HEC MONTRÉAL
École affiliée à l'Université de Montréal

Three essays on survival forests

par

Hoora Moradian

Thèse présentée en vue de l'obtention du grade de Ph. D. en administration
(option Méthodes quantitatives de gestion)

Décembre 2016

HEC MONTRÉAL

École affiliée à l'Université de Montréal

Cette thèse intitulée :

Three essays on survival forests

Présentée par :

Hoora Moradian

a été évaluée par un jury composé des personnes suivantes :

Marc Fredette
HEC Montréal
Président-rapporteur

Denis Larocque
HEC Montréal
Directeur de recherche

François Bellavance
HEC Montréal
Directeur de recherche

Jean-François Plante
HEC Montréal
Membre du jury

Lajmi Lakhal-Chaieb
Université LAVAL
Examinateur externe

Nicolas Sahuguet
HEC Montréal
Représentant du directeur de HEC Montréal

Résumé

Les méthodes basées sur les arbres sont versatiles et utiles pour l'analyse de données de survie avec censure à droite. Les forêts de survie, c'est-à-dire des combinaisons ensemblistes d'arbre pour données de survie, sont des méthodes performantes qui sont populaires auprès des analystes. Les trois articles de cette thèse proposent des solutions à trois limites des implantations actuelles des forêts de survie. La première limite est que la plupart d'entre elles utilisent le test log-rank comme critère de partitionnement. Ce test perd de l'efficacité lorsque l'hypothèse de proportionnalité des fonctions de risques n'est pas vérifiée. Le premier article propose d'utiliser comme critère de partitionnement l'intégrale de la différence en valeur absolue entre les fonctions de survie des deux nœuds enfants. La deuxième limite des implantations actuelles des forêts de survie est qu'elles ne fournissent pas de prévisions dynamiques lorsqu'il y a des variables explicatives qui varient dans le temps. Le deuxième article propose différentes façons de faire pour obtenir des estimations dynamiques de la fonction de risque avec des données de survie à temps discret. Finalement, les implantations actuelles des forêts de survie sont valides sous l'hypothèse que le temps de survie et le temps de censure sont indépendants, étant donné les variables explicatives. Le troisième article propose deux approches pour traiter le problème de censure dépendante avec des forêts de survie. La première approche consiste à utiliser une estimation finale de la fonction de survie qui corrige pour la censure dépendante. La deuxième consiste à utiliser un critère de partitionnement qui ne dépend pas de l'hypothèse de censure indépendante.

Mots clés: Analyse de survie à temps discret; variables explicatives qui varient dans le temps; censure dépendante; copula-graphique; données de survie; données censurées à droite; méthodes ensemblistes; forêts aléatoires; forêts de survie.

Méthodes de recherche: Exploitation de données

Abstract

Tree-based methods are versatile and useful tools for analysing survival data with right-censoring. Random survival forests, that are ensembles of trees for time-to-event data, are powerful methods and are popular among practitioners. The three articles of this thesis propose solutions for three limitations of current implementations of random survival forests. The first limitation is that most of them use the log-rank test as the splitting rule which loses power when the proportional hazards assumption is violated. The first article proposes the use of the integrated absolute difference between the two children nodes survival functions as the splitting rule. The second limitation of current random survival forest techniques is that they do not provide dynamic predictions in presence of time-varying covariates. The second article proposes different ways to obtain dynamic estimations of the hazard function with random forests with discrete-time survival data. Lastly, current implementations of random survival forests work under the assumption that the event time and the censoring time are independent, given the covariates. The third article proposes two approaches to tackle the problem of dependent censoring with random forests. The first approach is to use a final estimate of the survival function that corrects for dependent censoring. The second one is to use a splitting rule which does not rely on the independent censoring assumption.

Keywords: Conditional discrete-time survival analysis; time-varying covariates; dependent-censoring; copula-graphic; survival data; right-censored data; ensemble methods; random forests; survival forests.

Research methods: Data mining

Contents

Résumé iii

Abstract iv

List of Figures viii

List of Tables x

Acknowledgements xiii

Introduction 1

Chapter 1 L_1 Splitting Rules in Survival Forests 9

1.1 Abstract 9

1.2 Introduction 10

1.3 Survival Forest Approach and Splitting Criterion 13

1.4 Simulation Study 15

1.4.1 Simulation Design 16

1.4.2 Simulation Results 19

1.4.3 Additional Simulations 21

1.5 Real Data Sets 22

1.6 Discussion and Concluding Remarks 30

1.7 Acknowledgement 32

Bibliography 32

1.8 Appendix 35

1.8.1 Hazard functions used in the simulation for DGP 1 and DGP 2 35

1.8.2 Simplification of the Lin and Xu (2010) statistic leading to the L_1^* splitting rule. 36

1.8.3 Motivating example 37
1.8.4 Complete simulations results 38
1.8.5 Complete results for the additional simulations 45

Chapter 2 Dynamic Predictions with Random Forests for Discrete-Time Survival Data 54

2.1 Abstract 54
2.2 Introduction 55
2.3 Description of the methods 57
2.3.1 Description of existing methods for discrete-time survival data 58
2.3.2 Methods investigated in this paper 60
2.4 Simulation Study 67
2.4.1 Simulation Design 67
2.4.2 Simulation Results 71
2.5 Data Example 73
2.6 Concluding Remarks 76
2.7 Acknowledgement 79
Bibliography 79

Chapter 3 Survival Forests for Data with Dependent Censoring 82

3.1 Abstract 82
3.2 Introduction 83
3.3 Settings and Methods 85
3.3.1 Data Description 85
3.3.2 Copula-Graphic Estimator 85
3.3.3 Modification of Existing Methods 86
3.3.4 Point-Wise Forest 88
3.4 Simulation Study 90
3.4.1 Methods Compared 90
3.4.2 Simulation Design 92
3.4.3 Simulation Results 93
3.5 Discussion and Concluding Remarks 110
3.6 Acknowledgement 111

Bibliography 111

Conclusion **113**

References **x**

Bibliography x

List of Figures

1.1 The two survival curves in the terminal nodes of the tree for DGP 1 17

1.2 The four survival curves in the terminal nodes of the tree for DGP 2 18

1.3 IAE of methods for DGP 1, DGP 3 and DGP 4 at 10% and 40% censoring proportions 20

1.4 Integrated Brier score, across 20 runs of 10-–fold cross-validation, for the real data sets 25

1.5 C-index, across 20 runs of 10-–fold cross-validation, for the real data sets 27

1.6 Predicted and observed survival curves by prognosis group in GBC data 28

1.7 IAE for DGP 1 across different censoring proportions 40

1.8 IAE for DGP 2 across different censoring proportions 41

1.9 IAE for DGP 3 across different censoring proportions 42

1.10 IAE for DGP 4 across different censoring proportions 43

1.11 IAE for DGP 5 across different censoring proportions 44

1.12 IAE for DGP 1 across different censoring proportions with small n and large p 46

1.13 IAE for DGP 2 across different censoring proportions with small n and large p 47

1.14 IAE for DGP 3 across different censoring proportions with small n and large p 48

1.15 IAE for DGP 4 across different censoring proportions with small n and large p 49

1.16 IAE for DGP 5 across different censoring proportions with small n and large p 50

1.17 IAE for DGP 2 across different censoring proportions with binary X_1 and continuous X_2 52

1.18 IAE for DGP 2 across different censoring proportions with binary X_1 and X_2 53

2.1 Hazard functions for a typical sample of 50 subjects 70

2.2 Simulation results comparing the distribution of absolute log odds ratio on test sets across methods 72

2.3 Average cumulative gains curves for the 12 scenarios with an estimation horizon of 3 77

2.4 Average cumulative gains curves for the 18 scenarios with an estimation horizon greater than 3 78

3.1 % increase in IAE of all methods with respect to the best one over the 4500 runs on the top and a zoom in of the lower part at the bottom 99

3.2 IAE of methods for DGP 1 at 20% censoring proportion 100

3.3 IAE of methods for DGP 1 at 40% censoring proportion 101

3.4 IAE of methods for DGP 1 at 60% censoring proportion 102

3.5 IAE of methods for DGP 2 at 20% censoring proportion 103

3.6 IAE of methods for DGP 2 at 40% censoring proportion 104

3.7 IAE of methods for DGP 2 at 60% censoring proportion 105

3.8 IAE of methods for DGP 3 at 20% censoring proportion 106

3.9 IAE of methods for DGP 3 at 40% censoring proportion 107

3.10 IAE of methods for DGP 3 at 60% censoring proportion 108

3.11 Comparisons of IAE results for data generated with the Clayton versus the Frank Copula , Clayton Copula as working copula in both cases, with a censoring proportion of 60% and a true of τ of 0.8 109

List of Tables

1.1 Average ranks (smaller is better), according to the IAE criterion, of the nine methods over all individual data sets (12,500) with censoring. 20

1.2 Average ranks (smaller is better), according to the IAE criterion, of the nine methods over all individual data sets (12,500) with censoring, for the modified DGPs (smaller n and larger p). 21

1.3 Average ranks (smaller is better), according to the IAE criterion, of the nine methods over all individual data sets (2,500) with censoring for DGP 2. The original setup and two variations are presented. 22

1.4 Description of the data sets 23

1.5 Ranking of methods based on % increase of median IBS with respect to best method 24

1.6 Ranking of methods based on % decrease of median C-index with respect to best method 26

2.1 A generic data set with ten observations and two covariates 61

2.2 Person-period data set. Only the first six subjects (up to id=6) are shown to save space. 61

2.3 The 15 different estimating problems when $T = 5$ 62

2.4 Data set used with the first approach, separate forests for all combinations of (t, u), when $(t, u) = (2, 4)$ 63

2.5 Data set used with the Bou-hamad et al. (2009) and Schmid et al. (2016) method when $t = 2$ 64

2.6 Data set used with the third approach, i.e. a super person-period data set used for all combinations of (t, u). Only the first four subjects (up to id=4) are shown to save space. 66

2.7 The 10 different estimation problems for the data example 74

2.8 Summary of the data over the 30 scenarios 75

2.9 Average AUC with respect to the estimation horizon 76

3.1 Global results. Average and median of the % increase in IAE with respect to the best globally for the 4500 runs. 95

To my lovely sweet daughter, Eliana.

Acknowledgements

There are a number of wonderful people without whom this thesis might not have been written, and to whom I am indebted forever.

First, I would like to express my deepest gratitude to my supervisor, Professor Denis Larocque, for his support, patient guidance, enthusiasm and many learning opportunities he has provided throughout my studies. Thank you for always being available and willing to respond to my questions so patiently and promptly despite your busy schedule. You have set me an example of excellence as a researcher, mentor, and role model.

I extend a warm thank you to my co-supervisor, Professor François Bellavance, for the attention he has given me, for his advice and listening, unswerving moral and financial support. Had I not taken your course four years ago, I might have lost my interest. I am were I am today because of you. You are the main reason I am getting a PhD in an area I love.

Denis and François, I have been extremely lucky to have you two as my supervisors.

I also have to thank the members of my PhD committee, Professors Marc Fredette, Jean-François Plante and Lajmi Lakhal-Chaieb for their insightful feedback and suggestions. Your comments have been absolutely invaluable.

I also thank all the great people in the department of Decision Sciences and GERAD as well as in other departments for always being so helpful and friendly.

My special recognition goes out to my family, for their support, encouragement and patience during my studies. I thank all of you for your understanding and love you more than you will ever know.

To my mother, Talieh, who continues to learn, grow and develop and who has been a source of inspiration to me throughout my life. And also for the myriad of ways in which, you have supported me in my determination to realise my potential, to follow my dreams, and to make this contribution to the world.

To my father, Iraj, for always believing in me, always encouraging me, and always loving me unconditionally.

To my dear husband Mohammad-Reza, who inspired me and provided constant encouragement during the entire process. A very special thank you for always being there for me throughout my ups and downs. My words of thanks cannot compensate your contributions. You sacrificed a lot along the way.

To my sweet daughter, Eliana, who missed out on a lot of Mommy time while I sought personal development and intellectual enlightenment. Always chase your dreams! This work is dedicated to you and your journeys in learning to thrive.

Last but not least, to the loving memories of my grandmother, Shamseh-Iran and father-in-law, Hafez. We miss you everyday.

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by le fonds de recherche du Québec – Nature et technologies (FRQNT), bourse Édouard-Montpetit – Manuvie (Ph.D.), bourse Daniel-Brosseau et Peter-Letko, J. Armand Bombardier bourse of academic excellence and by HEC Montréal.

Introduction

Survival analysis answers the question of how long it takes for an event of interest to happen. It studies time-to-event data where the true time is observed for some subjects and only partial information about the time is available for other subjects, that is these observations are incomplete since the event has not yet occurred at the time of data collection. The presence of these censored observations is what distinguishes survival analysis from an ordinary regression analysis.

In general, there are three types of censoring in survival analysis; right censoring, left censoring and interval censoring. Left censoring happens when the event of interest has already occurred before the study begins. Interval censoring happens when the event time is known only to be within an interval instead of being observed exactly.

Right censoring is the most commonly encountered form of censoring. When a subject still has not experienced the event of interest at the end of the study, it is said to be right censored. Survival analysis has become a growing field of interest in many research areas. The duration of a client-firm relationship in subscription-based businesses and time to a customer's next purchase in retails are some examples of application of survival analysis in business.

To produce valid inference and optimize the prediction model for survival data, we must adequately incorporate all available information (complete and incomplete). The traditional methods for analyzing survival data, the parametric (Gamma, Weibull) and semi-parametric (Cox) models, can be useful and have been discussed in details in the literature (Hosmer Jr et al., 2011). However, they are restricted to the situation where the link between the covariates and the time response must be specified in advance. Nonparametric models have the advantage of relaxing the restrictive structure assumptions. These methods are more data-driven and less assumption-driven and are often preferred as prediction tools. Tree-based methods are one of the widely popular classes of nonparametric models among practitioners. Tree-mechanism is mainly a recursive procedure where the tree is recursively partitioned into homogeneous terminal nodes through binary splits. The original tree—based methods were developed to model the relation between covariates and either a categorical or a continuous outcome. The Classification and Regression Tree paradigm (CART) is well-established today (Breiman et al., 1984). Loh (2014) provides a recent and comprehensive survey about classification and regression trees.

Gordon and Olshen (1985) was the first to adapt the CART paradigm to right censored data and intro- duced survival trees (Leblanc and Crowley, 1993; Segal, 1988). Many splitting rules have been suggested so far to find best split that makes the two children nodes the most different according to a given criterion. First, Gordon and Olshen (1985) explored the idea of maximizing within-node homogeneity by minizing the Wasserstein distance between the estimated Kaplan|Meier of each node and a point-mass function. Then, Ciampi et al. (1988) suggested Wilcoxon-Gehan and Kolmogorov-Smirnov tests as measures to maximize the heterogeneity between two children nodes. However, the log-rank statistic (Ciampi et al., 1986) has become the most popular splitting rule.

Single trees are generally powerful descriptive tools. However, it is well-known that they are unstable predictive tools. Ensemble methods with trees as base learners such as random forests (Breiman, 2001) often provide better predictive accuracies than a single tree. Similarly, a combination of survival trees, called a survival forest (e.g., Hothorn et al., 2006; Ishwaran et al., 2008; Zhu and Kosorok, 2012), is often preferable than a single survival tree. Bou-Hamad et al. (2011b) present a general overview of the vast literature on survival trees and forests. In the following, we will focus on survival forests. First, Hothorn et al. (2006) suggested a random forest method for the log-transformed survival time. In this survival ensemble, inverse probability of censoring (IPC) are used as the estimated weights for observations in each bootstrap sample. A regression tree is grown for each bootstrap sample with log-transformed time as the outcome. Hothorn et al. (2006) used a "similarity" weighting scheme defined as the number of times each test point falls into the same terminal node of each tree as the i^{th} observation in the training sample. The ensemble prediction is in the form of the mean log-transformed survival time, obtained as a weighted average, over all trees. By far, the most popular random forest technique for survival data is Random Survival Forest (RSF) proposed by Ishwaran et al. (2008). The main output of this method is an ensemble of cumulative hazard function computed by averaging the Nelson—Aalen cumulative hazard function of each tree. This technique is implemented in the R package `randomForestSRC` (Ishwaran and Kogalur, 2014). The package user has the option of making a customized splitting rule or using one of the three available splitting rules for right censored data: the log-rank statistic (Segal, 1988), which is the default one, the log-rank score (Hothorn and Lausen, 2003) or the random splitting rule (Cutler and Zhao, 2001; Lin and Jeon, 2006). Some other features available in the package are variable selection and missing data imputation for covariates and outcomes. Recursively Imputed Survival Tree (RIST) is a more recent random forest technique proposed by Zhu and Kosorok (2012) for right-censored data. In this method, information on censored observations is retained through computation of conditional survival distribution and recursive imputation. The log-rank test statistic is used as the splitting rule.

As apparent from the above discussion, the current implementations of survival forest mostly use the log-rank test as the splitting criterion.

However, the log-rank test may have a significant loss of power in situations where the proportionality assumption is violated that is when the hazard functions or when the survival functions cross each other in the two compared groups (Lin and Wang, 2004; Lin and Xu, 2010).

Therefore, using the log-rank test as the splitting rule may lead to inaccurate estimations of the conditional survival function. The first article of this thesis provides a solution to this restraint. We suggest the use of the integrated absolute difference between the two children nodes survival functions as the splitting rule. The effectiveness of survival forests built with this rule, that we call L_1-forest, is investigated through simulations studies and applications to real data sets. The results show that L_1-forest produces better results compared to forests built with the log-rank splitting rule in many cases, even when the survival functions of different subjects do not cross each other.

The first studies on survival trees with time-dependent covariates were done by Bacchetti and Segal (1995) and Huang et al. (1998). Bacchetti and Segal (1995) proposed a survival tree method capable of handling time-varying covariates through replacement of each subject by "pseudo-subjects" and using smooth nonparametric hazard estimates as node summaries. They used a two-sample rank statistic as the splitting rule. Huang et al. (1998) suggested a piecewise exponential structure and were the first ones to include time as a time-dependent covariate in their model. More recently, Bertolet et al. (2012) proposed a method that combines a time-varying Cox model with CART and controls for confounding variables through time-varying indicators. This method offers two types of splitting procedures: recursive splitting and forward stepwise splitting. Pointing out that this technique is not able to handle time-dependent covariates that change non-monotonically though time, Wallace (2014) suggested a technique that overcomes this problem. This method also corrects for the usual selection bias towards time-varying covariates through permutation test of independence at each node for each covariate to select the variables with the highest association with the outcome. This makes computational intensity one of the limitations of this technique. Furthermore, given the use of a mixed-effects model to create permutations for each time-dependent covariate, the accuracy of the permutation test depends on the accuracy of the time-varying covariate(s) prediction model.

The studies presented above assume that the time-to-event is measured continuously. In many situations however, such as annual surveys or for a truly discrete process, the observed time is measured on a discrete scale. The reader can refer toTutz and Schmid (2016) for a recent, in-depth overview on discrete-time survival analysis. Bou-hamad et al. (2009) proposed a survival-tree method for discrete-time outcomes. They used a splitting rule based on maximum likelihood that rests on a discrete-time proportional odds model described in Willett and Singer (1993). This method accommodates time independent covariates with time-varying effects. Bou-Hamad et al. (2011a) extended this technique to the case of discrete-time survival forests with time-varying covariates through the person-period data structure suggested by Willett and Singer (1993) where, for each subject, there is one line of observation per period for which the subject is at risk. With this method, subjects can be split across different nodes. The periods where the splitting rule is true would go to one node and the periods where it is false would go to the other node. This approach is basically a discrete version of the Bacchetti and Segal (1995) "pseudo-subjects" idea. In case of only time invariant covariates, their splitting rule reduces to the splitting criterion suggested by Bou-hamad et al. (2009). The `R` package `DStree` (Mayer et al., 2014) implements the method of Bou-hamad et al. (2009). Schmid et al. (2016) proposed a similar method with time-varying covariates. The main difference is that they include time itself as a potential covariate. Hence, subjects can be split apart even if no time-varying covariates (besides time itself) are used and may produce more accurate estimations in some circumstances.

Unlike the low number of studies on survival-tree techniques capable of handling time-varying covariate, the literature for non-tree-based methods is rich. "Dynamic predictions" has become a growing area of interest. As inferred from the term, when some time-varying covariates are present, estimates of survival probabilities are updated for each subject as new longitudinal information becomes available. At a certain time t , the goal is to provide estimates of survival probabilities at time u ($u > t$) using covariates information history up to t. The two main approaches are joint modeling and landmark analysis. The first one estimates the survival probabilities through joint modelling of the time-varying covariates processes and the event time data(Henderson et al., 2000). This approach relies on a correct specification of the model for time-varying covariates trajectories which can be problematic when the number of time-varying covariates is large. The reader can refer to Tsiatis and Davidian (2004) and Rizopoulos (2012) for an in-dept treatment of the joint modeling approach. The second approach is landmark analysis (Anderson et al., 1983; Madsen et al., 1983). The main idea is to build models, usually Cox, at different landmark times t, using the covariate information available up to t and only the subjects still at risk of experiencing the event at t. Van Houwelingen (2007) and van Houwelingen and Putter (2011) provide comprehensive overviews of this approach. Van Houwelingen and Putter (2011) suggest stacking data sets from many landmark times and fitting a single model to reduce the variability. The landmark approach is more appropriate when the number of time-varying covariates is low since the model gets more complex as their number goes up. One challenge with the landmark approach is that all subjects must have measurements for the time-varying covariate(s) on all landmark points (Huang et al., 2016). Two techniques built around the landmark approach are the partly conditional survival model of Zheng and Heagerty (2005) and the two-stage approach of Huang et al. (2016). Zheng and Heagerty (2005) pointed out that their technique overcomes the "lack of parsimony" often assigned to the fully stratified landmarking approach through explicit modeling of measurement time. Huang et al. (2016) recently proposed a "two-stage" method for dynamic predictions useful when continuous predictions over time is desired. As an advantage over joint models, this technique does not model time-varying covariates trajectories but only the trajectories of their effects. At the first stage, baseline predictions are made based on baseline covariates information only. At the second stage, this method provides the conditional hazard function at any future time point. In this stage, the effect of each time-varying covariate on future survival is the parameter to be estimated. In order to ensure its smoothness over time, the effect of each time-varying covariate is assumed to be a fractional polynomial. Despite overcoming some limitations of the first two approaches, this technique still gets computationally intense when the number of time-varying covariates goes up since it requires estimating parameters of one polynomial for each time-varying covariate plus one for each interaction term between covariates if such interactions are needed. Elgmati et al. (2015) suggested a penalized Aalen additive model for dynamic predictions for discrete-time recurrent event data, but the method is limited to one-step ahead predictions. From the above discussion, we see that no tree-based methods have addressed the problem of dynamic predictions. Indeed, Loh(2014) considers the problem of dealing with time-varying covariates in the survival context as one of the yet remaining challenges despite half a century of progress made in tree-based methods.

The second article of this thesis investigates ways to produce dynamic predictions with discrete-time survival data. When time-varying covariates are present, predictions have to be updated each time new information becomes available. Assuming that T is the largest event time possible (or available in the data set), if the subject is currently alive at time

$t \in \{0, 1, \ldots, T-1\}$, then we are interested in estimating its hazard function at time u for $u = t+1, \ldots, T$. The outcome can be seen as a binary variable y which has a value of 1 if the event occurred at time u and 0 otherwise. We explore three main approaches where random forests can be used for dynamic estimation of the hazard function. The first approach uses only local information builds separate ordinary forests with a binary outcome for each unique combination of (t, u). The second approach pools all u values together for a given t. That is, we build one forest for each value of t and get the whole hazard function for $u > t$ at once. For this approach, two ways of building trees and forests are explored based on the ideas of Bou-Hamad et al. (2011a) and Schmid et al. (2016). The third approach, inspired by the super data set used in landmark analysis (van Houwelingen, 2007; van Houwelingen and Putter, 2011), pools all the information for all combinations of (t, u) at once. This way only one ordinary forest with binary outcome is built for all combinations of (t, u). The forest is fitted to a super person-period data set created by stacking the data sets from all values of t. The results from simulation studies and an application to a real data set show that a simple average of the estimated hazard functions from all these methods works well in most cases.

All present implementations of survival forests are only valid under the independent censoring assumption. This means that they either explicitly state that their methods' underlying assumption is that the true event time and the true censoring time are independent given the covariates, or implicitly by using a splitting rule (e.g. log-rank test) or an estimation method (e.g. Kaplan-Meier, Nelson-Aalen) that works under the independent censoring assumption. None of them can properly estimate the survival function in a dependent censoring context. The third article of this thesis addresses this problem and suggests different ways to build survival forests in this case. Different ways to account for dependent censoring are proposed and investigated. The first one is to use an appropriate method when computing the final survival function estimation with the forest. The second one is to modify the splitting rule. The "copula-graphic" estimator of the survival function introduced by Zheng and Klein (1995), and its closedform expression provided by Rivest and Wells (2001), are used as the basis for these adaptations. Further, a new method, called point wise forest or simply "p-forest", for building survival forests when dependent censoring is suspected is proposed. This method can also be used as a new survival forest method in general. The results from a simulation study indicate that these modifications improve greatly the estimation of the survival forest in situations of dependent censoring.

Chapter 1

L_1 Splitting Rules in Survival Forests

1.1 Abstract

The log-rank test is used as the split function in many commonly used survival trees and forests algorithms. However, the log-rank test may have a significant loss of power in some circumstances, especially when the hazard functions or when the survival functions cross each other in the two compared groups. We investigate the use of the integrated absolute difference between the two children nodes survival functions as the splitting rule. Simulations studies and applications to real data sets show that forests built with this rule produce very good results in general, and that they are often better compared to forests built with the log-rank splitting rule.

Keywords: Survival data; right censored data; ensemble methods; random forests; survival forests.

1.2 Introduction

There have been numerous studies on time-to-event data in a wide range of research areas. One important feature of survival data is that some observations are censored, that is, these observations are incomplete since the event has not yet occurred at the time of data collection. In these situations, we must adequately incorporate all the available information to optimize the prediction models. Parametric models (Gamma, Weibull, etc...), and semi-parametric ones such as the Cox proportional hazard model, can be useful and have been discussed in details in the literature (Hosmer Jr et al., 2011). However, (semi)parametric models have the important limitation that the functional link between the survival time and the covariates must be specified in advance. This is why more flexible nonparametric methods, like survival forests, that let the data automatically find the structure of the model, are useful alternatives (e.g., Hothorn et al., 2006a; Ishwaran et al., 2008; Zhu and Kosorok, 2012).

Tree-based methods were originally developed to model the relation between covariates and either a categorical or a continuous outcome. The Classification and Regression Tree paradigm (CART) is widely popular (Breiman et al., 1984). Survival trees, introduced by Gordon and Olshen (1985), are an adaptation of the tree paradigm to right censored data. A variety of splitting rules have been suggested for survival trees so far. First, Gordon and Olshen (1985) used the idea of imposing homogeneity in each node through the use of a Wasserstein distance between the Kaplan–Meier estimators of the two survival functions. Even though this test does not require any underling assumption, it has not been used much in later works. Wilcoxon-–Gehan statistics and Kolmogorov-–Smirnov test are other metrics to maximize the heterogeneity between two children nodes that were proposed (Ciampi et al., 1988). However, the log-rank statistic proposed by Ciampi et al. (1986) gained the most popularity. The reader can refer to Bou-Hamad et al. (2011) for a review on various splitting statistics proposed in the literature.

As is well known, single trees, despite being very powerful descriptive tools, may be unstable predictive tools. Ensemble methods constructed from trees as base learners such as random forests (Breiman, 2001) can improve the predictive performance through additional randomization. The reader can refer to Siroky (2009) and Verikas et al. (2011) that provided recent surveys on random forests or to Rokach (2009) for a discussion on ensemble methods, in general. A similar argument can be applied to the survival data settings; a combination of survival trees generally leads to higher predictive accuracy. For more in-depth discussions on survival trees and forests, the reader can refer to Bou-Hamad et al. (2011) for a comprehensive overview, and to Boulesteix et al. (2012) and Chen and Ishwaran (2012) for an overview of the subject in genomics and bioinformatics.

Perhaps, the most popular random forest technique for survival data is the one proposed by Ishwaran et al. (2008), called random survival forest (RSF). It is implemented in their R package `randomForestSRC` (Ishwaran and Kogalur, 2014). In this method, an ensemble of cumulative hazard function is built by averaging the Nelson-Aalen cumulative hazard function of each survival tree. It uses the log-rank statistic (Leblanc and Crowley, 1993; Segal, 1988) as the default splitting rule. Other available splitting rules in `randomForestSRC` are log-rank score (Hothorn and Lausen, 2003) and random splitting rule (Cutler and Zhao, 2001; Lin and Jeon, 2006). According to Ishwaran et al. (2008), RSF is the only survival forest technique that adheres to all random forest principles introduced by Breiman (2001). They also provided a built-in new missing data handling algorithm which deals with two problems not addressed by the previous missing data methods for forests: i) the biasedness of out-of-bag estimates of prediction error and ii) the inability to predict on test data sets including missing values. Further, Ishwaran and Kogalur (2010) proved uniform consistency of RSF under the assumption that all variables are categorical. Ishwaran et al. (2010) came up with a regularization strategy for RSF, applicable to survival data where the sample size is small and the number of covariates is large. In their method, the importance of a covariate is measured by the tree depth at which the first split on that covariate happens, a concept called "the minimal depth of a maximal subtree". This technique is useful for both variable selection and variable importance ranking. Ishwaran et al. (2011) provided a follow-up for the use of this method through the `randomSurvivalForest` package, the older version of `randomForestSRC`. They discussed ways to select the tuning parameters of random forest as well as a weighted variable selection technique in order to better regularize the forest. Chen and Ishwaran (2013) studied the use of the minimal depth concept through the `randomSurvivalForest` package in high-dimensional genomic data for effective pathway selection and suggested a "pathway hunting" algorithm for extremely high-dimensional data. Recently, Zhu and Kosorok (2012) proposed a nonparametric regression technique called recursively imputed survival tree (RIST) suitable for right-censored data. In this method, through calculation of the conditional survival distribution, censoring information of observations is retained and then, through recursive imputation and refitting steps, conditional failure information is constantly updated leading to higher predictive accuracy of the final model. The authors suggest three to five steps of imputation to get the best performance. In their implementation, the best split is again obtained through the log-rank test statistic.

From the above discussion, it appears that the log-rank test is routinely used in the various implementations of survival forests. This means that the best split is chosen as the one that makes the two children nodes the most significantly different according to this test. However, it is well known that the log-rank test may have a significant loss of power in some circumstances, especially when the hazard functions or when the survival functions cross each other in the two compared groups (Lin and Wang, 2004; Lin and Xu, 2010). This means that if the goal is to accurately estimate the conditional survival function, then using the log-rank test as splitting criterion may not be adequate. This is why we propose and investigate a splitting rule which works directly with the survival function, defined by:

$$L_1 = (n_L n_R) \int_t |\hat{S}_R(t) - \hat{S}_L(t)| dt, \tag{1.1}$$

where $\hat{S}_L(t)$, $\hat{S}_R(t)$, n_R and n_L are the Kaplan-Meier survival function estimates and the number of observations in the left and right node, respectively. We call it L_1 splitting rule. The L_1 splitting rule is related to the test statistic proposed by Lin and Xu (2010).

The rest of the paper is organized as follows. Section 1.3 describes the data setting and the proposed method. The results from a simulation study are presented in Section 1.4. It aims at comparing the proposed method to traditional and popular methods. Section 1.5 pursues the comparison with real data sets. Section 1.6 concludes and provides directions for further work. Appendices and Supplemental Materials are presented in Section 1.8.

1.3 Survival Forest Approach and Splitting Criterion

We have data on N independent subjects. For subject i, observations are in the form of (τ_i, δ_i, x_i) where τ_i is the observed survival time, δ_i is the censoring index which takes a value of 0 if i is right censored and a value of 1 if i has experienced the event of interest, and x_i is a vector of covariates. Note that only time-invariant covariates are considered in this paper. The true time-to-event and the true censoring times for subject i are denoted by U_i and V_i, respectively. We have τ_i=min(U_i,V_i) and assume that U_i and V_i are independent given x_i. The survival function for subject i is denoted by $S_i(t) = P(U_i > t)$. We use this simplified notation but it should be obvious that τ_i, δ_i, U_i and V_i depend on x_i.

We assume that the reader is familiar with the CART paradigm (Breiman et al., 1984) and the basic random forest method (Breiman, 2001). Basically, a forest is a collection of large unpruned trees built on bootstrap samples from the original data. Moreover, at each node of any tree, a random subset of the predictors are selected at random and the best split is obtained from them. The final forest prediction is the average of the predictions from the individual trees.

The main focus of this paper is to investigate the use of a new splitting criterion to build the trees in the forest. Assume that we are at a given node of a tree and we want to split it in two children nodes. If x is continuous (or at least ordinal), the possible splits take the form $C = I(x \leq c)$. If x is categorical, the possible splits take the form $C = x \in \{c_1, \ldots, c_q\}$ where $\{c_1, \ldots, c_q\}$ is a subset of the possible values of x. Once the best split is found, observations with $C = 0$ go to the left node and the ones with $C = 1$ go to the right node. We saw in the introduction that, typically, the log-rank test with the two children nodes acting as the two samples is used as the splitting criterion. However, the log-rank test has a low power for detecting differences between the two groups in some situations. For the testing problem, Lin and Xu (2010) proposed a new method that has greater power than the log-rank test under a variety of situations. To avoid introducing unnecessary notation, suppose we want to test the equality of the survival functions in two children nodes, L (left) and R (right). Their test is based on

$$\begin{aligned}\Delta &= \int_0^\tau |\hat{S}_L(t) - \hat{S}_R(t)|dt \\ &= \sum_{j|t_j<\tau} |\hat{S}_L(t_j) - \hat{S}_R(t_j)|(t_{j+1} - t_j)\end{aligned}$$

where S_L and S_R are the Kaplan-Meier estimators of the survival function in nodes L and R, $t_1 < t_2 < \cdots < t_k$ are the pooled distinct event times, and τ is the last time point by which the areas under the survival curves can be calculated for both groups. To perform a formal test, Lin and Xu (2010) use the standardized statistic $\Delta^* = (\Delta - \hat{E}(\Delta))/(\widehat{Var}(\Delta))^{1/2}$ where $\hat{E}(\Delta)$ and $\widehat{Var}(\Delta)$ are suitable estimates of the mean and variance of Δ. However, since the splitting criterion has to be evaluated a large number of times when building a forest, we proceed to a simplification, detailed in the Appendix, to speed up computations. With this simplification, we consider

$$L_1^* = \sqrt{n_L n_R}\Delta = \sqrt{n_L n_R}\int_t |\hat{S}_L(t) - \hat{S}_R(t)|dt \tag{1.2}$$

as the splitting criterion, where n_L and n_R are the left and right node sizes. We call it the L_1^* splitting criterion, as opposed to the L_1 splitting criterion given by $L_1 = (n_L n_R)\Delta$ and introduced in (1.1). As we will see in the next sections, both versions provide good results but the L_1 criterion was slightly better in the cases considered in this paper. For completeness, we will report the results for both splitting criteria. A more complete discussion about these two criteria appears in the concluding remarks section. To use any of these criteria for tree building, we simply compute it with the two groups formed by the left and right nodes for each candidate binary split. The one with the maximal value is the best split.

The forest algorithm can be described as follows:

1. Draw B bootstrap samples from the original data.

2. For each bootstrap sample, grow a tree with the L_1 (or L_1^*) splitting criterion. At each node, randomly select k out of p covariates where $k \leq p$ and is a user-specified

parameter. Splitting ends when a stopping criterion is reached; for instance, when a node has less than a predetermined number of observations. No pruning is performed.

3. To compute the estimated survival function of an observation with covariate vector x, use a "similarity" weighting scheme as in Hothorn et al. (2006a). More precisely, send x in all of the B trees and collect all the observations that end in the same terminal nodes. Note that some observations may appear more than one time. $\hat{S}(t|x)$ is then the Kaplan-Meier estimate of this pooled set of observations.

Evaluating the performance of any model on a given data set with survival data is not a straightforward task because of the censoring. One approach is to use the Brier score (Graf et al., 1999) and other criteria derived from it. The `R` package `pec` (Mogensen et al., 2012) can be useful for that matter. We will use one of them, the integrated Brier score, in Section 4 when we analyse real data sets.

1.4 Simulation Study

In this section, we investigate the performance of our proposed method through a simulation study. Nine methods are compared, seven forest methods and two benchmarks. They are:

A forest where trees are build with the proposed L_1 splitting rule. It is denoted by L_1-forest. A forest where trees are build with the proposed L_1^* splitting rule. It is denoted by L_1^*-forest. A forest where trees are build with the log-rank splitting rule. It is denoted by RFsrc1. A forest where trees are build with the log-rank score splitting rule. It is denoted by RFsrc2. A forest where trees are build with the random splitting rule. It is denoted by RFsrc3. A forest built with three-step imputation in the RIST method (Zhu and Kosorok, 2012). It is denoted by RIST3. A forest built with five-step imputation in the RIST method (Zhu and Kosorok, 2012). It is denoted by RIST5. A Cox model where the covariates are entered linearly with main effects only. This is the first benchmark. It is denoted by Cox. A Kaplan-Meier estimator, which does not use the covariates. This is the second benchmark. It is denoted by KM.

The L_1-forest and L_1^*-forest are implemented in Fortran and callable from `R` (R Development Core Team, 2014). The `R` package `randomForestSRC` (Ishwaran and Kogalur, 2014) was used for RFsrc1, RFsrc2 and RFsrc3. The R code generously made available by the authors Zhu and Kosorok (2012) was used for RIST3 and RIST5. Note that in case of no censoring in the data, no imputation is required with the RIST method. Therefore, a simple forest was used in these scenarios. We denote it by RIST0. The `R` package `survival` (Therneau, 2014) was used for both the Cox model and the Kaplan-Meier estimator.

In all seven forest methods, 100 trees are grown and the number of covariates tried at each split is set to the integer part of $\sqrt{p}$, as suggested by Ishwaran et al. (2008). As a stopping criterion, the minimum number of observations in a terminal node is 3, the default value in `randomForestSRC`.

To evaluate the performance of these methods, two commonly used criteria were employed to measure how well the survival function is estimated. Assume that S is the true survival function and that $\hat{S}$ is the estimated survival function. The two criterion are the Integrated Absolute Error (IAE) and the Integrated Square Error (ISE) defined by:

$$\text{IAE} = \int_t |S(t) - \hat{S}(t)| dt \tag{1.3}$$

and

$$\text{ISE} = \int_t (S(t) - \hat{S}(t))^2 dt. \tag{1.4}$$

Since the results for the ISE were very similar, only the ones for the IAE are reported.

1.4.1 Simulation Design

In the main simulation study, five Data Generating Processes (DGPs) are used to generate artificial data. For each DGP, six different censoring proportion ranging from 0% to 50% are considered, namely, 0%, 10%, 20%, 30%, 40% and 50%. Thus, overall 30 scenarios are investigated.

Each model is fitted with a training sample of size 500. Then the performance of the fitted models is evaluated with an independent test set of size 1000. Each simulation is repeated 500 times. Here are a detailed description of the DGPs. In all cases, the parameter α controls the proportion of censoring. The values of α which produce the desired censoring proportions were found empirically.

DGP 1. The model is a tree with two equally likely terminal nodes with respective survival functions illustrated in Figure 1.1. Ten iid uniform covariates on the interval (0,1) are available, $X_1, \ldots, X_{10}$, but the response is only related to X_1. The censoring times are uniformly distributed on the interval $(0,\alpha)$. The hazard function is presented in the Appendix.

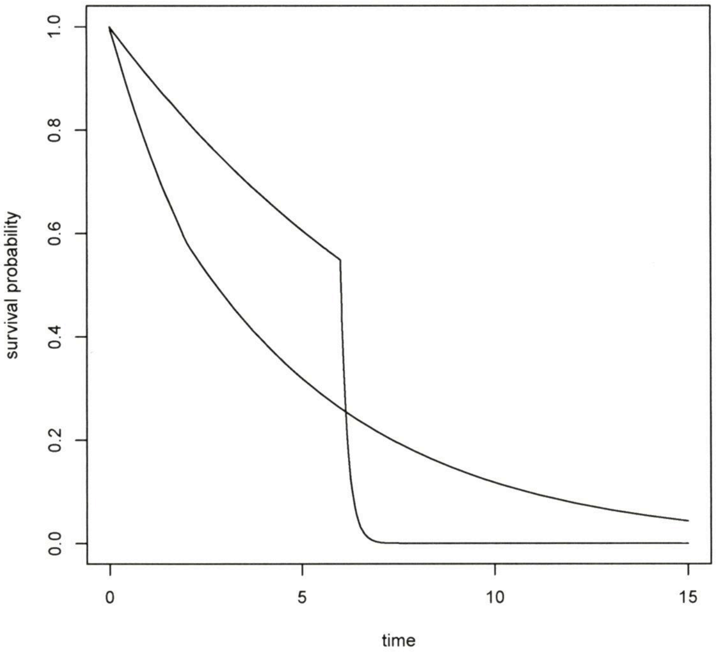

Figure 1.1: The two survival curves in the terminal nodes of the tree for DGP 1

DGP 2: This DGP is slightly more complex than DGP 1. It is a tree with four equally likely terminal nodes with respective survival functions illustrated in Figure 1.2. Again, ten iid uniform covariates on the interval (0,1) are available, $X_1, \ldots, X_{10}$. The response is related to X_1 and X_2. The censoring times are uniformly distributed on the interval $(0,\alpha)$. The hazard function is presented in the Appendix.

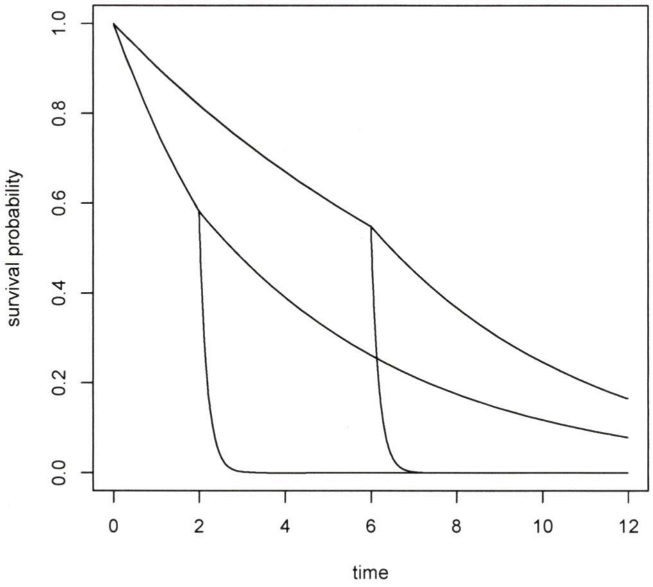

Figure 1.2: The four survival curves in the terminal nodes of the tree for DGP 2

DGP 3. It is an altered version of scenario 2 from section 4.1 of Zhu and Kosorok (2012). Ten iid uniform covariates on the interval (0,1) are available, $X_1, \ldots, X_{10}$. Survival times are drawn from an exponential distribution with mean μ where $\mu = 10|\sin(X_1\pi - 1)| + 3|X_2 - 0.5| + X_3$. The censoring times are uniformly distributed on the interval $(0,\alpha)$.

DGP 4. It is adapted from scenario 3 in section 4.1 of Zhu and Kosorok (2012). Twenty-five covariates $X_1, \ldots, X_{25}$ are generated from a multivariate normal distribution with covariance matrix $\sigma_{ij} = 0.75^{|i-j|}$. The survival time follows a gamma distribution with shape parameter $\mu = 0.5 + 0.3|\sum_{i=11}^{15} X_i|$ and scale parameter of 2. The censoring times are uniformly distributed on the interval $(0,\alpha)$.

DGP 5. This is a dependent censoring DGP. It is adapted from scenario 1 in section 4.1 of Zhu and Kosorok (2012). Twenty-five covariates $X_1, \ldots, X_{25}$ are generated from a multivariate normal distribution with covariance matrix $\sigma_{ij} = 0.9^{|i-j|}$. The survival time follows an exponential distribution with mean of $\mu = 0.1|\sum_{i=11}^{20} X_i|$. The censoring times are drawn from an exponential distribution with mean μ/α.

1.4.2 Simulation Results

We first present a global summary of the results in Table 1.1. For each scenario with some censoring (that is, excluding the 0% censoring case), we are comparing nine methods. For each individual data set, we ranked these methods from 1 to 9 with respect to the IAE criterion (1.3) evaluated on the test set. The rank of one was given to the method with the lowest value of the IAE, hence the best one for this data set. Table 1.1 reports the average ranks over all 12,500 simulation runs with censoring. Namely, over the 500 repetitions × 5 proportions of censoring × 5 DGPs. We see that the L_1 and L_1^* forests obtained the best results overall. The L_1-forest came in 2.13^{th} place among the nine methods, on average, while the L_1^*-forest came in 2.78^{th} place. The two RIST methods have the next best average ranks, followed by RFsrc1. Not surprisingly, the KM method which does not use the covariates, comes in last.

The detailed results for all DGPs are summarized in Figures 7 to 11 of the Supplementary Material. As an overall summary, only results for DGP 1, DGP 3 and DGP 4 at 10% and 40% censoring proportions are presented in Figure 1.3. As expected, the L_1-type forests are the best performing method in terms of IAE in the first two DGPs with crossing survival functions. RIST comes in second place for DGP 1 while RIST and RFsrc1 also perform well for DGP 2. The first two DGPs were designed explicitly to exhibit the advantage of the proposed methods when the survival functions are crossing. But what is even more interesting is that the L_1-type forests also do very well for the other DGPs, that do not involve crossing survival functions. For DGP 3, the RFsrc1 is the best for censoring proportions up to 20%, then the L_1^*-forest does better when the censoring proportion reaches 40%. For DGP 4 and 5, RIST performs best followed closely by the L_1-type forests. Hence the L_1-type forests seem to be generally competitive in a wide variety of situations.

Name	Cox	KM	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
Average rank	5.60	7.80	5.15	7.84	6.66	3.64	3.37	2.12	2.78

Table 1.1: Average ranks (smaller is better), according to the IAE criterion, of the nine methods over all individual data sets (12,500) with censoring.

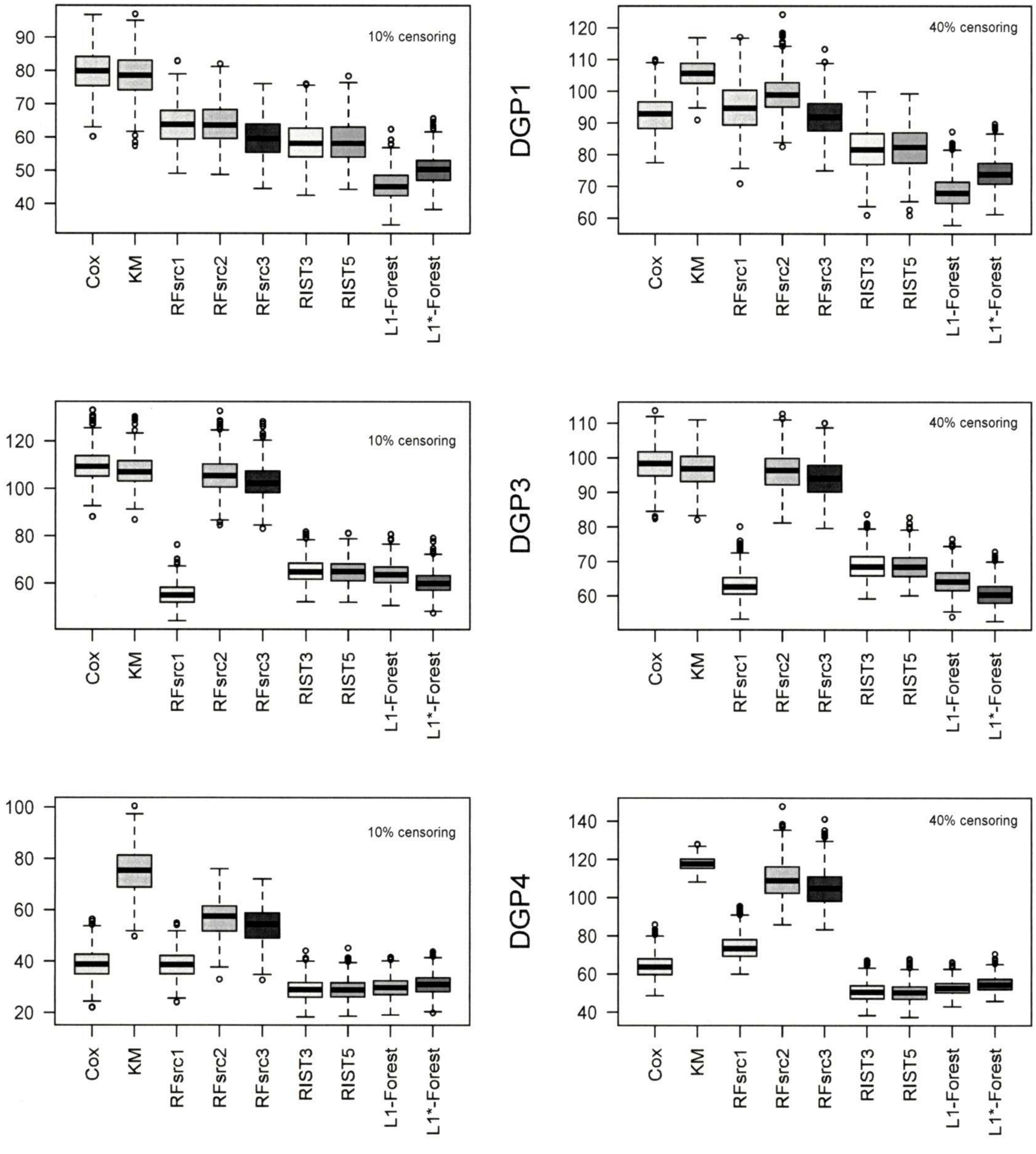

Figure 1.3: IAE of methods for DGP 1, DGP 3 and DGP 4 at 10% and 40% censoring proportions

1.4.3 Additional Simulations

We present briefly the results of some additional simulations following reviewers suggestions. Complete results with seven additional figures are available in the Supplementary Material.

Firstly, the performance of the methods is investigated with a smaller training sample and with additional noise covariates. The same 30 scenarios are investigated. But the training sample size is divided by two (hence it is 250) and the number (which depends on the DGP) of noise covariates is multiplied by 5 in each DGP. The added noise covariates are iid uniform covariates on the interval (0,1). Table 1.2 presents the average ranks of the nine methods for these modified DGPs. The results are very similar to before. Again, L_1 type forests obtained the best results overall, followed by the two RIST methods and then by RFsrc1.

Method	Cox	KM	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
Average rank	7.90	6.26	4.57	7.39	6.85	3.76	3.71	2.26	2.30

Table 1.2: Average ranks (smaller is better), according to the IAE criterion, of the nine methods over all individual data sets (12,500) with censoring, for the modified DGPs (smaller n and larger p).

Secondly, since all covariates are continuous in the main simulation, a few scenarios with binary covariates are investigated here. Only DGP 2 is considered. Recall that only two covariates, X_1 and X_2, are related to the response. In the main simulation, X_1 and X_2 are uniformly distributed on (0,1). In the first variation, X_2 is still uniformly distributed on (0,1) but X_1 is now a binary covariate taking values 0 and 1 with probability $1/2$. In the second variation, both X_1 and X_2 are binary covariates taking values 0 and 1 with probability $1/2$. Table 1.3 presents the average ranks, as before, for DGP 2 only, for these three situations. For the original setup where both X_1 and X_2 are continuous (top part of the table), the L_1 type forests are the best but the Cox model comes in third place, followed by the two RIST methods. Then, when X_1 is binary and X_2 continuous (middle part), the Cox model comes in between the L_1 type forests, followed again by the two RIST methods. Finally, when both X_1 and X_2 are binary (bottom part), then the Cox model has the best performance followed by the L_1-forest. But this time both RIST perform better than the L_1^*-forest. Again, more complete results are presented in the Supplementary Material which suggest that the IAE of the forest methods did not really degrade when moving from the continuous to the binary covariates. Rather, it is the performance of the Cox model that improved.

X_1 and X_2 continuous (original setup)									
Name	Cox	KM	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
Ranking	3.75	7.18	5.42	7.58	6.77	5.17	5.19	1.65	2.24

X_1 binary and X_2 continuous									
Name	Cox	KM	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
Ranking	2.63	7.25	6.69	7.70	6.39	4.39	4.50	2.39	3.04

X_1 and X_2 binary									
Name	Cox	KM	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
Ranking	2.42	6.76	7.15	7.57	5.65	4.07	4.08	2.99	4.31

Table 1.3: Average ranks (smaller is better), according to the IAE criterion, of the nine methods over all individual data sets (2,500) with censoring for DGP 2. The original setup and two variations are presented.

1.5 Real Data Sets

In this section, we compare the performance of the same methods used in the simulation study with six real data sets: The Primary Biliary Cirrhosis (PBC) data, the CSL liver chirrosis data, the German Breast Cancer (GBC) Study Group data, the Wisconsin Breast Cancer Prognostic (WPBC) data, the Veteran data, and the National Wilm's Tumor Study (NWTCO) data. A brief description of these data sets is presented in Table 1.4.

The PBC data is described in the monograph by Fleming and Harrington (1991). We use all twelve covariates used by Bou-Hamad et al. (2011) plus copper, sgot and stage. The same 312 patients who participated in the randomized trial are used here. Missing values are replaced by the median as in Bou-Hamad et al. (2011) and Fleming and Harrington (1991). The CSL data was obtained by Schlichting et al. (1983) and is provided in the `timereg` package (Scheike et al., 2009). In this example, we only use the six time-invariant covariates. Records are grouped by id variable so the number of observations used is 446. The GBC data (Schumacher et al., 1994) is obtained from the package `mfp` (Ambler and Benner, 2014). The data contains 686 observations and eight covariates. There is no missing data. The WPBC data is available in the UCI machine learning repository (Bache and Lichman, 2013). There

Name	# Covariates	Sample Size	% Censoring	Source
PBC	15	312	60%	(Fleming and Harrington, 1991)
CSL	6	446	39%	(Schlichting et al., 1983) in `timereg` package (Scheike et al., 2009)
GBC	8	686	56%	(Schumacher et al., 1994) in `mfp` package (Ambler and Benner, 2014)
WPBC	32	198	76%	(Bache and Lichman, 2013)
Veteran	6	137	7%	(Kalbfleisch and Prentice, 1980) in `randomForestSRC` package (Ishwaran and Kogalur, 2014)
NWTCO	4	4088	85%	(Breslow and Chatterjee, 1999) in `survival` package (Therneau, 2014)

Table 1.4: Description of the data sets

are 198 observations in the data. However, four missing values are replaced by the median as in the PBC data. Thirty-two covariates are used in this example. The Veteran data (Kalbfleisch and Prentice, 1980) is obtained from the `randomForestSRC` package (Ishwaran and Kogalur, 2014). There are 137 observations with no missing values. It contains six covariates. Finally, the NWTCO data (Breslow and Chatterjee, 1999) is available in the package `survival` (Therneau, 2014). The four relevant covariates, instit, histol, age and stage, are used here. The data consists of 4088 observations and no missing values.

We use the same parameters as in the simulation section to build the forests. Namely, 100 trees are grown, the number of covariates tried at each split is set to the integer part of $\sqrt{p}$, and the minimum number of observations in a terminal node is 3.

Since the true survival function is not known, we can not compute the IAE (or ISE) as we did with the artificial data sets. Instead, our primary criterion is the integrated Brier score (Graf et al., 1999). Let $\hat{S}(t|x)$ denote the estimated survival function, estimated by any model, at time t for a subject with covariate vector x. Let $\hat{G}$ denote the Kaplan-Meier estimate of the censoring distribution. The Brier score at any time t is computed as

$$\mathrm{BS}(t) = \frac{1}{N}\sum_{i=1}^{n}\Bigg((\hat{S}(t|x_i)^2 I(\tau_i \leq t \quad and \quad \delta_i = 1)\hat{G}^{-1}(\tau_i) + (1-\hat{S}(t|x_i))^2 I(\tau_i > t)\hat{G}^{-1}(t)\Bigg).$$

The integrated Brier score is given by

$$\mathrm{IBS} = \frac{1}{max(\tau_i)}\int_0^{max(\tau_i)} BS(t)dt.$$

Lower values of IBS indicate better performances. Basically, the IBS is an integrated weighted squared distance between the estimated survival function and the empirical survival curve. For each subject, censoring time does not depend on the covariate vector. The inverse weighting scheme is used to adjust for censoring. It is thus similar in spirit to the IAE used in the previous section.

Name	Cox	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
PBC	5.08	3.73	12.55	8.89	1.78	2.29	1.66	0.00
Veteran	6.72	0.39	2.47	3.38	0.00	0.77	1.05	0.22
GBC	0.12	3.92	4.26	4.58	0.96	0.46	1.20	0.00
CSL	0.00	13.72	11.73	6.41	8.46	8.22	5.32	5.43
NWTCO	1.95	7.62	4.96	0.62	1.02	1.07	0.00	0.71
WPBC	19.30	6.27	3.78	2.45	0.04	0.00	0.90	2.33
Average	5.52	5.90	6.62	4.36	2.03	2.14	1.69	1.44

Table 1.5: Ranking of methods based on % increase of median IBS with respect to best method

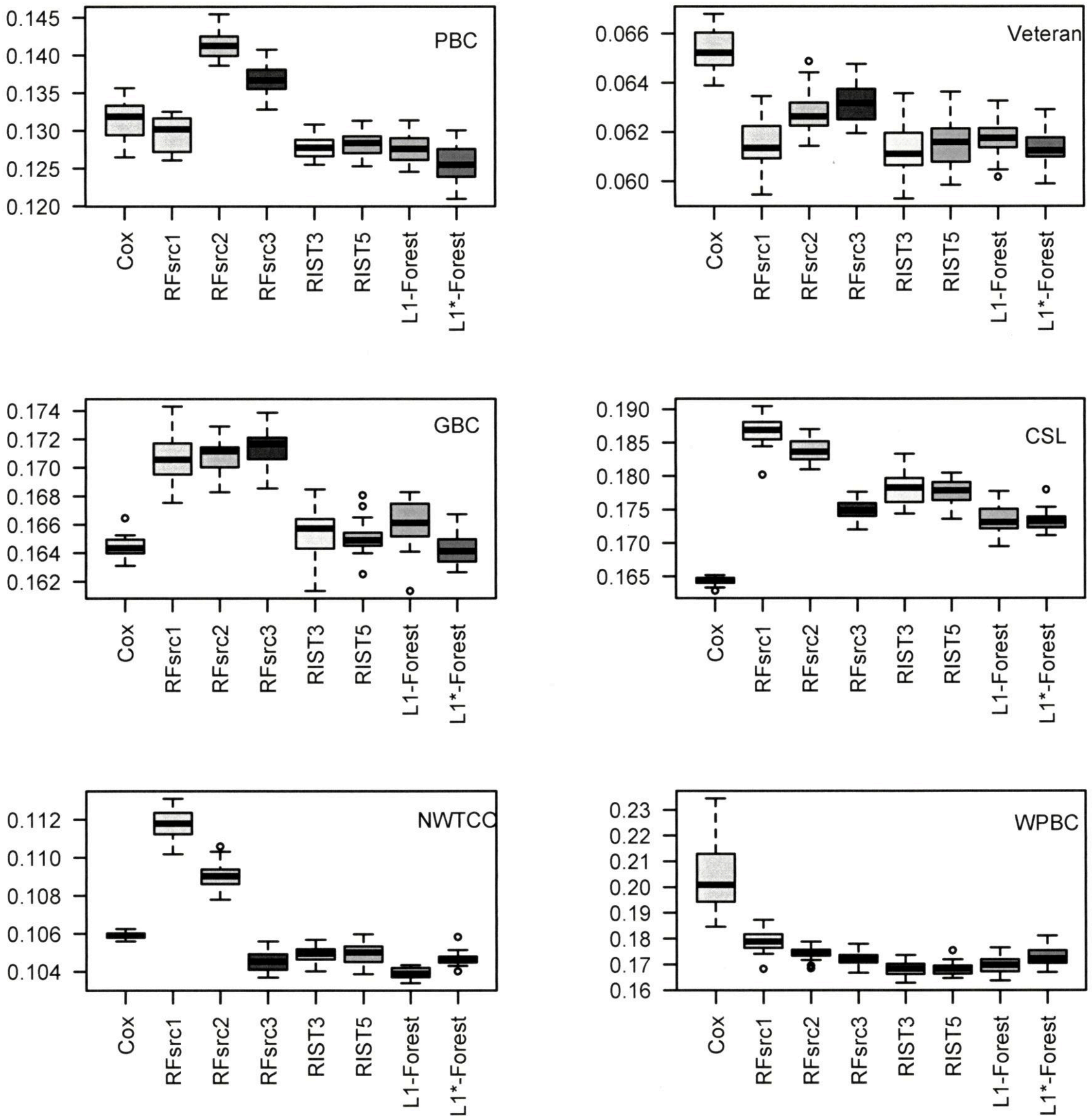

Figure 1.4: Integrated Brier score, across 20 runs of 10-–fold cross-validation, for the real data sets

Figure 1.4 illustrate the results for the IBS across 20 runs of 10-fold cross-validation for each data set. Note that the results for the Kaplan-Meier method are not included because it is so much worse than the other methods that the plots would have been distorted. Table 1.6 provides another look at these results by computing the % increase of the median IBS with respect to best method for each data set. The median here is the one over the 20 runs of 10-–fold cross-validation. We see in Table 1.5 that the L_1 type forests are globally the best according to the average % increase in IBS. Indeed, the IBS of L_1^*-forest is only 1.44% greater than the IBS of the best method, on average. It is even the best one for two of the six data sets. Moreover, the IBS of L_1-forest is 1.69% greater than the IBS of the best method, on average. It is also the best method for the NWTCO data. The two RIST methods have the next best performance with average percent increases just above 2%. Hence, we get the same top four methods as the ones we obtained with the artificial data sets. The fact that the L_1 type forests are always competitive for all data sets is clearly seen in Figure 1.4.

Name	Cox	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
PBC	0.94	1.21	1.05	0.00	0.35	0.52	1.09	0.86
Veteran	0.00	0.37	1.72	4.22	0.54	0.17	0.92	1.26
GBC	1.81	0.36	0.00	0.05	0.15	0.09	0.90	0.07
CSL	0.00	4.40	4.17	4.09	4.64	4.46	4.48	4.25
NWTCO	0.57	8.64	7.28	3.83	2.48	2.52	0.00	0.09
WPBC	4.36	7.49	0.27	0.00	0.04	1.05	2.80	5.08
Average	1.28	3.74	2.41	2.03	1.37	1.47	1.70	1.94

Table 1.6: Ranking of methods based on % decrease of median C-index with respect to best method

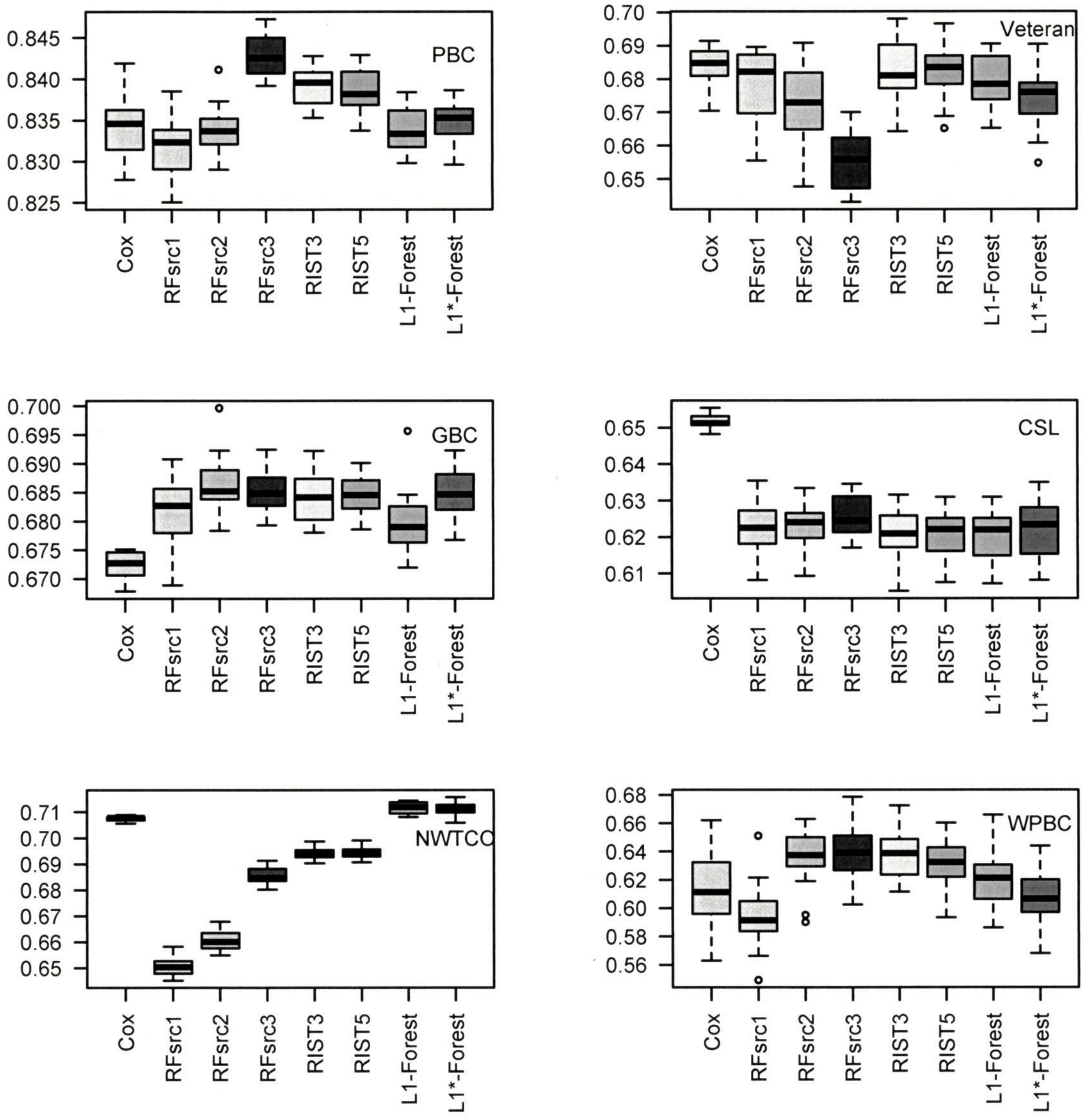

Figure 1.5: C-index, across 20 runs of 10--fold cross-validation, for the real data sets

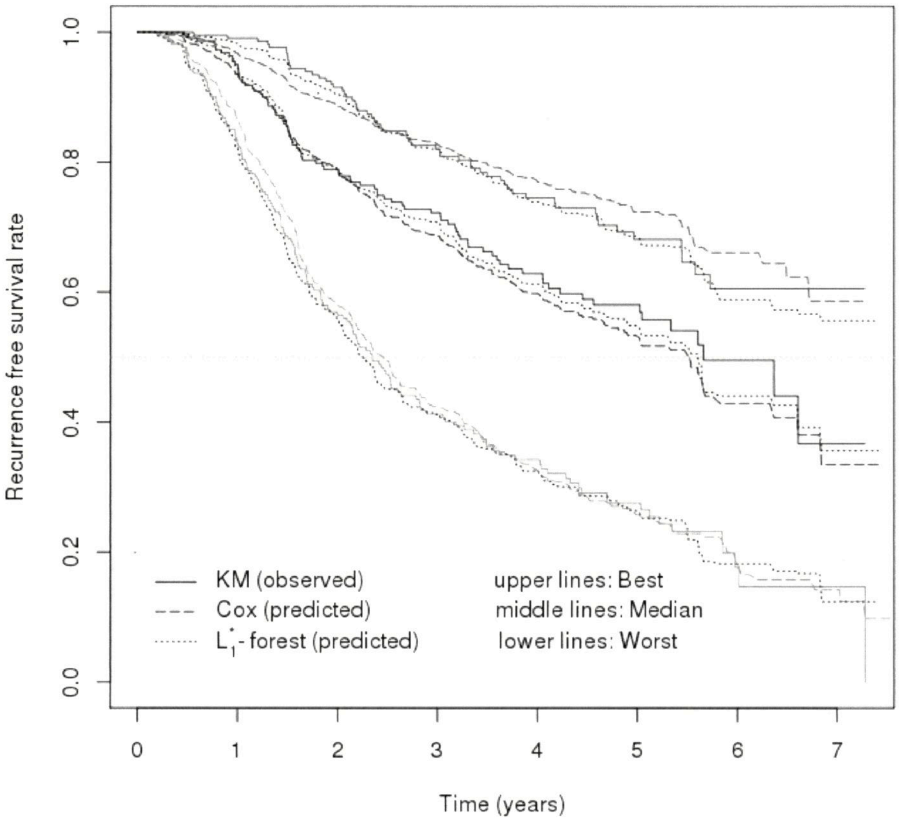

Figure 1.6: Predicted and observed survival curves by prognosis group in GBC data

Another popular criterion to evaluate a model with survival data is the C-index (Harrell et al., 1982). We use it as a complement measure here because we think the IBS is more appropriate when the goal is to estimate the survival function. The C-index is a concordance measure that evaluates if the predictions from a model are ranked in the same way as the observed times. Following Ishwaran et al. (2008), let $\hat{H}(t|x)$ denote the estimated cumulative hazard function, estimated by any model, at time t for a subject with covariate vector x. Let $t_1 < t_2 < \cdots < t_m$ be the distinct event times and let $\hat{H}_i = \sum_{l=1}^{m} \hat{H}(t_l|x_i)$. The C-index, using the usable pairs, is computed as following

$$\mathrm{CI} = \frac{\sum_{i<j}(I(t_i < t_j)I(\hat{H}_i > \hat{H}_j)\delta_i + I(t_i > t_j)I(\hat{H}_i < \hat{H}_j)\delta_j)}{\sum_{i<j}(I(t_i < t_j)\delta_i + I(t_i > t_j)\delta_j)}.$$

Higher values of CI indicate better performances. Figure 1.5 illustrate the results for the CI across 20 runs of 10-–fold cross-validation for each data set. Table 1.6 reports the % decrease of the median CI with respect to best method for each data set. We see in Table 1.6 that the Cox model has the best performance according to the CI, followed by the two RIST methods and the L_1 type forests. The average percent decrease in CI of these methods are all below 2%. As a matter of fact, all methods do fairly well expect maybe for RFsrc1 which is a bit further apart. Hence, it seems that the CI is a less discriminating criterion compared to the IBS, for these data sets. Zhu and Kosorok (2012) also report that the C-index is not as sensitive as other measurements and even that its interpretability is sometimes unclear. This may be partly explained by the fact that the C-index is uniquely a discrimination measure. That is, it measures if the predicted survival times are in the right order. The IBS is a discrimination and calibration measure. The calibration aspect measures the similarity between the actual and predicted survival curves; see Riccardo et al. (2014). Hence for these data sets, it seems that all methods do fairly well in terms of discrimination, but the L_1 type forests perform better in terms of calibration.

To conclude this section, we will take a closer look at the GBC data analysed in details in Sauerbrei and Royston (1999). We will refer to this paper by SR for simplicity. After a careful analysis and using fractional polynomials, SR came up with a seemingly good

Cox model based on the covariates $(X_1/50)^{-2}$, $(X_1/50)^{-0.5}$, $I(X_4 \geq 2)$, $\exp(-0.12X_5)$, and $(X_6+1)^{0.5}$; see Model III in Table 4 of SR. Still following SR, we divide the sample into three groups of nearly equal sizes based on the prognostic index $\mathrm{PI}_i = x_i\hat{\beta}$, where $\hat{\beta}$ are the estimates of the parameters in the above Cox model. The 228 subjects with the lowest PI scores are assigned to the best prognosis group, the following 229 subjects are assigned to the median prognosis group, and the remaining 229 with the highest PI scores go into the worse prognosis group. Note however that SR used another simpler model to define the three groups. But that model had only 27 different covariates patterns which caused the group sizes to be imbalanced. Figure 4 of SR shows the Kaplan-Meier curves of each group. Fleming and Harrington (1991) present a similar plot (see their Figure 4.6.13 on page 195) in their analysis of the PBC data set. However, they also plotted the average estimated survival curves of each group in order to visually inspect the goodness-of-fit of their model. Figure 6 is such a plot where the average survival curves of the Cox model and of the L_1^*-forest are depicted. We can observe that the Cox model seems to fit the data fairly well in all groups. But strikingly, the L_1^*-forest seems to fit the data slightly better, even though the groups are derived from the Cox model. This simple example illustrates the fact that, sometimes, a good off-the-shelf method like a forest do as well as a carefully crafted parametric model. When confronted between two models with similar performance, the choice should often be the one which is easier to interpret. In this case, it would be the Cox model even though interpreting the effect of the transformed covariates is not straightforward. Thus, a forest can serve as a benchmark to evaluate if an easier to interpret parametric model fits well enough.

1.6 Discussion and Concluding Remarks

The log-rank test is commonly used as the splitting rule in the various implementations of survival forests within the CART paradigm, such as in Ishwaran et al. (2008) and Zhu and Kosorok (2012). However, the log-rank test is not designed to detect all possible differences between two survival curves. For instance, the log-rank test is inadequate to detect a difference between two groups when the hazard or survival functions cross each other in the two compared groups. Consequently, if the goal is to accurately estimate the conditional survival function, then using the log-rank test as splitting criterion may not be optimal. This was never thoroughly investigated for survival forests. At a time where more refinements and features are added to existing packages, it seemed that going back to "basics" was in order. In this paper, it was showed that forests built with a simple splitting rule, based on the integrated absolute difference between the two children nodes survival functions, are very competitive compared to forests built with the log-rank splitting rule. Indeed, these forests often got the best performance in the cases considered, either with simulated data or with real data sets. Hence, it is certainly worthwhile to consider the proposed methods as potential competitors. It would certainly be helpful if some well established, very comprehensive and useful package like `randomForestSRC` could incorporate L_1-type splitting rules.

The two splitting rules investigated are $L_1 = (n_L n_R)\Delta$ and $L_1^* = \sqrt{n_L n_R}\Delta$ where $\Delta = \int_t |\hat{S}_L(t) - \hat{S}_R(t)|dt$. The factors $(n_L n_R)$ and $\sqrt{n_L n_R}$ can be seen as penalization factors that favor splits with children nodes of nearly equal sizes. For instance, if we have two potential splits with the same value of Δ, then the one with the largest value of $n_L n_R$ should be favored because the two $\hat{S}$ in Δ are obtained from larger sample sizes. As an extreme example, assume that $n_L + n_R = 100$ and that two splits have an equal value of Δ. We would be more confident with a split with $n_L = 50$ and $n_R = 50$, than one with $n_L = 5$ and $n_R = 95$ because, in the second case, there is a lot of variability in the estimation of $\hat{S}_L$ which induces a larger variance for Δ. In fact, the Appendix establishes that, under some simplifying assumptions, the factor $\sqrt{n_L n_R}$ is the one producing a test statistic which is asymptotically normally distributed. The factor $\sqrt{n_L n_R}$ favors more heavily splits with nearly equal size children nodes than the factor $n_L n_R$. But in practice, at a given node, we do not know if the best split is located more towards the center (with respect to the children nodes sizes) or not. So we do not know how much we should favor splits towards the center. Even if the factor $\sqrt{n_L n_R}$ has a theoretical justification, the factor $n_L n_R$ had globally a slightly better performance with the data generating processes considered in the simulation study. But with the real data sets, the factor $\sqrt{n_L n_R}$ had a slightly better performance according to the integrated Brier score and was slightly worse according to the C-index. In practice, one possibility would be to try forests built with both splitting rules and select the one according to the integrated Brier score under a cross-validation scheme. More research on this aspect could be interesting.

The scope of this work is limited to traditional forests built within the CART paradigm. Other paradigms including "unbiased" trees are also available; see Loh (2002) and Loh (2013) for the GUIDE approach and Hothorn et al. (2006b) for the ctree approach. For example, the GUIDE approach fits different types of proportional hazards regression models for censored data. Hence, it would be interesting to investigate the robustness of this method when the proportionality assumption is not met. Moreover, we saw in the additional simulations that the covariates' types can have an impact on the performance. Hence, developing forests built with unbiased trees using L_1 types criteria for variable and split selections might be interesting.

1.7 Acknowledgement

The authors would like to thank the Associate Editor and two reviewers whose comments helped in preparing an improved version of this article. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by Le Fonds québécois de la recherche sur la nature et les technologies (FRQNT).

Bibliography

G. Ambler and A. Benner. *mfp: Multivariable Fractional Polynomials*, 2014. URL `http://CRAN.R-project.org/package=mfp`. R package version 1.5.0.

K. Bache and M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics surveys*, 5:44–71, 2011.

A.L. Boulesteix, S. Janitza, J. Kruppa, and I.R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics.

Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6):493–507, 2012.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and regression trees. wadsworth & brooks. *Monterey, CA*, 1984.

N.E. Breslow and N. Chatterjee. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(4):457–468, 1999.

X. Chen and H. Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6): 323–329, 2012.

X. Chen and H. Ishwaran. Pathway hunting by random survival forests. *Bioinformatics*, 29 (1):99–105, 2013.

A. Ciampi, J. Thiffault, J.P. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational statistics & data analysis*, 4(3):185–204, 1986.

A. Ciampi, S.A. Hogg, S. McKinney, and J. Thiffault. Recpam: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. i. methods and program features. *Computer Methods and Programs in Biomedicine*, 26(3):239–256, 1988.

A. Cutler and G. Zhao. Pert-perfect random tree ensembles. *Computing Science and Statistics*, 33:490–497, 2001.

T.R. Fleming and D.P. Harrington. Counting processes and survival analysis. *John Wiley&Sons, Hoboken, NJ, USA*, 1991.

L. Gordon and R.A. Olshen. Tree-structured survival analysis. *Cancer treatment reports*, 69 (10):1065, 1985.

E. Graf, C. Schmoor, W. Sauerbrei, and M. Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

F.E. Harrell, R.M. Califf, D.B. Pryor, K.L. Lee, and R.A. Rosati. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

D.W. Hosmer Jr, S. Lemeshow, and S. May. *Applied survival analysis: regression modeling of time to event data*, volume 618. Wiley. com, 2011.

T. Hothorn and B. Lausen. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2):121–137, 2003.

T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M.J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006a.

T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006b.

H. Ishwaran and U.B. Kogalur. Consistency of random survival forests. *Statistics & probability letters*, 80(13):1056–1064, 2010.

H. Ishwaran and U.B. Kogalur. Random forests for survival, regression and classification (rf-src). 2014. URL `http://cran.r-project.org/web/packages/randomForestSRC/`. R package version 1.5.5.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

H. Ishwaran, U.B. Kogalur, E.Z. Gorodeski, A.J. Minn, and M.S. Lauer. High-dimensional variable selection for survival data. *Journal of the American Statistical Association*, 105 (489):205–217, 2010.

H. Ishwaran, U.B. Kogalur, X. Chen, and A.J. Minn. Random survival forests for high-dimensional data. *Statistical analysis and data mining*, 4(1):115–132, 2011.

J.D. Kalbfleisch and R.L. Prentice. The statistical analysis of failure time data. *Wiley series in probability and mathematical statistics*, 1980.

M. Leblanc and J. Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467, 1993.

X. Lin and H. Wang. A new testing approach for comparing the overall homogeneity of survival curves. *Biometrical Journal*, 46(5):489–496, 2004.

X. Lin and Q. Xu. A new method for the comparison of survival distributions. *Pharmaceutical statistics*, 9(1):67–76, 2010.

Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

W.Y. Loh. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12(2):361–386, 2002.

W.Y. Loh. Guide classification and regression trees user manual for version 15. 2013.

U.B. Mogensen, H. Ishwaran, and T.A. Gerds. Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software*, 50(11):1, 2012.

D.B. Riccardo, W. Sauerbrei, and A.L. Boulesteix. Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. *Statistics in medicine*, 33(30):5310–5329, 2014.

L. Rokach. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. *Computational Statistics & Data Analysis*, 53(12):4046–4072, 2009.

W. Sauerbrei and P. Royston. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):71–94, 1999.

T. Scheike, T. Martinussen, and J. Silver. timereg: timereg package for flexible regression models for survival data. *R package version*, pages 1–2, 2009.

P. Schlichting, E. Christensen, P.K. Andersen, L. Fauerholdt, E. Juhl, H. Poulsen, and N. Tygstrup. Prognostic factors in cirrhosis identified by cox's regression model. *Hepatology*, 3(6):889–895, 1983.

M. Schumacher, G. Bastert, H. Bojar, K. Huebner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R.L. Neumann, and H.F. Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *Journal of Clinical Oncology*, 12(10): 2086–2093, 1994.

M.R. Segal. Regression trees for censored data. *Biometrics*, pages 35–47, 1988.

D.S. Siroky. Navigating random forests and related advances in algorithmic modeling. *Statistics Surveys*, 3:147–163, 2009.

T.M. Therneau. *A Package for Survival Analysis in S*, 2014. URL `http://CRAN.R-project.org/package=survival`. R package version 2.37-7.

A. Verikas, A. Gelzinis, and M. Bacauskiene. Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349, 2011.

R. Zhu and M.R. Kosorok. Recursively imputed survival trees. *Journal of the American Statistical Association*, 107(497):331–340, 2012.

1.8 Appendix

1.8.1 Hazard functions used in the simulation for DGP 1 and DGP 2

Hazard function formula for DGP 1.

$$\begin{cases} 0.27t & x_1 \leq 0.5, t \leq 2 \\ 0.2(t-2)+5.4 & x_1 \leq 0.5, t > 2 \\ 0.1t & x_1 > 0.5, t \leq 6 \\ 5.5(t-6)+0.6 & x_1 > 0.5, t > 6. \end{cases}$$

Hazard function formula for DGP 2.

$$\begin{cases} 0.27t & x_1 \leq 0.5, x_2 \leq 0.5, t \leq 2 \\ 0.2(t-2)+5.4 & x_1 \leq 0.5, x_2 \leq 0.5, t > 2 \\ 0.27t & x_1 \leq 0.5, x_2 > 0.5, t \leq 2 \\ 5.5(t-2)+5.4 & x_1 \leq 0.5, x_2 > 0.5, t > 2 \\ 0.1t & x_1 > 0.5, x_2 \leq 0.5, t \leq 6 \\ 0.2(t-6)+0.6 & x_1 > 0.5, x_2 \leq 0.5, t > 6 \\ 0.1t & x_1 > 0.5, x_2 > 0.5, t \leq 6 \\ 5.5(t-6)+0.6 & x_1 > 0.5, x_2 > 0.5, t > 6. \end{cases}$$

1.8.2 Simplification of the Lin and Xu (2010) statistic leading to the L_1^* splitting rule.

For $i = L, R$, the left and right nodes, denote by $\hat{\sigma}_i^2(t)$ the estimated variance of $\hat{S}_i(t)$ at t from Greenwood's formula. To perform a formal test of the equality of the survival functions in the left and right nodes, Lin and Xu (2010) propose the statistic

$$\Delta^* = \frac{\Delta - \hat{E}(\Delta)}{\sqrt{\widehat{Var}(\Delta)}}$$

where

$$\hat{E}(\Delta) = \sum_{j|t_j<\tau} \left[2/\pi\{\hat{\sigma}_L^2(t_j) + \hat{\sigma}_R^2(t_j)\}\right]^{1/2} (t_{j+1} - t_j)$$

and

$$\begin{aligned}\widehat{Var}(\Delta) &= \sum_{j|t_j<\tau} (t_{j+1}-t_j)^2(1-2/\pi)\{\hat{\sigma}_L^2(t_j)+\hat{\sigma}_R^2(t_j)\} \\ &+ \sum_{j<j'|t_j,t_{j'}<\tau} (t_{j+1}-t_j)(t_{j'+1}-t_{j'})(1-2/\pi) \\ &\times \left[\{\hat{\sigma}_L^2(t_j)+\hat{\sigma}_R^2(t_j)\}\{\hat{\sigma}_L^2(t_{j'})+\hat{\sigma}_R^2(t_{j'})\}\right]^{1/2}\end{aligned}$$

are estimates of $E(\Delta)$ and $Var(\Delta)$. These estimates arise from a normal approximation for $\hat{S}_L(t) - \hat{S}_R(t)$, and the test statistic Δ^* is asymptotically normally distributed under the null hypothesis of equality of the two survival functions. To simplify this statistic in order to speed up computations for tree building, assume that all observations are from the same population with survival function $S(t)$, that is we are under the null hypothesis and there is no censoring. Then $Var(\hat{S}_i(t)) = S(t)(1-S(t))/n_i$, for $i = L, R$. In that case,

$$\hat{E}(\Delta) = \sqrt{2/\pi}\sqrt{(n_L+n_R)/(n_Ln_R)} \sum_{j|t_j<\tau} (S(t_j)(1-S(t_j)))^{1/2}(t_{j+1}-t_j) = c_1/\sqrt{n_Ln_R}$$

where c_1 is the same constant for all candidate splits. Similarly, $\widehat{Var}(\Delta) = c_2^2/(n_Ln_R)$ where c_2 is the same constant for all candidate splits. Hence,

$$\frac{\Delta - \hat{E}(\Delta)}{\sqrt{\widehat{Var}(\Delta)}} = \frac{\sqrt{n_Ln_R}\Delta}{c_2} - \frac{c_1}{c_2}.$$

But using this last expression is equivalent to using $\sqrt{n_Ln_R}\Delta$ as the splitting criterion.

1.8.3 Motivating example

In this section, a motivating example is used to illustrate the fact that the log-rank test may have a significant loss of power when the survival functions cross each other in the two compared groups. We first investigate the ability of the log-rank splitting rule to identify the only covariate related with the response. In this example, data are generated according to DGP 1 in the article without censoring. Hence, there are ten covariates $X_1, \ldots, X_{10}$ that are independently and identically distributed uniformly on the interval (0,1). Only the first one is related with the response. The nine others are noise covariates. In this example, there are only two survival patterns depicted in Figure 1 of the article. If $X_1 \leq 0.5$, then the true survival function is the smooth exponential one, while it is the one with the sharp drop when $X_1 > 0.5$. The sample size is 200 and the results are based on 500 simulation runs. For each sample, we built a tree with a single split by using two different splitting rules. The first splitting rule is the log-rank test. Strikingly, the log-rank splitting rule could only detect the right covariate 14.4% of the times, which is just a bit higher than the random selection probability of 10%. The second splitting rule is the L_1 splitting rule which was able to detect the right covariate (X_1) 100% of times.

However, in practice, we are not building a single tree with a single split, but a forest. Forests are well-known to be able to adapt themselves to many situations. It might be the case that even if the log-rank splitting rule is not able to find the right covariate in the first split, it could be able to detect the structure somehow deeper in the trees and the forest could still provide a good performance at the end. But this is exactly what was investigated in the simulation study and it was found that the forest paradigm is not able to fully compensate for the fact that the splitting rule is not appropriate. The complete simulation results are presented in the next two sections.

1.8.4 Complete simulations results

This section provides complete details about the results of the main simulations presented in Sections 1.4.1 of the thesis. The detailed results for all DGPs are summarized in Figures 1.7 to 1.11 at six different censoring proportion ranging from 0% to 50%. Note that in computation of the L_1 and L_1^* splitting rules, if the survival curve in either right or left nodes does not take a value of zero at the last observed time, an Exponential tail is glued to the curve. For fair comparison among methods, the IAE and ISE criteria are computed over 1001 equidistance (0.001 quantiles plus $t = 0$) of the true survival function of each point in the test set where the maximum point at each run is the .999 quantile of this function. Survival estimates of each method is then interpolated at these 1001 points. The sum is not divided by 1001.

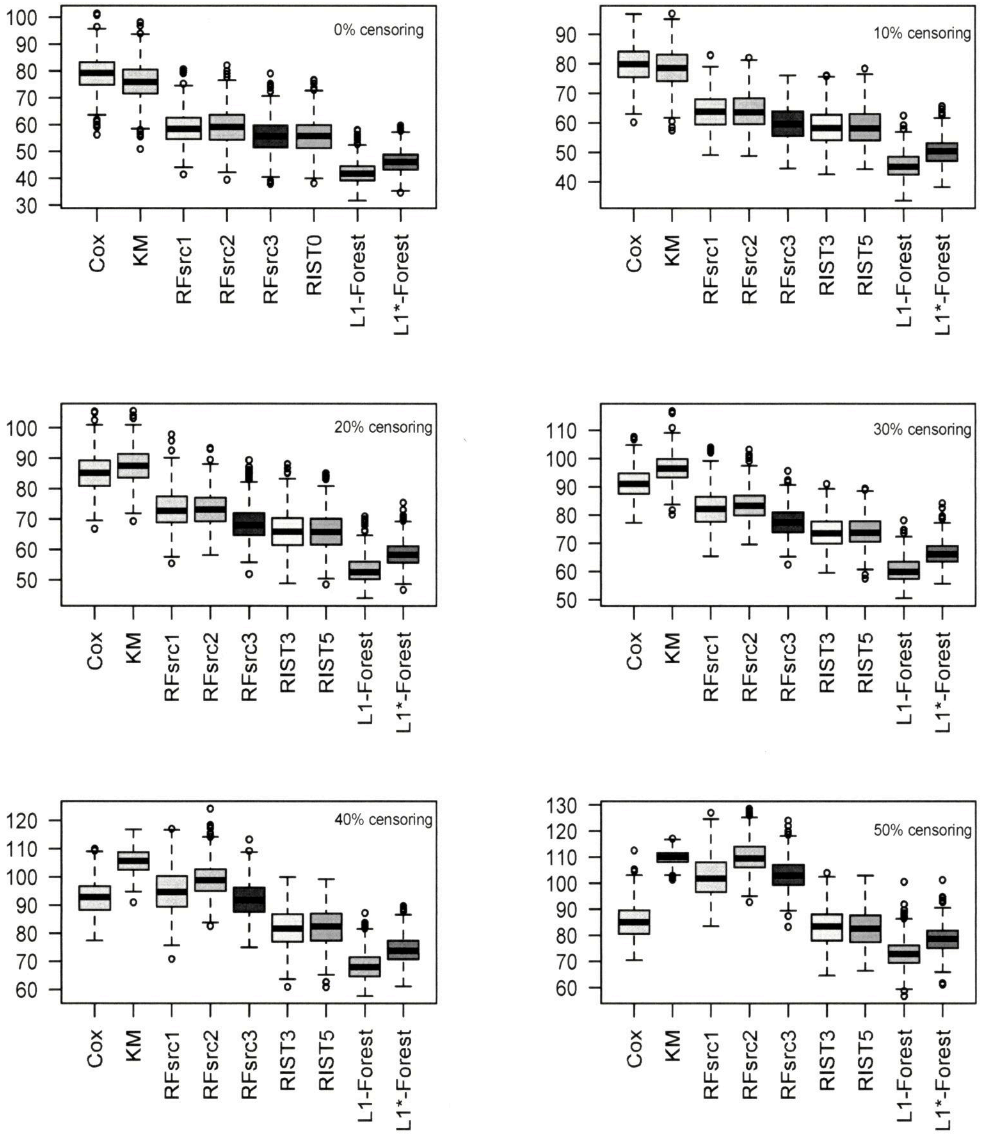

Figure 1.7: IAE for DGP 1 across different censoring proportions

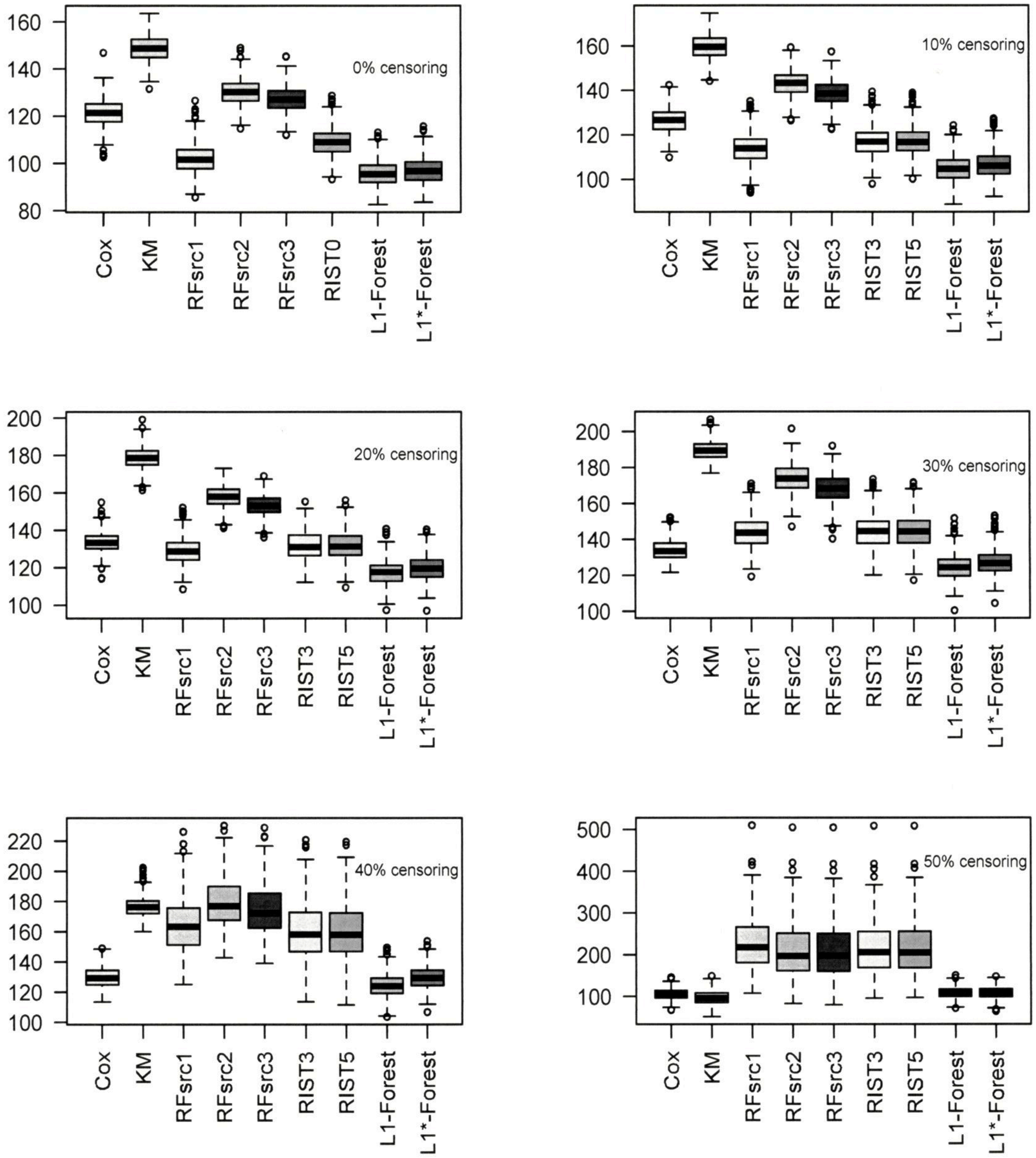

Figure 1.8: IAE for DGP 2 across different censoring proportions

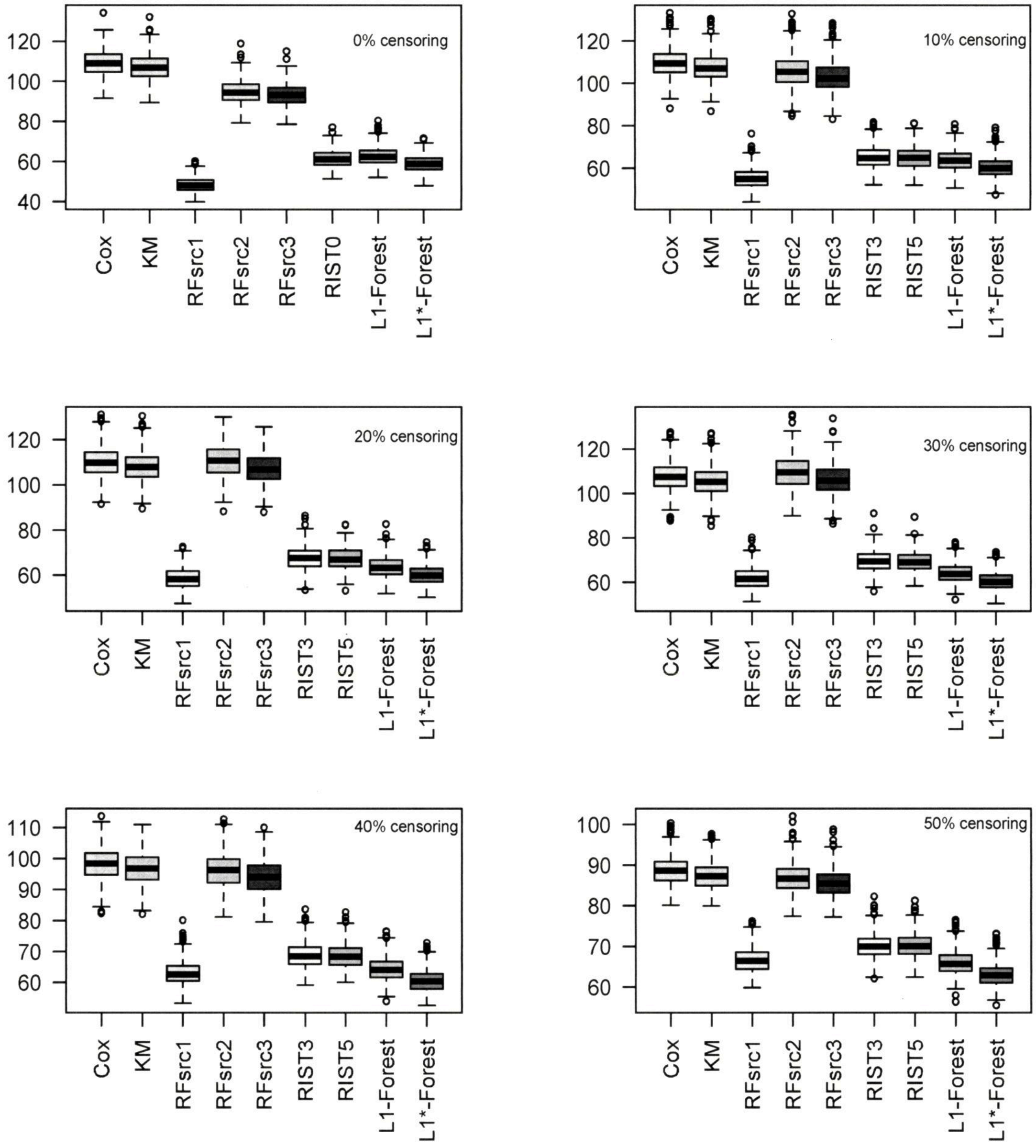

Figure 1.9: IAE for DGP 3 across different censoring proportions

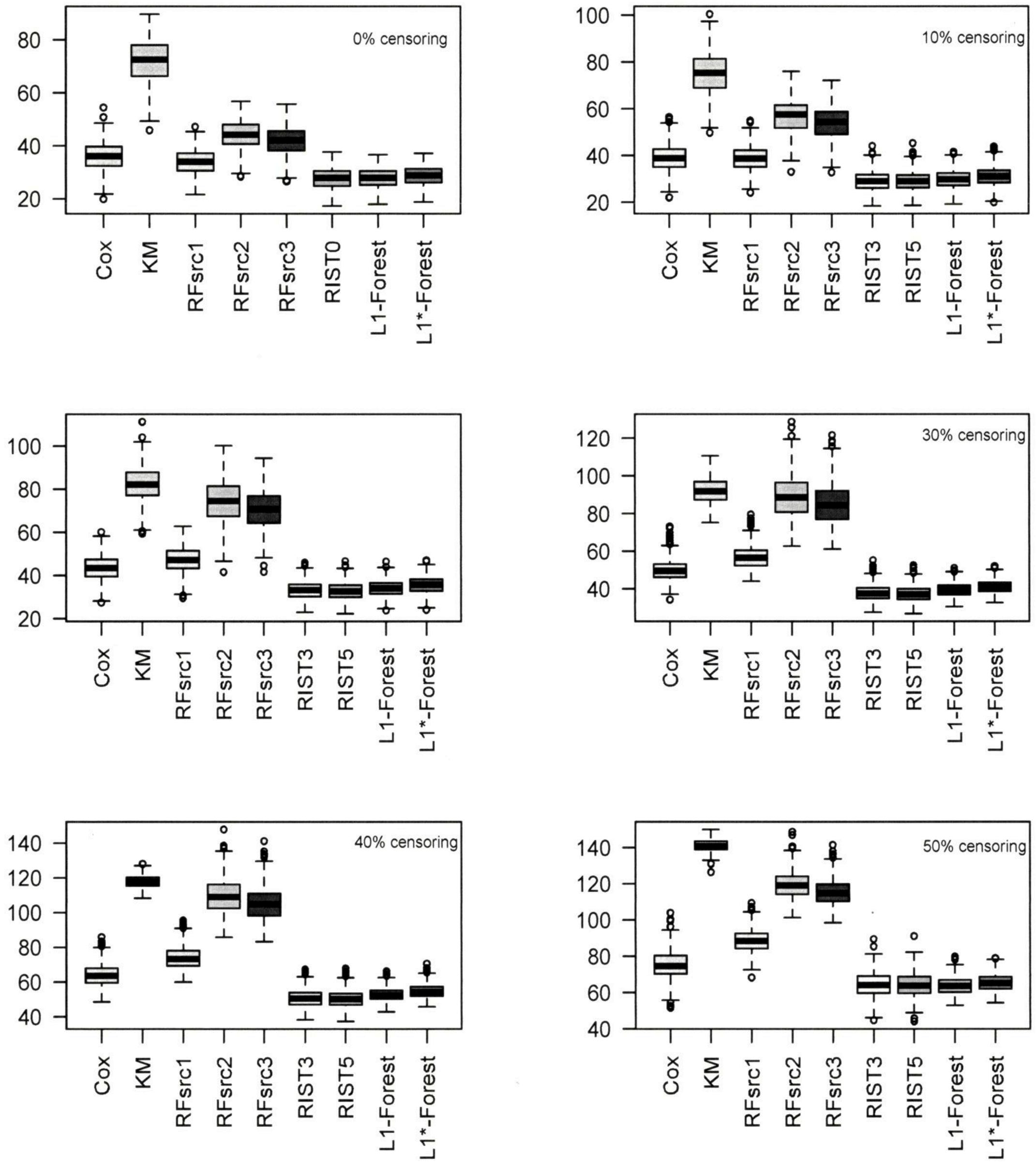

Figure 1.10: IAE for DGP 4 across different censoring proportions

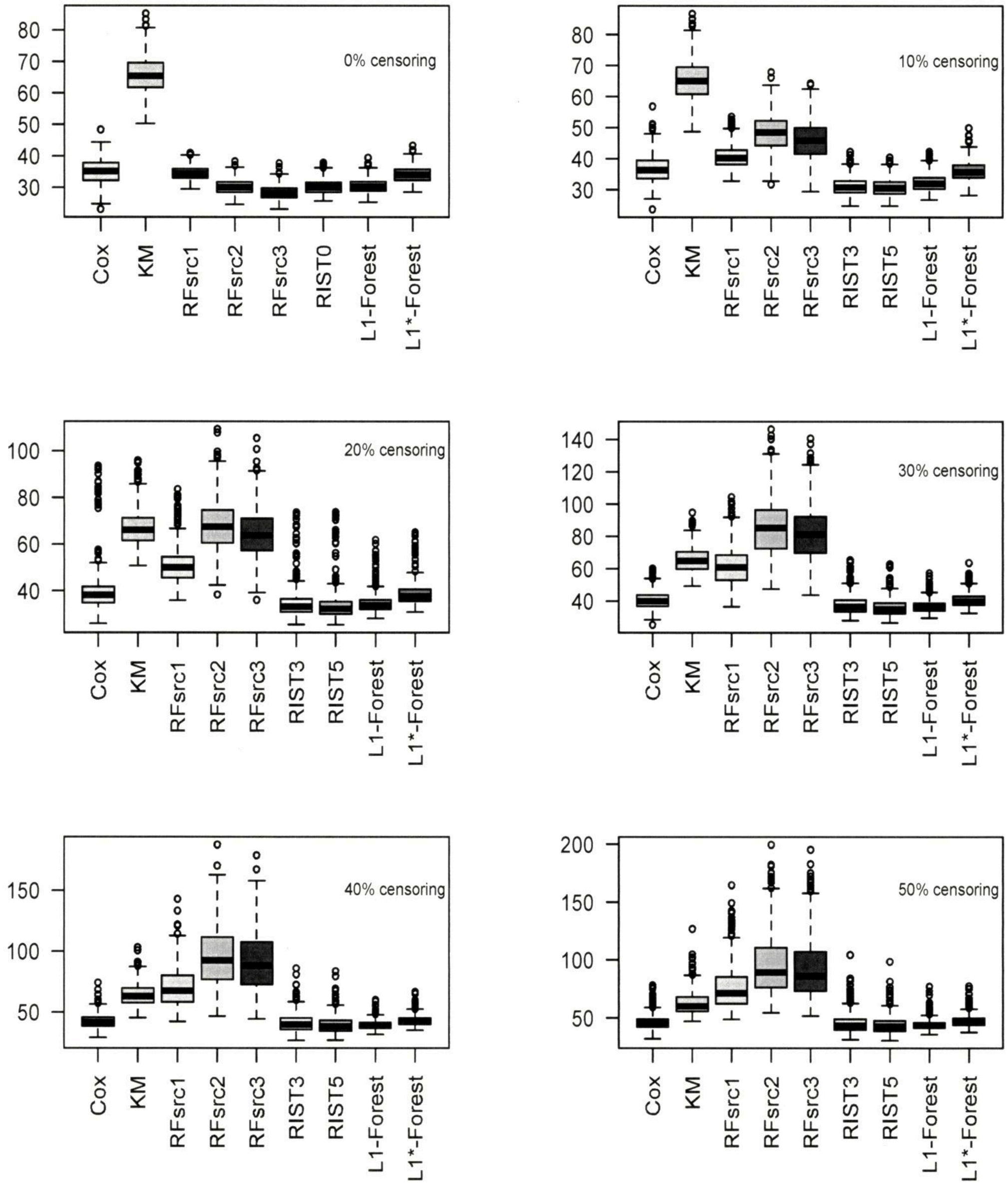

Figure 1.11: IAE for DGP 5 across different censoring proportions

1.8.5 Complete results for the additional simulations

This section provides complete details about the results of the additional simulations presented in Section 1.4.3.

Small sample size and large number of covariates

The main simulations are described in sections 1.4.1 and 1.4.2. Here, using the same DGPs as for the main simulation, the performance of the methods is investigated with a smaller training sample ($n = 250$) and with five times more noise covariates, as explained in Section 1.4.3. The detailed results for this simulation study are summarized in Figures 1.12 to 1.16. The results are very similar to the ones from the main simulations. As expected, the L_1-type forests are the best performing methods in terms of IAE in the first two DGPs with crossing survival functions. The only exception is for DGP 1 with 50% censoring where the KM is the best one. For DGP 3, the RFsrc1 performs better but the gap between L_1-type forests and RFsrc1 narrows down as the censoring proportion goes up. For DGPs 4 and 5, L_1-type forests and the RIST are very close to each other and perform the best. So again, the L_1-type forests are very competitive in the last three DGPs that does not involve crossing survival functions.

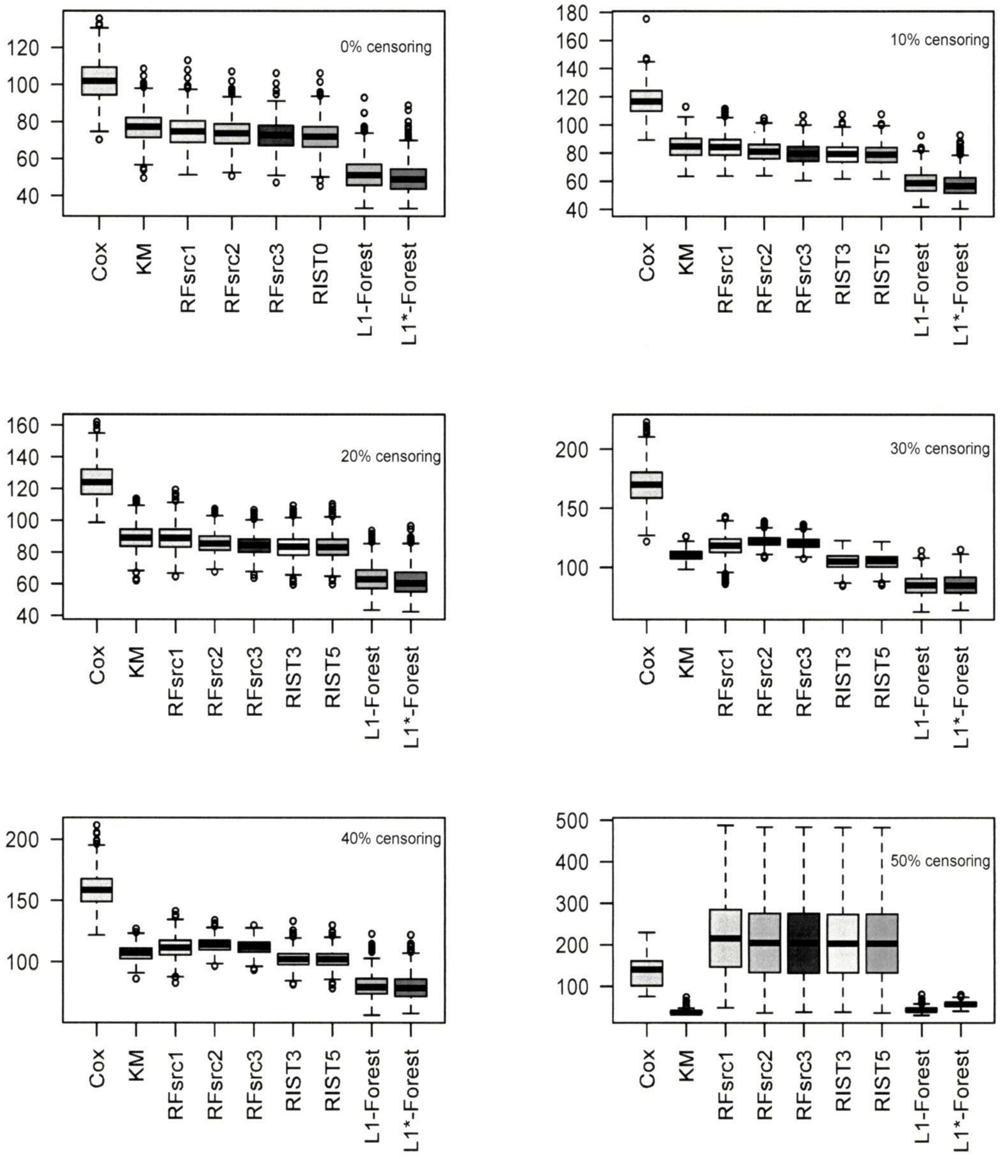

Figure 1.12: IAE for DGP 1 across different censoring proportions with small n and large p

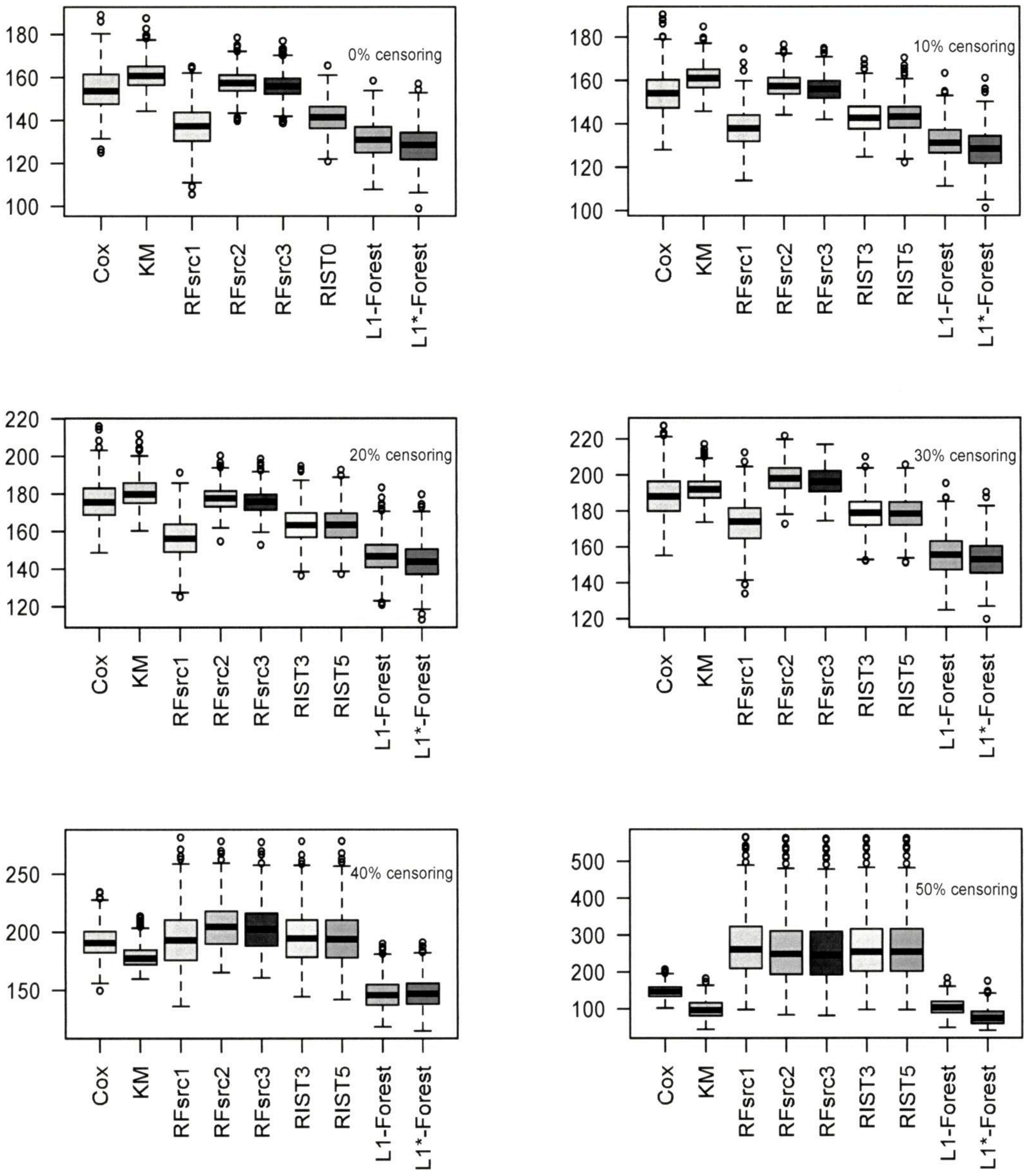

Figure 1.13: IAE for DGP 2 across different censoring proportions with small n and large p

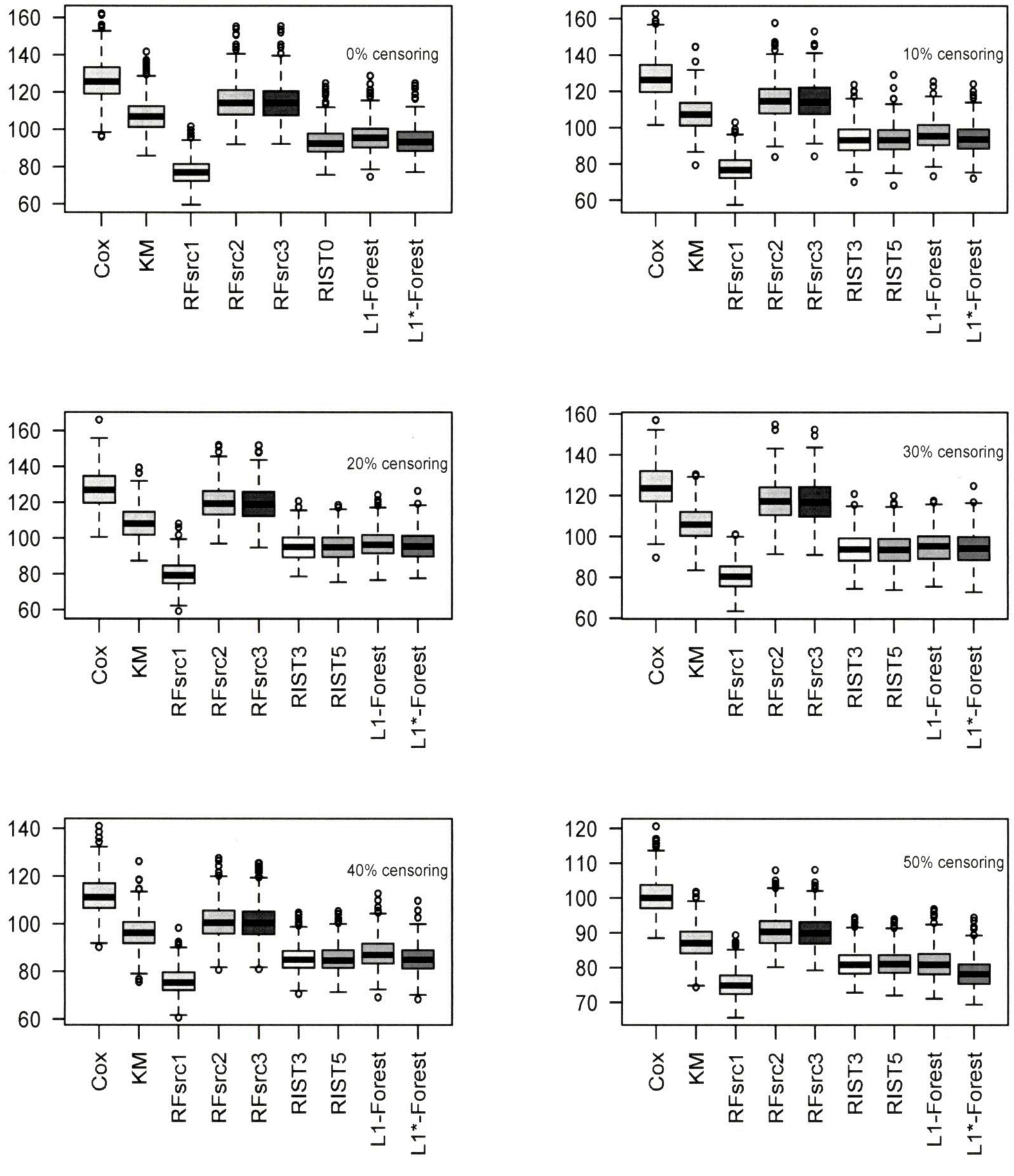

Figure 1.14: IAE for DGP 3 across different censoring proportions with small n and large p

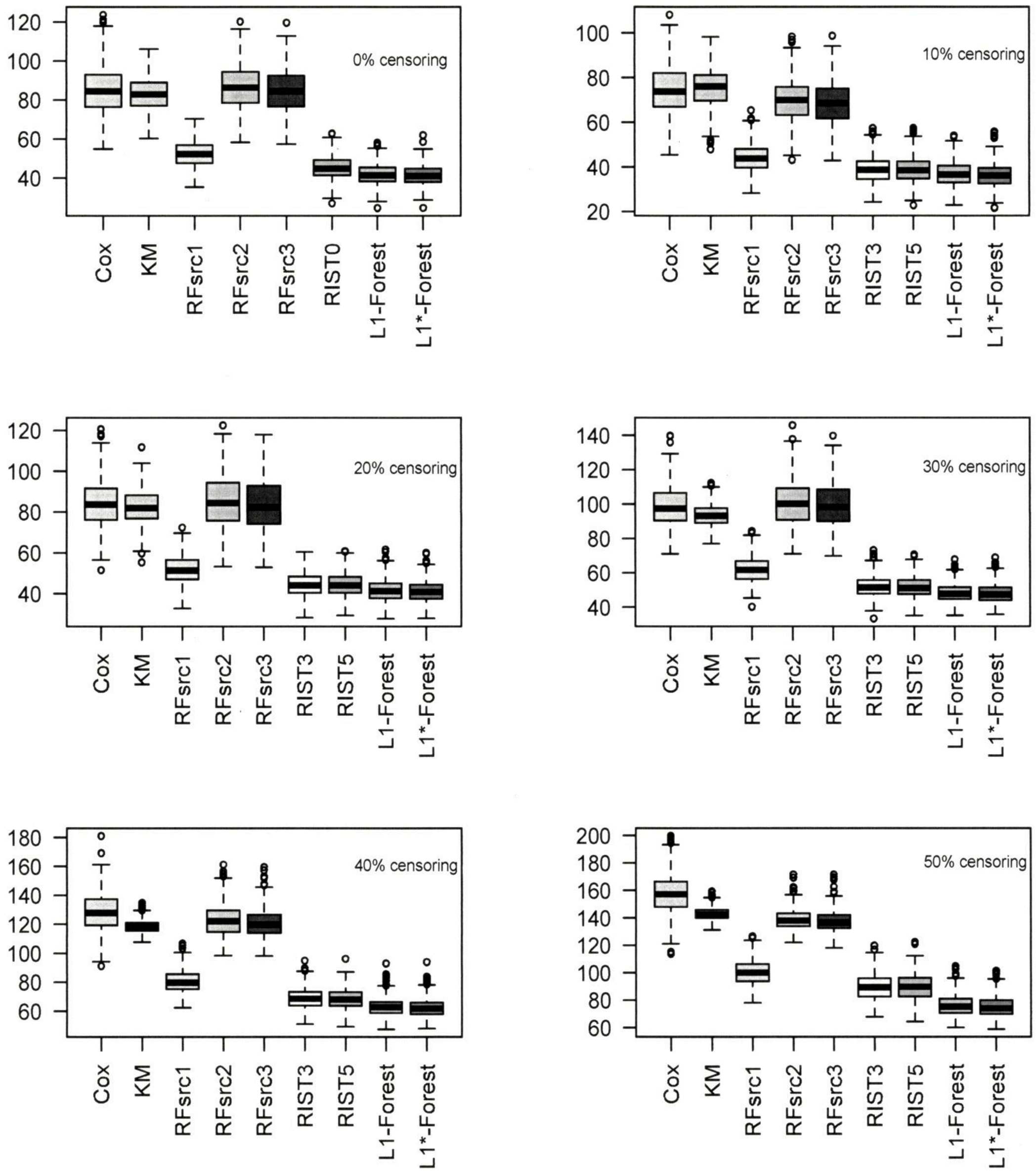

Figure 1.15: IAE for DGP 4 across different censoring proportions with small n and large p

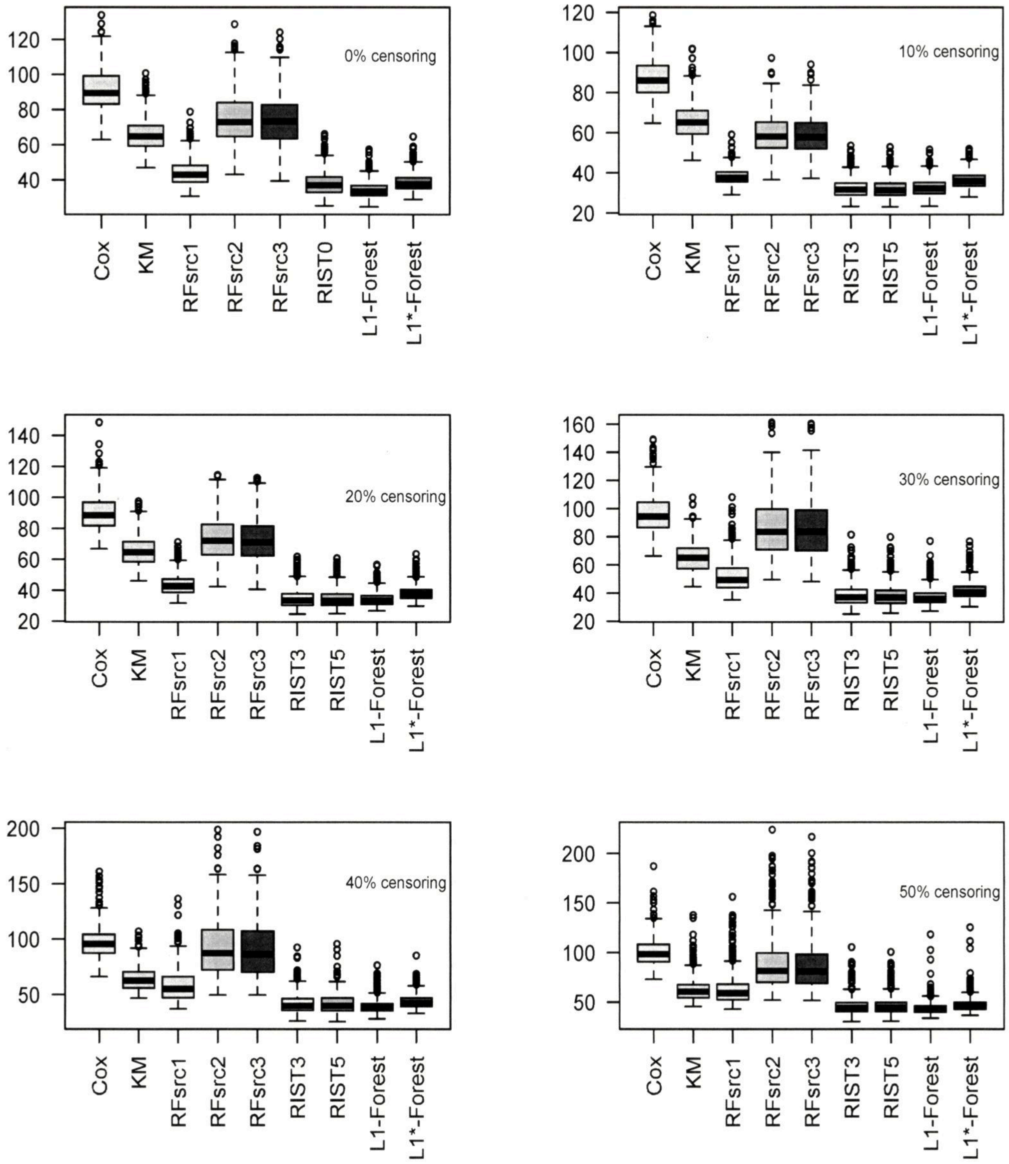

Figure 1.16: IAE for DGP 5 across different censoring proportions with small n and large p

Categorical covariates

Since all covariates are continuous in the main simulation, a few scenarios with binary covariates are investigated here. As explained in Section 1.4.3, only DGP 2 is considered. Only two covariates, X_1 and X_2, are related to the response. In the main simulation, X_1 and X_2 are uniformly distributed on (0,1). In the first variation, X_2 is still uniformly distributed on (0,1) but X_1 is now a binary covariate taking values 0 and 1 with probability 1/2. In the second variation, both X_1 and X_2 are binary covariates taking values 0 and 1 with probability 1/2. Figure 1.17 shows the results of the simulation study for the IAE criterion for the first variation while Figure 1.18 shows the results for the second variation. Figure 1.8 shows the results for the original simulation (X_1 and X_2 both continuous). We can see that the IAE of the forests methods are quite similar across the three scenarios. However, the performance of the Cox model improves when we move from the original scenario, to the one with X_1 only binary, and then to the one when both X_1 and X_2 are binary.

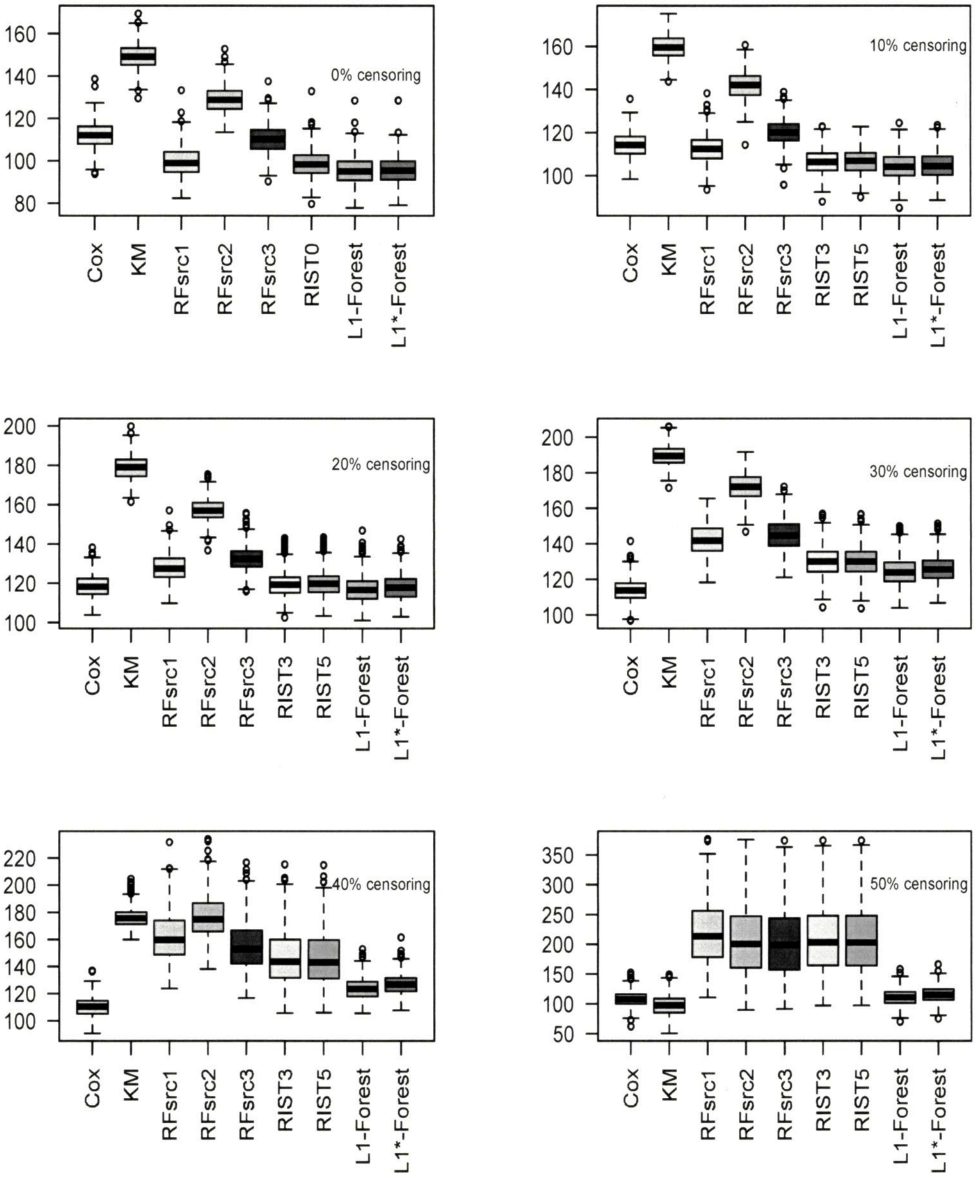

Figure 1.17: IAE for DGP 2 across different censoring proportions with binary X_1 and continuous X_2

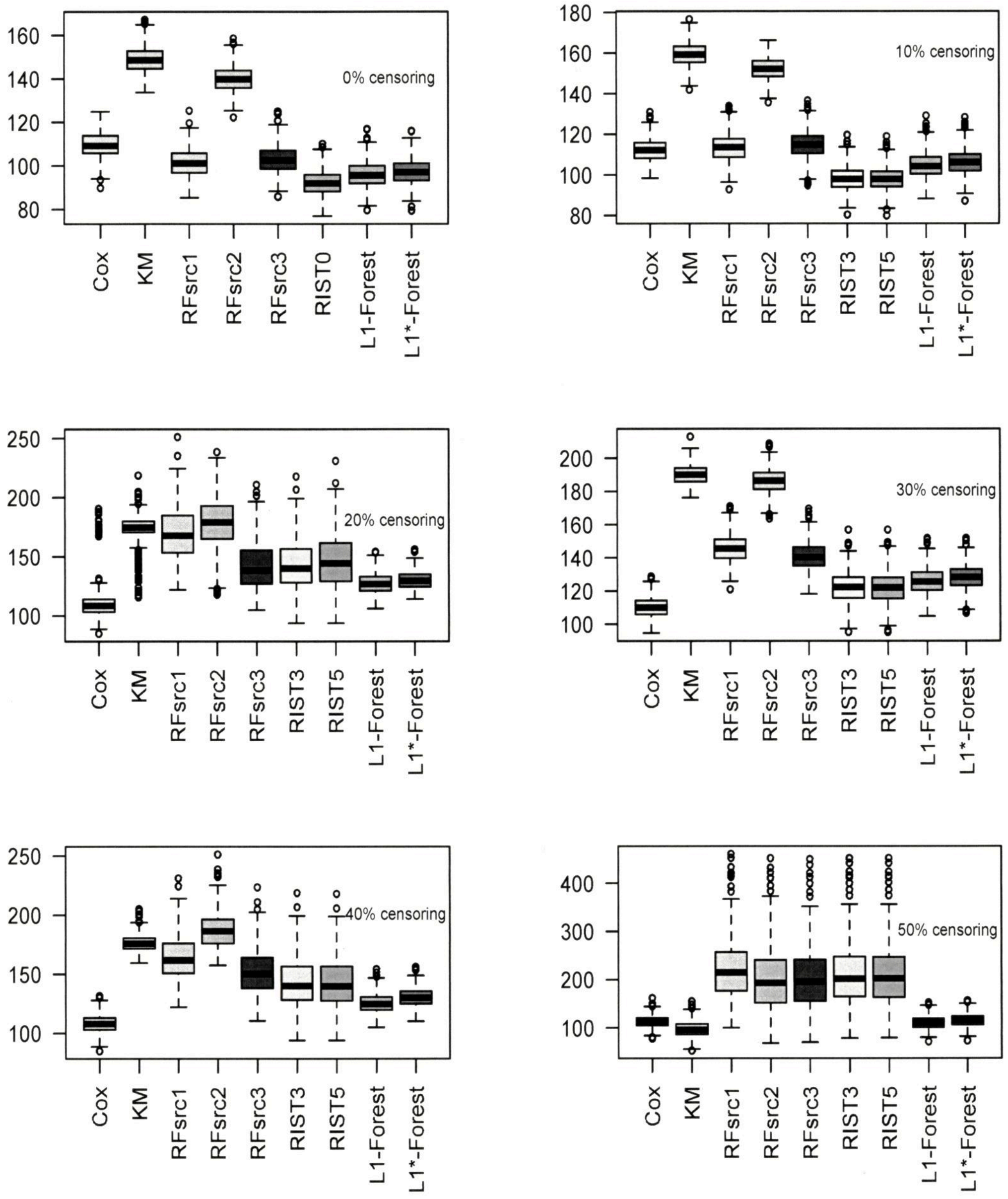

Figure 1.18: IAE for DGP 2 across different censoring proportions with binary X_1 and X_2

Chapter 2

Dynamic Predictions with Random Forests for Discrete-Time Survival Data

2.1 Abstract

Time-varying covariates are often available in survival studies and predictions need to be updated as new information becomes available. In this paper, we investigate different ways that random forests can be used for dynamic estimation of the hazard function with discrete-time survival data. The results from a simulation study and from a real data example about firm bankruptcies indicate that all methods can perform well, that none dominate the others, and that taking a simple average of the estimated hazard functions is a good way to get good results in most cases.

Keywords: Discrete-time survival analysis; time-varying covariates; landmark analysis; random forests; survival forests.

2.2 Introduction

Survival analysis studies with time-to-event data have applications in many research areas. Typically, the true time is observed only for some of the subjects and only partial information about the time is available for other subjects. The most common situation is right-censoring, when only a lower bound on the true time is observed. Many textbooks (e.g., Hosmer Jr et al., 2011) are dedicated to the modeling of such data. The traditional methods for analysing continuous survival data rely on parametric (e.g. Weibull) and semi-parametric (e.g. Cox) models. These models can be efficient if the link between the covariates and the time response is adequately modeled but this is sometimes a difficult task. Alternatively, more flexible models can be useful. These models let data automatically find relevant structures instead of imposing them a priori. One class of such models is tree-based methods which are now well-established among practitioners.

Tree-based methods were first developed for a categorical or continuous outcome. Breiman et al. (1984) is the earliest monograph about trees and details the Classification and Regression Tree (CART) paradigm. However, it is well-known that ensemble of trees often provide better predictive performance than a single tree. One popular and efficient ensemble method is the random forest introduced by Breiman (2001). Gordon and Olshen (1985) extended the tree paradigm to survival data and introduced survival trees (Leblanc and Crowley, 1993; Segal, 1988). Likewise, ensemble of survival trees called survival forests were proposed by Hothorn et al. (2006), Ishwaran et al. (2008), and Zhu and Kosorok (2012). There is vast literature on survival trees and forests and Bou-Hamad et al. (2011b) presents a general overview.

In many studies, the estimation of the survival curve for a subject is obtained at time 0 using only the covariates information at baseline. However, when some time-varying covariates are present, it is often preferable to update the estimates of survival probabilities as new longitudinal information becomes available. This is the topic of "dynamic prediction" which is a growing area of interest. There are mainly two approaches to build dynamic predictions in this context: 1) landmark analysis, and 2) joint modeling. The basic idea of landmark analysis (Anderson et al., 1983; Madsen et al., 1983) is to built models, usually Cox, at different landmark times t using the covariates information available up to t and only the subjects still at risk of experiencing the event at t. Comprehensive treatments of this approach are given in van Houwelingen (2007) and van Houwelingen and Putter (2011). In order to reduce the variability in the estimated parameters by borrowing information from other time points, one method consists in stacking the data sets from many landmark times and fitting a single model (van Houwelingen and Putter, 2011). The partly conditional survival model of Zheng and Heagerty (2005) and the two-stage approach of Huang et al. (2016) are two related methods. The second approach to dynamic predictions is to estimate the survival probabilities through joint modelling of the time-varying covariates processes and the event time data (Henderson et al., 2000). This approach depends on the correct specification of the model for the time-varying covariates trajectories, and this problem amplifies as the number of time-varying covariates increases. The reader can refer to Tsiatis and Davidian (2004) and Rizopoulos (2012) for in-dept treatments of the joint modeling approach.

Most of the research, including the works presented above, assume that the time-to-event is measured continuously. However, in many cases, it is measured on a discrete scale. This can happen with grouped data where the event occurs in an interval of time, which are not necessarily of the same length. For example, in yearly surveys, it may be known that the event occurred during a specific year but not exactly when. Alternatively, the observed time may come from a truly discrete process. For example, the number of trials before reaching a specific goal. Tutz and Schmid (2016) provide a recent and comprehensive treatment of discrete-time survival analysis. Survival trees and forests designed specifically for discrete-time responses were developed by Bou-hamad et al. (2009), Bou-Hamad et al. (2011a) and Schmid et al. (2016). The `R` package `DStree` (Mayer et al., 2014) implements the method of Bou-hamad et al. (2009). Section 2.1 provides a description of these methods since they are central to this article. Elgmati et al. (2015) propose a penalized Aalen additive model for dynamic predictions for discrete-time recurrent event data, but the method is limited to one-step ahead predictions.

From the above discussion, we see that no tree-based methods have addressed the problem of dynamic predictions either with continuous or discrete survival responses. This paper will study the later problem. We investigate different ways that random forests can be used for dynamic predictions with discrete-time survival data.

The rest of the paper is organized as follows. Section 2.3 describes the data setting and the proposed methods. The results from a simulation study are presented in Section 2.4. A real data example with bankruptcy data is presented in Section 2.5. Section 2.6 concludes and provides directions for future work.

2.3 Description of the methods

We have data on N independent subjects. For each subject i, observations are in the form of $(\tau_i,\, \delta_i,\, x_i)$ where $\tau_i \in \{1, 2, ..., T\}$ is the discrete survival time and T is the maximum observed time in the data set, δ_i is the censoring index which takes a value of 0 if subject i is right censored and a value of 1 if subject i has experienced the event of interest, and x_i is a vector of covariates, some of which can be time-varying and some time-independent. We will denote by $x_{ki}(t)$ the value of the k^{th} covariate, $k \in \{1, 2, \ldots, p\}$, at time $t \in \{0, 1, \ldots, T\}$ for subject i. Hence, $t = 0$ corresponds to the baseline values of the covariates. For simplicity, we will use this notation for all covariates, time-varying or not. Hence $x_{ki}(t)$ remains constant for all t for a time-independent covariate. The values of the true time-to-event and the true censoring times for subject i are denoted by U_i and V_i, respectively. Hence we have τ_i=min$(U_i$,$V_i)$ and we assume that U_i and V_i are independent given x_i. The hazard function for subject i is denoted by $h_i(t) = P(U_i = t \mid U_i \geq t)$ for simplicity but it is obvious that τ_i, δ_i, U_i and V_i depend on x_i. Similarly, the survival function for subject i is $S_i(t) = P(U_i > t)$, and the probability that the event occurs at time t is $\pi_i(t) = P(U_i = t)$. These two functions can be obtained from the hazard function with the following recursive formulas: $S_i(t) = S_i(t-1)(1-h_i(t))$ and $\pi_i(t) = S_i(t-1) - S_i(t)$, with $S_i(0) = 1$. Hence, it is sufficient to model the hazard function (or any one of the other two function) to recover the other ones.

2.3.1 Description of existing methods for discrete-time survival data

Since they are central to what will follow, we give a description of the tree building methods for discrete-time survival data of Bou-hamad et al. (2009), Bou-Hamad et al. (2011a) and Schmid et al. (2016). In general, the log-likelihood function of a discrete-time survival model can be written as

$$\mathrm{LL} = \sum_{i=1}^{N} \delta_i \ln(\pi_i(\tau_i)) + (1-\delta_i)\ln(S_i(\tau_i)). \tag{2.1}$$

In this paper, we take the view of estimating the hazard of a subject at a given time t using all the available covariates information up to time $t-1$. This is more realistic in a prediction context where the prediction for time t might be needed as soon as we know that the subject is alive at time $t-1$. This is in contrast with correlational studies where the relations between the contemporary values of the covariates and the response might be of interest. One popular model is the discrete-time proportional odds (DTPO) model

$$\log\left(\frac{h_i(t)}{1-h_i(t)}\right) = \alpha_1 D_{1i}(t) + \cdots + \alpha_T D_{Ti}(t) + \beta_1 x_{1i}(t-1) + ... + \beta_p x_{pi}(t-1), \tag{2.2}$$

for $i = 1, \ldots, N$ and $t = 1, \ldots, T$, where the $D_{ri}(t)$'s are indicator variables indexing the time periods that are defined by $D_{ri}(t) = 1$ if $r = t$ and 0 otherwise. The intercept parameters $\alpha_1, \ldots, \alpha_T$ define the baseline of hazard in each time period and the β coefficients describe the effects of covariates on the baseline hazard function. This model was described in details in Willett and Singer (1993) and became widely used. The key idea is that the likelihood function (2.2) is equivalent to the one of an independent Bernoulli trials model with transformed (augmented) data with a logistic dependence on the covariates (see Willett and Singer, 1993, p. 171). The augmented data set is often called the "person-period" data set. It has one line per time period where the subject is at risk of experiencing the event and the response y equals 1 if the event occurred at that time and 0 otherwise. To fix ideas, a generic data set with ten observations and two covariates is given in Table 2.1. Assume X_1 is time-varying and X_2 is time-independent. For instance, the first subject experienced the event at period 2 and thus values of the time-varying covariate are only available up to period 2, with NA's for the other periods. The person-period data set obtained from this generic data set is given in Table 2.2. Only the first six subjects (up to id=6) are shown to save space. Practically, the equivalence in the likelihood functions mentioned above means that we can obtain the estimations of the parameters of model (2.2) by fitting a logistic regression to the person-period data set.

We assume that the reader is familiar with the CART paradigm (Breiman et al., 1984) and the basic random forest method (Breiman, 2001). The methods in Bou-hamad et al. (2009) and Bou-Hamad et al. (2011a) are based on the same idea except that the first one is limited to time-independent covariates and the second one allows the use of time-varying covariates. Basically, the idea is to use a splitting rule based on the log-likelihood where the hazard function in each node is estimated from a saturated model, that is a model where each time period has its own free parameter. In that case, the maximum likelihood estimator of the hazard function in a node is

$$\hat{h}(t) = e(t)/r(t), \tag{2.3}$$

where $e(t)$ is the number of subjects that experienced the event at period t in the node and $r(t)$ is the number of subjects at risk of experiencing the event at period t in the node. These estimators are obtained separately in both children nodes and the best split is the one such that the sum of the log-likelihood (2.1), in the left and right nodes, is maximum among all allowable splits. This splitting rule can also be obtained from the DTPO model. Consider a binary random variable $C(t) \in \{0, 1\}$ defining an allowable split using one of the covariates, at time t. The DTPO model with this covariate and all its interactions with the time indicators is

$$\log\left(\frac{h_i(t)}{1-h_i(t)}\right) = \alpha_1 D_{1i}(t) + \cdots + \alpha_T D_{iT}(t) + \beta_1 C_i(t-1) D_{1i}(t) + \ldots + \beta_p C_i(t-1) D_{iT}(t). \tag{2.4}$$

Fitting this model is equivalent to fitting a saturated model separately in the two nodes defined by $C(t)$. Hence, one way to build a tree with this approach and easily keep track of the subjects, is to use the person-period data set. To fix ideas, with the data set in Table 2.2, it would mean building a classification tree with the LL-based splitting rule with y as the response and using the two covariates X_1, X_2. It is easy to see that when a split occurs on a time-independent covariate, then all lines corresponding to a given subject in the person-period data set go into the same node. However, when a split occurs on a time-varying covariate, then the lines of a given subject may be splitted apart. This is similar to the "pseudo-subjects" idea of Bacchetti and Segal (1995). The hazards estimates in a terminal node are given by (2.3). To build a forest, Bou-hamad et al. (2009) proposed to use the average of these hazard estimates obtained from many trees built from bootstrap samples of the original data. However, it is also possible to start directly from the person-period data set and use boostrap samples from it.

Schmid et al. (2016) proposed a similar method. They basically add the period itself as a new ordinal covariate, on top of the other covariates, and use the Gini splitting rule with the person-period data set (called augmented data in their paper) with y as the response. The choice of the Gini criterion is based on the fact that it is asymptotically equivalent to the Brier score. To fix ideas, with the data set in Table 2.2, it would mean building a classification tree with y as the response using the three covariates X_1, X_2, and the variable period. Using the variable period means that the subjects can also be splitted apart even if no time-varying covariates are present among the original covariates because period itself is a time-varying covariate. In a terminal node, the estimate of the hazard is the proportion of 1 (events) in the node. Note that Schmid et al. (2016) apply a Laplace correction to these proportions. Hence, all time periods present in the node receive the same hazard estimate. But since splits on the variable period are allowed, it is possible to have different hazard estimates for different periods. To build a forest, it is again possible to bootstrap from the original data or from the person-period data set. Schmid et al. (2016) use the person-period data set.

2.3.2 Methods investigated in this paper

Using a training data set, we are interested in deriving dynamic estimations of the hazard function for current and new subjects. Some subjects barely begin the process, some are

Table 2.1: A generic data set with ten observations and two covariates

id	τ	δ	$X_1(0)$	$X_1(1)$	$X_1(2)$	$X_1(3)$	$X_1(4)$	$X_1(5)$	X_2
1	2	1	$x_{11}(0)$	$x_{11}(1)$	$x_{11}(2)$	NA	NA	NA	x_{21}
2	5	1	$x_{12}(0)$	$x_{12}(1)$	$x_{12}(2)$	$x_{12}(3)$	$x_{12}(4)$	$x_{12}(5)$	x_{22}
3	3	0	$x_{13}(0)$	$x_{13}(1)$	$x_{13}(2)$	$x_{13}(3)$	NA	NA	x_{23}
4	1	0	$x_{14}(0)$	$x_{14}(1)$	NA	NA	NA	NA	x_{24}
5	5	1	$x_{15}(0)$	$x_{15}(1)$	$x_{15}(2)$	$x_{15}(3)$	$x_{15}(4)$	$x_{15}(5)$	x_{25}
6	5	0	$x_{16}(0)$	$x_{16}(1)$	$x_{16}(2)$	$x_{16}(3)$	$x_{16}(4)$	$x_{16}(5)$	x_{26}
7	2	0	$x_{17}(0)$	$x_{17}(1)$	$x_{17}(2)$	NA	NA	NA	x_{27}
8	4	1	$x_{18}(0)$	$x_{18}(1)$	$x_{18}(2)$	$x_{18}(3)$	$x_{18}(4)$	NA	x_{28}
9	3	1	$x_{19}(0)$	$x_{19}(1)$	$x_{19}(2)$	$x_{19}(3)$	NA	NA	x_{29}
10	4	0	$x_{110}(0)$	$x_{110}(1)$	$x_{110}(2)$	$x_{110}(3)$	$x_{110}(4)$	NA	x_{210}

Table 2.2: Person-period data set. Only the first six subjects (up to id=6) are shown to save space.

id	y	period	D_1	D_2	D_3	D_4	D_5	X_1	X_2
1	0	1	1	0	0	0	0	$x_{11}(0)$	x_{21}
1	1	2	0	1	0	0	0	$x_{11}(1)$	x_{21}
2	0	1	1	0	0	0	0	$x_{12}(0)$	x_{22}
2	0	2	0	1	0	0	0	$x_{12}(1)$	x_{22}
2	0	3	0	0	1	0	0	$x_{12}(2)$	x_{22}
2	0	4	0	0	0	1	0	$x_{12}(3)$	x_{22}
2	1	5	0	0	0	0	1	$x_{12}(4)$	x_{22}
3	0	1	1	0	0	0	0	$x_{13}(0)$	x_{23}
3	0	2	0	1	0	0	0	$x_{13}(1)$	x_{23}
3	0	3	0	0	1	0	0	$x_{13}(2)$	x_{23}
4	0	1	1	0	0	0	0	$x_{14}(0)$	x_{24}
5	0	1	1	0	0	0	0	$x_{15}(0)$	x_{25}
5	0	2	0	1	0	0	0	$x_{15}(1)$	x_{25}
5	0	3	0	0	1	0	0	$x_{15}(2)$	x_{25}
5	0	4	0	0	0	1	0	$x_{15}(3)$	x_{25}
5	1	5	0	0	0	0	1	$x_{15}(4)$	x_{25}
6	0	1	1	0	0	0	0	$x_{16}(0)$	x_{26}
6	0	2	0	1	0	0	0	$x_{16}(1)$	x_{26}
6	0	3	0	0	1	0	0	$x_{16}(2)$	x_{26}
6	0	4	0	0	0	1	0	$x_{16}(3)$	x_{26}
6	0	5	0	0	0	0	1	$x_{16}(4)$	x_{26}

still alive at time 1, others are still alive at time 2 and so on up to time $T-1$. More precisely, if the subject is currently alive at time $t \in \{0, 1, \ldots, T-1\}$, then we are interested in estimating its hazard function at time u for $u = t+1, \ldots, T$. To fix ideas, Table 2.3 illustrates the possible combinations of t and u when $T = 5$.

Table 2.3: The 15 different estimating problems when $T = 5$

t	u				
Value	1	2	3	4	5
0	✓	✓	✓	✓	✓
1		✓	✓	✓	✓
2			✓	✓	✓
3				✓	✓
4					✓

For instance, the line $t = 0$ corresponds to a new subject. Hence we are interested in estimating its hazard function at times $1, \ldots, 5$. The line $t = 1$ corresponds to a subject who has survived up to time 1. Hence we are interested in estimating its hazard function at times $2, \ldots, 5$, and so on. The diagonal in Table 2.3 corresponds to the one-step ahead problems, i.e. given a subject is alive at time t, we want to estimate its hazard function at time $t+1$. The cases above the diagonal correspond to more than one-step ahead estimation problems. Moreover, we assume that measurements for all covariates are available at $0, 1, \ldots, t$ and the models are entitled to use all that information. Hence all covariates information up to time t is used to derive the estimation of the hazard function at u for $u = t+1, \ldots, T$.

We investigate different methods to get these estimations based on random forests. They can be divided into three main approaches to address the problem.

A first direct approach is to only use local information at each unique combination of (t, u). In this case, we build separate forests for each of the $T * (T+1)/2$ possible combination of t and u. For example, when $T = 5$ as in Table 2.3, we would have 15 forests. More precisely, for a given pair (t, u), we use only the subjects that are still alive and not censored at time $u - 1$ to build the forest to estimate the hazard at time u. This way they are at risk of experiencing the event at time u. However, we only use all the covariates information available up to time t to build the forest. To illustrate this approach, we use the generic data set given in Table 2.1. To build the forest for the pair $(t, u) = (2, 4)$, the data set to

Table 2.4: Data set used with the first approach, separate forests for all combinations of (t, u), when $(t, u) = (2, 4)$

id	y	$X_1(0)$	$X_1(1)$	$X_1(2)$	X_2
2	0	$x_{12}(0)$	$x_{12}(1)$	$x_{12}(2)$	x_{22}
5	0	$x_{15}(0)$	$x_{15}(1)$	$x_{15}(2)$	x_{25}
6	0	$x_{16}(0)$	$x_{16}(1)$	$x_{16}(2)$	x_{26}
8	1	$x_{18}(0)$	$x_{18}(1)$	$x_{18}(2)$	x_{28}
10	0	$x_{110}(0)$	$x_{110}(1)$	$x_{110}(2)$	x_{210}

use would be the one given in Table 2.4. Note that only the subjects that are still alive and not censored at time $u - 1 = 3$ are included. The outcome y has a value of 1 if the event occurred at time $u = 4$ and 0 otherwise. Three potential covariates are available: $X_1(1)$, $X_1(2)$, and X_2. They are all measurements of the time-varying covariate X_1 up to $t = 2$, and the unique value of the time-independent covariate X_2. Each subject appears on at most one line. An ordinary forest for a binary outcome can then be built using this data set. The predictions from this forest provide direct estimates of the hazard at time $u = 4$, using all the covariates information up to time $t = 2$. Using separate models might be good if the hazards at different times are related to different covariate patterns. However, this approach will likely lose efficiency when the hazards are related to similar covariate patterns because of the variability induced by using separate models.

The second approach works by pooling all u values together for a given t. In this case, we build one forest for each value of t and get the whole hazard function for $u > t$ at once. More precisely, for a given t, we use all the covariates information available up to time t to build a forest to get estimations of the hazard function at $u = t + 1, t + 2, ..., T$. Overall, we have to build T forests to get the predictions for all possible combinations of t and u. Two ways of building trees and forests based on this idea have been proposed by Bou-hamad et al. (2009) and Bou-Hamad et al. (2011a) on one hand, and by Schmid et al. (2016) on the other hand, and were presented in the last section. See also chapter 6 of Tutz and Schmid (2016). We investigate both methods. An illustration of the structure of the data set used by the Bou-hamad et al. (2009) method for $t = 2$, again starting with the generic data set of Table 2.1, is given in Table 2.5. This data set is in person-period format and only the

Table 2.5: Data set used with the Bou-hamad et al. (2009) and Schmid et al. (2016) method when $t = 2$

id	y	period	$X_1(0)$	$X_1(1)$	$X_1(2)$	X_2
2	0	3	$x_{12}(0)$	$x_{12}(1)$	$x_{12}(2)$	x_{22}
2	0	4	$x_{12}(0)$	$x_{12}(1)$	$x_{12}(2)$	x_{22}
2	1	5	$x_{12}(0)$	$x_{12}(1)$	$x_{12}(2)$	x_{22}
3	0	3	$x_{13}(0)$	$x_{13}(1)$	$x_{13}(2)$	x_{23}
5	0	3	$x_{15}(0)$	$x_{15}(1)$	$x_{15}(2)$	x_{25}
5	0	4	$x_{15}(0)$	$x_{15}(1)$	$x_{15}(2)$	x_{25}
5	1	5	$x_{15}(0)$	$x_{15}(1)$	$x_{15}(2)$	x_{25}
6	0	3	$x_{16}(0)$	$x_{16}(1)$	$x_{16}(2)$	x_{26}
6	0	4	$x_{16}(0)$	$x_{16}(1)$	$x_{16}(2)$	x_{26}
6	0	5	$x_{16}(0)$	$x_{16}(1)$	$x_{16}(2)$	x_{26}
8	0	3	$x_{18}(0)$	$x_{18}(1)$	$x_{18}(2)$	x_{28}
8	1	4	$x_{18}(0)$	$x_{18}(1)$	$x_{18}(2)$	x_{28}
9	1	3	$x_{19}(0)$	$x_{19}(1)$	$x_{19}(2)$	x_{29}
10	0	3	$x_{110}(0)$	$x_{110}(1)$	$x_{110}(2)$	x_{210}
10	0	4	$x_{110}(0)$	$x_{110}(1)$	$x_{110}(2)$	x_{210}

subjects that are still alive and not censored at time $t = 2$ are included. The outcome y has a value of 1 if the event occurred at the corresponding period and 0 otherwise The LL-based splitting rule described in the last section is used and the covariates are $X_1(1), X_1(2), X_2$. An illustration of the structure of the data set used by the Schmid et al. (2016) method for the same problem is also given in Table 2.5. The same covariates as for the preceding method are available except that the period itself is also considered as a covariate. This time, the Gini splitting rule is used. Unlike the first approach which builds separate forests for all combinations of (t, u), this one pools all the information from the different estimation horizons for a given t and may reduce the variability when the hazards are related to similar covariate patterns.

The third approach is inspired from the supermodel based on stacked data used in landmark analysis (van Houwelingen, 2007; van Houwelingen and Putter, 2011). Instead of only pooling the information from the different estimation horizons for a given t, as in the second approach, we can go a step further and pool all the information for all combinations of (t, u) at once. The idea is to borrow information from different values of t, in addition to the one from the different estimation horizons. Consequently, a single forest is built and provides estimations of the hazards for all combinations of (t, u). A super person-period data set is created by stacking the data sets from all values of t in the Bou-hamad et al. (2009) or the Schmid et al. (2016) method described above. An illustration of the super person-period data set created for the data in Table 2.1 is given in Table 2.6. Only the first four subjects (up to id=4) are shown in the table to save space. Each subject has one line for each pair (t, u) where it was at risk of experiencing the event. This time, both the estimation horizon u and the value of t are potential covariates, in addition to the two other covariates. Note that only the current values, at t, of the time-varying covariates are used. An ordinary forest for a binary outcome with the Gini criterion can be built using this data set. The estimation of the hazard function for a pair (t, u) is obtained by getting the prediction from the forest using the covariate values at time t and the pair (t, u) itself.

A summary of the four methods is given below. They will be compared in the simulation study.

1. Separate forests for all combinations of (t, u). Number of forests$=T(T+1)/2$. This method will be called "Separate" from now on.

2. Separate forests for all values of t with the Bou-hamad et al. (2009) method. The bootstrap is performed on the original data. Number of forests$=T$. This method will be called "BLB" from now on.

3. Separate forests for all values of t with the Schmid et al. (2016) method. The bootstrap is performed on the person-period data. Number of forests$=T$. This method will be called "SKHT" from now on.

4. Super person-period forest. The bootstrap is performed on the stacked data. Number of forest=1. This method will be called "Superpp" from now on.

A fifth method consisting on taking the simple average of the hazard estimates from the first four methods, i.e. Separate, BLB, SKHT and Superpp, will also be added. This method will be called "Average" from now on.

Table 2.6: Data set used with the third approach, i.e. a super person-period data set used for all combinations of (t, u). Only the first four subjects (up to id=4) are shown to save space.

id	y	u	t	X_1	X_2
1	0	1	0	$x_{11}(0)$	x_{21}
1	1	2	0	$x_{11}(0)$	x_{21}
1	1	2	1	$x_{11}(1)$	x_{21}
2	0	1	0	$x_{12}(0)$	x_{22}
2	0	2	0	$x_{12}(0)$	x_{22}
2	0	2	1	$x_{12}(1)$	x_{22}
2	0	3	0	$x_{12}(0)$	x_{22}
2	0	3	1	$x_{12}(1)$	x_{22}
2	0	3	2	$x_{12}(2)$	x_{22}
2	0	4	0	$x_{12}(0)$	x_{22}
2	0	4	1	$x_{12}(1)$	x_{22}
2	0	4	2	$x_{12}(2)$	x_{22}
2	0	4	3	$x_{12}(3)$	x_{22}
2	1	5	0	$x_{12}(0)$	x_{22}
2	1	5	1	$x_{12}(1)$	x_{22}
2	1	5	2	$x_{12}(2)$	x_{22}
2	1	5	3	$x_{12}(3)$	x_{22}
2	1	5	4	$x_{12}(4)$	x_{22}
3	0	1	0	$x_{13}(0)$	x_{23}
3	0	2	0	$x_{13}(0)$	x_{23}
3	0	2	1	$x_{13}(1)$	x_{23}
3	0	3	0	$x_{13}(0)$	x_{23}
3	0	3	1	$x_{13}(1)$	x_{23}
3	0	3	2	$x_{13}(2)$	x_{23}
4	0	1	0	$x_{14}(0)$	x_{24}

Note that we also investigated a version of methods 1 to 3 where only the last measurement of the time-varying covariates is used. This version was almost always inferior to the one using the whole covariate history described above. Hence only the results for this version are reported. Finally, in the simulation study, the following benchmark method that do not use the covariates at all is also included in the simulation study. At each period, it is the ratio of the number of events to the total number of subjects at risk of experiencing the event.

2.4 Simulation Study

The `R` software (R Core Team, 2014) was used to perform the simulation study. The package `randomForestSRC` (Ishwaran and Kogalur, 2014) was used to build the forests for methods 1, 3 and 4 (methods Separate, SKHT and Superpp), i.e., all methods that require a classification forest. Method 2, the discrete survival forest by Bou-Hamad et al. (2011a) (method BLB), is implemented in `Fortran` and callable from `R` (R Core Team, 2014). The number of trees in all forests is 100.

2.4.1 Simulation Design

The data generating process (DGP) is inspired by models I and II used in the simulation study in Huang et al. (2016). The number of periods is 5. Let $(\beta_0, \beta_1, \beta_2)' = AZ + b$, where Z is a three-dimensional random vector from the multivariate normal distribution with 0 mean vector and identity covariance matrix, A is a 3×3 matrix and b is a three-variate vector. The parameters A and b are parameters specified below. For $t = 1, \ldots, 5$, the hazard function for a given subject is $h(t) = (1 + \exp(-(f(t)))^{-1}$ where $f(t) = \beta_0 + \beta_1 t + \beta_2 t^2$. Three time-independent and four time-varying covariates are available. The three time-independent covariates are generated as $(X_1, X_2, X_3)' = Z + \epsilon_F$ where ϵ_F has a three-variate normal distribution with a mean vector of 0 and a covariance matrix of Σ_F specified below. The four time-varying covariates are denoted by $X_k(t)$ for $k = 4, \ldots, 7$, and $t = 0, \ldots, 4$. They are generated independently for different values of t. For a given t, $(X_4(t), X_5(t), X_6(t), X_7(t))$

is generated from a multivariate normal distribution with mean vector $\mu_k(t)$ and covariance matrix Σ_V. The matrix Σ_V is also specified below and the mean functions are given by, for $t = 1, \ldots, 5$,

$$\mu_1(t-1) = \begin{cases} h(t) & h(t) \leq 0.4 \\ 0.4 & h(t) > 0.4 \end{cases}$$

$$\mu_2(t-1) = \begin{cases} 0 & h(t) \leq 0.2 \\ h(t) - 0.2 & 0.2 < h(t) \leq 0.6 \\ 0.4 & h(t) > 0.6 \end{cases}$$

$$\mu_3(t-1) = \begin{cases} 0 & h(t) \leq 0.4 \\ h(t) - 0.4 & 0.4 < h(t) \leq 0.8 \\ 0.4 & h(t) > 0.8 \end{cases}$$

$$\mu_4(t-1) = \begin{cases} 0 & h(t) \leq 0.6 \\ h(t) - 0.6 & h(t) > 0.6 \end{cases}$$

Hence, contrarily to the setup in Huang et al. (2016) where a single time-varying covariate contains all the information about the hazard function, the seven covariates used in this DGP all possess only partial information about the hazard function. Finally, additional probabilities of censoring were added independently of the covariates. The vector (c_1, c_2, c_3, c_4), with $0 \leq c_i \leq 1$, gives the probability of being censored at times 1 to 4. Note that the censoring occurs after the event. Consequently, if a subject has not experienced the event at time i ($i = 1, \ldots, 4$), then we check if it is censored by generating a Bernoulli random variable with probability of success c_i. All subjects that have not experienced the event at time 5 are automatically censored at 5.

The parameters we used are $b = (-3.0, 1.5, -0.15)$,

$$A = \begin{pmatrix} 0.7461 & 0.07571 & 0.009966 \\ 0.07571 & 0.3671 & 0.008611 \\ 0.009966 & 0.008611 & 0.03511 \end{pmatrix},$$

$$\Sigma_F = \begin{pmatrix} 1.125 & 0.1687 & 0.01688 \\ 0.01688 & 0.2812 & 0.008437 \\ 0.01687500 & 0.0.008437 & 0.002813 \end{pmatrix},$$

$$\Sigma_V = \begin{pmatrix} 0.010 & 0.003 & 0.003 & 0.003 \\ 0.003 & 0.010 & 0.003 & 0.003 \\ 0.003 & 0.003 & 0.010 & 0.003 \\ 0.003 & 0.003 & 0.003 & 0.010 \end{pmatrix}.$$

With this matrix A, the vector $(\beta_0, \beta_1, \beta_2)$ has a covariance matrix of

$$\Sigma_A = \begin{pmatrix} 0.5625 & 0.08437 & 0.008438 \\ 0.08437 & 0.1406 & 0.004218 \\ 0.008438 & 0.004218 & 0.001406 \end{pmatrix}.$$

The correlation matrices of Σ_A, Σ_F and Σ_V are all of the equicorrelation type with a correlation of 0.3. Finally, the probabilities of censoring are $(c_1, c_2, c_3, c_4) = (0.3, 0.2, 0.1, 0.1)$. With this DGP, the proportion of censoring is about 42%.

Figure 2.1 shows the typical hazard functions for a sample of 50 subjects generated with this DGP. The average hazard function is the one with the dots.

Each model is fitted with a training sample of size 1000. Then the performance of the fitted models is evaluated with five independent test sets of size 1000 each. The first test set includes only the subjects that are still at risk at $u = 1$. So it can be used when $(t, u) = (0, 1)$. The second test set includes the subjects that are still at risk at $u = 2$. So it can be used when $(t, u) = (0, 2)$ and $(t, u) = (1, 2)$. The third test set includes the subjects that are still

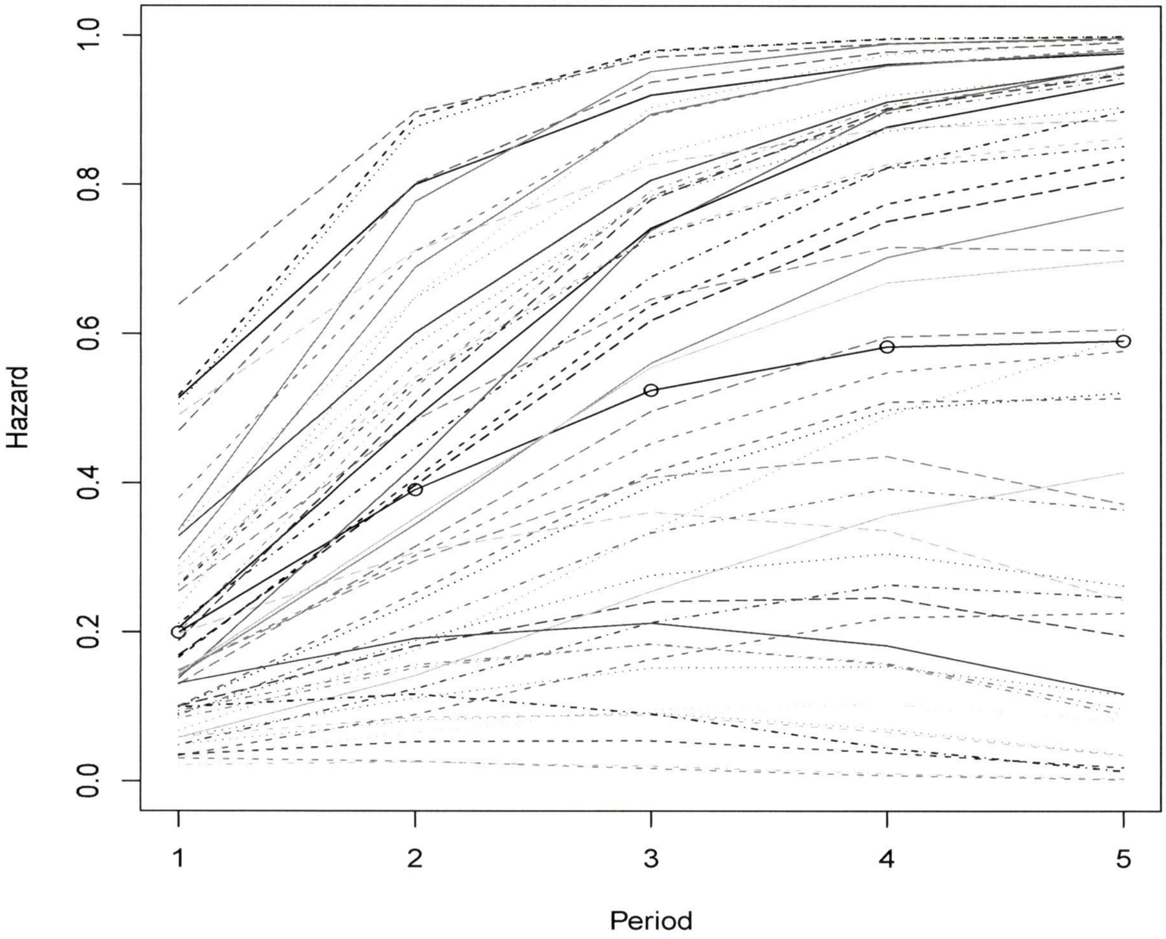

Figure 2.1: Hazard functions for a typical sample of 50 subjects

at risk at $u = 3$ and is used when $(t, u) = (0, 3)$, $(t, u) = (1, 3)$ and $(t, u) = (2, 3)$. Similarly, the fourth and fifth test sets include the subjects that are still at risk at $u = 4$ and $u = 5$, respectively. To ensure that there are 1000 qualified points in each test set, a large sample is generated first. Then five samples of size 1000 with the mentioned conditions are drawn from it. Each simulation is repeated 100 times.

2.4.2 Simulation Results

The criterion for evaluating the accuracy of the methods is the absolute log odds ratio (ALOR) defined by $|\ln((\hat{h}(1-h))/((1-\hat{h})h))|$ where $\hat{h}$ and h are the estimated and the true hazards. The ALOR takes a minimum value of 0 when $\hat{h} = h$. Note that the absolute difference $|\hat{h} - h|$ was also used but the conclusions are essentially the same and are not reported. The ALOR was chosen because it has the advantage of taking the magnitude of h and $\hat{h}$ into account. Namely, $|0.5 - 0.49|$ is treated as the same error as $|0.02 - 0.01|$ with the absolute difference, which can be questionable. The ALOR for the same two situations are 0.04 and 0.70, respectively.

Figure 2.2 provides the box-plots for the 100 simulations runs of each method for each combination (t, u). The first plot on the top left corner is for the case $t = 0$, the one on the top right corner is for the case $t = 1$, and so on. The first finding is that all methods perform better than the benchmark that does not use the covariates information. The second finding is that, for a given t (i.e. for a given plot), the performance worsen as u increases. This is obviously expected since it is more difficult to estimate the hazard for horizons further away. The third finding is that, for a given u, the performance of the methods get better as t increases. This is again expected because, as t increases, more information from the time-varying covariates becomes available. For example, if we look at the case $u = 5$, the last part of each plot, we see that the ALOR decreases from around 1.4 (when $t = 0$) to around 1.1 (when $t = 4$).

The four individual methods, Separate, SKHT, BLB and Superpp, have generally similar performances although there are some differences. It is interesting to see that which one is the best depends on the combination (t, u). For example, looking at the upper left plot

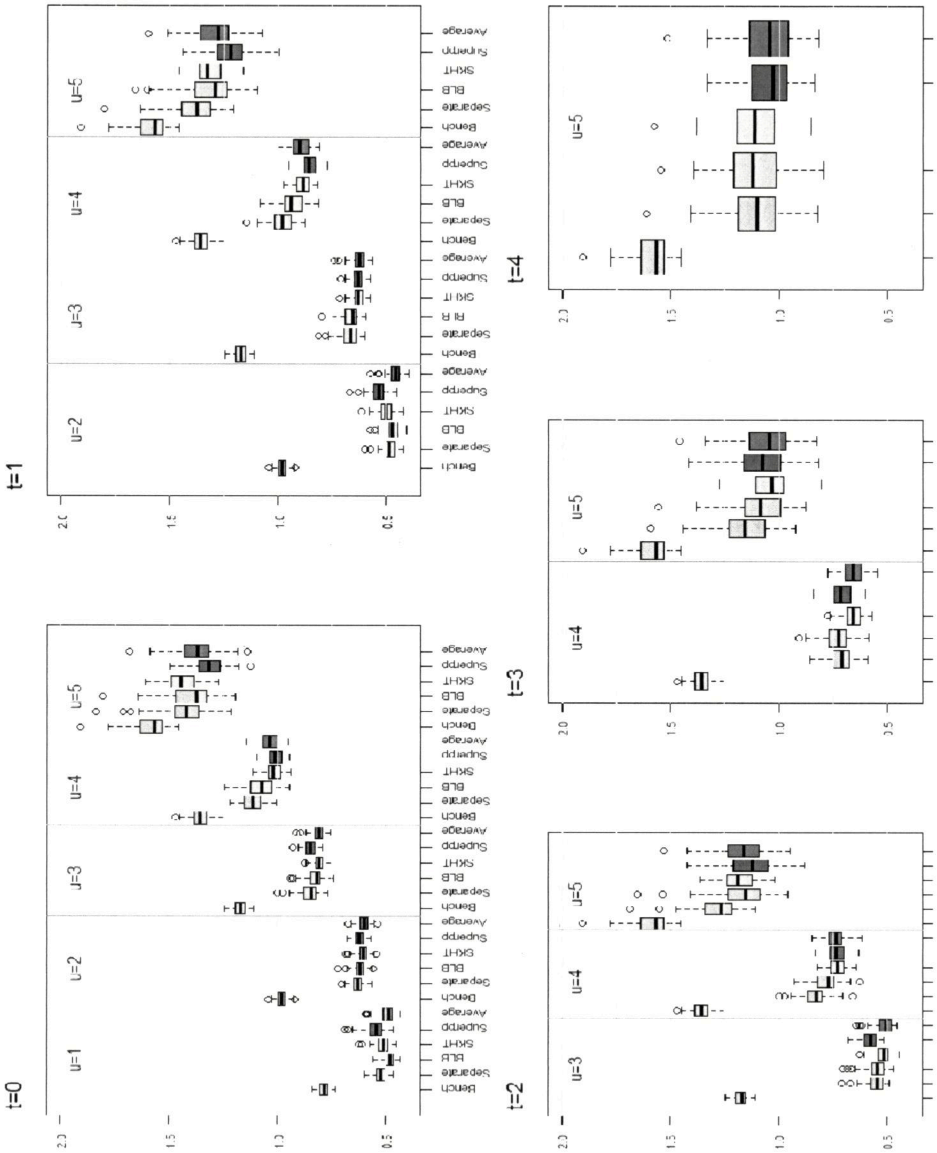

Figure 2.2: Simulation results comparing the distribution of absolute log odds ratio on test sets across methods

($t = 0$), we see that the BLB method is slightly better than the other ones when $u = 1$, then SKHT has a slight edge when $u = 2, 3$, then both SKHT and Superpp take the lead when $u = 4$, and finally, Superpp ends up on top when $u = 5$. So even for a given t, the relative performance of the four methods may depend on the estimation horizon. The SKHT method seems to do rather well throughout as it is generally the best method or close to the best one. However, choosing the best one might not be required in practice. Indeed, we see that using the simple average of the four individual methods (the last box-plot in each subplot) is generally very good and very close to the best one. It is even the best method in some cases. Note that we tried using more complex combination methods like stacking (Breiman, 1996; Wey et al., 2015) but the results were worse than the ones of the simple average.

2.5 Data Example

In this section, we revisit the bankruptcy data used in Section 3 of Bou-Hamad et al. (2011a) about United States firms that conducted IPOs (Initial Public Offerings) between 1990 and 1999. The data comes from COMPUSTAT and each firm is followed yearly starting from its IPO until 2006. The target variable is the bankruptcy, measured in number of years after the IPO. All firms that filed for bankruptcy under Chapter 7 or 11 are considered bankrupt. The data has 1143 firms, of which 189 went bankrupt during the study period. The time-varying covariates are the following financial ratios collected yearly:

- X_1 = Current Assets/Current Liabilities
- X_2 = Net Working Capital/Total Assets
- X_3 = Sales/Net Working Capital
- X_4 = Sales/Net Fixed Assets
- X_5 = Sales/Total Assets
- X_6 = Total Debt/Total Assets
- X_7 = Long-Term Debt/(Long-Term Debt+Total Equity)

- X_8 = Net Income/Total Assets
- X_9 = Net Income/Total Equity
- X_{10} = Market Value of Equity/Book Value of Total Debt.

In Bou-Hamad et al. (2011a), estimations at an horizon of three years were considered and only aggregate results were reported. This time, to make the investigation more complete and closer to a real practical application, we incorporate the time factor and include other estimation horizons. More precisely, we suppose that we are in a given year (here it will be 1997, 1998, or 1999), and then we want to estimate the hazard function, for the firms that are still currently alive, for horizons between three to six years. Table 2.7 describe the scenarios that we investigate.

Table 2.7: The 10 different estimation problems for the data example

Year after IPO	Predictions needed for (u)			
Current (t)	3	4	5	6
0	✓	✓	✓	✓
1		✓	✓	✓
2			✓	✓
3				✓

For instance, suppose we are in 1997, then some companies are just starting (0 years after the initial IPO). We want to estimate their hazard functions for years 3, 4, 5 and 6 after the IPO, using the initial ratio covariates (i.e., at time 0). Still in 1997, some companies are at their year 1 after the IPO and still alive. We want to estimate their hazard functions for years 4, 5 and 6 after the IPO, using the ratio covariates at time 0 and 1. The same logic remains for the companies that are at their year 2 or 3 after the IPO. These 10 pairs form the test data for 1997. Still for 1997, the training data set is comprised of all the information up to 1997. This is realistic because if we are in 1997, we do have all the covariates and outcome informations up to 1997. Hence, the outcomes used for the training data set are strictly prior to the ones used in the test data set. However, a given company can be into both the training and test data sets. For example, if a company had its IPO in 1994 and is still alive in 1997 (year 3 after its IPO), then it is part of the training data set because the

Table 2.8: Summary of the data over the 30 scenarios

	Trainnig data			Test data		
	Number of firms at risk	Number of bankruptcies	Proportion of bankruptcies (%)	Number of firms at risk	Number of bankruptcies	Proportion of bankruptcies (%)
Mean	315.5	10.6	3.23	131.1	3.53	2.69
Min	95	2	2.10	93	1	0.68
Max	730	23	4.20	163	11	6.75

covariates in 1994 can be used to estimate the hazard in 1997, which is one of the problem that interests us, ie. it is the (t=0, u=3) problem in table 2.7. The same company is in the test data set because now we want to estimate its hazard in 2000, which is the problem (t=3, u=6) in table 2.7. The important issue is that this setup respects the chronological order and these analyses could all be performed in a real application.

We repeat this process for 1998 and 1999. Consequently, we have 30 scenarios, 3 years × 10 pairs. Table 2.8 presents summary statistics for the number of firms at risk, the number of bankruptcies, and the percentage of bankruptcies over the 30 scenarios for the training and test data sets. The fact that bankruptcies is a rare event is clearly apparent which makes it challenging for both estimating the models and estimating their performances. The same four forests based methods are used to estimate the hazard functions and their average is also included as a fifth method. To be clear, for a given year (e.g. 1997), 10 forests are built for the Separate method (one for each pair (t,u)), four forests are built for the BLB and SKHT methods (one for each t), and one forest is built for the Superpp method. The same process is repeated for the three years 1997, 1998, and 1999.

Obviously, the true hazard functions are unknown with these data. Cumulative gains charts and Area Under the ROC Curve (AUC), obtained from the test data, will be used to evaluate the performance. A cumulative gains chart (and the related lift chart) provides more practical insight then a ROC curve. But AUC can still provide a general summary of the discriminative power of a model in a single number. A cumulative gains chart plots the percentage of events (here bankruptcies) detected if the p% of the observations with the highest estimated probabilities are classified as events, as a function of p. Hence it shows how much more a model is able to detect events compared to random assignment. The

Table 2.9: Average AUC with respect to the estimation horizon

	Method				
Horizon	Separate	BLB	SKHT	Superpp	Average
3	0.740	0.712	0.738	0.814	0.784
4 or more	0.723	0.735	0.765	0.756	0.764

cumulative gains computed on the test data are computed separately for each methods and each of the 30 scenarios. Figures 2.3 and 2.4 provide the average cumulative gains curves for the 12 scenarios with an estimation horizon of 3 on one hand and for the 18 scenarios with a horizon greater than 3 on the other hand. Table 2.9 provides the average AUC for the same two groups of scenarios.

When the estimation horizon is 3, Figure 2.3 shows that all methods perform better than random assignment (the diagonal line), but the Superpp method has a slight edge. For instance, if we target 20% of the firms with the highest probability of bankruptcies, then we detect 68.4% of the bankruptcies on average. This is much higher than the best of the four other methods (the method Average) that detects 51.6% of bankruptcies. Moreover, the average AUC of the Superpp method is 0.814, again much higher than for the other methods. When the estimation horizon is 4 or more, again all methods perform better than random assignment (see Figure 2.4). But this time, no method stands out compared to the others, as apparent from the plot and the average AUC. The cumulative gains chart indicate that, overall, the performance of the methods are slightly worse than for the scenarios with a horizon of 3 since the curves are closer to the diagonal line. This is less apparent by looking at the AUC since the average AUC is smaller for only three of the five methods.

2.6 Concluding Remarks

This paper provides an investigation of different discrete-time survival forest methods for dynamic predictions with time-varying covariates. All methods investigated can be easily implemented using existing *R* packages, except the Bou-Hamad et al. (2011a) method since it requires a specific splitting rule. The results show that all methods perform well and none surpasses the others. Taking the average of the estimated hazard functions from all methods

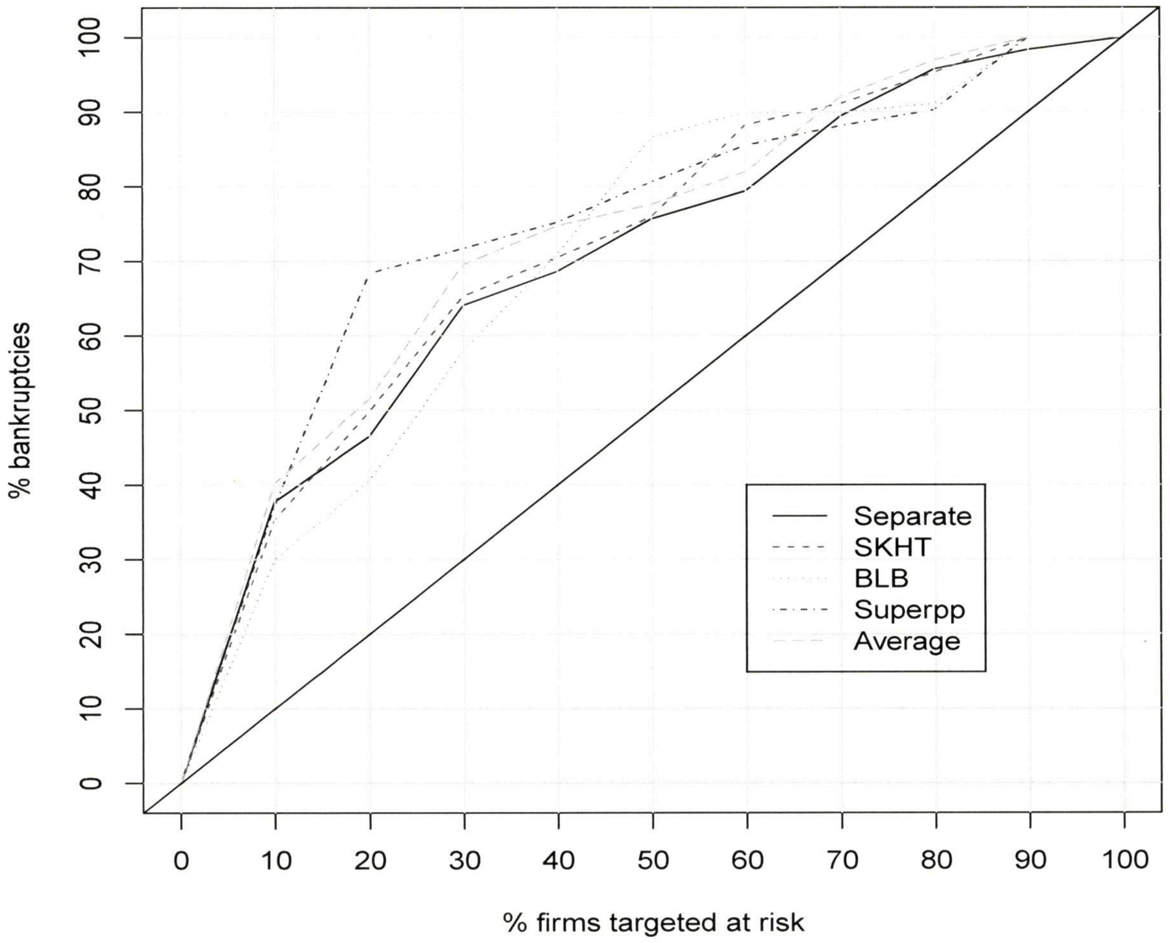

Figure 2.3: Average cumulative gains curves for the 12 scenarios with an estimation horizon of 3

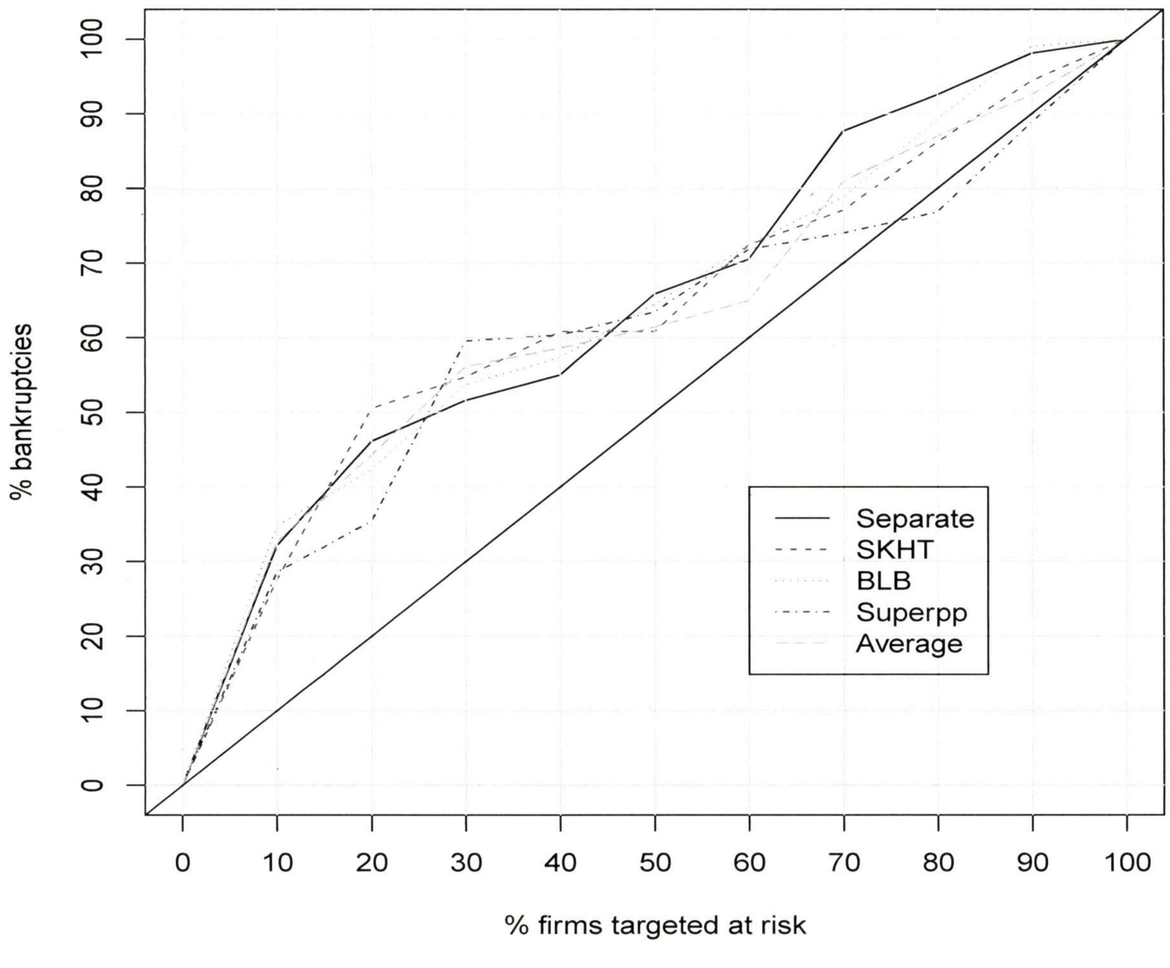

Figure 2.4: Average cumulative gains curves for the 18 scenarios with an estimation horizon greater than 3

provides a simple way to get a good result in most circumstances.

The methods investigated in this paper are basically landmark type analyses, where the trajectories of the time-varying covariates are not modelled explicitly. One possibility for future research would be to develop methods based on the joint modelling approach for the same problem. The challenge would be to have a practical way to handle situations with many time-varying covariates.

2.7 Acknowledgement

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by HEC Montreal.

Bibliography

J.R. Anderson, K.C. Cain, and R.D. Gelber. Analysis of survival by tumor response. *Journal of Clinical Oncology*, 1(11):710–719, 1983.

P. Bacchetti and M.R. Segal. Survival trees with time-dependent covariates: application to estimating changes in the incubation period of aids. *Lifetime data analysis*, 1(1):35–47, 1995.

I. Bou-hamad, D. Larocque, H. Ben-Ameur, L.C Mâsse, F. Vitaro, and R.E Tremblay. Discrete-time survival trees. *Canadian Journal of Statistics*, 37(1):17–32, 2009.

I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. Discrete-time survival trees and forests with time-varying covariates application to bankruptcy data. *Statistical Modelling*, 11(5): 429–446, 2011a.

I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics surveys*, 5:44–71, 2011b.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and regression trees. wadsworth & brooks. *Monterey, CA*, 1984.

Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

E. Elgmati, R.L Fiaccone, R Henderson, and J.NS Matthews. Penalised logistic regression and dynamic prediction for discrete-time recurrent event data. *Lifetime data analysis*, 21 (4):542–560, 2015.

L. Gordon and R.A. Olshen. Tree-structured survival analysis. *Cancer treatment reports*, 69 (10):1065, 1985.

R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

D.W. Hosmer Jr, S. Lemeshow, and S. May. *Applied survival analysis: regression modeling of time to event data*, volume 618. Wiley. com, 2011.

T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M.J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

X. Huang, F. Yan, J. Ning, Z. Feng, S. Choi, and J. Cortes. A two-stage approach for dynamic prediction of time-to-event distributions. *Statistics in medicine*, 2016.

H. Ishwaran and U.B. Kogalur. Random forests for survival, regression and classification (rf-src). 2014. URL `http://cran.r-project.org/web/packages/randomForestSRC/`. R package version 1.5.5.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

M. Leblanc and J. Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467, 1993.

E.B. Madsen, P. Hougaard, and E. Gilpin. Dynamic evaluation of prognosis from time-dependent variables in acute myocardial infarction. *The American journal of cardiology*, 51(10):1579–1583, 1983.

P Mayer, D Larocque, and M Schmid. Dstree: Recursive partitioning for discrete-time survival trees, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, A., 2014. URL `http://www.R-project.org/`.

D. Rizopoulos. *Joint models for longitudinal and time-to-event data: With applications in R*. CRC Press, 2012.

M. Schmid, H. Küchenhoff, A. Hoerauf, and G. Tutz. A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. *Statistics in Medicine*, 35(5):734–751, 2016. ISSN 1097-0258. doi: 10.1002/sim.6729. URL `http://dx.doi.org/10.1002/sim.6729`. sim.6729.

M.R. Segal. Regression trees for censored data. *Biometrics*, pages 35–47, 1988.

A.A Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

G. Tutz and M. Schmid. Modeling discrete time-to-event data, 2016.

H. van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007.

H. van Houwelingen and H. Putter. *Dynamic prediction in clinical survival analysis*. CRC Press, 2011.

Andrew Wey, John Connett, and Kyle Rudser. Combining parametric, semi-parametric, and non-parametric survival models with stacked survival models. *Biostatistics*, 16(3):537–549, 2015.

J.B Willett and J.D Singer. Investigating onset, cessation, relapse, and recovery: why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of consulting and clinical psychology*, 61(6):952, 1993.

Y. Zheng and P.J. Heagerty. Partly conditional survival models for longitudinal data. *Biometrics*, 61(2):379–391, 2005.

R. Zhu and M.R. Kosorok. Recursively imputed survival trees. *Journal of the American Statistical Association*, 107(497):331–340, 2012.

Chapter 3

Survival Forests for Data with Dependent Censoring

3.1 Abstract

Tree-based methods are very powerful and popular tools for analysing survival data with right-censoring. The existing methods assume that the true time-to-event and the censoring times are independent given the covariates. We propose different ways to build survival forests when dependent censoring is suspected, by using an appropriate estimator of the survival function when aggregating the individual trees and/or by modifying the splitting rule. The appropriate estimator used in this paper is the copula-graphic estimator. We also propose a new method for building survival forests, called point-wise forest, that may be used not only when dependent censoring is suspected, but also as a new survival forest method in general. The results from a simulation study indicate that these modifications improve greatly the estimation of the survival forest in situations of dependent censoring.

Keywords: Survival data; survival forest; dependent-censoring; ensemble methods; copula-graphic; right—censored data; random forest.

3.2 Introduction

Analysis of time-to-event data has become a growing field of interest in many research areas. What makes survival analysis different from ordinary regression analyses is the presence of censored observations which makes the analysis more complicated. Parametric (gamma,weibull) and semi-parametric (Cox) models are well developed and very useful when the functional link between the survival time and the covariates is known or can be adequately specified in advance. However, nonparametric methods like survival forests provide alternatives that do not need a rigid specification of a model beforehand. These methods are in a way more data-driven and require less assumptions.

Survival trees, introduced by Gordon and Olshen (1985), are an adaptation of the Classification and Regression Tree paradigm (CART) to right censored data (Breiman et al., 1984). Many other tree-building algorithms have been proposed since then; see Bou-Hamad et al. (2011) for an overview. However, it is now well-known that combining many trees, that is using an ensemble of trees, often provides a better predictive performance. Breiman (2001) introduced the random forest (RF) method for the classification problem and it became one of the most popular ensemble method.

Ensemble of survival trees were first proposed by Hothorn et al. (2006) but the random survival forest (RSF) method of Ishwaran et al. (2008) is the one which is the most closely related to the original RF method. The main output of this method is an estimated cumulative hazard function computed by averaging the Nelson—Aalen cumulative hazard function of each tree. This method is implemented in the R package `randomForestSRC` (Ishwaran and Kogalur, 2014; R Core Team, 2014). The log-rank statistic (Segal, 1988) is the default splitting rule to built individual trees in this package but other ones are available (e.g. log-rank score) and the user can even make a customized splitting rule.

Another random forest technique for right-censored data is the recursively imputed survival tree (RIST) proposed by Zhu and Kosorok (2012). In this method, information on censored observations is retained through computation of conditional survival distribution and recursive imputation. The log-rank test statistic is used as the default splitting rule.

Finally, L_1-forest is another random forest method for right-censored data (Moradian et al., 2016). Using the integrated absolute difference between survival functions' Kaplan-Meier estimates of two children nodes as the splitting rule makes this method less sensitive to the proportional hazards assumption, which is somewhat implicitly made when using the log-rank test as the splitting rule.

All aforementioned implementations of survival forests either explicitly state that their methods' underlying assumption is that the true event time and the true censoring time are independent given the covariates, or implicitly by using a splitting rule (e.g. log-rank test) or an estimation method (e.g. Kaplan-Meier, Nelson—Aalen) that works under the independent censoring assumption. However, this assumption is not always verified. For example, Li et al. (2007) consider the survival of patients diagnosed with prostate cancer. Patients dying from other causes are considered censored, with cardiovascular diseases being the major cause. Since prostate cancer and cardiovascular diseases share common risk factors, it is reasonable to suspect that the event and censoring times are dependent. None of the current implementations of survival forests can properly estimate the survival function in the dependent censoring context. The goal of this paper is to propose different ways to build survival forests when dependent censoring is suspected. More precisely, we will investigate the following two aspects of a survival forest to account for dependent censoring: 1) use an appropriate final estimator, and /or 2) modify the splitting rule. Moreover, we will present a new method for building survival forests, called point-wise forest or simply "p-forest", that may be used not only when dependent censoring is suspected, but also as a new survival forest method in general.

These modifications will be built around the "copula-graphic" estimator of the survival function introduced by Zheng and Klein (1995) and using the closed-form expression derived in Rivest and Wells (2001). A description of this estimator will be given in the next section for completeness.

This paper is organized as follows. The data setting and the proposed methods are described in Section 3.3. The results from a simulation study are presented in Section 3.4. Conclusion and directions for future research are discussed in Section 3.5.

3.3 Settings and Methods

3.3.1 Data Description

Observations for each subject i ($i = 1, \ldots, N$) are in the form of (τ_i, δ_i, x_i) where τ_i is the observed survival time for subject i, δ_i is the censoring index which takes a value of 0 if it is right censored and a value of 1 if it has experienced the event of interest, and x_i is the vector of covariates. Only time-invariant covariates are considered. Let U_i and V_i be the true time-to-event and the true censoring times for subject i. Hence, the observed time is τ_i=min(U_i,V_i), and $\delta_i = I(U_i \leq V_i)$. Obviously, τ_i, δ_i, U_i and V_i are all conditional on x_i but this dependence is suppressed to simplify the notation. Let $S(t|x_i) = S_i(t) = P(U_i > t) = P(U > t|x_i)$ be the survival function for subject i. The goal is to estimate the survival function for a new subject with vector of covariate x, that is $S(t|x)$.

The assumption of independent censoring is usually made. This means that U_i and V_i are independent, i.e. that the true time-to-event and the true censoring time are independent given the covariates. However, in this paper, we assume that U_i and V_i may be dependent, i.e. that the true time-to-event and the true censoring time may be dependent even after the effect of the covariates is taken into account.

3.3.2 Copula-Graphic Estimator

To estimate the survival function of a sample of survival times without covariates but with possible dependent censoring, Zheng and Klein (1995) proposed a nonparametric estimator called the "copula-graphic". It is based on copulas, first introduced by Sklar (1959), that allows to model the dependence relationship without any particular specification of the marginal distributions. In this paper, we focus on a special class of copula called the Archimedean copulas. Let U and V be two random variables (think about the time-to-event and censoring time) with respective marginal distribution functions $S(u)$ and $C(v)$. An Archimedean copula assumes that the joint distribution of U and V, $H(u, v)$, is constructed through a φ-generator and takes the form $H(u, v) = \varphi^{-1}[\varphi(S(u)) + \varphi(C(v))]$, where $\varphi(.)$ is a decreasing function defined on (0,1] with $\varphi(1) = 0$ and $\varphi(0) = \infty$. Two well-known families of Archimedean copulas are the Clayton's family (Clayton, 1978) with $\varphi_\alpha(t) = (t^{-\alpha} - 1)/\alpha$,and Frank's family (Frank, 1979) with $\varphi_\alpha(t) = -\ln[\{1 - \exp(-\alpha t)\}/\{1 - \exp(-\alpha)\}]$. In both cases, α is the dependence parameter. The reader can refer to Nelsen (1999) for more details on the generators of Archimedean copulas.

Rivest and Wells (2001) provide the following closed form expression for the estimation of the survival function with the copula-graphic estimator for Archimedean copulas :

$$\hat{S}_\alpha(t) = \varphi_\alpha^{-1}\Big(- \sum_{\tau_i \leq t, \delta_i = 1} \{\varphi_\alpha(\hat{\pi}(\tau_i)) - \varphi_\alpha(\hat{\pi}(\tau_i) - \frac{d_i}{N})\}\Big), \tag{3.1}$$

where $\hat{\pi}(\tau_i) = \frac{\sum_{i=1}^{N}(I(\tau_i > t))}{N}$ and d_i is the number of observations that experienced the event at τ_i. As stated by Rivest and Wells (2001) with an Archimedean copula for the dependency between censoring and survival and the dependence function given by $\varphi(.)$, (3.1) provides a net estimate of the marginal survival function.

3.3.3 Modification of Existing Methods

In this subsection, we present different possibilities for modifying existing survival forests methods to account for dependent censoring. We assume that the reader is familiar with the CART paradigm (Breiman et al., 1984) and the basic random forest method (Breiman, 2001). As mentioned in the Introduction, if dependent censoring is suspected, we could

1. use an appropriate estimator to compute the final estimator of the survival function and/or

2. modify the splitting rule to account for dependent censoring.

The easiest practical way to do it is to use a well-established package like `randomForestSRC` and only modify the way to combine the trees, that is, only use approach 1) above. This package allows to easily extract the set of observations ending in each terminal node of each tree in a forest. Hence, we can build a forest of trees as usual (e.g. with the log-rank statistic as splitting rule). To estimate the survival function of a new subject, the usual way would be to take the average over the trees of the Nelson—Aalen cumulative hazard function (the output from `randomForestSRC`) and then deriving the survival function from it (with $S(t) = \exp(-\Lambda(t))$ where Λ is the cumulative hazard function). But instead of that, we can also pool together all the observations that end up in the terminal nodes of all trees for this new observation and compute the copula-graphic estimator with them. This was the aggregation method used in Hothorn et al. (2006). Note that any given observation can appear more than once in the pooled set of observations. To make things more formal, let F be a survival forest (built with any method) and let x be the covariates vector of the observation for which we need an estimate of the survival function. Define by $T_F(x)$ the $1 \times n$ vector giving the number of times that each observation in the training data set ends up in a terminal node of F when x is the covariates vector. For example, with $n = 5$ observations in the training data set and a forest F formed by 3 trees, $T_F(x) = (3, 0, 1, 2, 2)$ means that x is in the same terminal node as the first observation in the training set in all 3 trees of the forest, that it is never in the same terminal node as the second observation, that it is in the same terminal node as the third observation in one of the 3 trees, and so on. Hence it is simply a weight vector. The estimated survival function is then the copula-graphic estimator computed with the set of observations $T_F(x)$. In the above example, it is the copula-graphic estimator computed with 8 observations (three times the first one, 1 time the third one, 2 times the fourth one and 2 times the fifth one).

Building the forest itself without correcting for dependent censoring and making a correction at the time of combining the information in the trees, as the method above, will most likely produce a better estimate of the survival function. However, it may be that using a modified splitting rule improves things even more. A survival forest that lends itself to this end is the L_1-forest introduced in Moradian et al. (2016). The usual splitting rule used to build survival forests is based on the log-rank statistic. However, the log-rank test may have a significant loss of power in some circumstances, for instance when the hazard functions or when the survival functions cross each other in the two compared groups. This is why Moradian et al. (2016) proposed using the following L_1 splitting rule:

$$(n_L n_R) \int_t |\hat{S}_L(t) - \hat{S}_R(t)| dt, \tag{3.2}$$

where $\hat{S}_L(t)$, $\hat{S}_R(t)$, n_L and n_R are the Kaplan-Meier survival function estimates and the number of observations in the left and right node, respectively. This splitting rule is designed to detect any type of differences between the two nodes. Obviously, if dependent censoring is present, then using the Kaplan-Meier estimator is not necessarily a good strategy. This is why we will investigate replacing it by a copula-graphic estimator giving the splitting rule

$$(n_L n_R) \int_t |\hat{S}_{\alpha L}(t) - \hat{S}_{\alpha R}(t)| dt, \tag{3.3}$$

where $\hat{S}_\alpha$ is defined in (3.1). To sum up, with this approach, both the splitting rule and the combination of the trees are modified to account for possible dependent censoring.

3.3.4 Point-Wise Forest

In the last subsection, we describe two approaches for dealing with dependent censoring with a survival forest. We can build a forest using a common algorithm and only apply a correction at the end when combining the information from the trees. But we can also apply a correction at the time of tree building by modifying the splitting rule and proposed to replace the Kaplan-Meier estimator by a copula-graphic estimator in the L_1 splitting rule of Moradian et al. (2016). This approach requires a choice of copula and a choice of dependency parameter (the α above) at the moment of tree building. If a sensitivity analysis is required by varying α, then a different forest would be required for each of them and could require a lot of computation time. In this subsection, we describe another approach that tries to account for dependent censoring at tree building time while not requiring the choice of a dependency parameter. This method, that we call the Point-wise forest or just p-forest for simplicity, uses a splitting rule which is free of the independent censoring assumption.

We can define the status of an observation at a given time point t as follows:

$$Y_i(t) = \begin{cases} 0 & \text{if subject } i \text{ is alive at } t\text{, that is, if } \tau_i > t \\ 1 & \text{if subject } i \text{ is dead at } t\text{, that is, if } \tau_i \leq t \text{ and } \delta_i = 1 \\ 2 & \text{if subject } i \text{ is censored at } t\text{, that is, if } \tau_i \leq t \text{ and } \delta_i = 0 \end{cases}$$

The "dead" and "alive" terms are used metaphorically, meaning the subject has observed or has not observed the event, respectively. The idea behind the p-forest is to build a random forest using the status $Y_i(t)$ as the outcome. It provides a predictive model for the status at time t which is free of the independent censoring assumption. The idea is to use the information from many forests built at different times t to estimate the survival function of new observations. Here is the detailed description of the p-forest method.

1. Select the number of forests M, the number of trees in a forest B, and the minimal number of forests required for an observation to count m ($m \in \{1, 2, \ldots, M\}$). In this paper, we selected $M = 9$, $B = 100$, and $m = 5$, or 9.

2. Select the points $t_1, t_2, \ldots, t_M$ where to compute the forests. In this paper, we use $M = 9$ and these points are the quantiles $0.1, 0.2, \ldots, 0.9$ of the Kaplan-Meier estimator of the survival function of the training data set.

3. For each $j = 1, 2, \ldots, M$, build a classification forest with the Gini criterion using $Y_i(t_j)$ as the outcome and all the covariates. We have M forests $F_1, F_2, \ldots, F_M$.

4. Let x be the vector of covariates of a new observation. Let $I(x, j)$ be a $1 \times n$ vector of indicator variable (0-1). The i^{th} element of $I(x, j)$, $I(x, j)_i$, is 1 if and only if the i^{th} observation in the training data is in the same terminal node as the new observation with vector of covariates x in at least one of the trees of the forest F_j.

5. Let $I^m(x)$ be a $1 \times n$ vector of indicator variable. The i^{th} element of $I^m(x)$, $I^m(x)_i$, is 1 if and only if the i^{th} observation in the training data is in the same terminal node as the new observation with covariate vector x in at least m forests. That is, $I^m(x)_i = 1$ if and only if $I(x, 1)_i + I(x, 2)_i + \cdots, I(x, M)_i \geq m$.

6. Let $W^m(x)$ be a $1 \times n$ vector giving the final weight of the training data used to compute

the estimate of the survival function for x. The i^{th} element of $W^m(x)$, $W^m(x)_i$, is 0 if $I^m(x)_i = 0$. If $I^m(x)_i = 1$, it is equal to the number of trees among all $B * M$ trees in all forests where the i^{th} observation is in the same terminal node as the new observation with covariate vector x.

7. Compute the estimate of the survival function (Kaplan-Meier or copula-graphic) for x using the training data with the weights $W^m(x)$.

The idea is that a training observation is used to estimate the survival function of a new observation x only if it used in enough (m) individual forests, that is if it contributes in a sufficient part of the time range. Once it is selected, an observation receives a weight to reflect its relative importance and the estimate of the survival function can be computed with these weights. Choosing $M = 9$ and $m = 5$, requires that an observation looks important in at least more than half the forests, that is more than half the time points selected, in order to be used. Choosing $M = 9$ and $m = 9$, requires that an observation looks important in all the forests in order to be used. Note that the set of observations used and the weights vary with x, that is, each new observation gets its own weight vector to compute the estimated survival function.

3.4 Simulation Study

3.4.1 Methods Compared

In this section, we investigate the performance of the proposed methods through a simulation study. Here is a description of the methods compared in the study. We use the Clayton copula for all methods. The link function between the dependency parameter α and Kendall's τ in Clayton's copula is $\tau = \alpha/(\alpha + 2)$. In what follows, we abbreviate Kaplan-Meier by KM and copula-graphic by CG.

Random survival forest (RSF) of Ishwaran et al. (2008) and some modifications

The log-rank test is used as the splitting rule for all RSF methods.

- A first benchmark method is the original RSF. In this case $\hat{S}(t) = \exp(-\hat{\Lambda}(t))$ where $\hat{\Lambda}$

is the estimated cumulative hazard function computed by averaging the Nelson—Aalen cumulative hazard function of each tree. It is denoted by RSF.

- Instead of that, we can compute an estimated survival curve using the pool of observations that end up in the terminal nodes of all trees as explained in Section 2.3. The copula-graphic estimator with a Kendall's τ of 0, 0.2, 0.4, 0.6, and 0.8 are used. Note that the case $\tau = 0$ reduces to the KM estimator. These five methods are denoted by RSF-KM, RSF-CG2, RSF-CG4, RSF-CG6 and RSF-CG8.

The forests are built with the `randomForestSRC` package.

Recursively imputed survival tree (RIST) of Zhu and Kosorok (2012)

This is the second benchmark method. We use it with three imputation steps. The log-rank test is used as the splitting rule. It is denoted by RIST3. The R code generously made available by the authors Zhu and Kosorok (2012) is used for RIST3.

L_1-forest of (Moradian et al., 2016) and some modifications

- A third benchmark method is the original L_1-forest. In this case, the trees are built with the L_1 splitting rule using KM, that is (3.2). The final estimator with the pooled set of observations is also computed with KM. It is denoted by L1-KM.

- As a first modification, we can still use KM in the splitting rule but use CG for the final estimator with the pooled set of observations. Again, we use CG with a Kendall's τ of 0.2, 0.4, 0.6, and 0.8. These four methods are denoted by L1-CG2, L1-CG4, L1-CG6 and L1-CG8.

- We can go a step further and use CG in the splitting rule, that is (3.3). We then use CG also for the final estimator, with the same dependency parameter as the one in the splitting rule. These four methods are denoted by L1-S-CG2, L1-S-CG4, L1-S-CG6 and L1-S-CG8.

These methods are implemented by the authors in Fortran and callable from `R`.

Point-wise forest

In all cases we use $M = 9$ points, that is we build nine forests. They are selected as equispaced quantiles $(0.1, 0.2, \ldots, 0.9)$ of the KM of the training data. Two values of m (the minimal number of forests required for an observation to count in the final estimator) are used, $m = 5$ and 9. Finally we use the copula-graphic estimator with a Kendall's τ of 0, 0.2, 0.4, 0.6, and 0.8 to compute the final estimate with the pooled set of observations. These 10 methods are denoted by pf5-KM, pf5-CG2, pf5-CG4, pf5-CG6, pf5-CG8, pf9-KM, pf9-CG2, pf9-CG4, pf9-CG6, and pf9-CG8.

These methods are implemented by the authors in `R`.

Consequently, there are 26 methods that are investigated. In all forests, 100 trees are grown and the number of covariates tried at each split is set to the integer part of $\sqrt{p}$ (p being the number of covariates), as suggested by Ishwaran et al. (2008). As a stopping criterion, the minimum number of observations in a terminal node is 3, the default value in `randomForestSRC`.

3.4.2 Simulation Design

Three Data Generating Processes (DGPs) are used to generate artificial data. For each DGP, five different Kendall's τ values are used to incorporate a dependence between the true time-to-event U and the censoring time V. They are 0, 0.2, 0.4, 0.6, and 0.8. A Kendall's τ of 0 corresponds to the case of independent censoring, given the covariates. Three different censoring proportions are considered, namely, 20%, 40% and 60%. Overall, 45 scenarios are investigated. All models are fitted with a training sample of size 500 and their performances are evaluated with an independent test set of size 1000. Each simulation is repeated 100 times. In all cases, the parameter β controls the proportion of censoring. The values of β which produce the desired censoring proportions were found empirically. The description of the marginal distributions of U and V are given below. The vector (U, V) is generated with Clayton's copula with the aforementioned values of Kendall's τ.

DGP 1. It is an altered version of scenario 2 from Section 4.1 of Zhu and Kosorok (2012). Ten independent and uniformly distributed (iid) uniform covariates on the interval (0,1) are available, $X_1, \ldots, X_{10}$. The marginal survival time is drawn from an exponential distribution with mean $\mu_U = 10|\sin(X_1\pi - 1)| + 3|X_2 - 0.5| + X_3$. The marginal censoring time is uniformly distributed on the interval $(0,\beta)$. Hence, only the survival time is related to the covariates. Moreover, only the first three covariates are related to the survival time and the others are noise covariates.

In the next two DGPs, the survival and censoring times depend on the same covariates and are linked to each other and they also follow a joint Clayton's copula distribution.

DGP 2: It is a tree with four terminal nodes. Ten iid standard normal covariates are available, $X_1, \ldots, X_{10}$. The marginal survival time is generated from the exponential distribution with mean μ_U given by:

$$\mu_U = \begin{cases} 1 & \text{if } X_1 \leq 0, X_2 \leq 0 \\ 1/2 & \text{if } X_1 \leq 0, X_2 > 0 \\ 1/3 & \text{if } X_1 > 0, X_3 \leq 0 \\ 1/4 & \text{if } X_1 > 0, X_3 > 0. \end{cases}$$

The censoring time is generated also from the exponential distribution with mean $\mu_V = \beta/(1-\beta)(\mu_U)$. Hence, only the first three covariates are related to the survival and censoring times and the others are noise covariates.

DGP 3. This is another complex dependent censoring DGP. It is adapted from scenario 1 in Section 4.1 of Zhu and Kosorok (2012). The 25 covariates $X_1, \ldots, X_{25}$ are generated from a multivariate normal distribution with covariance matrix $\sigma_{ij} = 0.9^{|i-j|}$. The marginal survival time follows an exponential distribution with mean of $\mu_U = 0.1|\sum_{i=11}^{20} X_i|$. The marginal censoring time also have an exponential distribution with mean $\mu_V = \mu_U/\beta$.

3.4.3 Simulation Results

To evaluate how well the survival function is estimated for each method, we use the Integrated Absolute Error (IAE). Let S be the true survival function and $\hat{S}$ be an estimated survival function. The IAE is defined by:

$$\text{IAE} = \int_t |S(t) - \hat{S}(t)| dt. \tag{3.4}$$

Obviously, smaller values of the IAE are better. The box plots of the IAE for the 100 runs for all methods in all scenarios are provided in figures 3.3 to 3.11. But before looking at these detailed results, we give an overall view in figures 3.1 and 3.2. There are 4500 simulation runs in all (100 runs × 45 scenarios). In a given run, we can compute the % increase in IAE with respect to the best performer for this run, for all 26 methods. More precisely, for a given run, let IAE_i ($i = 1, 2, \ldots, 26$) be the IAE for the i^{th} method and $\text{IAE}_{\text{min}} = \min\{\text{IAE}_1, \ldots, \text{IAE}_{26}\}$ be the best (smallest) IAE for this run. Then the % increase in IAE of method i with respect to the best is given by $(\text{IAE}_i/\text{IAE}_{\text{min}}\text{-}1)\times 100$. Hence, for a given run, if there are no tied IAE, then one and only one method will get a % increase of 0 (the best one) and the others will get positive values. Using this criterion makes it possible to compare the methods across all scenarios at once.

In the following discussion, the "working" τ is the value of Kendall τ used by the estimation method and the "true" τ is the value of Kendall τ used to generate the data. Figure 3.1 presents the box plots of the % increase in IAE with respect to the best method for the 4500 runs for all methods. But to see more clearly, a zoom in is given at the bottom of Figure 3.1. In complement, Table 3.1 presents the average and median of the % increase in IAE with respect to the best globally for the 4500 runs. Looking first at the global results with Figure 3.1 and Table 3.1, we see that RSF-CG6, pf5-CG6, pf5-CG4, and RSF-CG8 have the four lowest % increase averages with respective values 14.9%, 16.3%, 16.4%, and 17.3%. The same four also have the lowest medians but in the different order of 10.4% (pf5-CG4), 10.7% (RSF-CG8), 11.4% (pf5-CG6), and 11.6% (RSF-CG6). As seen in Figure 3.1, the same four also have the lowest third quartile. More generally, all methods that use CG as the final estimator of the survival function also have a good performance and are not that far from the top four. The globally worst methods are the ones that do not use CG as the final estimator. These methods are RSF, RIST3, RSF-KM, L1-KM, pf5-KM, and pf9-KM who all use KM as the final estimator except RSF and RIST3. The method RSF uses an average (over the trees) of Nelson—Aalen cumulative hazard functions and then we transform it to get the survival function. Since it is based on imputation and no censored observations remains, the method RIST3 uses an empirical type estimator (similar to KM) in the terminal nodes

Method	Mean	Median
RIST3	68.7	49.6
RSF	119.8	91.5
RSF-KM	75.8	45.8
RSF-CG2	37.6	24.0
RSF-CG4	19.2	13.7
RSF-CG6	14.9	11.6
RSF-CG8	17.3	10.7
L1-KM	51.7	36.1
L1-CG2	27.8	20.3
L1-CG4	21.6	18.1
L1-CG6	25.9	20.8
L1-CG8	26.0	20.7
L1-S-CG2	27.8	20.3
L1-S-CG4	22.5	19.1
L1-S-CG6	24.1	20.9
L1-S-CG8	28.2	24.2
pf5-KM	72.4	24.2
pf5-CG2	31.9	12.9
pf5-CG4	16.4	10.4
pf5-CG6	16.3	11.4
pf5-CG8	22.0	15.1
pf9-KM	86.4	33.7
pf9-CG2	47.1	22.1
pf9-CG4	24.6	17.0
pf9-CG6	19.1	15.2
pf9-CG8	21.6	15.4

Table 3.1: Global results. Average and median of the % increase in IAE with respect to the best globally for the 4500 runs.

of each tree and averages them. Thus, like KM, it does not try to account for dependent censoring. Consequently, the first general finding is that trying to account for dependent censoring by using CG, even if the dependency parameter is wrong, is much better than doing nothing.

The second general finding comes by noting that the two approaches L1-CGx and L1-S-CGx for a given value of x, that is working τ, have a very similar performance. Both approaches use CG as the final estimator, but L1-S-CGx also uses CG in the splitting rule. Hence, it seems that modifying the splitting rule to account for possible dependent censoring is not required if the final estimator does account for it, at least with the L_1 approach. It is possible that the biases in the estimation of the survival function in both nodes somewhat cancel out and that the basic L_1 splitting rule with KM does a fairly good job at finding reasonable splits, at least compared to using the L_1 splitting rule with CG. This is a nice finding in practice because it means that we may use existing survival tree building algorithms as we only need to modify the final estimator of the survival function when combining the information from the trees in the forest.

The third general finding is the fact that RSF-KM is better than RSF, at least in the scenarios considered. This means that, to compute the final estimate of the survival function, it is preferable to use the pooling method described in Section 2.3 instead of taking the average of the Nelson—Aalen cumulative hazard function that comes out of `randomForestSRC` and transform it. The same trees are built in both cases, only the way to combine the information in the end changes. A more detail comparison is given below when we look at the different scenarios separately.

Figures 3.2 to 3.10 present detailed results for the individual scenarios. Each figure presents the results for one DGP, one proportion of censoring, and for all values of the true τ. The first finding was expected. For a given method, all things being equal, the performance worsens as the proportion of censoring increases. We can see that by looking at the corresponding plots in figures 3.2, 3.3 and 3.4, for DGP 1, at the ones in figures 3.5, 3.6 and 3.7, for DGP 2, and at the ones in figures 3.8, 3.9 and 3.10, for DGP 3. We also see that the variability between the methods increases as the proportion of censoring increases.

Another finding is that for a given approach (i.e RSFxx, L1xx, L1-Sxx, pf5xx, or pf9xx), the value of the working τ that produces the best result tends to increase with the true value of the true τ, all other things being equal. To give a specific example, take Figure 3.4 (DGP 1 with 60% censoring) and look at the pf5 approach in the next to last portion of the plots (the 5 box plots just before the last 5 box plots in each plot). When the true τ is 0 (plot in the upper left corner), then pf5-CG4 is the best among the pf5 methods, that is, the best one is the one that uses a working $\tau = 0.4$ to compute the final estimator when the true τ is 0. When the true τ is 0.2 (plot in the upper right corner), then pf5-CG6 is the best. When the true τ is 0.4 (plot in the middle left), then pf5-CG6 is again the best, followed closely by pf5-CG8. When the true τ is 0.6 (plot in the middle right), then pf5-CG8 is the best.

Finally, when the true τ is 0.8 (lower plot), then pf5-CG8 is again the best. This tendency occurs in most cases. This was expected. However, the value of working τ that produces the best result has a tendency to be greater or equal than the true τ.

The global results discussed before showed that RSF-KM is better than RSF. In fact, we see that RSF-KM is better than RSF for 28 of the 45 scenarios and the opposite is true for the other 17 scenarios. The RSF method does better for DGP 1 while RSF-KM is generally better for the other two DGPs.

From these 9 figures, we can see that the value of the working τ used to compute the estimate of the survival function has generally a much bigger impact on the performance than the choice of the particular approach (i.e RSFxx, L1xx, L1-Sxx, pf5xx, or pf9xx). Moreover, as expected, the impact of the choice of the working τ increases as the proportion of censoring increases.

For DGP 1, the pf9xx approach is the best when the censoring proportion is 20% (Figure 3.2). When the censoring proportion is 40% (Figure 3.3), the pf5xx approach has a slight edge over the other, but the RSFxx and pf9xx approaches can be very close to it. When the censoring proportion is 60% (Figure 3.4), the best method in each plot comes from the pf5xx approach. However, this method can also have a poor performance if the value of the working τ used is not adequate. The L1xx approach is more stable (not the best but never bad) when the true τ is greater or equal to 0.2.

For DGP 2, all methods offer a similar performance when the censoring proportion is 20% (Figure 3.5), except RSF and RIST3 which are worse. More differences occur when the censoring proportion is 40% (Figure 3.6). But when they use the same value of working τ, all methods perform similarly, except RSF and RIST3 which again are worse than the others. Even more differences occur when the censoring proportion is 60% (Figure 3.7). In this case, we see that the methods that use KM perform particularly badly and it gets worse as the true value of τ increases.

For DGP 3 and 20% of censoring (Figure 3.8), the pf5xx approach is the best one up to a true τ value of 0.4, then it is joined by the pf9 approach for greater values of the true τ. When the proportion of censoring is 40% or 60% (figures 3.9 and 3.10), we see that the value of the working τ is the most important driver of the performance. Any approach can do well if the working τ is well selected.

In all the simulations so far, the methods use Clayton copula (the working copula) and the data were also generated with the Clayton copula (the true copula). We ran some additional simulations to investigate how robust are the methods if the working copula is misspecified. The same three DGPs are used but this time the data are generated with the Frank copula and only the most extreme case is treated, that is the one with 60% of censoring and a value of α giving a true τ of 0.8 for the Clayton copula. Figure 3.11 presents the results. The plots in the left column are the same as the lower ones in figures 3.4, 3.7, and 3.10 and are reproduced to ease the comparison. The plots in the right column are the new ones when the data are generated with the Frank copula. Note that the methods still use Clayton as the working copula. Comparing the plots line by line in Figure 3.11, we see that the methods seem robust to the choice of the working copula as the results are very similar. In fact, the results seem slightly better when the data are generated with the Frank copula, at least for the methods that use KM or CG with a small working τ as the final estimator.

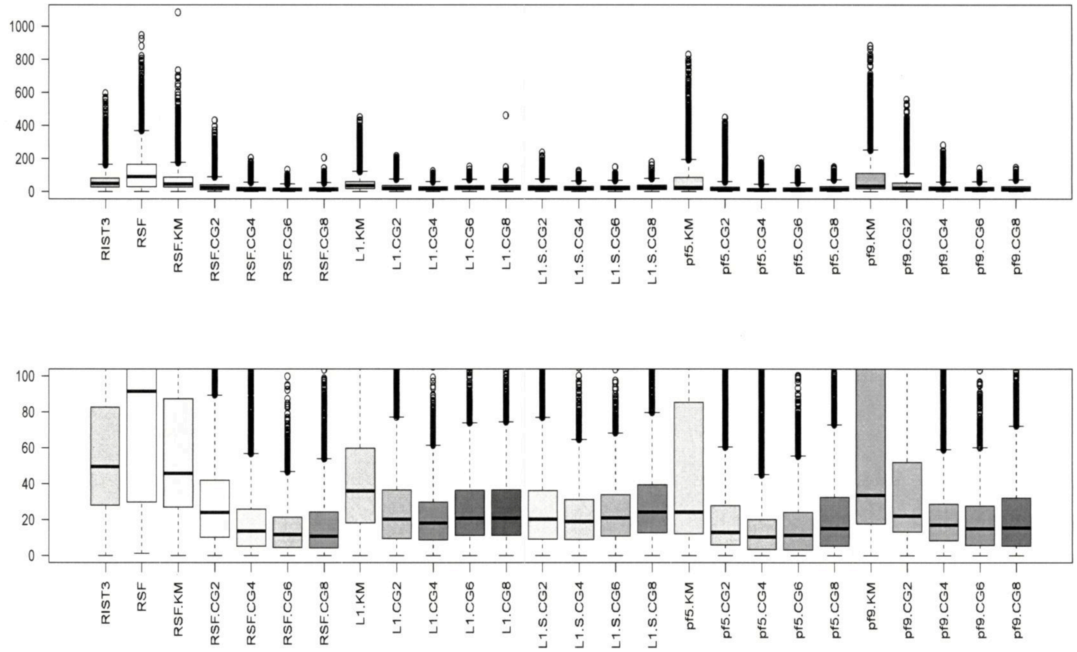

Figure 3.1: % increase in IAE of all methods with respect to the best one over the 4500 runs on the top and a zoom in of the lower part at the bottom

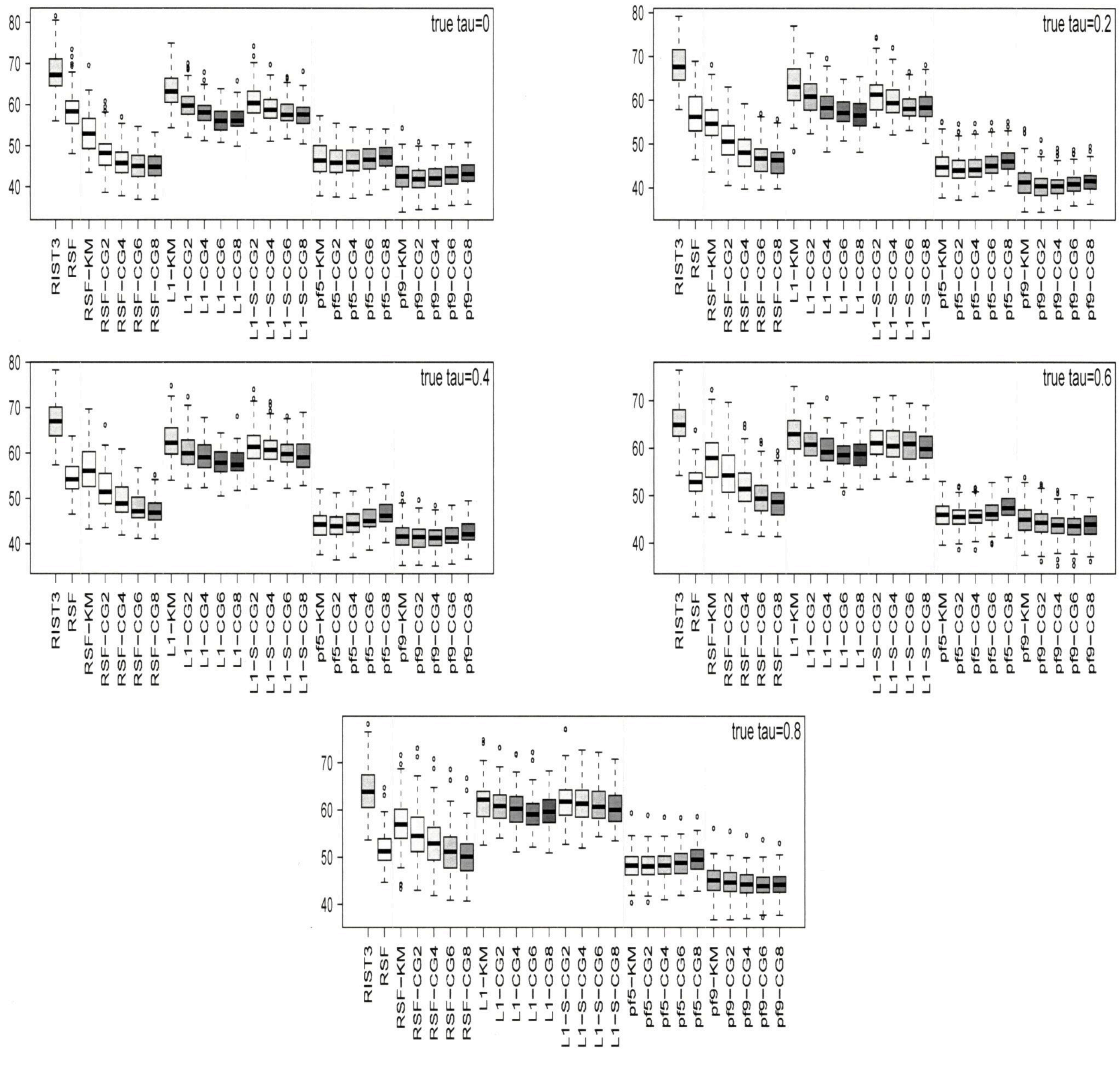

Figure 3.2: IAE of methods for DGP 1 at 20% censoring proportion

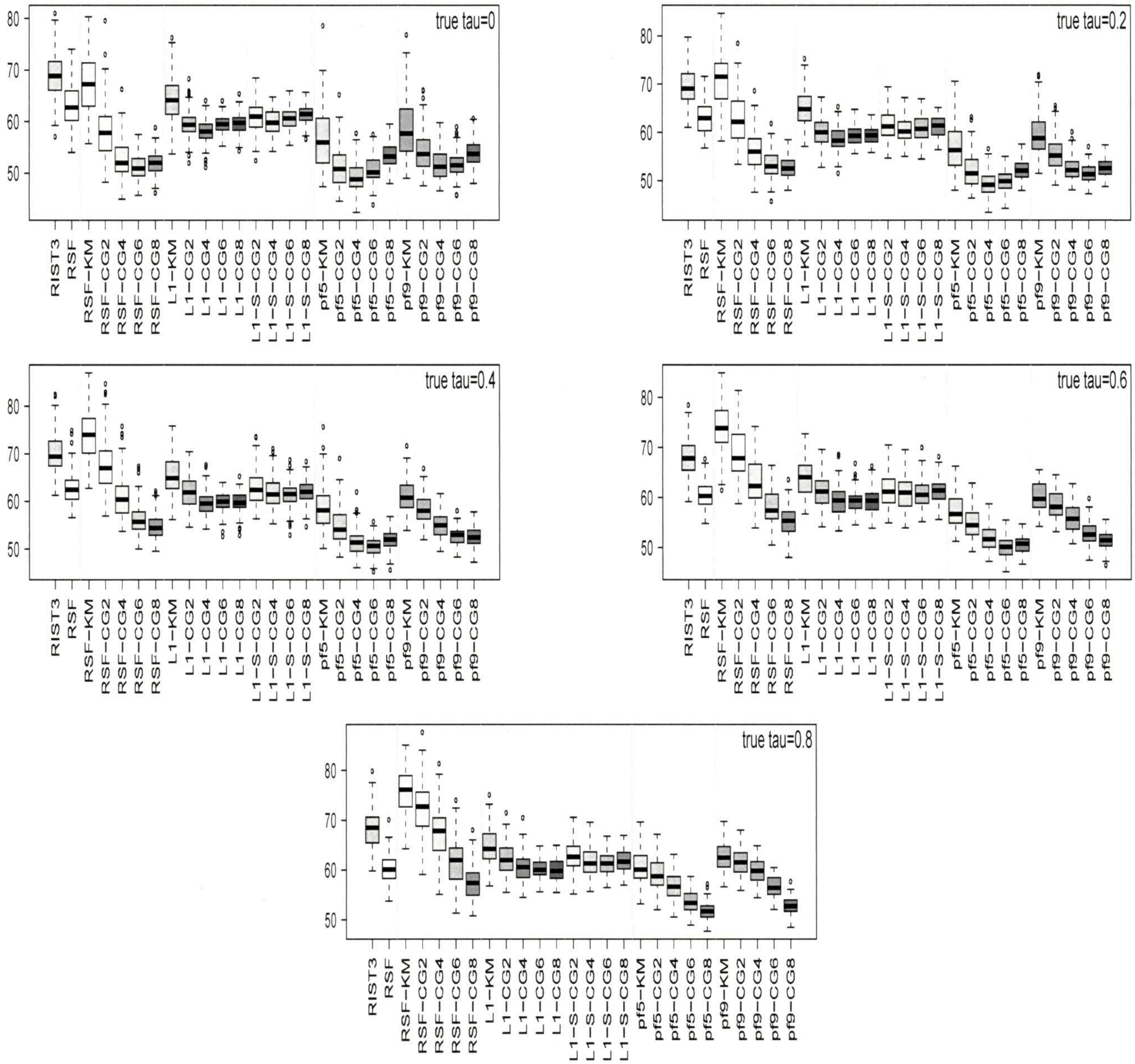

Figure 3.3: IAE of methods for DGP 1 at 40% censoring proportion

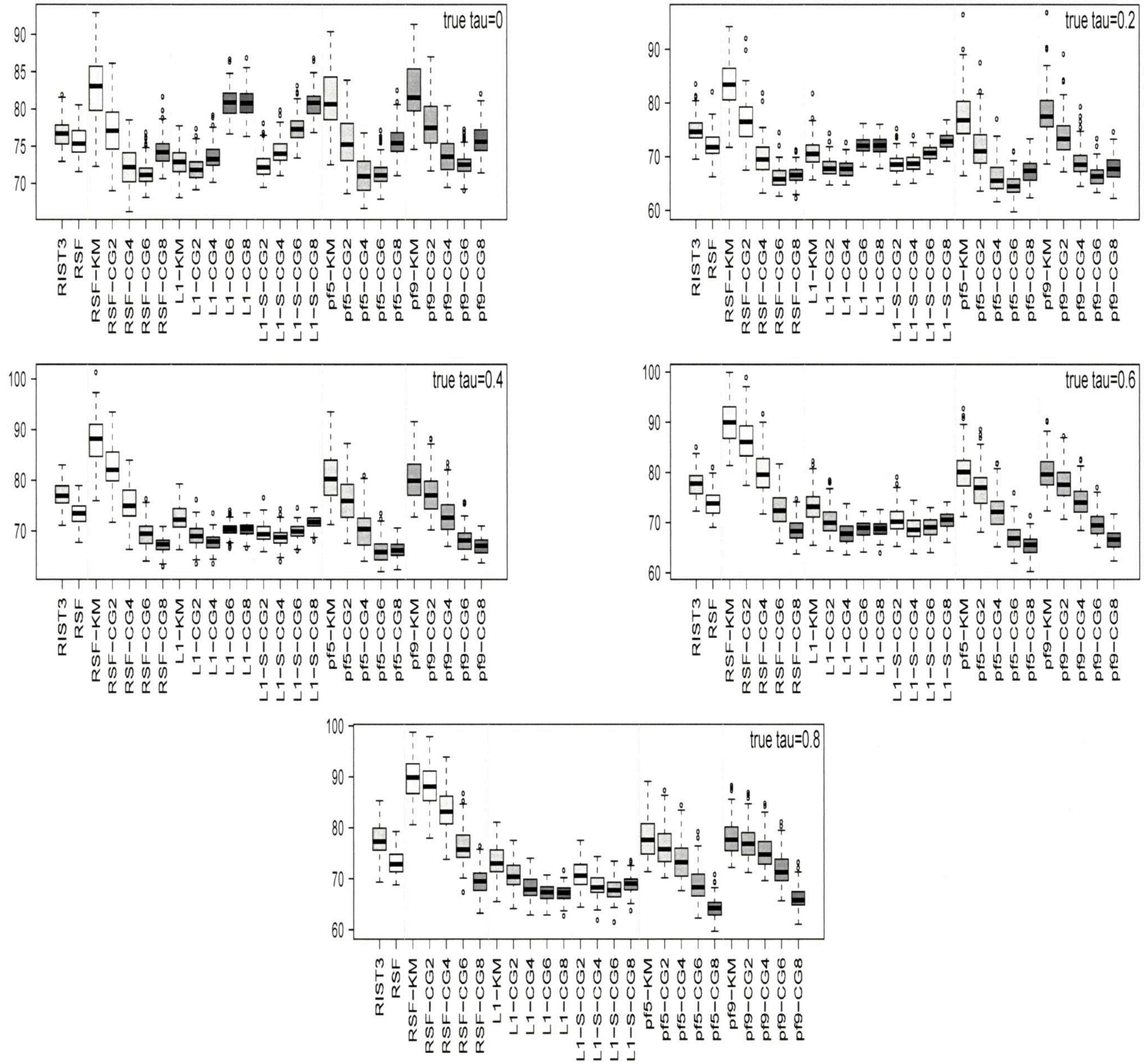

Figure 3.4: IAE of methods for DGP 1 at 60% censoring proportion

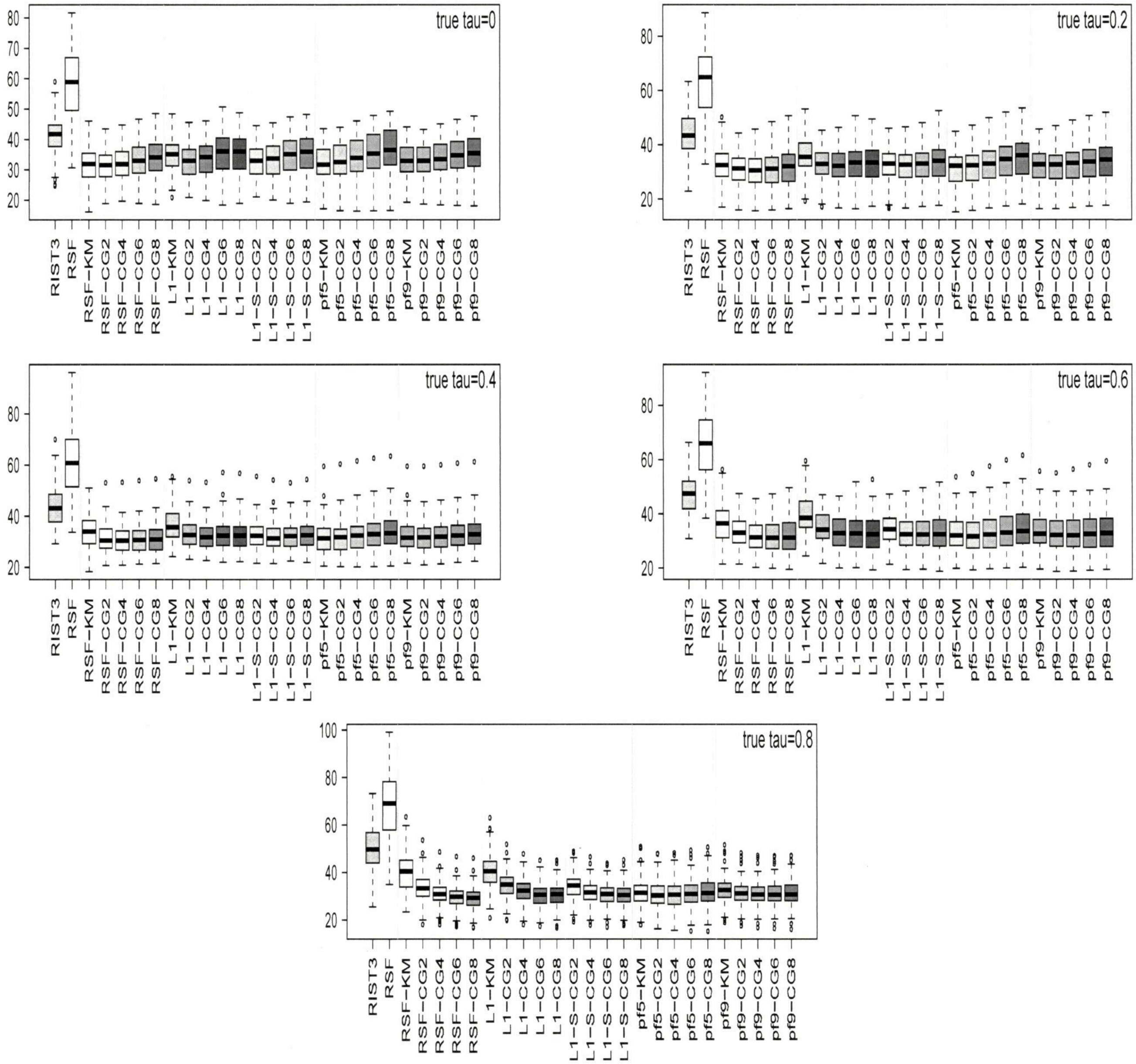

Figure 3.5: IAE of methods for DGP 2 at 20% censoring proportion

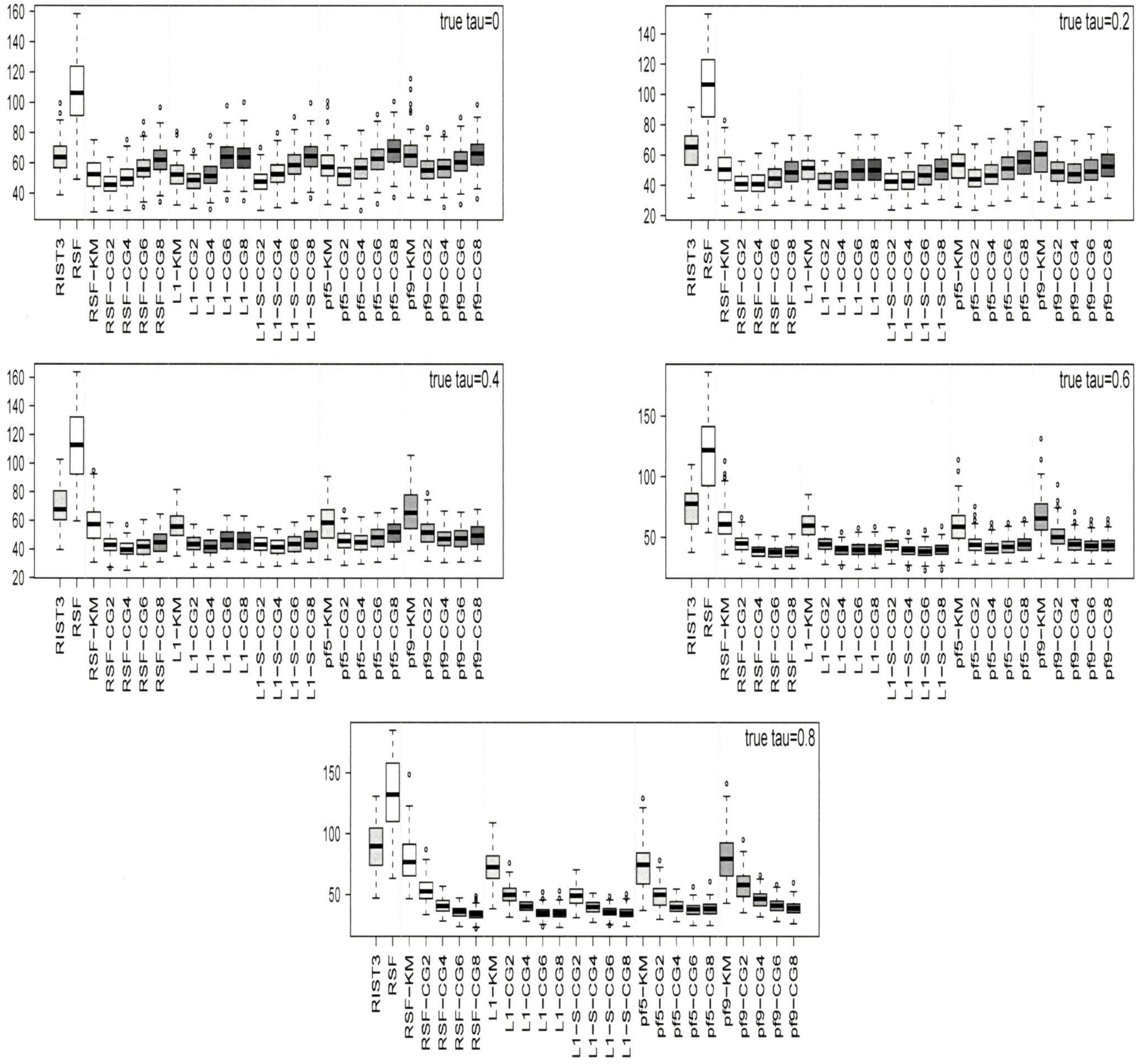

Figure 3.6: IAE of methods for DGP 2 at 40% censoring proportion

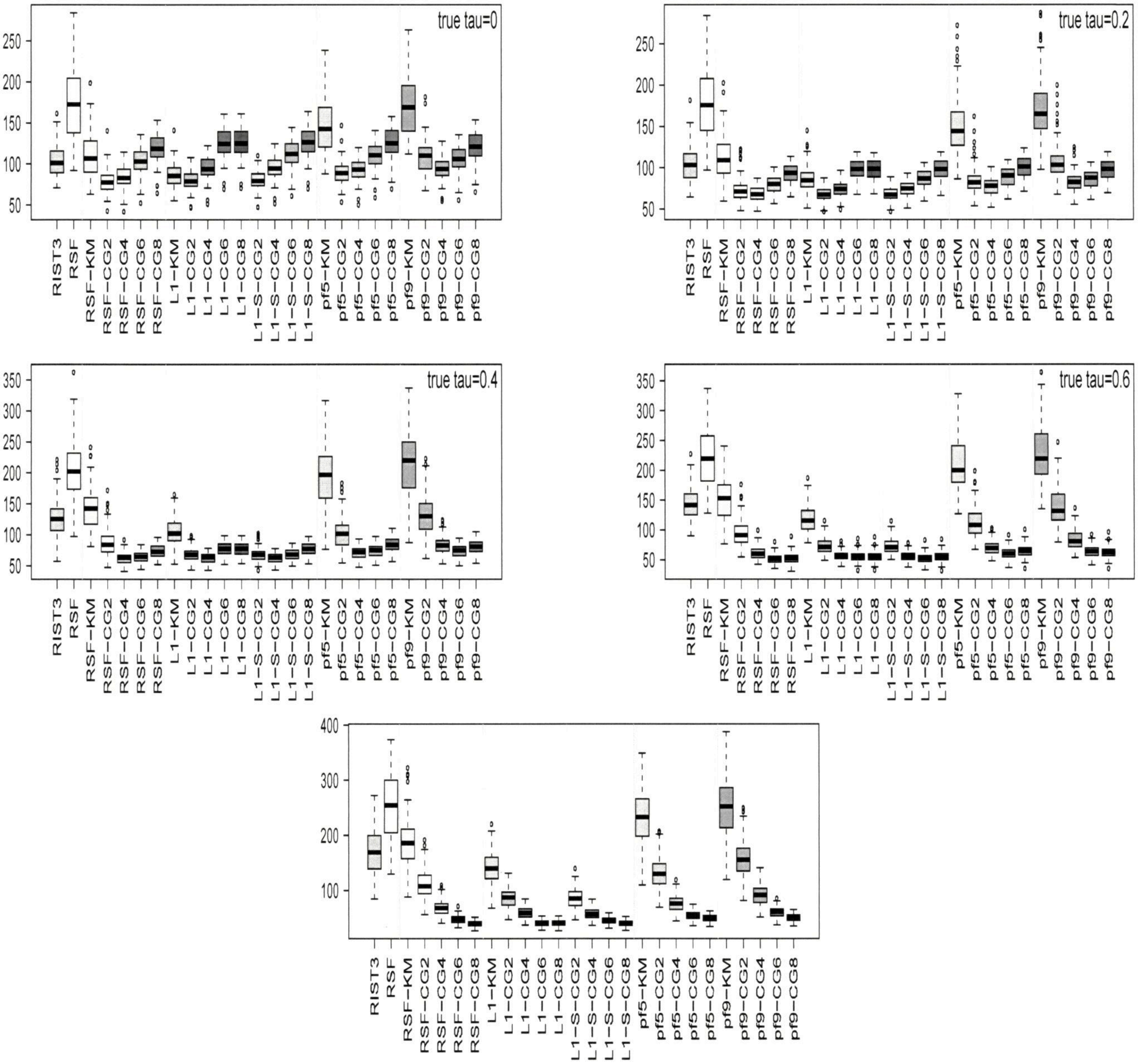

Figure 3.7: IAE of methods for DGP 2 at 60% censoring proportion

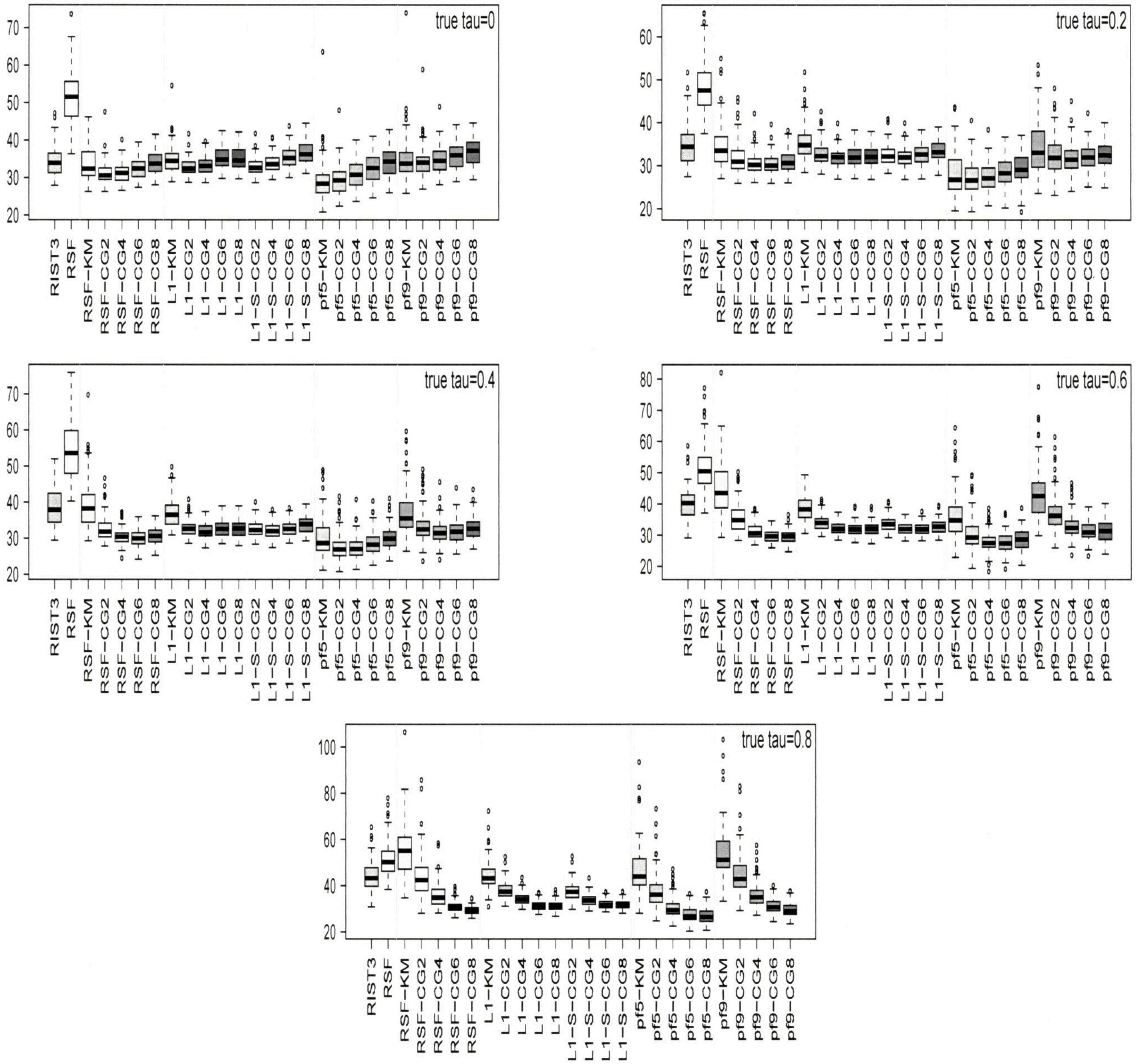

Figure 3.8: IAE of methods for DGP 3 at 20% censoring proportion

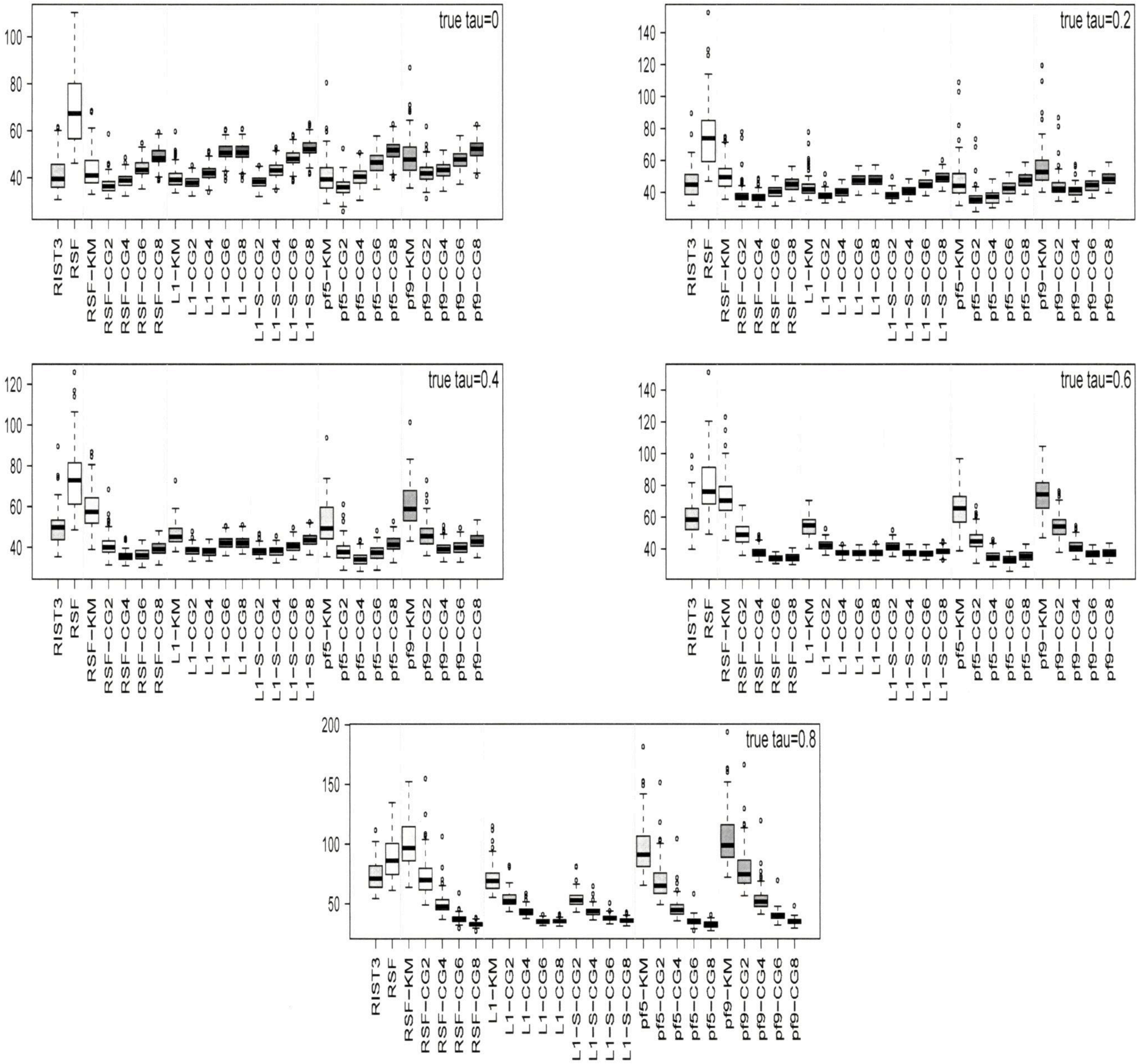

Figure 3.9: IAE of methods for DGP 3 at 40% censoring proportion

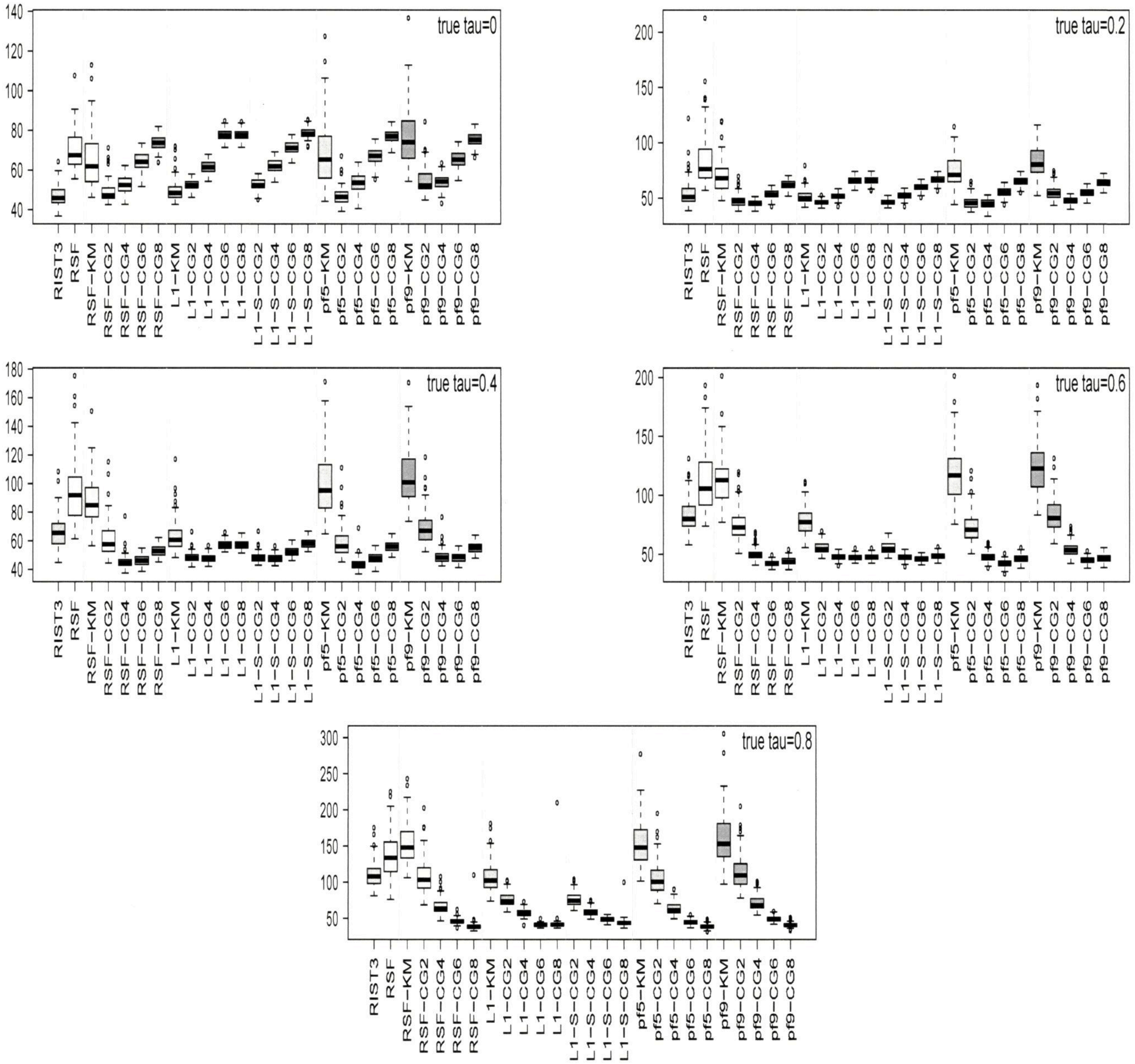

Figure 3.10: IAE of methods for DGP 3 at 60% censoring proportion

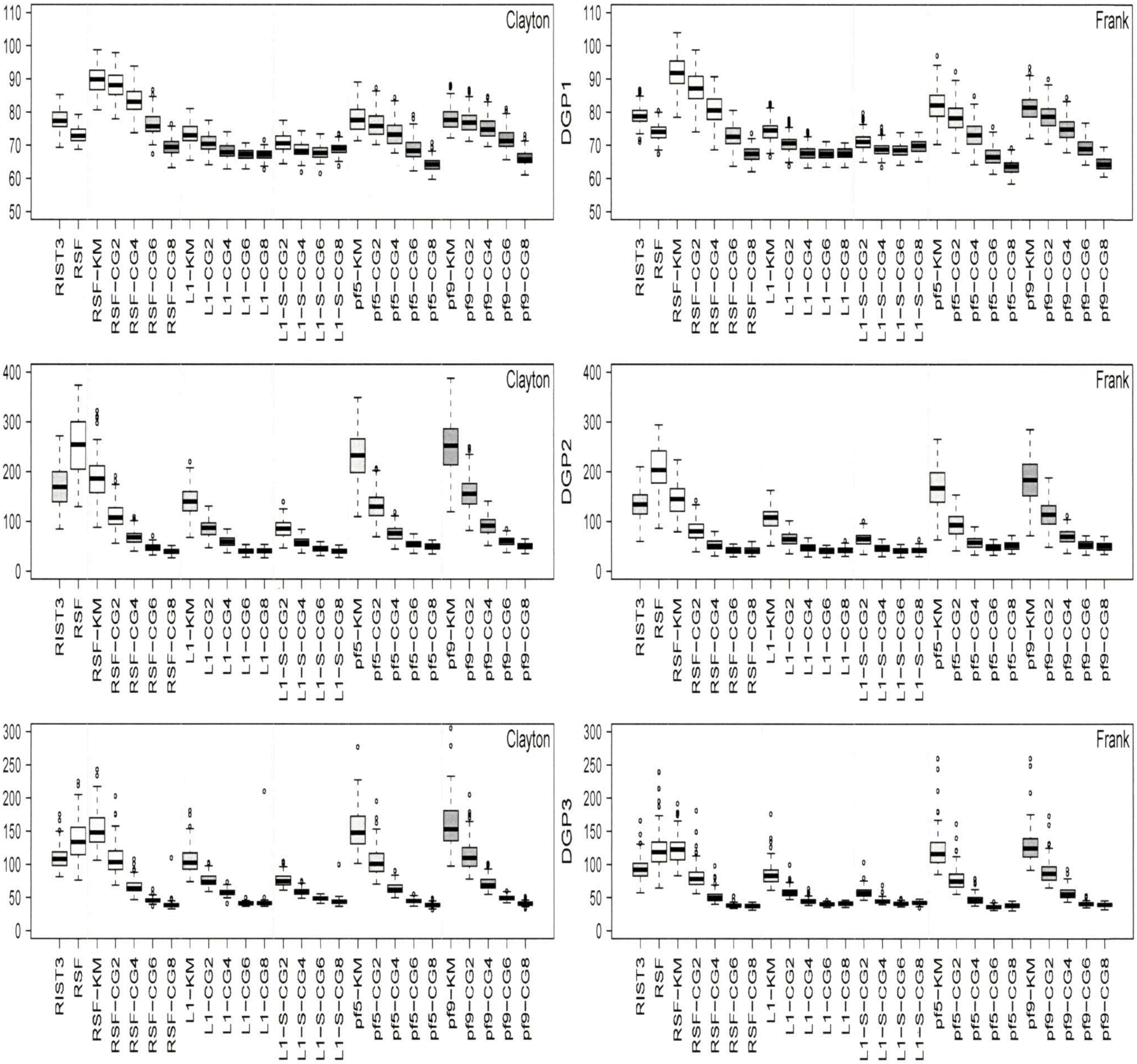

Figure 3.11: Comparisons of IAE results for data generated with the Clayton versus the Frank Copula , Clayton Copula as working copula in both cases, with a censoring proportion of 60% and a true of τ of 0.8

3.5 Discussion and Concluding Remarks

In this paper, we proposed different ways to account for dependent censoring given the covariates when building survival forests. They are: 1) use an appropriate final estimator, and /or 2) modify the splitting rule. We investigated using the copula-graphic instead of the Kaplan-Meier as the final estimator, and also use it in the splitting rule. We also proposed a new way to build survival forests called point-wise forest. The first conclusion is that all these approaches work. However, the main driver of the performance of the methods is using an adequate value of Kendall τ to compute the copula-graphic estimator. The problem is that it is not possible to estimate the dependency parameter without making stronger assumptions on the data (Wang et al., 2015). For example, in some situations, a maximum length of follow-up time is assumed for all subjects. In this case, the censoring time for all subjects is available making the dependency parameter identifiable.

From the limited simulation study presented here, it seems that using the Clayton copula with a value of τ between 0.4 and 0.6 might be a good compromise if dependent censoring is suspected. Indeed, even when the true τ is 0, the best method was most often one that used a value of τ greater than 0. More generally, the best value of τ to use was most often greater or equal than the true value of τ. However, it would be interesting to investigate if it is possible, by adding minimal assumptions, to get a practical estimator of the dependency parameter that would lead to an improved final estimator of the survival function.

One interesting finding concerns the method RSF in the package `randomForestSRC`. By default, the function `rfsrc` in the package returns an estimated cumulative hazard function computed by averaging the Nelson—Aalen cumulative hazard function of each tree. Let $\Lambda(t)$ be this function. Then we can obtain the survival function with $S(t) = \exp(-\Lambda(t))$. Our results show that combining the information from the individual trees by pooling all the observations ending in the terminal nodes of each tree and computing the Kaplan-Meier (or another estimator like the copula-graphic) often produces better results. Note that the same forests are built in both cases. Only the aggregation of the information in the trees is different. It is straightforward to use this combination method with `randomForestSRC`.

Hence we can benefit from all the features available in `randomForestSRC` and possibly get an improved estimation of the survival function. This certainly deserves more investigation and needs to be validated with other simulations studies.

The point-wise forest method introduced in this paper also deserves to be investigated more deeply. The main idea is to use a splitting rule that accounts for dependent censoring without committing to a particular value of a dependency parameter at tree-building time. But our results show that the point-wise forest seems to be a good forest building method in its own right, and not only for the case of dependent censoring. In this paper, the number of points (forests) and number of forests needed for an observation to count in the final estimation were arbitrarily fixed at 9 for the first parameter and 5 or 9 for the second. But it would be interesting to investigate how the performance is sensible to these parameters and if it is possible to get an all around good performance by selecting them automatically with a data-driven procedure.

3.6 Acknowledgement

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), by le Fonds de recherche du Québec Nature et technologies (FRQNT), and by HEC Montréal.

Bibliography

I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics surveys*, 5:44–71, 2011.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and regression trees. wadsworth & brooks. *Monterey, CA*, 1984.

D.G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1): 141–151, 1978.

M.J. Frank. On the simultaneous associativity off (x, y) andx+y- f (x, y). *Aequationes mathematicae*, 19(1):194–226, 1979.

L. Gordon and R.A. Olshen. Tree-structured survival analysis. *Cancer treatment reports*, 69 (10):1065, 1985.

T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M.J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

H. Ishwaran and U.B. Kogalur. Random forests for survival, regression and classification (rf-src). 2014. URL `http://cran.r-project.org/web/packages/randomForestSRC/`. R package version 1.5.5.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

Y. Li, R.C. Tiwari, and S. Guha. Mixture cure survival models with dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):285–306, 2007.

H. Moradian, D. Larocque, and F. Bellavance. L_1 splitting rules in survival forests. *Lifetime Data Analysis*, pages 1–21, 2016. doi: 10.1007/s10985-016-9372-1.

R.B. Nelsen. *An introduction to copulas*, volume 139. Springer Science & Business Media, 1999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, A., 2014. URL `http://www.R-project.org/`.

L.P. Rivest and M.T. Wells. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79(1): 138–155, 2001.

M.R. Segal. Regression trees for censored data. *Biometrics*, 44:35–47, 1988.

M Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

A. Wang, K. Chandra, R. Xu, and J. Sun. The identifiability of dependent competing risks models induced by bivariate frailty models. *Scandinavian Journal of Statistics*, 42(2): 427–437, 2015.

M. Zheng and J.P. Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.

R. Zhu and M.R. Kosorok. Recursively imputed survival trees. *Journal of the American Statistical Association*, 107(497):331–340, 2012.

Conclusion

The three articles of this thesis propose solutions for three of the limitations of current implementations of survival forests. The first limitation is that they use the log-rank test as the splitting rule which only works when the proportionality assumption is met. Due to the loss of power of log-rank test when the proportional hazards assumption is violated, that is the hazard (survival) functions in the two compared groups cross each other, the first article proposes the use of the integrated absolute difference between the two children nodes survival functions as the splitting rule. The forests built with this splitting rule, the L_1-forest, provide competitive results compared to their counterparts even in general circumstances where the proportionality assumption is met. The L_1-forest often got the best performance with simulated data and with real data sets. The second limitation of presently available implementations of random survival forest is that they do not provide dynamic predictions in presence of time-varying covariates. The second article of this thesis studies settings in which time-varying covariates are present. We investigate different ways that random forests can be used for dynamic predictions with discrete-time survival data. The results show that all considered methods perform well and the average of the estimated hazard functions from all methods provides a simple way to get a good result in most circumstances. Lastly, available implementations of survival forest work under the assumption that the event and censoring times are independent, given the covariates. The third article of this thesis investigates different ways that random forests can tackle the problem of dependent censoring. The first one is to use a final estimate of the survival function that corrects for dependent censoring. The second one is to use a splitting rule which does not rely on the independent censoring assumption. The results of simulations show that all methods using the copula-graphic as the final estimator of the survival function have a good performance. So there is no need for a special splitting rule as long as the dependent censoring is taken care of through the final estimator. The worst methods are the ones that do not use the copula-graphic as the final estimator like RSF and RIST3. Another general finding, from the considered scenarios, is the fact that RSF-KM is better than RSF. This means that, to compute the final estimate of the survival function, it is preferable to use the pooling method described in the first article instead of taking the average of the Nelson—Aalen cumulative hazard function that comes out of `randomForestSRC` and transform it. We also proposed a new survival forest technique called point-wise forest or in short p-forest. The main idea is to use a splitting rule that accounts for dependent censoring without relying on a particular value of a dependency parameter at tree-building time. The p-forest deserves to be investigated more deeply since the simulations showed that it is very competitive even in independent censoring scenarios.

This thesis answered many questions about survival forests but also paved the way for potential future research. The two splitting rules investigated in the first article are $L_1 = (n_L n_R)\Delta$ and $L_1^* = \sqrt{n_L n_R}\Delta$ where $\Delta = \int_t |\hat{S}_L(t) - \hat{S}_R(t)|dt$. The factors $(n_L n_R)$ and $\sqrt{n_L n_R}$ can be seen as penalization factors that favor splits with children nodes of nearly equal sizes. There was no clear winner between the two in the simulation study and with the real data sets. It is clear that, for a given data set, the choice could be based on cross-validation with a performance criterion like the Brier score. But it is also possible that the best factor varies from node to node in a tree. Hence, a method that selects the splitting criterion adaptively at each node could entail better performance. The drawback would be a higher computational cost however. The dynamic prediction method in the second article is aimed at a discrete-time survival response. It would be interesting to develop a similar method for a continuous response and the "pseudo-subjects" idea of Bacchetti and Segal(1995) could be adapted to this task. In the third article, given that the main driver of the performance of the methods is using an appropriate value for the working Kendall τ to compute the copula-graphic estimator, one future direction for research would be to find a way to estimate the dependency parameter by adding minimal assumptions. In all this thesis, CART was the main paradigm used to build the trees. Other general approaches are available including the so-called "unbiased" trees (Hothorn et al., 2006; Loh, 2013). Hence, it would be interesting to investigate how to implement the methods proposed in this thesis with these tree-building methods. Finally, the settings of this thesis involved only right-censoring. The general literature on survival trees and forests is almost entirely aimed at this situation. Hence there are many opportunities to develop methods for other types of mechanism like left truncated data and interval censored data.

Bibliography

J.R. Anderson, K.C. Cain, and R.D. Gelber. Analysis of survival by tumor response. *Journal of Clinical Oncology*, 1(11):710–719, 1983.

P. Bacchetti and M.R. Segal. Survival trees with time-dependent covariates: application to estimating changes in the incubation period of aids. *Lifetime data analysis*, 1(1):35–47, 1995.

M. Bertolet, M.M Brooks, and V. Bittner. Tree-based identification of subgroups for time-varying covariate survival data. *Statistical methods in medical research*, page 0962280212460442, 2012.

I. Bou-hamad, D. Larocque, H. Ben-Ameur, L.C Mâsse, F. Vitaro, and R.E Tremblay. Discrete-time survival trees. *Canadian Journal of Statistics*, 37(1):17–32, 2009.

I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. Discrete-time survival trees and forests with time-varying covariates application to bankruptcy data. *Statistical Modelling*, 11(5): 429–446, 2011a.

I. Bou-Hamad, D. Larocque, and H. Ben-Ameur. A review of survival trees. *Statistics surveys*, 5:44–71, 2011b.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and regression trees. wadsworth & brooks. *Monterey, CA*, 1984.

A. Ciampi, J. Thiffault, J.P. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational statistics & data analysis*, 4(3):185–204, 1986.

A. Ciampi, S.A. Hogg, S. McKinney, and J. Thiffault. Recpam: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. i. methods and program features. *Computer Methods and Programs in Biomedicine*, 26(3):239–256, 1988.

A. Cutler and G. Zhao. Pert-perfect random tree ensembles. *Computing Science and Statistics*, 33:490–497, 2001.

E. Elgmati, R.L Fiaccone, R Henderson, and J.NS Matthews. Penalised logistic regression and dynamic prediction for discrete-time recurrent event data. *Lifetime data analysis*, 21 (4):542–560, 2015.

L. Gordon and R.A. Olshen. Tree-structured survival analysis. *Cancer treatment reports*, 69 (10):1065, 1985.

R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480, 2000.

D.W. Hosmer Jr, S. Lemeshow, and S. May. *Applied survival analysis: regression modeling of time to event data*, volume 618. Wiley. com, 2011.

T. Hothorn and B. Lausen. On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis*, 43(2):121–137, 2003.

T. Hothorn, P. Bühlmann, S. Dudoit, A. Molinaro, and M.J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2006.

T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.

X. Huang, S. Chen, and SJ Soong. Piecewise exponential survival trees with time-dependent covariates. *Biometrics*, pages 1420–1433, 1998.

X. Huang, F. Yan, J. Ning, Z. Feng, S. Choi, and J. Cortes. A two-stage approach for dynamic prediction of time-to-event distributions. *Statistics in medicine*, 2016.

H. Ishwaran and U.B. Kogalur. Random forests for survival, regression and classification (rf-src). 2014. URL `http://cran.r-project.org/web/packages/randomForestSRC/`. R package version 1.5.5.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *The Annals of Applied Statistics*, pages 841–860, 2008.

M. Leblanc and J. Crowley. Survival trees by goodness of split. *Journal of the American Statistical Association*, 88(422):457–467, 1993.

X. Lin and H. Wang. A new testing approach for comparing the overall homogeneity of survival curves. *Biometrical Journal*, 46(5):489–496, 2004.

X. Lin and Q. Xu. A new method for the comparison of survival distributions. *Pharmaceutical statistics*, 9(1):67–76, 2010.

Y. Lin and Y. Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

W.Y. Loh. Guide classification and regression trees user manual for version 15. 2013.

W.Y. Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014.

E.B. Madsen, P. Hougaard, and E. Gilpin. Dynamic evaluation of prognosis from time-dependent variables in acute myocardial infarction. *The American journal of cardiology*, 51(10):1579–1583, 1983.

P Mayer, D Larocque, and M Schmid. Dstree: Recursive partitioning for discrete-time survival trees, 2014.

LP Rivest and M.T Wells. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79(1): 138–155, 2001.

D. Rizopoulos. *Joint models for longitudinal and time-to-event data: With applications in R.* CRC Press, 2012.

M. Schmid, H. Küchenhoff, A. Hoerauf, and G. Tutz. A survival tree method for the analysis of discrete event times in clinical and epidemiological studies. *Statistics in Medicine*, 35(5):734–751, 2016. ISSN 1097-0258. doi: 10.1002/sim.6729. URL `http://dx.doi.org/10.1002/sim.6729`. sim.6729.

M.R. Segal. Regression trees for censored data. *Biometrics*, pages 35–47, 1988.

A.A Tsiatis and M. Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

G. Tutz and M. Schmid. Modeling discrete time-to-event data, 2016.

H. van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007.

H. van Houwelingen and H. Putter. *Dynamic prediction in clinical survival analysis.* CRC Press, 2011.

ML Wallace. Time-dependent tree-structured survival analysis with unbiased variable selection through permutation tests. *Statistics in medicine*, 33(27):4790–4804, 2014.

J.B Willett and J.D Singer. Investigating onset, cessation, relapse, and recovery: why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of consulting and clinical psychology*, 61(6):952, 1993.

M. Zheng and J.P Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82(1):127–138, 1995.

Y. Zheng and P.J. Heagerty. Partly conditional survival models for longitudinal data. *Biometrics*, 61(2):379–391, 2005.

R. Zhu and M.R. Kosorok. Recursively imputed survival trees. *Journal of the American Statistical Association*, 107(497):331–340, 2012.