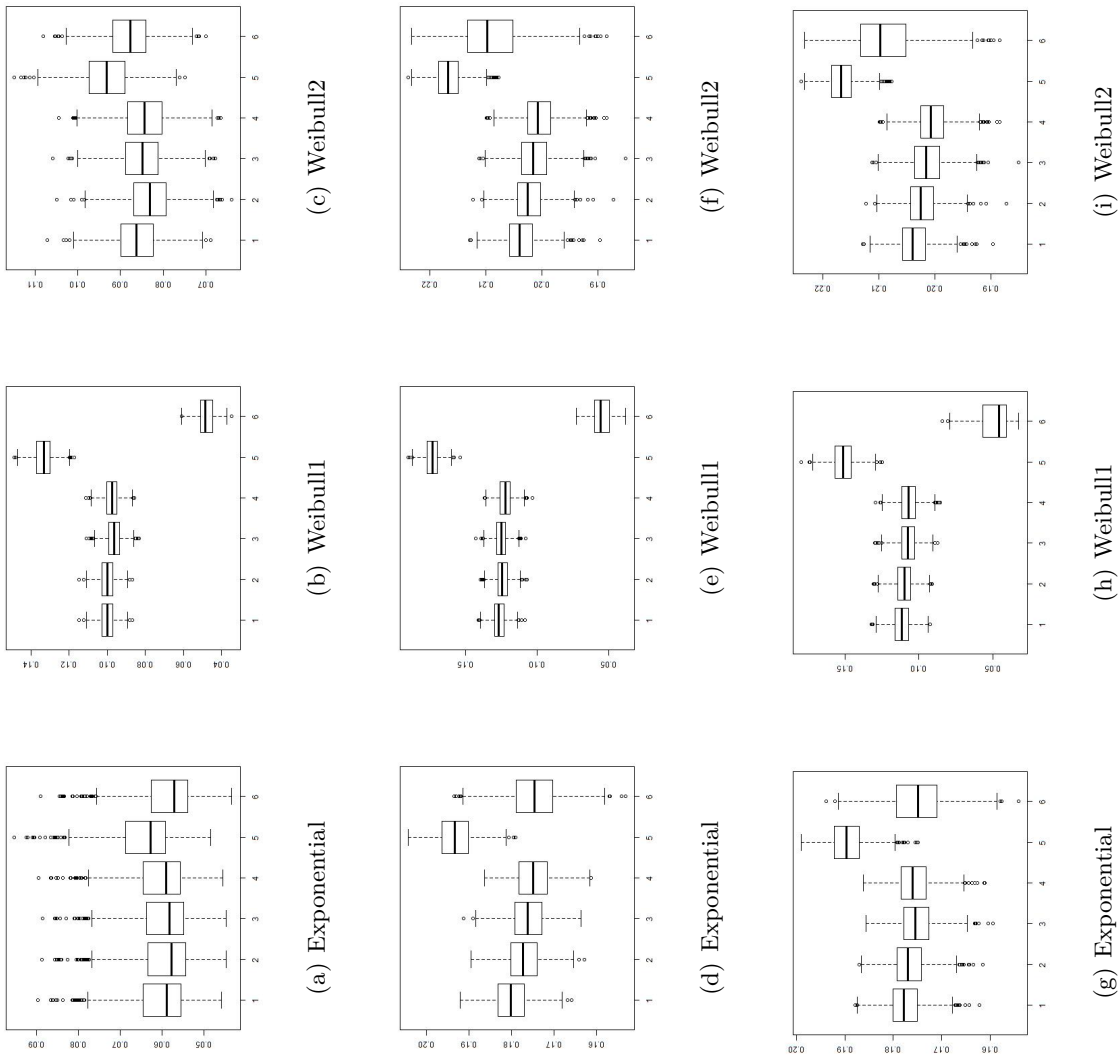


Figure 4.7: IBS boxplots with size 200 of the test data. First row corresponds to the result of 0% right censoring rate, second row corresponds to 20% right censoring rate and third row corresps to 40% right censoring rate.



(a) Exponential (b) Weibull1 (c) Weibull2

(d) Exponential (e) Weibull1 (f) Weibull2

(g) Exponential (h) Weibull1 (i) Weibull2

Figure 4.8: IBS boxplots with size $N = 500$ of the test data. First row corresponds to the result of 0% right censoring rate, second row corresponds to 20% right censoring rate and third row corresponds to 40% right censoring rate.

Chapter 5

Applications

In this chapter we present applications of the survival tree models we proposed in a variety of disciplines including medicine and industry. What's more, an application in the economic field is discussed, where we propose a survival-based model for the credit scores of small and medium-sized enterprises (SMEs) using survival techniques and machine learning methods. Computational experiments are also carried out.

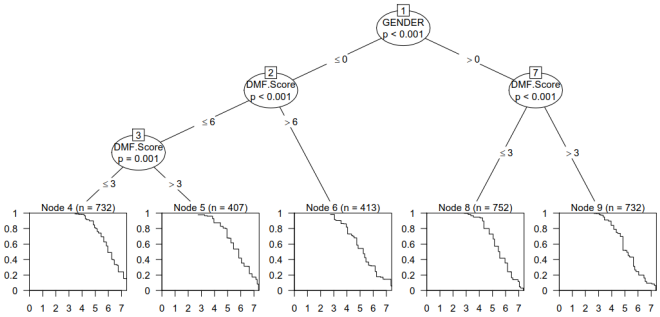
5.1 Survival tree model for emergence of permanent teeth

A real interval-censored dataset is provided by the Signal *Tandmobiel*® study [231]. The study carried out a longitudinal survey of 4468 primary school pupils in Flanders (North of Belgium) regarding the time of emergence of permanent teeth and caries development from 1996 to 2001. The dataset is available in the R package bayesSurv containing the information on 28 teeth. The dataset is interval-censored because the event of interest in this study is the emergence of the permanent teeth. Annual examinations for each child only record the interval to which the time of emergence belongs, rather than the exact time of emergence of permanent teeth. The time to the emergence of permanent tooth is considered as the survival time. We shift the time origin to 5 years of age, as proposed in [149]. Covariates include

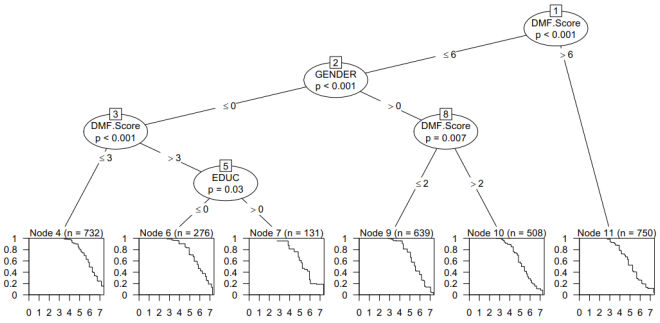
- geographical factor (province of residence);
- gender;
- use of fluoride;

- type of education system;
- starting age of brushing teeth;
- total number of deciduous teeth extracted due to orthodontic reasons (denoted by BAD);
- total number of decayed,filled or missing deciduous teeth due to caries (denoted by DMF.Score).

More details about the dataset can be seen in [231, 132].



(a) $ICS_1^{0.05,0.05}tree$ and $ICS_2^{0.69,0.5}tree$

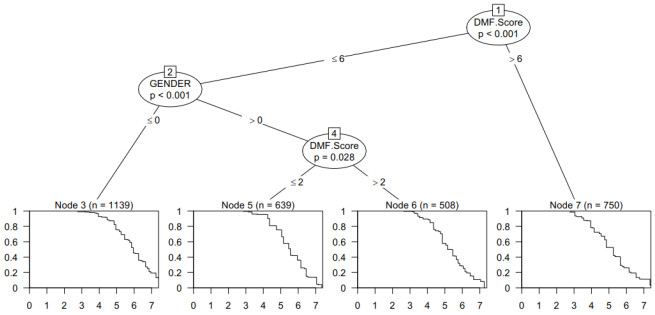


(b) $ICS_1^{0.69,0.03}tree$ and $ICS_2^{0.69,0.03}tree$

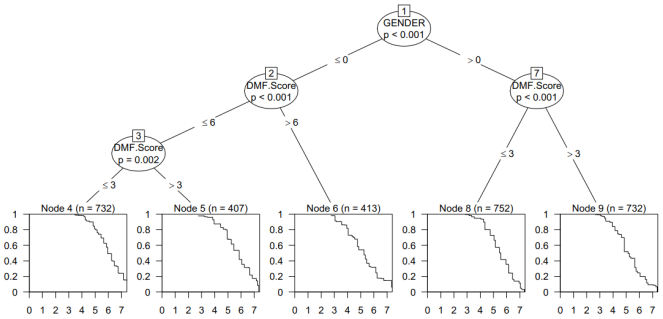
Figure 5.1: Interval-censored trees with recommended hyperparameter values are used for analyzing the significant covariates associated with the time of emergence of the permanent first upper right premolar (tooth 14).

Figures 5.1 and 5.2 show how the instances are classified into several categories by the tree models and the important covariates selected for the survival time. We conclude that both gender and DMF.Score are more important than other covariates. We utilize

$(\rho, \gamma) = (0.05,\ 0.05)$ for $ICS_1^{\rho,\gamma}tree$, $(\rho, \gamma) = (0.69,\ 0.5)$ for $ICS_2^{\rho,\gamma}tree$ and $(\rho, \gamma) = (0.01,\ 0.01)$ for $ICG^{\rho,\lambda}tree$. These hyperparameter values are the recommended values as discussed in Subsection 4.3.1. For $ICS_1^{0.05,0.05}$ and $ICS_2^{0.69,0.5}$, gender is more important than DMF.Score, which is consistent with the results from $ICtree$. Moreover, we also consider $(\rho, \gamma) = (0.69,\ 0.03)$ for both $ICS_1^{\rho,\gamma}tree$ and $ICS_2^{\rho,\gamma}tree$. Here gender is a critical factor only when DMF.Score is less than or equal to 6 in this case.



(a) $ICWtree$



(b) $ICG^{0.01,0.01}tree$

Figure 5.2: Interval-censored trees with recommended hyperparameter values are used for analyzing the significant covariates associated with the time of emergence of the permanent first upper right premolar (tooth 14).

Based on the survival curves in terminal nodes for the $ICS_1^{0.69,0.03}tree$ (Figure 5.1), we conclude that the time to emergence of permanent tooth tends to occur at a younger age for girls (Gender = 1) than boys (Gender = 0). Additionally, children with a higher total number of decayed or missing deciduous teeth due to caries (DMF.Score > 6) have earlier emergence of their permanent teeth compared to others. These results are also applicable for the $ICWtree$ and $ICG^{0.01,0.01}tree$, as can be seen from Figure 5.2. $ICWtree$ grows with fewer splits on DMF.Score in the right branch and the number of terminal nodes is

also reduced. As far as $ICG^{0.01,0.01}tree$ is concerned, DMF.Score=4 is the splitting rule of root node which is quite different from the others.

The results reported here are relative to the permanent second premolar (tooth 15-45 in European dental notation), the results of other teeth are similar. The size of the dataset is N=3036. We split it into two parts by 2:1 as a training set and a test set respectively. Algorithm 7 is applied to the training set for hyperparameter tuning using the measurement

$$\phi(\cdot) = d_{mean} = \frac{\sum_{j=1}^{N} d_j}{\sum_{j=1}^{N} I(\tilde{T}_j \notin (L_j, R_j])},$$

as criterion proposed in [249], where $\tilde{T}_j$ is the predicted time for the jth observation, the true survival time T_j is within the observed interval $(L_j, R_j]$, and

$$d_j = \begin{cases} \min\{|\tilde{T}_j - L_j|, |\tilde{T}_j - R_j|\}, & \text{if } \tilde{T}_j \notin (L_j, R_j]; \\ 0, & \text{if } \tilde{T}_j \in (L_j, R_j]. \end{cases}$$

In a word, d_{mean} represents the average absolute distance away from the observed intervals when the predicted time fall outside of observed intervals. The smaller d_{mean} is, the better the predictive performance is.

The recommended hyperparameter values (ρ, γ) are (0.51, 0.03),(0.43, 0.01) and (0.05, 0.21) for $ICS_1^{\rho,\gamma}tree$, (0.5, 0.03) and (0.69, 0.21) for $ICS_2^{\rho,\gamma}tree$ under different cases of right censoring rate. These values differ significantly from the values obtained with the synthetic dataset in Chapter 4. A possible explanation is the right censoring rate is very high (over 58%), and larger than the one discussed in previous simulations. We selected (ρ, γ) =(0.05, 0.21), this is in line with the recommended value when right censoring rate is over 40% in Subsection 4.3.1.

Table 5.1: Evaluation on permanent 2^{nd} premolar data in Signal $Tandmobiel^{®}$ study

Models	$Hyperparameter$	15 (58%)	25 (57%)	35 (59%)	45 (59%)
	(ρ, γ)	$\bar{d}_{mean}$	$\bar{d}_{mean}$	$\bar{d}_{mean}$	$\bar{d}_{mean}$
$ICtree$	-	0.6696	0.6985	0.6797	0.6803
$ICPH$	-	0.6982	453.35*	0.7154	0.7069
$ICS_1^{\rho,\gamma}tree$	(0.43,0.01)	0.6673*	0.6798*	0.6772*	0.6724*
	(0.51,0.03)	0.6673*	0.6806*	0.6772*	0.6593*
	(0.7,0.02)	0.6689*	0.6773*	0.6760*	0.6615*
	(0.03,0.09)	0.6712	0.6958	0.6786*	0.6898
	(0.05,0.21)	0.6673*	0.6932*	0.6785*	0.6806
$ICWtree$	-	0.7066	0.6628*	0.6410*	0.6442*
$ICG^{\rho,\gamma}tree$	(0.03,0.03)	0.6675*	0.6793*	0.6797	0.6581*
	(0.27,0.01)	0.6696	0.6760*	0.6789*	0.6581*
	(0.01,0.01)	0.6696	0.6769*	0.6797*	0.6581*
$ICS_2^{\rho,\gamma}tree$	(0.51,0.03)	0.6699	0.6819*	0.6771*	0.6593*
	(0.71,0.02)	0.6718	0.6735*	0.6760*	0.6615*
	(0.69,0.21)	0.6731	0.6846*	0.6789*	0.6769*

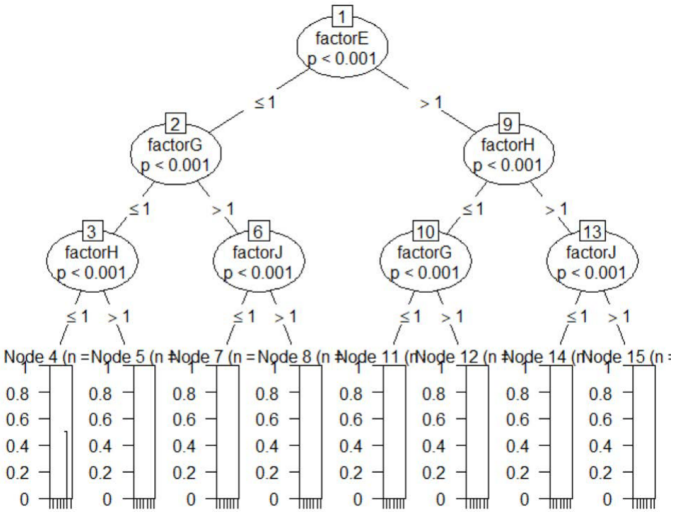
To evaluate the predictive performance of tree models with the selected hyperparameters, a 5-fold cross validation is carried out in the test set. The results are reported in Table 5.1. In this table, 57%, 58% and 59% correspond to the right censoring rate, $\bar{d}_{mean}$ denotes mean value of d_{mean} obtained by 5-fold cross validation, and '*' corresponds to smaller values of $\bar{d}_{mean}$ of the proposed tree methods than $ICtree$. As can be seen from Table 5.1, $ICS_1^{\rho,\gamma}tree$ shows the best predictive accuracy. Even when the hyperparameters are not the optimal ones, the proposed tree methods can adapt well to the data and demonstrate good performance in prediction.

We note that the performance of $ICWtree$ and $ICG^{\rho,\lambda}tree$ is also better than $ICtree$ for this real dataset, but are still inferior to $ICS_1^{\rho,\gamma}tree$. More important, the value of $\bar{d}_{mean}$ from the various trees is smaller than the one from $ICPH$ method introduced in Subsection 4.3.2. An unexpected particularly bad result for $ICPH$ (we mark it with boldface) is also obtained. A possible explanation is the following: if the jth interval is a finite interval, meaning that the observation value falling in the interval is not right censored, and one of the jth predicted values of the $ICPH$ is far from the endpoints of the jth interval but falls in the right censoring area, then the predicted time is not within the jth interval, resulting in a very large distance between them. This leads to a terribly bad prediction for $ICPH$.

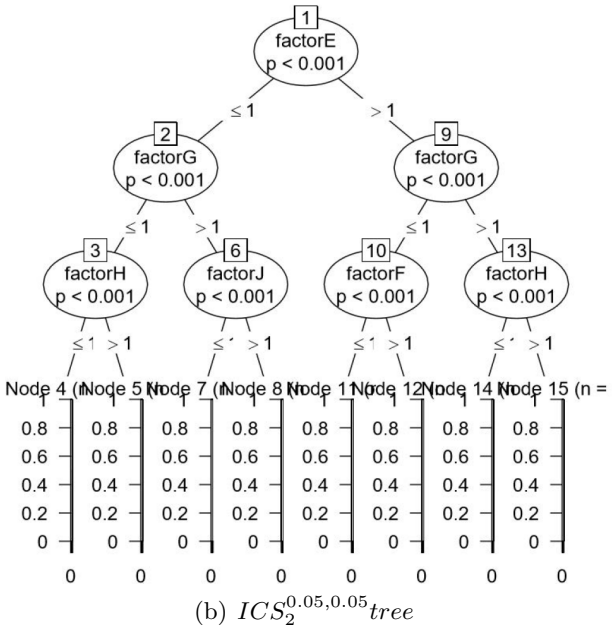
5.2 Survival tree for heat exchanger life

An experiment was designed to study how ten two-level factors(A - H, J, K) affect a heat exchanger's reliability [99]. The experiment utilized the 12 run Plackett-Burman designs. The event of interest is the failure of the unit. The design and parts of interval-censored data are shown in Table 5.2. "A unit fails when it develops a tube wall crack with lifetime measured in ($\times$ 100) cycles" [99]. For example, if the unit's lifetime is 3 (x 100) cycles, it would mean that the unit fails after 300 cycles. Each run was inspected after cycles 42, 56.5, 71, 82, 93.5, 105, 116 and was stopped after 128 cycles. The exact time of occurrence of the event of interest is unknown; it is only known within the observed intervals. Thus the dataset is interval-censored, and the observed time serves as the endpoint of the censored intervals. The sample size is 96 and the right censoring rate 6%.

We utilize $ICS_1^{0.5,0.03}tree$ to identify the important effects on survival time among 10 factors (Figure 5.3). Factor E is much more important than other factors. Factor G and H also influence the survival time and reliability. The results consistent with the conclusion in [99] where factor E appears to be important, and factor G and factor H are also important when considering "potentially important interactions between E and the other nine factors". For $ICS_1^{0.05,0.05}tree$ some slight difference in the effects of various factors can be observed.



(a) $ICS_1^{0.5,0.03}tree$



(b) $ICS_2^{0.05,0.05}tree$

Figure 5.3: Interval-censored trees for heat exchanger life data.

Table 5.2: Design and data for the heat exchanger experiment

Run	*factor*										data
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*G*	*K*	
1	1	1	1	1	1	1	1	1	1	1	(93.5,105]
2	1	1	1	1	1	2	2	2	2	2	(42,56.5]
3	1	1	2	2	2	1	1	2	2	2	$(128, +\infty]$
4	1	2	1	2	2	2	2	1	1	2	(56.5,71]
5	1	2	2	1	2	1	2	1	2	1	(56.5,71]
6	1	2	2	2	1	2	1	2	1	1	(0,42]
7	2	1	2	2	1	2	2	1	2	1	(56.5,71]
8	2	1	2	1	2	2	1	1	1	2	(42,56.5]
9	2	2	2	1	1	1	2	2	1	1	(82,93.5]
10	2	2	2	1	1	1	2	2	1	2	(82,93.5]
11	2	2	1	2	1	1	1	1	2	2	(82,93.5]
12	2	2	1	1	2	2	1	2	2	1	(42,56.5]

5.3 A survival-based model for credit risk assessment of SMEs

5.3.1 Background

Credit risk assessment of an enterprise is essential for a bank to make a reasonable credit decision. Thus the study of reliable credit risk modelling is important since it provides a useful tool for risk management and improves the efficiency of capital allocation. Credit risk assessment systems for small and medium-sized enterprises (SMEs) attract great attention because the information of SMEs are less opaque. For example, there is a lack of a credit history, a shortage of valuable collateral and a lack of reliable credit ratings, etc.

Various credit risk models for SMEs have been proposed in recent years. A comprehensive review of credit risk modellings for SMEs was provided by Lin [154]. The evaluation of the discriminative power of credit rating systems was obtained by AUC, the area below the Receiver Operating Characteristic (ROC) curve [71]. Statistical tools such as histograms, Q-Q graphs, goodness-of-fit tests have been applied to evaluate the important credit risk factors for SMEs [67]. Moreover, Logistic regression is a common technique in SMEs credit risk assessment and the probability of default can also be calculated by it [5].

Recently, Machine Learning methods have been developed as alternatives to logistic regression. Zhao [253] showed that Decision Trees provide a good classification method for the credit rating of SMEs. Ko [130] combined multivariate adaptive regression splines with regression tree to predict the credit performance of SMEs. Lang [142] applied Support Vector Machine (SVM) to accurately divide SMEs into default or non-default. Artificial Neural Network (ANN) has been also applied for credit risk estimation [126]. Yang [248] evaluated the credit risk of SMEs based on Random Forest (RF). The credit rating of

SMEs was predicted by RF with higher accuracy than logistic regression [235].

What's more, Albaz [4] utilized Principal Component Analysis (PCA) to avoid multi-collinearity of predictor variable and improves prediction accuracy of risk level for SMEs in Saudi Arabia. A Data Envelopment Analysis (DEA)-cluster cross model was proposed to assess the relative credit scores and credit rating of SMEs [213], which combined the DEA cross model with cluster analysis.

Popular approaches of survival analysis are utilized to model credit risk of SMEs. Narain first used techniques of survival analysis for credit scoring model [174]. This idea was further developed by Thomas et al. [221], and they proved that competing risks approach of survival analysis was useful to deal both with defaults and early repayment. M. Stepaova. et.al. [210] applied the Cox proportional hazards model to personal loan data and proposed a new method for classification. A nonparametric approach based on Random Survival Forests (RSF) was proposed for credit risk measurement for SMEs in [76]. Log-logistic hazard models were estimated separately for micro-enterprises and SMEs to study in detail the survival of newly established manufacturing firms by [105]. Several parametric models (Exponential, Log-normal, Weibull, Log-logistic, Gompertz, Gamma distribution) and non-parametric models (Kaplan-Meier, Nelson-Aalen) were also used in credit risk assessment based on progressive right censored data [116, 252].

A parametric Weibull mixture cure model was considered in [160] to analyze the time to default on a personal loan portfolio in the presence of disproportionate hazard rate and it was evidenced that this mixture cure model was appropriate for non-proportional scenarios. In [63], the default time was predicted based on 10 actual credit dataset by accelerated failure time models, Cox proportional hazards regression models and Cox-PH with splines in the hazard function model respectively, where default was the event of interest and mature or early repayment were labeled as censored data.

The existing literatures consider default as the event of interest and the data is right-censored, which occurs when a SME does not experiment default at the moment of data gathering. The censored cases include early repayment, repayment on time or the case that the predefined end date is not reached. However, the credit rating of SMEs with early repayment is much higher than those repaying loans on time. Therefore, banks should also grant more loan quotas and more favorable policies to such enterprises when making credit decisions. Moreover, if a bank does risk evaluation and makes a loan decision for the SMEs without bank loan experience, the bank can only predict whether the enterprise can repay the loan before the predefined time, i.e., the bank only can predict default or not for a SME. We consider this case and assume that the repayment occurs within an interval and the intervals are given by TOPSIS method [113]. Thus the dataset is Case II interval-censored. The techniques of interval-censored data are applied to predict the probability of early

repayment and repayment on time, respectively. Based on the existing literatures, three key issues need to be considered in the credit risk assessment system of SMEs:

- selection of features for evaluation;
- prediction of credit rating;
- prediction of the time of loan repayment and probability of loan repayment varying with time.

In [43], a model for a credit risk assessment system for SMEs to address these three issues is proposed. The results presented in Subsection 5.3 are based on [43].

5.3.2 Characteristics of the original dataset

As far as credit policies for Chinese SMEs are concerned, banks usually provide loans to competitive enterprises with stable supply and sales channels and offer preferential interest rates to enterprises with high reputation and low credit risk. Various pieces of information about a SME are required by banks to evaluate the credit risk based on its strength and reputation.

We estimate the credit risk of 425 Chinese SMEs based on the dataset available in CUMCM-2020[1]. Of those, 123 SMEs have loans with the bank, while other 302 SMEs have not received yet a loan from a bank and just applied for the loan. The information available includes:

- the enterprise code,
- enterprise name,
- the information of the purchase invoices,
- the information of the sales invoices.

For the 123 SMEs that received a bank loan, the information about "Credit Rating" and "Default Or Not" is available. For the other 302 SMEs, since the bank has no credit records for them, the information about "Credit Rating" and "Default Or Not" is not known.

The purchase invoices information specifically show the details of transactions of raw materials between 425 Chinese SMEs and 13,972 upstream companies for 3 years (from 01/2017 to 12/2019). The total number of purchase invoices is 540,836. The sales invoices

[1] http://en.mcm.edu.cn/html_en/node/c8a5ea02d468b1a8e7282ce0cef1dd7e.html

for the 425 SMEs show the details of transactions of goods sold to 24,450 downstream companies over a 3-year period (from 01/2017 to 12/2019). The total number of sales invoices is 555,176. The available information for the invoices includes:

1. the invoice number,
2. issued date of invoice,
3. validity of invoice (valid/invalid),
4. ID of upstream companies (i.e.,supplier's ID)/ID of downstream companies (i.e.,buyer's ID)
5. quantities of goods,
6. transactions tax,
7. total agreed prices for products and tax.

The original information provided in the dataset is too scattered and the operating situation and strength of a enterprise is hardly obtained based on the current information. So a data cleaning process is necessary to extract useful features.

5.3.3 Feature extraction

In our study we consider more than 1 million of invoices for the 425 SMEs over a period of 3 years. However, some invoices are invalid. Since the ratio of invalid invoices in the dataset is less than 1.84%, we decide to remove them from the original dataset. To model credit risk assessment we extract features corresponding to the following aspects: enterprise size, the state of the business operation, the stability of channels of supply and demand. At the end of this phase, 23 variables have been extracted as listed in Table 5.3.

We utilize the total agreed prices for products and tax for sales to characterize the scale of an enterprise. Moreover, the capital flow which is used to buy raw materials also indicates the size of an enterprise.

For SMEs, sales performance are quite different between off-season and peak season. And they submit financial statements on a quarterly basis according to financial policy in China. Therefore, considering each MSE, we calculate the average total prices for sales and average total prices with tax for purchasing raw materials on a quarterly basis within a span of 3 years, according to the information from sale invoices and purchase invoices respectively. More specifically, let

- S_{ij} be the average total prices for sales for the ith enterprise in the jth quarter within 3 years (Unit: RMB), $i = 1, \ldots, 425$, $j = 1, \ldots, 4$;
- P_{ij} be the average amount with tax for purchasing raw materials for the ith enterprise in the jth quarter within 3 years (Unit: RMB), $i = 1, \ldots, 425$, $j = 1, \ldots, 4$.

The operating revenue is an effective indicator of the operating performance of an enterprise. The annual growth rate of operation revenue shows the development trend. Two following features are also considered:

- $rev_{ij} = S_{ij} - P_{ij}$: the average operation revenue of the ith enterprise in the jth quarter within 3 years (Unit: RMB), $i = 1, \ldots, 425$, $j = 1, \ldots, 4$;
- rat_i: the annual growth rate of operation revenue, i.e., the ratio of the total operating revenue growth in current year to total quantity of the operating revenue of the previous year, $i = 1, \ldots, 425$.

Table 5.3: The 23 Extracted Features

Symbols	Meaning
classnum:	Credit rating
xiaoxiangav$j$:	Average sales amount in the jth quarter,j = 1,2,3,4
jinxiangav$j$:	Average purchasing amount with tax in the jth quarter,j = 1,2,3,4.
netprofit$j$:	Average operation revenue in the jth quarter,j = 1,2,3,4.
upflowyearnum$k$:	Number of upstream companies that have been cooperating for k years, k = 1,2,3.
upflowyearnum:	Total number of upstream companies that have cooperated within 3 years.
downflowyearnum$k$:	Number of downstream companies that have been cooperating for k years, k = 1,2,3.
downflowyearnum:	Total number of downstream companies that have cooperated within 3 years.
growrate:	Annual growth rate of operation revenue within 3 years.
default:	Default or not.

If a SME has a large number of upstream enterprises, then it will have stable supply chain for raw materials, while the number of downstream enterprises reflects the sale channel of a SME. The number of years of transactions between a SME and its upstream/-downstream enterprises can be a valuable index to describe the stability of the operation state of the SME. Thus the following features are also considered:

- up_{ij}: the number of upstream enterprises of the ith SME where j represents the number of years of transactions between the upstream enterprises and the ith SME, $i = 1, \ldots, 425$, $j = 1,\ 2,\ 3$;

- dw_{ij}: the number of downstream enterprises of the ith SME where j represents the number of years of transactions between the downstream enterprises and the ith SME, $i = 1, \ldots, 425$, $j = 1,\ 2,\ 3$.

5.3.4 Variable selection

The new dataset comprises 23 variables and the sample size is 425, with the values of the variables "default" and "classcum" being unknown for 320 SMEs. We first consider the variable "classcum", representing the meaning of credit rating, as the output variable in the credit rating prediction model. A subset of the remaining variables will be selected as input variables.

Variable selection aims to determine which subset of features should be used in the prediction model and which features are more significant in the prediction. This procedure improves the predictive accuracy and reduces the dimensionality of the data.

As stated in the previous chapters, Random Forest (RF) is a nonparametric method that has the advantage of analyzing and dealing with high dimensional non-linear data and it is also an efficient tool for variable selection. We apply RF to determine the top 20 most important variables. The hyperparameter values for RF are chosen as follows: the number of variables selected randomly is set to 7 and the number of classifiers in RF is 500. Figure 5.4 shows the most important top 20 variables as detected by RF based on the Mean Decrease Accuracy.

A 6-fold cross validation is repeated 10 times to search for peak of predictive accuracy for RF. Figure 5.5 shows that the average error of cross-validation is minimal when 9 variables are selected. Hence the top 9 variables in Figure 5.4 are considered as the input variable in credit rating - predicting models.

Next, we use variable "Default" as the output variable in a default-prediction model, and RF is utilized again to determine the most important variables. The outcome of a 6-fold cross validation repeated 10 times shows that the predictive model has the best performance when 10 input variables are selected.

5.3.5 Credit risk prediction

We utilize DRF to predict the output value "classcum" representing the meaning of credit rating. The results are reported in Table 5.4. The values of the hyperparameters for DRF

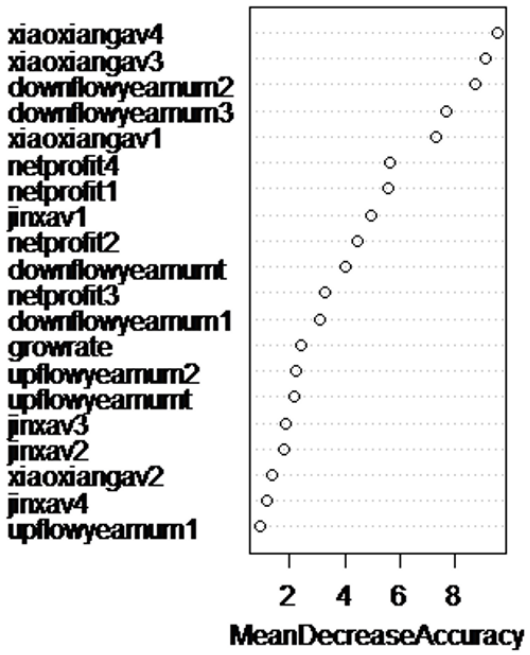


Figure 5.4: Importance Variable for credit rating-predicting Model

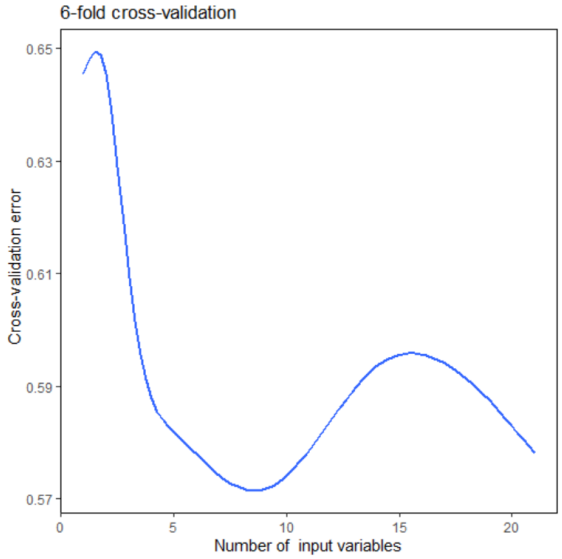


Figure 5.5: Relationship between the number of input variables and cross validation error

are the same as for RF in Subsection 5.3.3.

Table 5.4: Predictive results of credit rating

Credit Rating	A	B	C	D
Number of SMEs	74	140	41	47

We randomly split the entire dataset of 123 SMEs for which information on risk rating is available into 10 non-overlapping groups with approximately the same size. Then, 3 groups are randomly selected as a test set, and the remaining groups are considered as training set, on which the model is constructed. RF and SVM with radial basis kernel are also utilized for prediction. The predictive accuracy for DRF on the testing set is much higher than the one for DF and SVM, as shown by the confusion matrices in Table 5.5.

The performance of the RF and DRF is also evaluated using their ROC curves [78, 194], which are reported in Figures 5.6 - 5.8. For credit risk level A, B and C, the predictive accuracy of DRF is higher than that of RF. It can be observed that DRF outperforms the other two Machine Learning methods, as demonstrated by the confusion matrices and receiver operating characteristic (ROC) curves. Finally, DRF is applied to the default-prediction task, and the results are shown in Table 5.6.

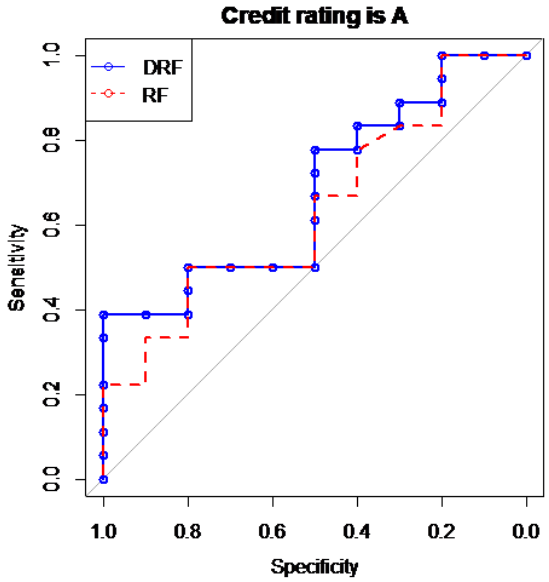


Figure 5.6: ROC Curve for the Random Forest and Double Random Forest models

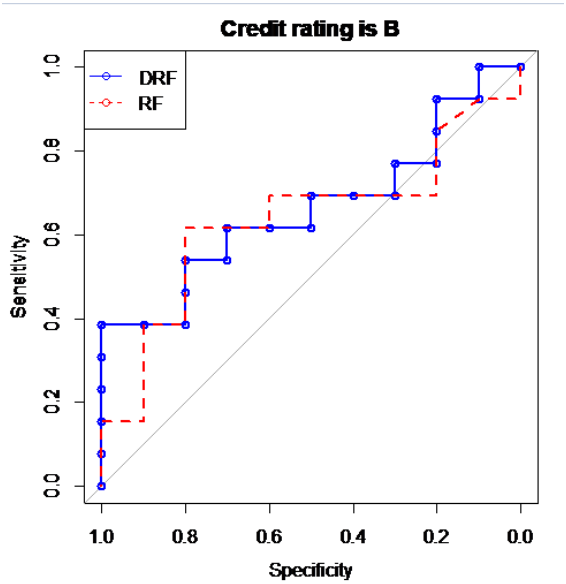


Figure 5.7: ROC Curve for the Random Forest and Double Random Forest models

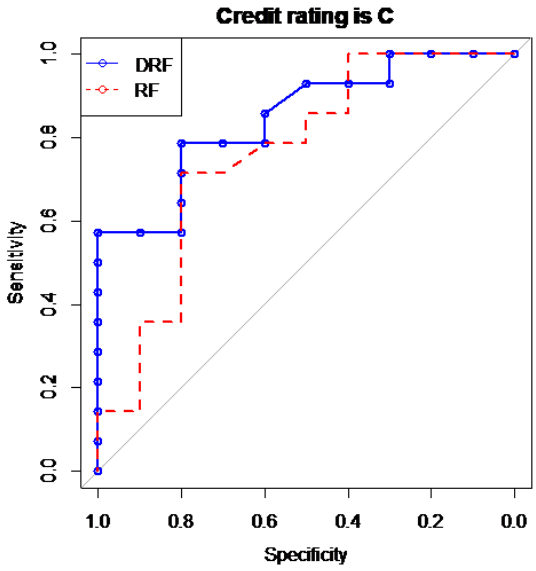


Figure 5.8: ROC Curve for the Random Forest and Double Random Forest models

Table 5.5: Confusion Matrix for Testing Data

Credit Rating	Predicted results by RF				Predicted results by DRF				Predicted results by SVM			
	A	*B*	*C*	*D*	*A*	*B*	*C*	*D*	*A*	*B*	*C*	*D*
True level A	2	4	4	0	2	5	3	0	2	1	9	0
True level B	0	9	8	1	0	8	9	1	2	2	16	0
True level C	2	4	5	2	2	4	5	2	1	1	9	0
True level D	0	0	11	3	0	0	8	6	0	0	12	0
Accuracy rate	34.55%				38.18%				24.07%			

Table 5.6: Predictive results for 'Default or Not'

Default or Not	Default	Not default
Number of SMEs	50	252

5.3.6 Credit risk assessment

Next we estimate how the probability of repaying the loan varies over time to assess the credit risk.

The loans were granted with terms of 365 days for 123 SMEs according to the loan contracts. A SME is considered as a defaulter if it had a period of 90 days without loan repayment. The information on repayment time is not precisely observed for all SMEs; instead, we assume that repayment time will fall into some interval which depends on three aspects: annual profit, stable channel of suppliers and sales, and growth rate of the enterprise. Here repayment time represents the time of event of interest (survival time), and the data is interval-censored. Therefore, the techniques of interval-censored data can be applied to assess credit risk.

Let T_i denote the time of repayment of loan for the ith SME, i = 1, ..., n. Let n_1 denote the number of SMEs with defaulting and n_2 denote the number of SMEs which are not defaulters ($n_1 + n_2 = n$). If the ith SME defaults, then T_i exceeds 455 days, leading to right censoring - a special case of interval censoring. If a SME is not a defaulter, we assume that it will repay the loan over a specific period of time according to its own operating conditions, i.e.,

$$T_i \in (L_i, R_i], i = 1, \ldots, n.$$

Observations are obtained at $\tau_0 < \tau_1 < \ldots < \tau_j < \ldots < \tau_m = 455$, where $\tau_j = \tau_{j-1} + \alpha_j$, α_j is randomly generated from a uniform distribution between 30 and 60, and $m = 7$. We first rank all the SMEs without defaulting in descending order by the TOPSIS method (Technique for Order Preference by Similarity to the Ideal Solution) introduced in [113], Then we divide the n_2 SMEs into m groups with repayment time T of the SMEs in jth group falling into the interval $(\tau_{j-1}, \tau_j]$, $j = 1, \ldots, m$. Next, we set $\tau_{m+1} = +\infty$. Therefore, T_i always belongs to some interval $(\tau_{j-1}, \tau_j]$, and $L_i = \tau_j$ and $R_i = \tau_{j+1}$.

We derive an estimation of the survival function $\hat{F}(t)$ by utilizing the approach discussed in the Subsection 2.3.2. The survival curves are drawn to describe the probability of repaying the loan over time. Moreover, we compare survival functions for different levels of credit rating A, B and C. SMEs with a credit rating D always default, and the credit risk is no longer under discussion. The survival curves are shown in Figures 5.9 - 5.11 which is based on the Case II interval-censored data of 425 SMEs according to groups of credit rating.

The survival curve of $\hat{S}(t)$ describes the relationship between of probability of repayment of the loan and time. It is evident that SMEs with A-level credit risk have a higher probability of repaying the loan earlier compared to those with B-level and C-level credit risk, as expected. For each SME, the probability of repayment of the loan can be obtained from the estimation of survival function. The bank can score each SME and make appropriate determination about credit strategies based on the credit risk assessment for it.

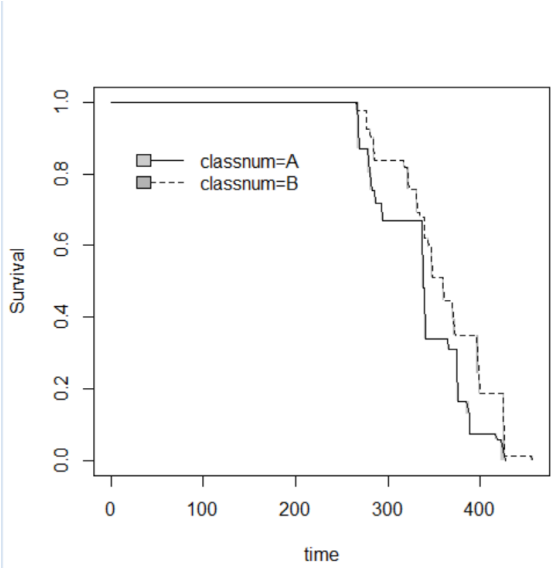


Figure 5.9: A comparison of survival curves between SMEs with A - level and B - level credit ratings.

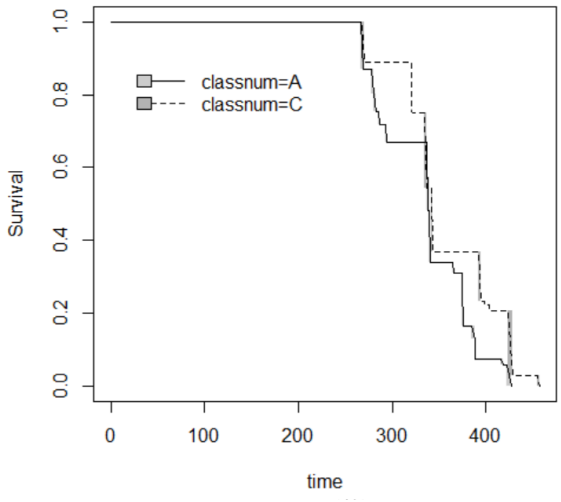


Figure 5.10: A comparison of survival curves between SMEs with A - level and C - level credit ratings.

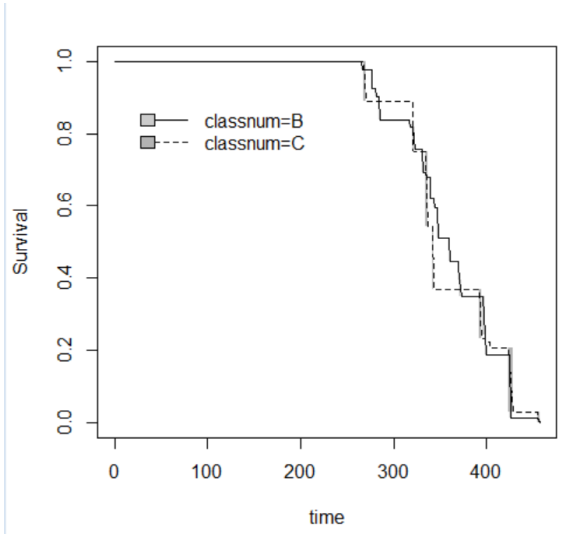


Figure 5.11: A comparison of survival curves between SMEs with B - level and C - level credit ratings.

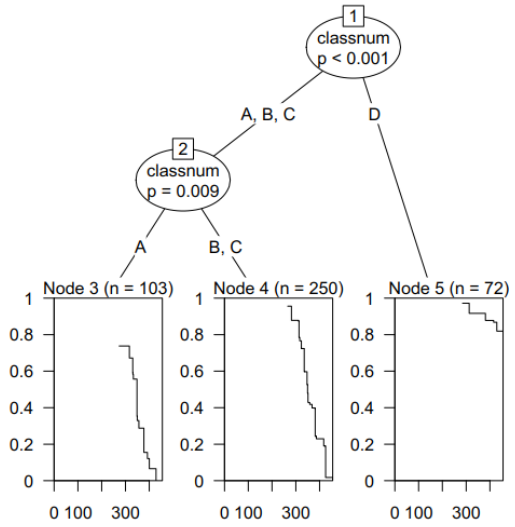


Figure 5.12: The assessment given by $ICS_1^{0.05,0.05}tree$

Survival tree can also be utilized to estimate the probability of repaying the loan over time. Figure 5.12 exhibits the credit risk of 425 SMEs drawn by the survival tree $ICS_1^{0.05,0.05}tree$. It is obviously inferred that credit rating is the most important factor for probability of repayment based on results of the survival tree. The probability of failing to repay their loans within 300 days for the SMEs with credit risk level A are the lowest, which is less 40%, while the probability for the SMEs with credit risk level B and C is around 60%, 80% for the SMEs with credit risk level D.

Conclusions and future research

Conclusions

The basic objective of this thesis was to tackle the main issues in Survival Analysis by utilizing approaches and theories from Machine Learning, especially tree methods and ensemble methods. To this end, the content involved two aspects: Survival Analysis and Machine Learning.

A brief introduction to Survival Analysis was reported at the beginning of the thesis, and the motivation and background in Survival Analysis were introduced. Then a comprehensive and extensive overview of literature on Survival Analysis was reported. We summarized the main issues that were concerned in Survival Analysis and explained why it was difficult to tackle the issues presented in this field. Next, some basic concepts and important data types in Survival Analysis were introduced in detail. In particular, survival functions describe the survival probability within a given time and hazard functions show survival hazard. However, censoring is a major feature of data in Survival Analysis. Statistical methods and models were applied earlier in Survival Analysis, and a variety of classic statistical methods were introduced. These methods are also useful tools to establish novel approaches based on Machine Learning for censored data. Some methods for the estimate of survival function were discussed. Nonparametric maximum likelihood estimate was a common method for various censored data. A series of parametric models were also introduced where the survival time was from a certain distribution. In addition, the well-known Cox-PH model was discussed. Then we discuss imputation method, which is a simple and useful technique and capable to tackle interval censored data by transforming interval censored data to right censored data.

A brief description of Machine Learning was also reported. We analyzed its historical background and the internal and extrinsic factors that promoted the rapid development of

Machine Learning. Some classical methods, especially tree methods and ensemble methods were outlined. What's more, the advantages of Machine Learning techniques for Survival Analysis were analyzed. An overview of a variety of modified Machine Learning techniques in Survival Analysis was outlined. Specifically, survival trees and survival ensembles were introduced in detail.

Next, the main results of our research were presented in detail. Novel survival trees for interval censored data have been constructed, where a new test statistics of GLRT was proposed and applied to conditional inference frame. The new survival trees were referred as $ICS_1^{\rho,\gamma}tree$ and $ICS_2^{\rho,\gamma}tree$ where ρ and γ are hyper-parameters. In various simulations, hyper-parameter tuning was explored, and we showed that the predictive power of $ICS_1^{\rho,\gamma}tree$ was improved by hyper-parameter tuning.

The optimal hyper-parameter values depend on the distribution and the right censoring rate of data. However, we determined some recommended values for real data according to the right censoring rate: the value $(\rho, \gamma) = (0.05, 0.05)$ was considered as default hyper-parameter value for lower right censoring rate ($\leq 40\%$), $(\rho, \gamma) = (0.51, 0.03)$ was for higher right censoring rate ($> 40\%$). The results of simulation show that $ICS_1^{\rho,\gamma}tree$ has an advantages in predictive accuracy with the recommended hyper-parameter values.

Although in some cases the performance of $ICWtree$ in predictive accuracy beats that of the other tree methods in tree-structured data without right censoring occurring, $ICS_1^{\rho,\gamma}tree$ was still more stable.

It is known that an ensemble method has powerful predictive performance due to its lower variance. Thus we considered $ICS_1^{\rho,\gamma}tree$ as base learners to construct a survival ensemble method referred as $ICS_1^{\rho,\gamma}forests$ for interval censored data. This survival ensemble method utilizes randomly selected subset of covariates and bootstrapped data for splitting and resorts to the average observation weights from Quantile Regression Forests to estimate the survival function.

The performance of $ICS_1^{\rho,\gamma}forests$ compared to $ICcforest$, Proportional Hazards model TPH and $ICS_1^{\rho,\gamma}tree$ was evaluated via an extensive simulation. The results showed that the predictive performance of $ICS_1^{\rho,\gamma}forests$ had a slight advantage compared to $ICcforest$ with the recommended hyper-parameter values of ρ and γ and outperformed the Proportional Hazards model TPH. What's more, the predictive effect is improved when the sample size increases. The code in R for $ICS_1^{\rho,\gamma}tree$ and $ICS_1^{\rho,\gamma}forests$ algorithm is provided in Appendix A.

Finally, various tree methods, ensemble methods and other machine learning techniques were utilized to tackle the practical problems in Survival Analysis. The applications involved in medicine, industrial and economic fields.

Future research

In this thesis, the proposed survival tree for interval censored data are based on conditional inference frame. However, from the point of view of censored data type and Machine Learning techniques, a lots of interesting work remain to be considered:

- construction of novel survival tree methods for more complex censored data, for example, double censored survival data, Case K interval censored data;
- construction of new random rotation survival forest for interval censored data or more more complex censored data;
- use of other Machine Learning methods for interval censored data or more complex censored data, for example, improved Support Vector Machine for double censored survival data.

Appendix A

R codes

The main R codes of the new tree algorithms are shown as follows.

```
1
2  library(partykit);
3  library(grid);
4  library(libcoin);
5  library(mvtnorm);
6 library(interval)
7  library(survival);
8 library(MLEcens);
9 library(perm);
10 library(Icens)###if (!requireNamespace("BiocManager", quietly = TRUE))
11                       ### install.packages("BiocManager")
12
13                        ###BiocManager::install("Icens")
14 library(LTRCtrees)
15 library(ICcforest)
16 library(FAdist)
17 library(icenReg)
18 library(Rcpp)
19 library(coda)
20 library(coin)
21
22 #run ctree.R
23 #run extree.R
24
25 testtype = "Bonferroni"
26 teststat = "quadratic"
27  splitstat = "quadratic"
28 splittest = FALSE
```

```
pargs = GenzBretz()
 nmax = c("yx" = Inf, "z" = Inf)
alpha = 0.05;
 mincriterion = 1 - alpha;
logmincriterion = log(mincriterion)
   minsplit = 20L
   minbucket = 7L
minprob = 0.01
 stump = FALSE
   MIA = FALSE
 lookahead = FALSE
 nresample = 9999L
    tol = sqrt(.Machine$double.eps)
    maxsurrogate = 0L
    numsurrogate = FALSE
    mtry = Inf
    maxdepth = Inf
    multiway = FALSE
    splittry = 2L
    intersplit = FALSE
    majority = FALSE
    caseweights = TRUE
    applyfun = NULL
    cores = NULL
    saveinfo = TRUE
    update = TRUE
splitflavour="ctree"


testtype <- match.arg("Bonferroni", c("Bonferroni", "MonteCarlo",
                 "Univariate", "Teststatistic"), several.ok = TRUE)


ttesttype <- testtype

 if (length(testtype) > 1) {
        stopifnot(all(testtype %in% c("Bonferroni", "MonteCarlo")))
        ttesttype <- "MonteCarlo"
    }


teststat <- match.arg("quadratic",c("quadratic", "maximum"))
 splitstat <- match.arg("quadratic",c("quadratic", "maximum"))

```

```
  splitflavour <- match.arg("ctree",c("ctree", "exhaustive"))

 control<-c(extree_control(criterion = ifelse("Teststatistic" %in%
    testtype,
                                        "statistic", "p.value"),
                      logmincriterion = logmincriterion, minsplit =
    minsplit,
                      minbucket = minbucket, minprob = minprob,
                      nmax = nmax, stump = stump, lookahead = lookahead,
                      mtry = mtry, maxdepth = maxdepth, multiway =
    multiway,
                      splittry = splittry, maxsurrogate = maxsurrogate,
                      numsurrogate = numsurrogate,
                      majority = majority, caseweights = caseweights,
                      applyfun = applyfun, saveinfo = saveinfo,  ###
    always
                      selectfun = .ctree_select(),
                      splitfun = if (splitflavour == "ctree") .ctree_split
    () else .objfun_test(),
                      svselectfun = .ctree_select(),
                      svsplitfun =.ctree_split(minbucket = 0),
                      bonferroni = "Bonferroni" %in% testtype,
                      update = update),
       list(teststat = teststat, splitstat = splitstat, splittest =
    splittest, pargs = pargs,
            testtype = ttesttype, nresample = nresample, tol = tol,
            intersplit = intersplit, MIA = MIA, splitflavour =
    splitflavour))



testtype = "Bonferroni"
teststat = "quadratic"
 splitstat = "quadratic"
splittest = FALSE
pargs = GenzBretz()
 nmax = c("yx" = Inf, "z" = Inf)
alpha = 0.05;
 mincriterion = 1 - alpha;
logmincriterion = log(mincriterion)
   minsplit = 20L
   minbucket = 7L
minprob = 0.01
```

```
 stump = FALSE
   MIA = FALSE
 lookahead = FALSE
 nresample = 9999L
    tol = sqrt(.Machine$double.eps)
    maxsurrogate = 0L;
    numsurrogate = FALSE;
    mtry = Inf;
    maxdepth = Inf
    multiway = FALSE
    splittry = 2L
    intersplit = FALSE
    majority = FALSE;
    caseweights = TRUE
    applyfun = NULL
    cores = NULL
    saveinfo = TRUE
    update = TRUE
splitflavour="ctree"


testtype <- match.arg("Bonferroni", c("Bonferroni", "MonteCarlo",
                 "Univariate", "Teststatistic"), several.ok = TRUE)


ttesttype <- testtype

 if (length(testtype) > 1) {
        stopifnot(all(testtype %in% c("Bonferroni", "MonteCarlo")))
        ttesttype <- "MonteCarlo"
    }


teststat <- match.arg("quadratic",c("quadratic", "maximum"))
 splitstat <- match.arg("quadratic",c("quadratic", "maximum"))

  splitflavour <- match.arg("ctree",c("ctree", "exhaustive"))

 control<-c(extree_control(criterion = ifelse("Teststatistic" %in%
   testtype,
                                      "statistic", "p.value"),
                     logmincriterion = logmincriterion, minsplit =
   minsplit,
                     minbucket = minbucket, minprob = minprob,
```

```
152                       nmax = nmax, stump = stump, lookahead = lookahead,
153                       mtry = mtry, maxdepth = maxdepth, multiway =
    multiway,
154                       splittry = splittry, maxsurrogate = maxsurrogate,
155                       numsurrogate = numsurrogate,
156                       majority = majority, caseweights = caseweights,
157                       applyfun = applyfun, saveinfo = saveinfo,  ###
    always
158                       selectfun = .ctree_select(),
159                       splitfun = if (splitflavour == "ctree") .ctree_split
    () else .objfun_test(),
160                       svselectfun = .ctree_select(),
161                       svsplitfun =.ctree_split(minbucket = 0),
162                       bonferroni = "Bonferroni" %in% testtype,
163                       update = update),
164       list(teststat = teststat, splitstat = splitstat, splittest =
    splittest, pargs = pargs,
165            testtype = ttesttype, nresample = nresample, tol = tol,
166            intersplit = intersplit, MIA = MIA, splitflavour =
    splitflavour)
167
168
169
170 extetionICtree1 <- function(Formula, data,rou,gama, Control = partykit::
    ctree_control()){
171   #library(partykit) -- in order to get partykit::extree_data function
172   requireNamespace("inum")
173
174   DATA = data
175   X <- DATA[,as.character(Formula[[2]][[2]])]
176   Y <- DATA[,as.character(Formula[[2]][[3]])]
177
178   if(sum(X==Y)){ ## add small noise to observed event case
179     epsilon <- min(diff(sort(unique(c(X,Y)))))/20
180     ID <- which(X==Y)
181     DATA[ID,as.character(Formula[[2]][[3]])] <- epsilon + DATA[ID,as.
    character(Formula[[2]][[3]])]
182   }
183
184   if(sum(X == Inf)){
185     stop("There are Infs in the left end of intervals, make sure the left
     end is bounded")
186   }
187
```

```
188   if(sum(Y == Inf)){
189     if(length(Y[Y!=Inf])==0){
190       stop("All Infs on the right end of censoring interval, unable to
      compute")
191     }
192     Right_end_impute <- max(Y[Y!=Inf])*100 # impute Inf with 100 times
      the max observed Y value
193     DATA[which(Y==Inf),as.character(Formula[[2]][[3]])] <- Right_end_
      impute
194   }
195
196
197   .logrank_trafo1 <- function(x2,Rou,Gama){
198     if(!(survival::is.Surv(x2) && isTRUE(attr(x2, "type") == "interval"))
      ){stop("Response must be a 'Survival' object with Surv(time1,time2,
      event) format")}
199
200
201     Curve <- icenReg::ic_np(x2[, 1:2])
202
203     Left <- 1 - icenReg::getFitEsts(Curve, q = x2[, 1])
204     Right <- 1 - icenReg::getFitEsts(Curve, q = x2[, 2])
205
206     Log_Left <- ifelse(Left<=0,0,Left*log(Left)*((Left^Rou)*(1-Left)^Gama
      ))
207     Log_Right<- ifelse(Right<=0,0,Right*log(Right)*((Right^Rou)*(1-Right)
      ^Gama))
208    result <- (Log_Left-Log_Right)/(Left-Right)
209
210     return(as.double(result))
211   }
212
213  Rou<-rou;Gama<-gama
214   h2 <- function(y, x, start = NULL, weights, offset, estfun = TRUE,
      object = FALSE, ...) {
215     if (is.null(weights)) weights <- rep(1, NROW(y))
216     s <- .logrank_trafo1(y[weights > 0,,drop = FALSE],Rou,Gama)
217     r <- rep(0, length(weights))
218     r[weights > 0] <- s
219     list(estfun = matrix(as.double(r), ncol = 1), converged = TRUE)
220   }
221 partykit::ctree(formula = Formula, data=DATA, ytrafo=h2, control =
      control)
222 }
```

```
223
224
225 extetionICtree2 <- function(Formula, data, Control = partykit::ctree_
      control()){
226   #library(partykit) -- in order to get partykit::extree_data function
227   requireNamespace("inum")
228
229   DATA = data
230   X <- DATA[,as.character(Formula[[2]][[2]])]
231   Y <- DATA[,as.character(Formula[[2]][[3]])]
232
233   if(sum(X==Y)){ ## add small noise to observed event case
234     epsilon <- min(diff(sort(unique(c(X,Y)))))/20
235     ID <- which(X==Y)
236     DATA[ID,as.character(Formula[[2]][[3]])] <- epsilon + DATA[ID,as.
      character(Formula[[2]][[3]])]
237   }
238
239   if(sum(X == Inf)){
240     stop("There are Infs in the left end of intervals, make sure the left
       end is bounded")
241   }
242
243   if(sum(Y == Inf)){
244     if(length(Y[Y!=Inf])==0){
245       stop("All Infs on the right end of censoring interval, unable to
      compute")
246     }
247     Right_end_impute <- max(Y[Y!=Inf])*100 # impute Inf with 100 times
      the max observed Y value
248     DATA[which(Y==Inf),as.character(Formula[[2]][[3]])] <- Right_end_
      impute
249   }
250
251   ## x2 is Surv(Left,right,type="interval2") object
252   .logrank_trafo <- function(x2){
253     if(!(survival::is.Surv(x2) && isTRUE(attr(x2, "type") == "interval"))
      ){stop("Response must be a 'Survival' object with Surv(time1,time2,
      event) format")}
254
255     # Fit IC survival curve
256     #--Curve <- interval::icfit(x2~1)
257     Curve <- icenReg::ic_np(x2[, 1:2])
258     ## get estimated survival
```

```
259     #--Left <- interval::getsurv(x2[,1], Curve)[[1]]$S
260     #--Right<- interval::getsurv(x2[,2], Curve)[[1]]$S
261     Left <- 1 - icenReg::getFitEsts(Curve, q = x2[, 1])
262     Right <- 1 - icenReg::getFitEsts(Curve, q = x2[, 2])
263 
264     Log_Left <- ifelse(Left<=0,0,log(Left))
265     Log_Right<- ifelse(Right<=0,0,log(Right))
266     result <- (Log_Left+Log_Right)-1
267 
268     return(as.double(result))
269   }
270 
271   h2 <- function(y, x, start = NULL, weights, offset, estfun = TRUE,
      object = FALSE, ...) {
272     if (is.null(weights)) weights <- rep(1, NROW(y))
273     s <- .logrank_trafo(y[weights > 0,,drop = FALSE])
274     r <- rep(0, length(weights))
275     r[weights > 0] <- s
276     list(estfun = matrix(as.double(r), ncol = 1), converged = TRUE)
277   }
278 partykit::ctree(formula = Formula, data=DATA, ytrafo=h2, control =
      control)
279 }
280 extetionICtree3 <- function(Formula, data, Rou,Lamta,Control = partykit::
      ctree_control()){
281   #library(partykit) -- in order to get partykit::extree_data function
282   requireNamespace("inum")
283 
284   DATA = data
285   X <- DATA[,as.character(Formula[[2]][[2]])]
286   Y <- DATA[,as.character(Formula[[2]][[3]])]
287 
288   if(sum(X==Y)){ ## add small noise to observed event case
289     epsilon <- min(diff(sort(unique(c(X,Y)))))/20
290     ID <- which(X==Y)
291     DATA[ID,as.character(Formula[[2]][[3]])] <- epsilon + DATA[ID,as.
      character(Formula[[2]][[3]])]
292   }
293 
294   if(sum(X == Inf)){
295     stop("There are Infs in the left end of intervals, make sure the left
       end is bounded")
296   }
297 
```

```
298   if(sum(Y == Inf)){
299     if(length(Y[Y!=Inf])==0){
300       stop("All Infs on the right end of censoring interval, unable to
      compute")
301     }
302     Right_end_impute <- max(Y[Y!=Inf])*100 # impute Inf with 100 times
      the max observed Y value
303     DATA[which(Y==Inf),as.character(Formula[[2]][[3]])] <- Right_end_
      impute
304   }
305
306   ## x2 is Surv(Left,right,type="interval2") object
307   .logrank_trafo <- function(x2,rou,lamta){
308     if(!(survival::is.Surv(x2) && isTRUE(attr(x2, "type") == "interval"))
      ){stop("Response must be a 'Survival' object with Surv(time1,time2,
      event) format")}
309
310     # Fit IC survival curve
311
312     Curve <- icenReg::ic_np(x2[, 1:2])
313
314     Left <- 1 - icenReg::getFitEsts(Curve, q = x2[, 1])
315     Right <- 1 - icenReg::getFitEsts(Curve, q = x2[, 2])
316 rou=Rou;lamta=Lamta;
317 ff<-function(x){(x^lamta)*((1-x)^(rou-1))}
318 trafo_integrate<- function(t){integrate(ff,0,1-t)}
319    Log_Left<-rep(0,length(Left));
320 Log_Right<-rep(0,length(Right));
321    for(i in 1:length(Left))
322 { Log_Left[i]<-ifelse(Left<=0||Left==1,0,trafo_integrate(Left[i])$value);
323     Log_Right[i]<- ifelse(Right<=0||Right==1,0,trafo_integrate(Right[i])$
      value)
324 };
325
326 result <-(Left*Log_Left-Right*Log_Right)/(Right-Left)
327
328
329
330
331     return(as.double(result))
332   }
333
334   h2 <- function(y, x, start = NULL, weights, offset, estfun = TRUE,
      object = FALSE, ...) {
```

```
335     if (is.null(weights)) weights <- rep(1, NROW(y))
336     s <- .logrank_trafo(y[weights > 0,,drop = FALSE])
337     r <- rep(0, length(weights))
338     r[weights > 0] <- s
339     list(estfun = matrix(as.double(r), ncol = 1), converged = TRUE)
340   }
341
342   partykit::ctree(formula = Formula, data=DATA, ytrafo=h2, control =
      Control)
343 }
344
345 extetionICtree4 <- function(Formula, data,rou,gama, Control = partykit::
      ctree_control()){
346   #library(partykit) -- in order to get partykit::extree_data function
347   requireNamespace("inum")
348
349   DATA = data
350   X <- DATA[,as.character(Formula[[2]][[2]])]
351   Y <- DATA[,as.character(Formula[[2]][[3]])]
352
353   if(sum(X==Y)){ ## add small noise to observed event case
354     epsilon <- min(diff(sort(unique(c(X,Y)))))/20
355     ID <- which(X==Y)
356     DATA[ID,as.character(Formula[[2]][[3]])] <- epsilon + DATA[ID,as.
      character(Formula[[2]][[3]])]
357   }
358
359   if(sum(X == Inf)){
360     stop("There are Infs in the left end of intervals, make sure the left
       end is bounded")
361   }
362
363   if(sum(Y == Inf)){
364     if(length(Y[Y!=Inf])==0){
365       stop("All Infs on the right end of censoring interval, unable to
      compute")
366     }
367     Right_end_impute <- max(Y[Y!=Inf])*100 # impute Inf with 100 times
      the max observed Y value
368     DATA[which(Y==Inf),as.character(Formula[[2]][[3]])] <- Right_end_
      impute
369   }
370
371
```

```
372   .logrank_trafo1 <- function(x2,Rou,Gama){
373     if(!(survival::is.Surv(x2) && isTRUE(attr(x2, "type") == "interval"))
      ){stop("Response must be a 'Survival' object with Surv(time1,time2,
      event) format")}
374
375
376     Curve <- icenReg::ic_np(x2[, 1:2])
377
378     Left <- 1 - icenReg::getFitEsts(Curve, q = x2[, 1])
379     Right <- 1 - icenReg::getFitEsts(Curve, q = x2[, 2])
380
381     Log_Left <- ifelse(Left<=0,0,Left*tan((pi/2)*(Left-1))*(Left^Rou)*(1-
      Left)^Gama)
382     Log_Right<- ifelse(Right<=0,0,Right*tan((pi/2)*(Right-1))*(Right^Rou)
      *(1-Right)^Gama)
383    result <- (Log_Left-Log_Right)/(Left-Right)
384
385     return(as.double(result))
386   }
387
388  Rou<-rou;Gama<-gama
389   h2 <- function(y, x, start = NULL, weights, offset, estfun = TRUE,
      object = FALSE, ...) {
390     if (is.null(weights)) weights <- rep(1, NROW(y))
391     s <- .logrank_trafo1(y[weights > 0,,drop = FALSE],Rou,Gama)
392     r <- rep(0, length(weights))
393     r[weights > 0] <- s
394     list(estfun = matrix(as.double(r), ncol = 1), converged = TRUE)
395   }
396 partykit::ctree(formula = Formula, data=DATA, ytrafo=h2, control =
      control)
397 }
```

Listing A.1: newtree.R

References

Bibliography

[1] O. Aalen. "A model for nonparametric regression analysis of counting processes". In: *Lecture Notes in Statistics* (1980), pp. 1–25.

[2] O. Aalen. "Nonparametric inference for a family of counting processes". In: *Annals of Statistics* 6 (1978), pp. 534–545.

[4] N. Albaz. *Modelling credit risk for SMEs in Saudi Arabia*. Ph.D. thesis, 2017.

[5] E. I. Altman and G. Sabato. "Modeling credit risk for SMEs: evidence from the US market". In: *A Journal of Accounting Finance and Business Studies* 19.6 (2006), pp. 716–723.

[6] Y. Amit and D. Geman. "Shape quantization and recognition with randomized trees." In: *Neural Computation* 9 (1997), pp. 1545–1588.

[7] C. Anderson-Bergman. "An Efficient Implementation of the EMICM Algorithm for the Interval Censored NPMLE". In: *Journal of Computational and Graphical Statistics* 26.2 (2017), pp. 463–467.

[9] J. Argaon and D. Eberly. "On convergence of convex minorant algorithms for distribution estimation with interval-censored data". In: *Journal of Computational and Graphical Statistics* 1.2 (1992), pp. 129–140.

[10] S. Athey, J. Tibshirani, and S. Wager. "Generalized Random Forests". In: *the Annals of Statistics* 47.2 (2019), pp. 1148–1178.

[11] J. S. Bae, C. H. Hwang, and J. Y. Shim. "Two-step LS-SVR for censored regression". In: *Journal of the Korean Data and Information Science Sociaty* 23.2 (2012), pp. 393–401.

[12] V. Bagdonaviius and M. Nikulin. *Accelerated Life Models: Modeling and Statistical Analysis*. Chapman & Hall/CRC, 2002.

[13] R. C. Barros et al. "A framework for bottom-up induction of oblique decision trees". In: *Neurocomputing* 135 (2014), pp. 3–12.

[14] G. F. Beadle et al. "Cosmetic results following primary radiation therapy for early breast cancer". In: *Cancer* 54 (1984), pp. 2911–2918.

[15] G. F. Beadle et al. "The effect of adjuvant chemotherapy on the cosmetic results after primary radiation treatment for early stage breast cancer". In: *International Journal of Radiation Oncology, Biology and Physics* 10 (1984), pp. 2131–2137.

[16] V. V. Belle et al. “Support vector methods for survival analysis: a comparison between ranking and regression approaches”. In: *Artificial Intelligence in Medicine* 53.2 (2011), pp. 107–118.

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

[18] S. Bennett. “Log-Logistic Regression Models for Survival Data”. In: *Journal of the Royal Statistical Society* 32.2 (1983), pp. 165–171.

[19] J. Berkson and R. P. Gage. “Survival curve for cancer patients following treatment”. In: *Journal of the American Statistical Association* 47 (1952), pp. 501–515.

[20] E. Biganzoli et al. “Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach”. In: *Statistics in Medicine* 17.10 (1998), pp. 1169–1186.

[21] J. W. Boag. “Maximum likelihood estimates of the proportion of patients cured by cancer therapy”. In: *Journal of the Royal Statistical Society. Series B: Methodological* 11.1 (1948), pp. 15–53.

[22] I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. “A review of survival trees”. In: *Statistics Surveys* 5 (2011), pp. 44–71.

[23] L. Breiman. “Bagging predictors”. In: *Machine Learning* 24 (1996), pp. 123–140.

[24] L. Breiman. “Heuristics of instability and stabilization in model selection”. In: *Annals of Statistics* 24.6 (1996), pp. 2350–2383.

[25] L. Breiman. “Random forests”. In: *Machine Learning* 45 (2001), pp. 5–32.

[26] L. Breiman et al. *Classification and regression trees*. CRC press, 1984.

[27] N. Breslow. “A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship”. In: *Biometrika* 57.3 (1970), pp. 579–594.

[28] N. E. Breslow. “Covariance analysis of censored survival data”. In: *Journal of the American Statistical Association* 30 (1974), pp. 89–99.

[29] G. W. Brown and M. M. Flood. “A statistical model for life-length of materials”. In: *Journal of the American Statistical Association* 53 (1958), pp. 151–160.

[30] G. W. Brown and M. M. Flood. “Tumbler mortality”. In: *Journal of the Royal Statistical Society. Series B: Methodological* 42.240 (1947), pp. 562–574.

[31] M. Buri and T. Hothorn. “Model-based random forests for ordinal regression”. In: *The International Journal of Biostatistics* 16.2 (2020).

[32] K. P. Burnham and D. R. Anderson. “Understanding AIC and BIC in model selection”. In: *Sociological Methods and Research* 33.2 (2004), pp. 261–301.

[33] D. P. Byar, R. Huse, and J. C. Bailar. “An Exponential Model Relating Censored Survival Data and Concomitant Information for Prostatic Cancer Patients”. In: *Journal of the National Cancer Institute* 52.2 (1974), pp. 321–326.

[34] C. Carter and J. Catlett. “Assessing credit card applications using machine learning”. In: *IEEE Expert* 2.3 (1987), pp. 71–79.

[35] I. Chang and C. A. Hsiung. *Counting Process Methods in Survival Analysis*. Encyclopedia of Biostatistics, 2005.

[36] P. Chaudhuri and W. Y. Loh. "Nonparametric estimation of conditional quantiles using quantile regression trees". In: *Bernoulli* 8.5 (2002), pp. 561–576.

[37] P. Chaudhuri et al. "Generalized regression trees". In: *Statistica Sinica* 5.2 (1995), pp. 641–666.

[38] P. Chaudhuri et al. "Piecewise-polynomial regression trees". In: *Statistica Sinica* 4.1 (1994), pp. 143–167.

[39] D. Chen and J. Sun. *Emerging topics in modeling interval-censored survival data*. Springer, 2022.

[40] D. Chen, J. Sun, and K. E. Peace. *Interval-censored time-to-event data:methods and applications*. Chapman and Hall/CRC, 2013.

[41] H. Y. Chen. "Weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the clase-cohort design". In: *Publications of the American Statistical Association* 96.456 (2001), pp. 1446–1457.

[42] J. Chen and R. De Leone. "A survival tree for interval-censored failure time data". In: *AIMS Mathematics* 7.10 (2022), pp. 18099–181261.

[43] J. Chen, C. Wang, and R. De Leone. "A model based on survival-based credit risk assessment system of SMEs". In: *The 14th International Conference on Computer Modeling and Simulation*. ACM, 2022, pp. 245–251.

[44] K. Chen, Z. Jin, and Z. Ying. "Semiparametric analysis of transformation models with censored data". In: *Biometrika* 89.3 (2002), pp. 659–668.

[45] L. Chen and J. Sun. "A multiple imputation approach to the analysis of interval censored failure time data with the additive hazards model". In: *Computational Statistics and Data Analysis* 103.4 (2016), pp. 242–249.

[46] S. C. Cheng, L. J. Wei, and Z. Ying. "Analysis of transformation models with censored data". In: *Biometrika* 82.4 (1995), pp. 835–845.

[47] S. C. Cheng and L. Ying. "Predicting Survival Probabilities With Semiparametric Transformation Models". In: *Journal of the American Statistical Association* 92.437 (1997), pp. 227–235.

[48] C. L. Chiang. "A stochastic study of the life table and its applications: I. probability distributions of the biometric functions". In: *Biometrics* 47 (1960), pp. 618–635.

[49] H. Cho, N. P. Jewell, and M. R. Kosorok. "Interval Censored Recursive Forests". In: *Journal of computational and graphical statistics: A joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 2 (2022), p. 31.

[50] C. F. Chung, P. Schmidt, and A. D. Witte. "Survival analysis: A survey". In: *Journal of Quantitative Criminology* 7.1 (1991), pp. 59–98.

[51] A. Ciampi, S. A. Hogg, and J. Thiffault. "RECPAM:A computer program for recursive partition and amalgamation for censored surveval data and other situations frequently occurring in biostatistics I methods and program features". In: *Computer Methods and Programs in Biomedicine* 26 (1988), pp. 239–256.

[52] A. Ciampi, S. A. Hogg, and J. Thiffault. "RECPAM:A computer program for recursive partition and amalgamation for censored surveval data and other situations frequently occurring in biostatistics II applications to data on small cell carcinoma of the lung (SCCL)". In: *Computer Methods and Programs in Biomedicine* 30 (1989), pp. 283–296.

[53] H. C. Co and R. Boosarawongse. "Forecasting Thailand's rice export: Statistical techniques vs. artificial neural networks". In: *Computers and Industrial Engineering* 53.4 (2007), pp. 610–627.

[54] C. Cortes and V. N. Vapnik. "Support-vector networks". In: *Machine learning* 20 (1995), pp. 273–297.

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

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

[57] D. R. Cox and E. J. Snell. *Analysis of Binary Data*. Crc Press, 1989.

[58] J. A. Cruz and D. S. Wishart. "Applications of Machine Learning in Cancer Prediction and Prognosis". In: *Cancer Informatics* 2 (2006), pp. 59–77.

[59] D. J. Davis. "An analysis of some failure data". In: *Journal of the American Statistical Association* 47 (1952), pp. 113–150.

[60] R. B. Davis and J. R. Anderson. "Exponential survival trees". In: *Statistics in Medicine* 8 (1989), pp. 947–961.

[61] L. Deng and X. Li. "Machine Learning Paradigms for Speech Recognition: An Overview". In: *Audio, Speech, and Language Processing, IEEE Transactions on* 21.5 (2013), pp. 1060–1089.

[62] E. Diday and J. V. Moreau. "Learning hierarchical clustering from examples - application to the adaptive construction of dissimilarity indices". In: *Pattern Recognition Letters* 2.6 (1984), pp. 365–378.

[63] L. Diric, G. Claeskens, and B. Baesens. "Time to default in credit scoring using survival analysis: a benchmark study". In: *The Journal of the Operational Research Society* 68.6 (2017), pp. 1–25.

[64] P. Doyle. "The use of automatic interaction detector and similar search procedures". In: *Operational Research Quarterly* 24 (1973), pp. 465–467.

[65] M. Du and J. Sun. "Statistical analysis of interval-censored failure time data." In: *Chinese Journal of Applied Probability and Statistics* 37 (2021), pp. 627–654.

[66] R. Duda and P. Hart. *Pattern classification and scene analysis*. John Wiley & Sons, 1973.

[67] J. Dvorský et al. "Evaluation of important credit risk factors in the SME segment". In: *Journal of International Studies* 11 (2018), pp. 204–216.

[68] B. Efron. "The efficiency of Cox's likelihood function for censored data". In: *Journal of the American Statistical Association* 72 (1977), pp. 557–565.

[69] B. Efron. "The efficiency of logistics regression compared to normal discriminant analysis". In: *Journal of the American Statistical Association* 70.352 (1975), pp. 892–898.

[70] B. Efron. "The two sample problem with censored data". In: *Proceedings of the Fifth Berkerley Symposium in Mathematical Statistics* 69 (1967), pp. 831–857.

[71] B. Engelmann, E. Hayden, and D. Tasche. "Measuring the discriminative power of rating systems". In: *PBanking and Financial Supervision:Deutsche Bundesbank Discussion paper N. 01/2003*. 2003.

[72] D. Enke and S. Thawornwong. "The use of data mining and neural networks for forecasting stock market returns". In: *Expert Systems with Applications* 29.4 (2005), pp. 927–940.

[73] B. Epstein. "The exponential distribution and its role in life testing". In: *Industrial Quality Control* 15 (1958), pp. 2–7.

[74] B. Epstein and M. Sobel. "Life testing". In: *Journal of the American Statistical Association* 48 (1953), pp. 486–502.

[75] A. V. Eye and E. Y. Mun. *Log-Linear Modeling: Concepts, Interpretation, and Application*. Wiley, 2013.

[76] D. Fantzzini and S. Figini. "Random Survival Forests Models for SME Credit Risk Measurement". In: *Methodology and Computing in Applied Probability* 11 (2009), pp. 29–45.

[77] D. Faraggi and R. Simon. "A Neural Network Model for Survival Data". In: *Statistics in Medicine* 14.1 (1995), pp. 73–82.

[78] T. Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27 (2006), pp. 861–874.

[79] P. Feigl and M. Zelen. "Estimation of exponential survival probabilities with concomitant information". In: *Biometrics* 21.4 (1965), pp. 826–838.

[80] M. Feinleib. "A method of analyzing log-normally Distributed Survival Data with Incomplete Follow-Up". In: *Journal of the American Statistical Association* 55.291 (1960), pp. 534–545.

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

[82] D. H. Fisher and K. B. Mckusick. "An Empirical Comparison of ID3 and Backpropagation". In: *Proceedings of the 11th International Joint Conference on Artificial Intelligence*. 1989, pp. 788–793.

[83] J. A. Fozard et al. *Techniques for Censored and Truncated Data*. Springer, 2011.

[84] J. Friedman. "On bias, variance, 0/1-loss, and the curse of dimensionality". In: *Data Mining and Knowledge Discovery* 1 (1997), pp. 55–77.

[85] K. Fukushima. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological Cybernetics* 36 (1980), pp. 193–202.

[86] M. A. Ganaie et al. “Ensembles of double random forest”. In: *arXiv:2111.02010v1* (2021).

[87] E. A. Gehan. “A generalized Wilcoxon test for comparing arbitrarily singly-censored samples”. In: *Biometrika* 52.1-2 (1965), pp. 203–223.

[88] E. A. Gehan. “A generalized Wilcoxon test for comparing arbitrarily singly-censored samples”. In: *Biometrika* 52 (1965), pp. 203–223.

[89] E. A. Gehan. “Estimating survior functions from the life table”. In: *Journal Chronic Diseases* 21 (1969), pp. 629–644.

[90] R. Genuer. “Variance reduction in purely random forests”. In: *Journal of Nonparametric Statistics* 24.3 (2012), pp. 543–562.

[91] P. Geurts, D. Ernst, and L. Wehenkel. “Extremely randomized trees”. In: *Machine Learning* 63.1 (2006), pp. 3–42.

[92] D. Glenn and k. E. Fabricius. “Classification and regression tree: a powerful yet simple technique for ecological data analysis”. In: *Ecology* 81.11 (2000), pp. 3178–3192.

[93] Y. Goldberg and M. R. Kosorok. “Support vector regression for right censored data”. In: *Institute of Mathematical Statistics* 11.1 (2017), pp. 532–569.

[95] B. Gompertz. “On the nature of the function expressive of the law of human mortality and on a new mode of determining the value of life contingencies”. In: *Proceedings of the Royal Society of London* (1825), pp. 513–582.

[96] L. Gordon and R. A. Olshen. “Tree-structured survival analysis”. In: *Cancer treatment reports* 69.10 (1985), pp. 1065–1069.

[97] E. Graf et al. “Assessment and comparison of prognostic classification schemes for survival data”. In: *Statistics in Medicine* 18 (1999), pp. 2529–2545.

[98] A. Hald. “Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point”. In: *Scandinavian Actuarial Journal* 1.32 (1949), pp. 119–134.

[99] M. Hamada and C. F. J. Wu. “Analysis of censored data from highly fractionated experiments”. In: *Technometrics* 33 (1991), pp. 25–38.

[100] B. A. Hamburg, H. C. Kraemer, and W. Jahnke. “A hierarchy of drug use in adolescence behavioral and attitudinal correlates of substantial drug use”. In: *American Journal of Psychiatry* 132 (1975), pp. 1155–1163.

[101] S. Han, H. Kim, and Y. Lee. “Double random forest”. In: *Machine Learning* (2020). DOI: 10.1007/s10994-020-05889-1.

[102] D. P. Harrington and R. Fleming T. “A class of rank test procedures for censored survival data”. In: *Biometrika* 69.3 (1982), pp. 553–566.

[103] D. Heath, S. Kasif, and S. Salzberg. “Induction of oblique decision tree”. In: *Journal of Artificial Intelligence Research* (1993), pp. 1–6.

[104] G. E. Hinton, S. Osindero, and Y. W. Teh. “A fast learning algorithm for deep belief nets”. In: *Neural Computation* 18.7 (2006), pp. 1527–1554.

[105] P. Holmes, A. Hunt, and I. Stone. "An analysis of new firm survival using a hazard function". In: *Applied Economics* 42 (2010), pp. 185–195.

[106] R. D. Horner. "Age at onset of Alzheimer's disease: clue to the relative importance of etiologic factors?" In: *American Journal of Epidemiology* 126 (1987), pp. 409–414.

[107] T Hothorn, K. Hornik, and A. Zeileis. "Unbiased recursive partitioning: a conditional inference framework". In: *Journal of Computational and Graphical Statistics* 15 (2006), pp. 651–674.

[108] T. Hothorn et al. "Bagging survival trees". In: *Statistics in Medicine* 23.1 (2004), pp. 77–91.

[109] T. Hothorn et al. "Survival ensembles". In: *Biostatistics* 7.3 (2006), pp. 355–373.

[110] J. C. van Houwelingen and S. le Cessie. "Logistic regression, a review". In: *Statistica Neerlandica* 42.4 (1988), pp. 216–232.

[111] C. J. Huang et al. "Frog classification using machine learning techniques". In: *Expert Systems with Applications* 36 (2009), pp. 3737–3743.

[112] J. Huang. "Efficient estimation of the partly linear additive Cox model". In: *The Annals of Statistics* 27 (1999), pp. 1536–1563.

[113] C. L. Hwang and K. Yoon. *Multiple attribute decision making: methods and applications*. Springer-Verlag, Berlin, 1981.

[114] H. Ishwaran. "Relative risk forests for exercise heart rate recovery as a predictor of mortality". In: *Publications of the American Statistical Association* 99.1 (2004), pp. 591–600.

[115] H. Ishwaran et al. "Random survival forests". In: *the Annals of Applied Statistics* 2.3 (2008), pp. 841–860.

[116] J. J. Jaber et al. "Credit risk assessment using survival analysis for progressive right-censored data: a case study in Jordan". In: *Journal of Internet Banking and Commerce* 22 (2017), pp. 1–18.

[117] Z. Jin et al. "Rank-based inference for the accelerated failure time model." In: *Biometrika* 90 (2003), pp. 341–353.

[118] H. Jing and A. J. Smola. "Neural Survival Recommender". In: *Proceedings of the 10th ACM International Conference on Web Search and Data Mining. ACM*. 2017, pp. 515–524.

[119] G. Jongbloed. "Iterative convex minorant algorithm for nonparametric estimation". In: *The iterative convex minorant algorithm for nonparametric estimation* 7 (1998), pp. 310–321.

[120] J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*. John Wiley and Sons, 2002.

[121] E. L. Kaplan and P. Meier. "Nonparametric estimation from incomplete observations". In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481.

[122] G. V. Kass. "An exploratory technique for investigating large quantities of categorical data". In: *Journal of the Royal Statistical Society* 29.2 (1980), pp. 119–127.

[124] S. Keles and M. R. Segal. "Residual-based tree-structured survival analysis". In: *Statistics in Medicine* 21.2 (2002), pp. 313–326.

[125] F. M. Khan and V. B. Zubek. "Support vector regression for censored data (SVRc): a novel tool for survival analysis". In: *eighth IEEE international conference on data mining*. 2008, pp. 863–868.

[126] A. Khashman. "Credit risk evaluation using neural networks: Emotional versus conventional models". In: *Applied Soft Computing* 11.6 (2011), pp. 5477–5484.

[127] J. Kim and S. Lee. "Two-sample goodness-of-fit tests for additive risk models with censored observations". In: *Biometrika* 85 (1998), pp. 593–603.

[128] J. P. Klein and M. L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. springer, 2003.

[129] D. G. Kleinbaum and M. Klein. *Survival analysis, a self-learning text*. Springer, 1998.

[130] M. Ko and O. Bryson. "Reexamining the impact of information technology investment on productivity using regression tree and multivariate adaptive regression splines". In: *Information Technology and Management* 9 (2008), pp. 285–299.

[131] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[133] M. Kretowska. "Dipolar regression trees in survival analysis". In: *Biocybernetics and Biomedical Engineering* 24.3 (2004), pp. 25–33.

[134] M. Kretowska. "Random forest of dipolar trees for survival prediction". In: *Lecture Notes in Computer Science* (2006), pp. 909–918.

[135] D. Krstajic et al. "Cross-validation pitfalls when selecting and assessing regression and classification models". In: *Journal of Cheminformatics* 6 (2014), pp. 1–15.

[136] W. H. Kruskal and W. A. Wallis. "Use of ranks in one-criterion variance analysis". In: *Journal of the American Statistical Association* 47.260 (1952), pp. 583–621.

[137] M. Kubat. *An introduction to machine learning*. Springer, 2015.

[138] M. Kulich and D. Y. Lin. "Additive Hazards Regression with Covariate Measurement Error". In: *Journal of the American Statistical Association* 95.449 (2000), pp. 238–248.

[139] E. Lakatos and K. Lan. "A comparison of sample size methods for the logrank test". In: *Statistics in Medicine* 11.2 (1992), pp. 179–191.

[140] G. H. Landeweerd et al. "Binary tree versus single level tree classification of white blood cells". In: *Pattern Recognition* 16.6 (1983), pp. 571–577.

[141] W. R. Lane, S. W. Looney, and J. W. Wansley. "An application of the cox proportional hazards model to bank failure". In: *Journal of Banking and Finance* 10.4 (1986), pp. 511–531.

[142] Z. Lang, H. Hu, and D. Zhang. "A credit risk assessment model based on SVM for small and medium enterprises in supply chain finance". In: *Financial Innovation* (2015), pp. 1–14.

[143] J. F. Lawless. *Statistical models and methods for lifetime data*. John Wiley, 2003.

[144] M. Leblance and J. Crowley. "Risk trees for censored survival data". In: *Biometrics* 48.2 (1992), pp. 411–425.

[145] M. Leblance and J. Crowley. "Survival trees by goodness of split". In: *Journal of the American Statistical Association* 88.422 (1993), pp. 457–467.

[146] Y. Lecun and Y. Bengio. "Convolutional networks for images, speech, and time series". In: *The Handbook of Brain Theory and Neural Networks* (1995).

[147] E. Lee and J. Wang. *Statistical Methods for Survival Data Analysis*. John Wiley & Sons, 2003.

[148] E. T. Lee, M. M. Desu, and E. A. Gehan. "A Monte Carlo study of the power of some two-sample tests". In: *Biometrika* 62.2 (1975), pp. 425–432.

[149] E. Lesaffre, A. Komárek, and D. Declerck. "An overview of methods for interval-censored data with an emphasis on applications in dentistry". In: *Statistical Methods in Medical Research* 14 (2005), pp. 539–552.

[150] K. C. Li, H. H. Lue, and C. H. Chen. "Interactive tree-structured regression via principal Hessian directions". In: *Journal of the American Statistical Association* 95 (2000), pp. 547–560.

[151] X. Li and R. C. Dubes. "Tree classifier design with a permutation statistic". In: *Pattern Recognition* 19.3 (1986), pp. 229–235.

[152] D. Y. Lin and Z. Ying. "Semiparametric analysis of general additive-multiplicative hazard models for counting processes". In: *Annals of Statistics* 23.5 (1995), pp. 1712–1734.

[153] D. Y. Lin and Z. Ying. "Semiparametric analysis of the additive risk model". In: *Biometrika* 81 (1994), pp. 61–71.

[154] S. M. Lin. *SMEs credit risk modelling for internal rating based approach in banking implementation of basel II requirement*. PhD thesis university of edinburgh, 2007.

[155] Y. Lin and Y. Jeon. "Random forests and adaptive nearest neighbors". In: *Journal of the American Statistical Association* 101.474 (2006), pp. 578–590.

[156] X. Liu et al. "Least squares support vector regression for complex censored data". In: *Artificial Intelligence in Medicine* (2023).

[157] W. Y. Loh. "Regression trees with unbiased variable selection and interaction detection". In: *Statistica Sinica* 12.2 (2002), pp. 361–386.

[158] W. Y. Loh. "Survival modeling through recursive stratification". In: *Computational Statistics and Data Analysis* 12.3 (1991), pp. 295–313.

[159] C. L. Loprinzi et al. "Prospective evaluation of prognostic variables from patient-completed questionnaires". In: *Journal of Clinical Oncology* 12.3 (1994), pp. 60–607.

[160] F. Louzada et al. "Modeling time to default on a personal loan portfolioin presence of disproportionate hazard rates". In: *Journal of Statistics Applications and Probability* 3 (2014), pp. 295–305.

[161] M. Luoma and E. K. Laitinen. "Survival Analysis as a Tool for Company Failure Prediction". In: *OMEGA International Journal of Management Science* 19.6 (1991), pp. 673–678.

[162] N. R. Mann, R. E. Schafer, and N. D. singpurvalla. *Methods for Survival Data Analysis of reliability and life data*. Wiley, 1974.

[163] N. Mantel. "Evaluation of survival data and two new rank order statistics arising in its consideration". In: *Cancer Chemother Rep* 50 (1966), pp. 163–170.

[164] M. Martinez-Arroyo and L. E. Sucar. "Learning an optimal naive bayes classifier". In: *18 International Conference on Pattern Recognition* (2006).

[165] D. McAlister. "The law of the geometric mean". In: *Proceedings of the Royal Society,* 29 (1879), pp. 367–376.

[166] N. Meinshausen and G. Ridgeway. "Quantile Regression Forests". In: *Journal of Machine Learning Research* 7.2 (2006), pp. 983–999.

[167] M. Merrell. "Time-specific life table contrasted with observed survivorship". In: *Biometrics* 3.3 (1947), pp. 129–136.

[168] T. M. Mitchell. *Machine learning*. McGraw-Hill, 1997.

[169] A. M. Molinaro, S. Dudoit, and M. J. Laan. "Tree-based multivariate regression and density estimation with right-censored data". In: *Journal of Multivariate Analysis* 90.1 (2004), pp. 154–177.

[170] J. N. Morgan and J.A. Sonquist. "Problems in the analysis of survey data, and a proposal". In: *Journal of the American Statistical Association* 58 (1963), pp. 415–434.

[171] S. A. Murphy, A. J. Rossini, and A. W. Van der Vaart. "Maximum likelihood estimation in the proportional odds model". In: *Journal of the American Statistical Association* 92 (1997), pp. 963–976.

[172] S. Murthy, S. Kasif, and S. Salzberg. "OC1: a randomized induction of oblique decision trees". In: *Proceedings of the 11th National Conference on Artificial Intelligence* (1993).

[173] S. K. Murthy, S. Kasif, and S. Salzberg. "A system for induction of oblique decision trees". In: *Journal of Artificial Intelligence Research* (1994), pp. 1–32.

[174] B. Narain. "Survival analysis and the credit granting decision". In: *Thomas LC, Crook JN and Edelman DB, editors, Credit Scoring and Credit Control*. Clarendon Press: Oxford, 1992, pp. 109–121.

[175] W. Nelson. "Hazard plotting for incomplete failure data". In: *Journal of Quality Technology* 1 (1969), pp. 27–52.

[176] W. Nelson. "Theory and applications of hazard plotting for censored failure data". In: *Technometrics* 14 (1972), pp. 945–965.

[177] H. J. Noh, T. H. Roh, and I. Han. "Prognostic personal credit risk model considering censored information". In: *Expert Systems with Applications* 28.4 (2005), pp. 753–762.

[178] J. ÓQuigley and L. Struthers. “Survival models based upon the logistic and log-logistic distributions”. In: *Computer Programs in Biomedicine* 15 (1982), pp. 3–12.

[179] E. E. Osgood and A. J. Seaman. “Treatment of chronic leukemias”. In: *Journal of the American Medical Association* 150 (1952), pp. 1372–1379.

[180] M. A. Paivafranco. “A log logistic model for survival time with covariates”. In: *Biometrika* 71.3 (1984), pp. 621–623.

[181] W. Pan. “Extending the iterative convex minorant algorithm to the Cox model for interval censored data”. In: *Journal of Computational and Graphical Statistics* 8.1 (1999), pp. 109–122.

[182] Y. Park and L. J. Wei. “Estimating subject-specific survival functions under the accelerated failure time model”. In: *Biometrika* 90.3 (2003), pp. 341–353.

[183] H. J. Payne and W. S. Meisel. “An algorithm for constructing optimal binary decision trees”. In: *IEEE Transactions on Computers* C-26.9 (1977), pp. 905–916.

[184] R. Peto and J. Peto. “Asymptotically efficient rank invariant test procedures”. In: *Journal of the Royal Statistical Society* 135 (1972), pp. 185–207.

[185] R. Peto et al. “Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples”. In: *Br. J. Cancer.* 35.1 (1977), pp. 1–39.

[186] M. C. Pike. “A method of analysis of certain class of experiments in carcinogenesis”. In: *Biometrics* 22.1 (1966), pp. 142–161.

[187] R. L. Prentice. “Linear rank tests with right-censored data”. In: *Biometrika* 65 (1978), pp. 167–179.

[188] J. R Quinlan. “Simplifying decision trees”. In: *International Journal of Human-Computer Studies* (1987), pp. 497–510.

[189] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers, 1993.

[190] J. R. Quinlan. “Induction of Decision Trees”. In: *Machine Learning* 1.1 (1986), pp. 81–106.

[192] B. D. Ripley. *Pattern Recongnition and neural networks*. Cambridge University Press, 1996.

[193] R. M. Ripley, A. L. Harris, and L. Tarassenko. “Non-linear survival analysis using neural networks”. In: *Statistics in Medicine* 23 (2004), pp. 825–842.

[194] X. Robin, N. Turck, and A. et al. Hainard. “pROC: an open-source package for R and S+ to analyze and compare ROC curves”. In: *BMC Bioinformatics* 7 (2011), p. 77.

[195] F. Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain”. In: *Psychological Review* 65 (1958), pp. 386–408.

[196] E. M. Rounds. “A combined nonparametric approach to feature selection and binary decision tree design”. In: *Pattern Recognition* 12.5 (1980), pp. 313–317.

[197] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley: New York, 2008.

[198] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning Internal Representaions by Error Propagation". In: *Parallel Distributed Processing-explorations in the microstructure of cognition* (1986).

[199] S. R. Safavian and D. Landgrebe. "A survey of decision tree classifier methodology". In: *IEEE Transactions on Systems, Man, and Cybernetics* 21.3 (1991), pp. 660–674.

[200] I. R. Savage. "Contributions to the theory of rank order statistics-the two sample case". In: *Annals of Mathematical Statistics* 27 (1956), pp. 590–615.

[201] V. M Schonberger and K. Cukier. *Big data: A revolutionthat will transform how we live,work, and think*. Houghton Mifflin Harcourt, 2013.

[202] G. Schwarz. "Estimating the Dimension of a Model". In: *Annals of Statistics* 6 (1978), pp. 461–464.

[203] M. R. Segal. "Regression trees for censored data". In: *Biometrics* 44 (1988), pp. 35–47.

[204] I. K. Sethi and B. Chatterjee. "Efficient decision tree design for discrete variable pattern recognition problems". In: *Pattern Recognition* 9.4 (1977), pp. 197–206.

[205] J. W. Shavlik, R. J. Mooney, and G. G. Towell. "Symbolic and neural learning algorithms: An experimental comparison". In: *Machine Learning* 6.2 (1998), pp. 111–143.

[206] P. K. Shivaswamy, W. Chu, and M. Jansche. "A support vector approach to censored targets". In: *Seventh IEEE international conference on data mining (ICDM)*. 2007, pp. 655–660.

[207] A. G. Singal et al. "Machine learning algorithms outperform conventional regression models in predicting development of hepatocellular carcinoma". In: *American Journal of Gastroenterology* 108.11 (2013), pp. 1723–1730.

[208] J. A. Steingrimsson et al. "Doubly robust survival trees". In: *Statistics in Medicine* (2016), pp. 3595–3612.

[209] O. A. Steingrimsson, L. Diao, and R. L. Strawderman. "Censoring unbiased regression trees and ensembles". In: *Journal of the American Statistical Association* 114.525 (2019), pp. 370–383.

[210] M. Stepanova and L. Thomas. "Survival analysis methods for personal loan data". In: *Operations Research* 50 (2002), pp. 277–289.

[211] R. Storrs et al. "The Effect of 6-Mercaptopurine on the duration of Steroid-induced Remissions in Acute Leukemia: A Model for Evaluation of Other Potentially Useful Therapy". In: *Blood* 21.6 (1963), pp. 699–716.

[212] H. Strasser and C. Weber. "On the asymptotic theory of permutation statistics". In: *Mathematical Methods of Statistics* 8 (1999), pp. 220–250.

[213] H. Sun. "Credit risk assessment of small and medium-sized enterprises based on a DEA-cluster cross model". In: *Journal of Changzhou Institute of Technology* 30 (2017), pp. 65–70.

[214] J. Sun. *The statistical analysis of interval-censored failure time data*. Springer, 2006.

[215] J. Sun, Q. Zhao, and X. Q. Zhao. "Generalized log-rank tests for interval-censored failure time data". In: *Scandinavian Journal of Statistics* 32 (2005), pp. 49–57.

[216] C. D. Sutton. "Classification and Regression Trees, Bagging, and Boosting". In: *Handbook of Statistics* 24.04 (2005), pp. 303–329.

[217] M. A. Tanner and W. H. Wong. "An application of imputation to an estimation problem in grouped lifetime analysis". In: *Technometrics* 29.1 (1987), pp. 23–32.

[218] M. A. Tanner and W. H. Wong. "Applications of multiple imputation to the analysis of censored regression data". In: *Biometrics* 47.4 (1991), pp. 1297–1309.

[219] R. Tarone and J. Ware. "On distribution-free tests for equality of survival distributions". In: *Biometrika* 64.1 (1977), pp. 156–160.

[220] S. Theodoridis. *Machine learning: a Bayesian and optimization perspective*. China Machine Press, 2017.

[221] L. C. Thomas, J. Banasik, and J. N. Crook. "Not if but when loans default". In: *Journal of the Operational Research Society* (1999), pp. 1185–1190.

[222] A. M. Ticlavilca, D. M. Feuz, and M. Mckee. "Forecasting agricultural commodity prices using multivariate bayesian machine learning regression". In: *Proceedings of the NCCC-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management* (2010), pp. 1–20.

[223] I. Todhunter. *A history of the mathematical theory of probability*. Chelse, 1949.

[224] A. Toffler. *The third wave*. A Bantam Book and William Morrow & Co., Inc., 1980.

[225] T. R. Tsai, S. H. Wu, and Y. Shen. "Model selection methods for reliability assessment based on interval-censored field failure samples". In: *International Journal of Reliability, Quality and Safety Engineering* 27 (2020), pp. 1–15.

[226] S. Tsouprou. *Measures of discrimination and predictive accuracy for interval censored survival data*. MA.D. thesis University Leiden, 2015.

[227] B. W. Turnbull. "Nonparametric estimation of a survivorship function with doubly truncated data". In: *Journal of the American Statistical Association* 69.345 (1974), pp. 169–173.

[228] B. W. Turnbull. "The distribution function with arbitrarily grouped, censored and truncated Data". In: *Journal of the Royal Statistical Society Series B* 38.3 (1976), pp. 290–295.

[229] B. W. Turnbull. "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 38.3 (1976), pp. 290–295.

[230] L. V. Utkin, E. D. Satyukov, and A. V. Konstantinov. "SurvNAM: the machine learning survival model explanation". In: *Neural Networks* 147 (2022), pp. 81–102.

[231] J. Vanobbergen et al. "The Signal-Tandmobiel project a longitudinal intervention health promotion study in Flanders (Belgium):baseline and first year results". In: *European Journal Of Paediatric Dentistry* 2 (2000), pp. 87–96.

[232] V. N. Vapnik and A. Y. Cervonenkis. "A class of algorithms for pattern recognition learning". In: *Avtomat.i Telemekh.* 25 (1964), pp. 937–945.

[233] D. Wang. *Modelling survival data in medical research*. Chapman and Hall/CRC, 1994.

[234] H. Wang and L. Zhou. "Random survival forest with space extensions for censored data". In: *Artificial Intelligence in Medicine* 79 (2017), pp. 52–61.

[235] L. Y. Wang and L. G. Zhu. "Research and application of credit risk of small and medium-sized enterprises based on random forest model". In: *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. 2021.

[236] P. Wang, Yan Li, and K. R. Chandan. "Machine Learning for survival Analysis :A Survey ." In: *ACM Computing Surveys* 51 (2019), pp. 110–145.

[237] F. Wei and J. S. Simonoff. "Survival trees for interval-censored survival data." In: *Statistics in medicine* 36 (2017), pp. 4831–4842.

[238] F. Wei and J. S. Simonoff. "Survival trees for left-truncated and right-censored data, with application to time-varying covariate data". In: *Biostatistics* 2 (2017), pp. 352–369.

[239] P Wei. "A comparison of some two-sample tests with interval censored data". In: *Journal of Nonparametric Statistics* 12 (1999), pp. 133–146.

[240] P Wei and J. E. Connett. "A multiple imputation approach to linear regression with clustered censored data". In: *Lifetime Data Analysis* 7.2 (2001), pp. 111–123.

[241] W. Weibull. "A Statistical distribution of wide applicability". In: *Journal of Applied Mechanics* 18 (1951), pp. 293–297.

[242] W. Weibull. "A statistical theory of the strength of materials". In: *Ingenioers Vetenskaps Akakemien Handlingar* 151 (1937), pp. 293–297.

[243] J. A. Wellner and Y. Zhan. "A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data". In: *Journal of the American Statistical Association* 92.439 (1997), pp. 945–959.

[244] D. C. Wickramarachchi et al. "HHCART: an oblique decision tree". In: *Computational Statistics and Data Analysis* 96 (2016), pp. 12–23.

[245] X. Z. Wu. *Applied regression and classification with R*. China Renmin University Press, 2016.

[246] B. B. Yang, S. Q. Shen, and W. Gao. "Weighted oblique decision trees". In: *The Thirty-Third AAAI Conference on Artifical Intelligence* (2019), pp. 5621–5627.

[247] G. Yang, Y. C. Cai, and C. K. Reddy. "Spatio-temporal check-in time prediction with recurrent neural network based survival analysis". In: *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 2018, pp. 2203–2211.

[248] M. Y. Yang. *Optimization of small and medium sized enterprises credit risk classification in N branch based on random forest*. M.D. thesis Guangxi University, 2020.

[249] W. Yao, F. Halina, and J. S. Simonoff. "An Ensemble Method for Interval-Censored Time-to-Event Data". In: *Biostatistics* 22.1 (2019), pp. 198–213.

[250] Y. M. Yin and S. J. Anderson. "Tree-structured modeling for interval-censored survival data". In: *Joint Statistical Meetings*. 2002, pp. 3877–3882.

[251] M. Zelen. "Application of exponential models to problems in cancer research". In: *Journal of the Royal Statistical Society. Series A* 129.3 (1966), pp. 368–398.

[252] J. Zhang and L. Thomas. "Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD". In: *International Journal of Forecasting* 18 (2012), pp. 204–215.

[253] L. Zhao. *Research on early warning of credit risk of small and medium enterprises based on decision tree*. PhD thesis Anhui university of finance and economics, 2008.

[254] L. Zhou, H. Wang, and Q. Xu. "Survival forest with partial least squares for high dimensional censored data". In: *Chemometrics and Intelligent Laboratory Systems* 179 (2018), pp. 12–21.

[255] Z. H. Zhou. "A Brief introduction to weakly supervised learning". In: *National Science Review* 1 (2017), p. 1.

[256] C. Zippin and P. Armitage. "Use of concomitant variables and incomplete survival information in the estimation of an exponential survival parameter". In: *Biometrics* 12 (1966), pp. 665–672.

Sitography

[3] R. Agarwal et al. *Neural Additive Models: Interpretable Machine Learning with Neural Nets*. 2020. URL: arXiv:2004.13912.

[8] C. Anderson-Bergman. *icenReg: Regression models for interval censored data. Version 2.0.15*. Accessed: 2020-10-03. URL: https://cran.r-project.org/web/packages/icenReg/index.html.

[94] G. Gomez and R. O. Pique. *A new class of rank tests for interval-censored data*. Accessed: 2008. URL: http://biostats.bepress.com/harvardbiostat/paper93.

[123] J. Katzman et al. *Deep survival: a deep cox proportional hazards network*. 2016. URL: arXiv:1606.00931.

[132] A. Komárek. *bayesSurv:Bayesian survival regression with flexible error and random effects distributions*. R package version3.3, 2020. URL: https://cran.r-project.org/web/packages/bayesSurv/index.html.

[191] R. Ranganath et al. *Deep Survival Analysis*. 2016. URL: arXiv:1608.02158.