OPTIMAL TREATMENT REGIMES ESTIMATION WITH CENSORED DATA AND RELATED TOPICS

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

SANHITA SENGUPTA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

June, 2021

Acknowledgments

I am grateful to my advisor Lan Wang for her guidance and support throughout the past 5 years. I am extremely fortunate to have Lan as a mentor. Her support, care, motivation and optimism has made this dissertation possible. She taught me to learn even from undesired results and encouraged me in all aspects of my life be it academics or outside of classes.

I would like to extend gratitude to my co-advisor Charles Doss, and my committee, David Vock, Xiaoou Li for their support and valuable suggestions which helped shape my dissertation. I owe my deepest gratitude to Yuhong Yang, Galin Jones, Taryn and the School of Statistics for being supportive, encouraging and providing with a great learning environment. I would also like to thank Gary Oehlert for the great learning experience which he provided as a mentor for my work as a student consultant. I thank PRISM and my collaborators for providing an opportunity to apply my skills and a warm atmosphere to learn from. I would like to offer my special thanks to my students over the years, who helped me grow both as a teacher and a Statistician.

My research and graduate school experience would be incomplete without mentioning my peers and friends, Riddhiman, Ziyue, Sarah, James, Haoran, Wenjing, Sakshi, Haema, Shannon, Yu Zhou, Yunan, Zhou, Martin, Georgia, Yuyoung who have supported me all along these past 5 years. Your feedback and support has been crucial in shaping my graduate life at the University of Minnesota. I would also like to thank my wonderful mathematician friends, Ipsita, Subham, Pinaki, Manasa whose support was undeterred by long distance or time difference throughout my academic journey for the past 10 years. I am also indebted to Diksha, Kerry, Bhavtosh, Purba, Kajari, Jash, Raghavendra, Saroni, Animesh, Koushiki, Dipnil, Manik, Sumitra, Dhanraj, Neha, Anita,

Nihita, Ari, Hitesh, Naveen, Deepa, Jackie, Anvita, Akshay, Anjali, Amanda, Gautham, Pratik whose constant care and encouragement has kept me motivated through ups and downs. I wish to extend special regards to SAATH for keeping the creativity in me alive and for being the home away from home.

Lastly I cannot imagine any of this without my parents and family who have cheered me on and been there to fall back on if needed. They have been crucial at every step of my life and career, never letting a stone unturned to ensure that I could pursue my interests and dreams.

Dedication

To my parents and grandparents.

Abstract

The thesis is divided in three sections of interconnected topics.

Motivated by applications from precision medicine, we consider the problem of estimating an optimal treatment regime (or individual optimal decision rule) based on right-censored survival data. We consider a non-parametric approach that maximizes the expected mean restricted survival time of the potential outcome distribution. Comparing with existing methods, our approach does not need to assume the decision rule belongs to a restricted class (e.g., class of index rules) and can accommodate high-dimensional covariates. We investigate the theory of the estimated optimal treatment regime. Monte Carlo studies and a real data example are used to demonstrate the performance of our proposed method.

Random forests are widely used today for various purposes such as regression classification, survival analysis however its theoretical properties are not yet explored completely. We propose a quantile random forest estimator which considers sub-sampling instead of complete bootstrap samples as in Meinshausen [2006]. We study the point wise asymptotics of quantile random forest estimator proposed by rendering it in the framework of U-statistics. We prove point-wise weak convergence to normality and also propose a consistent estimator of the variance. We further explore the asymptotic behavior of the proposed estimator via a simulation study.

Measuring the efficacy of a treatment or policy can involve data heterogeneity. In such cases, the entire conditional distributional impact of the treatment is important rather than just a discrete metric such as the average treatment effect. Quantiles inform more about the distribution than an average and multiple quantiles can be used together to get an idea about the entire distribution. In the context of survival analysis with censored data, we propose a

quantile regression model estimated using survival random forest. We further extend this to estimate quantile treatment effects under censoring. We show the efficacy of the proposed method via simulations. We also demonstrate using this method and interpreting quantile effect by analysing a colon cancer dataset.

Contents

1 Introduction **1**

1.1 Background 2

1.2 Overview 4

2 Optimal Treatment Regime to Maximize Restricted Mean Survival Time **7**

2.1 Introduction 7

2.2 Motivation 10

2.3 Background and Notations 12

2.3.1 Optimal Treatment Regime 13

2.3.2 Methodology 15

2.3.3 Properties of the Optimal Regime 16

2.4 Asymptotic Properties of the Optimal Regime Estimate 17

2.4.1 Notations and Assumptions 18

2.4.2 Consistency of Survival Function Estimate 20

2.4.3 Consistency of Optimal Regime 21

2.5 Simulations 21

2.5.1 Simulation Setting 22

2.5.2 Simulation Results 23

2.6 Real data analysis . . . 25
2.6.1 Estimates for Data Analysis . . . 25
2.6.2 Lung Cancer Data Analysis . . . 26
2.6.3 Colon Cancer Data Analysis . . . 27
2.7 Discussion . . . 29

3 Quantile Random Forest **31**
3.1 Introduction . . . 32
3.2 Review: U-statistic and Random Forest . . . 34
3.2.1 Decision Trees . . . 34
3.2.2 Random Forest . . . 36
3.2.3 Quantile Random Forest . . . 37
3.2.4 U-Statistic . . . 37
3.2.5 Subbagging . . . 38
3.3 Theoretical Framework . . . 39
3.3.1 Estimate of Conditional Cumulative Distribution Function . 40
3.3.2 Incomplete U-Statistic . . . 41
3.4 Asymptotic Properties . . . 42
3.4.1 Asymptotic Normality . . . 42
3.4.2 Consistent Estimator of Variance ζ_{1,k_n} . . . 48
3.5 Simulations . . . 50
3.5.1 Asymptotic Normality . . . 50
3.6 Discussion . . . 52
3.6.1 Alternate Method : Honest Trees . . . 52
3.6.2 Confidence Intervals and Tests of Significance . . . 55

4 Censored Quantile Regression **57**

4.1 Introduction . . . 58
4.1.1 Background and Notations . . . 60
4.2 Methodology . . . 62
4.3 Asymptotic Properties . . . 65
4.3.1 Consistency . . . 66
4.4 Quantile Treatment Effects under Censoring . . . 71
4.4.1 Preliminaries . . . 71
4.4.2 Methodology . . . 73
4.4.3 Consistency . . . 74
4.5 Simulations . . . 75
4.5.1 Simulations for Censored Quantile Regression . . . 75
4.5.2 Simulations for Quantile Treatment effect . . . 80
4.6 Real Data Analysis . . . 87
4.6.1 Colon Cancer Data Analysis . . . 87
4.7 Discussion . . . 93

5 Discussion and Areas for Future Work **95**
5.1 Conclusions . . . 96
5.2 Future Work . . . 97

References **98**

A Notations for Chapters 2 and 4 **110**

B Notations for Chapter 3 **112**

C Simulation Setting : Chapter 2 **113**

D **114**

D.1 Proof of Proposition 2.3.3 114

D.2 Proof of Theorem 2.3.4 115

E **118**

E.1 Proof of Proposition 2.4.8 118

E.2 Proof of Theorem 2.4.9 122

E.3 Proof of Theorem 2.4.10 123

F **127**

F.1 Proof of Theorem 4.4.1 127

List of Figures

3.1 Decision Tree partitions for Titanic Survival Data 35
3.2 Quantile Random Forest : 50% Quantile 51
3.3 Quantile Random Forest : 90% Quantile 51

4.1 Colon Cancer Survival : Treatment Groups 88
4.2 Colon Cancer Survival curves 89

List of Tables

2.1 Comparison of IPWE, tree and random forest estimates of optimal treatment regime in complete data case 11
2.2 Simulation Results 24
2.3 Lung cancer data analysis results 27
2.4 Colon cancer data analysis results 29

3.1 Model 1 : Comparison of GRF, QuantregForest, Honest forest . . . 54
3.2 Model 2 : Comparison of GRF, QuantregForest, Honest forest . . . 54

4.1 Setting 1 : Bias with $\tau = 0.5$ and $n = 1000$ 78
4.2 Setting 1 : Bias with $\tau = 0.7$ and $n = 1000$ 79
4.3 Censored Quantile Regression : Setting 1 80
4.4 Setting 2 : Bias with $\tau = 0.5$ and $n = 1000$ 81
4.5 Setting 2 : Bias with $\tau = 0.7$ and $n = 1000$ 82
4.6 Censored Quantile Regression : Setting 2 83
4.7 Setting 3 : Bias with $\tau = 0.5$ and $n = 1000$ 84
4.8 Setting 3 : Bias with $\tau = 0.7$ and $n = 1000$ 85
4.9 Censored Quantile Regression : Setting 3 86
4.10 Colon Cancer : Censored Quantile Regression 90

Chapter 1

Introduction

US President's Council of Advisors on Science and Technology 2008 report "Priorities for personalized medicine" defines : "*Precision Medicine refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology or prognosis of those diseases they may develop, or in their response to a specific treatment...*"

A personal one-on-one basis interaction with a doctor is the simplest form of personalized care. A diagnosis based on a patient's personal medical history, current vitals and family background would be much more effective than if there was to be a standard "One size fits all" care for all suffering from a specific ailment. Patient diversity may exist even within a certain target risk group of a disease, hence requiring the need for personalized approaches. Parallel to this but on a very different note, machine learning and AI are extensively used by companies today to improve consumer experience and make personalized recommendations/advertisements by real-time tracking and utilizing browsing

history, landing sites and interaction on the web.

1.1 Background

An inherent problem in various fields from E-commerce to healthcare is to identify optimal policy/treatment regime which maximizes an endpoint outcome. This necessitates identifying the causal effect of various treatments/policies assignments and comparing the possible outcomes from each. A possible way to assess treatments is to compare their effect across the entire population(such as average treatment effect). But different baseline characteristics of a patient or customer creates heterogeneity in their response to different treatments/policies. Hence a treatment/policy might be beneficial to only a subset of the population and assigning the same treatment to all on the basis of the population response may not be optimal. This substantiates the need of a personalized approach for inference.

When all possible treatment assignments cannot be tested on the same individual, especially in the case of single stage treatment assignment settings for example in healthcare and behavioral sciences, it gives rise to the potential outcome framework in the context of causal inference (discussed in Rubin [1974],Rubin [1977],Holland [1986] for both randomized studies and observational data). Potential outcome framework is used to address the problem of the counterfactual that we cannot observe all intervention outcomes at individual level. Although conducting a randomized experiment would be ideal to estimate the causal effect of treatments, intervening randomly is unfeasible in most situations(for example for diagnosing serious patients). Observational studies introduce a level of complication of confounding. Confounders are factors which can affect both the treatment assignment and the outcome hence tangling up the

causal effect estimate. Most literature(Rubin [1974],Holland [1986]) of causal inference aim to estimate the average causal effect of given treatments A on an observable outcome Y and set of confounders X. (*Note :* The term "treatment" throughout this document can refer to any policy/treatment/intervention/exposure assignment , not necessarily in the context of medicine but also psychology, economics, etc wherever such a setting can arise) Average treatment effects have been estimated and studied under a diverse set of possible data settings and confounding structures. As mentioned before randomized trials/experiments are the ideal cases, however confounding bias may arise in observational studies. We assume identifiability, conditional exchangeability and no unmeasured confounding in our causal framework. Methods such as difference in differences, matching, propensity score matching aim to mimic the randomized setting case, further model based approaches such as g-estimation and marginal structural models address confounding in different situations.

Instead of observing just an endpoint outcome, the lifetime of an individual might be of interest when selecting a treatment regime especially when making a sensitive choice of surgery say for a rare disease. Hence survival time of a patient after intervention is a natural outcome of interest. Survival analysis is a popular domain of Statistics which explores time up to an event of interest (which may be death). However following a subject for an entire lifetime is not feasible giving rise to the issue of censored data. In the case of censored observations we can no longer directly model the potential outcome by simple regression models. To find optimal treatment regime in such cases we need to address censoring and hence look beyond the usual causal inference in observed outcome setting.

Parametric approaches which have data assumptions such as Gaussian error or linearity limit the space of optimization, that being the key goal in personalized

approaches. Various machine learning algorithms such as support vector machines, decision trees, etc do not have such assumptions. Random forest, a supervised machine learning algorithm first introduced by Breiman [2001] is an ensemble of decision trees. Random forests are popular to address non-linearity and interaction effects, also returning variable importance measures to accommodate its lack of interpretability. Random forests can be used for classification, regression, missing data imputation, survival analysis , the algorithms of few of which have been unified into the R-package randomForestSRC Ishwaran et al. [2008].

1.2 Overview

Patient level heterogeneity impacts the entire distribution of their response to various treatments. In essence this affects more than just the mean or average conditional response. Doksum [1974] derive a complete treatment effect function using 2 marginal outcome distributions as an alternative to constant unconditional mean treatment effect. In chapter 2 we propose a method using survival random forests, to estimate treatment regimes which maximize restricted mean survival time of a patient in a single stage treatment assignment setting. This aims at maximizing patient's survival up to a clinically important time in the future. The statistic, restricted mean survival might be more consequential than other metrics such as survival probability at a particular point of time in the future or mean/median survival or hazard ratios while assigning treatments. A treatment regime for a single stage decision setting is a treatment assignment function dependent on covariates $g : X \to A$ where X are the covariates observed before intervention. If we assume that regimes are linear in X for example $\mathbb{I}\{\beta X > 0\}$ then maximizing within such a class of regimes Zhang et al. [2012] restricts the maximization and will yield a local maxima and not a global

maxima. Zhang et al. [2012] estimate the optimal treatment regime in the setting of a single stage decision; regimes which maximize observed outcome Y. But they search within a class of parametric regimes. Furthermore, the parametric form needs to be correctly specified, a misspecified regime may perform poorly. We use random forests to address these 2 drawbacks, but in the framework of survival times. We use simulations to demonstrate the added benefits of a random forest method when paramteric models can be misspecified. We explore few real datasets using the restricted survival metric at varying restrictions, which help uncover treatment effects over time.

Quantiles are known to assess the distributional impact of covariates better than an average especially for the cases of unknown and heteroskedastic error distributions. Quantiles in the context of regression had been first introduced in Koenker and Bassett [1978]. Powell [1984] proposed least absolute deviations estimation of parameters, replacing the least square estimator in the context of regression median when the dependent variable is censored. Meinshausen [2006] shows that random forests can inform of the entire conditional distribution of the outcome given covariates, and hence even quantiles. In chapter 3 we explore the asymptotic distribution of the quantile random forest, attempting to prove weak convergence.

Survival time data is often censored, we consider right censored survival data in a heterogeneous setting. Censoring becomes an added complication to quantile regression when we are interested in modeling conditional quantiles of survival times. Censored quantile regression is an alternative to the proportional hazards constraint when modeling heterogeneity in data. Koenker and Geling [2001] introduced quantile regression in survival analysis. In chapter 4, we extend the semiparametric censored regression model Wang and Wang [2009] in a potential high dimensional setting using survival random forests. Doksum [1974] and

Lehmann and D'Abrera [1998] introduce the idea of quantile treatment effect as the horizontal distance between 2 marginal cumulative outcome distributions. Quantile treatment effect(QTE) at various quantiles help explore beyond just the average tyreatment effect (ATE). We then use the regression framework proposed to estimate quantile treatment effects in a single stage treatment assignment setting. We use Monte Carlo simulations to exhibit the properties of the proposed method in case of error heterogeneity, and compare with Wang and Wang [2009] and Portnoy [2003]. We analyse the colon cancer dataset (Laurie et al. [1989]) to explore the distributional effects of the treatments on survival using quantile treatment effect estimates at various quantiles.

Chapter 2

Optimal Treatment Regime to Maximize Restricted Mean Survival Time

2.1 Introduction

It is a common problem to identify the causal effect of various treatment assignments on a specific outcome and compare. A possible way to assess treatments is to compare their effect across the entire population. But different baseline characteristics of a patient creates heterogeneity in their respones to different treatments. Hence a treatment might be beneficial to only a subset of the population and assigning the same treatment to all on the basis of the average treatment effect may not be optimal. This substantiates the need of a personalized approach for inference. The effect of a treatment on a patient depends on their baseline covariates.

Recent studies have explored various methods to find optimal treatment

regimes in case of both observational studies and clinical trials. Research has been done in both single stage treatment setting and sequential or dynamic treatment setting(e.g. Robins, Hernan, Brumback (2000), Murphy (2003)). The methods mainly involve regression model for the potential outcome to adjust for baseline covariates and treatment assignment or a propensity score model of the treatment assignment. But these models are highly dependent on being correctly specified.

Zhang et al. [2012] address the problem of estimating the optimal treatment regime in the setting of a single stage decision; regimes which maximize observed outcome Y. They compare regression approach, inverse probability weighting and an augmented form. In the case of uncensored or complete data when an observable outcome is to be maximized, this paper searches within a class of regimes. If the parametric form of the regime is unknown then a misspecified regime may be sub-optimal and hence perform poorly. A non-parametric approach such as a tree may address the problem of a misspecification and global optimality but a tree is known to have high variability. Especially in personalized medicine, when treatment assignment/regime is the final aim, an algorithm with highly variable regime is unfavourable. We consider random forests to approach these drawbacks, especially in the case of high dimensions.

Random forest, a supervised machine learning algorithm first introduced by Breiman [2001] is an ensemble of decision trees. Random forests are popular to address non-linearity and interaction effects, also returning variable importance measures to accommodate its lack of interpretability. Random forests can be used for classification, regression, missing data imputation, survival analysis , the algorithms of few of which have been unified into the R-package randomForestSRC.

In the case of censored observations we can no longer directly model the

potential outcome by simple regression models. To find optimal treatment regime in such cases we need to address censoring. Zhao et al. [2011] propose a Q-learning approach using support vector regression which can deal with censored data. They also consider multiple lines of treatment and estimate the best initial time for second-line treatment while taking into account the heterogeneity across patients, but they have not derived any theoretical properties. Goldberg and Kosorok [2012] extend this further to allow for flexible number of stages of treatment assignments as suited to patients' characteristics in order to maximize survival times.

We can also handle censoring via a missing data setting using propensity score of censoring. Zhao et al. [2014] develop a robust method for optimal individualized treatment regime which maximize mean survival for right censored survival data using an outcome weighted learning approach. Their proposed optimal rule is consistent whenever either the censoring model or the survival model can be misspecified, they also establish the theoretical properties such as fisher consistency.

Survival analysis is an important branch of statistics in which the data can be censored with the most common form being right-censored data. In that setting for a single stage decision, we look at treatment regimes which maximize restricted mean survival time of a patient. This aims at maximizing patient's survival up to a clinically important time in the future. The statistic, restricted mean survival might be more consequential than other metrics such as survival probability at a particular point of time in the future or mean/median survival or hazard ratios while assigning treatments.

We propose using random forest in a censored data setting so as to not restrict the class of regimes and derive a few properties of the proposed estimator. Survival random forest is a random forest method for the analysis of right-censored survival

data introduced by Ishwaran et al. [2008]. We propose an algorithm to find the optimal treatment regime which maximizes restricted mean survival time. To establish the consistency properties we consider a random splitting rule since data based splitting rules such as log-rank could be biased. Cui et al. [2017] show that data based splitting rules can falsely select a noise variable or ignore an important variable. We move on to establish consistency and providing a consistency rate under a given set of assumptions.

2.2 Motivation

If we assume that regimes are linear in X for example $\mathbb{I}\{\beta X > 0\}$ then maximizing within such a class of regimes [Zhang et al., 2012] restricts the maximization and will yield a local maxima and not a global maxima. For example, Zhang et al. [2012] search within a class of regimes. Moreover, the parametric form needs to be correctly specified, a misspecified regime may perform poorly. A non-parametric approach such as a tree may address the problem of a misspecification and global optimality but a tree is known to have high variability. Especially in personalized medicine, when treatment assignment/regime is the final aim, an algorithm with highly variable regime is unfavourable. Hence to overcome the high variance, we propose to use a random forest approach. We compare the parametric approach in Zhang et al. [2012] with trees and random forest.

We use the setting and notations as in [Zhang et al., 2012]. We first estimate propensity score $\pi(x) = \mathbb{P}[A = 1|X = x]$ as parametric model $\pi(X, \gamma)$. We assume in this case that the optimal regime belongs to the class of regimes which are of the form $g_\eta(x) = \mathbb{I}(\eta^T x > 0)$. For estimation at any covariate x, we find η^{opt} which maximizes $IPWE(\eta) = \hat{\mathbb{E}}[Y^*(g_\eta)]$.

In an outcome regression setup for example in the case of tree or random

forest, we estimate $\mathbb{E}[Y|A = a, X] = \mathbb{E}[Y^*(a)|X]$ for each treatment $a \in \{0, 1\}$. For any patient with covariates x_0 we get the estimated optimal regime to be $\hat{g}^{\text{opt}}(x_0) = \mathbb{I}(\hat{\mathbb{E}}[Y^*(1)|X = x_0] > \hat{\mathbb{E}}[Y^*(0)|X = x_0])$ for all $x_0 \in \mathcal{R}^p$ where p is the total number of covariates.

2.2.0.1 Simulation results

We take the sample size, n to be 500 and try various covariate dimensions $p = 2, 10, 20$. The results are from 100 simulation runs, with test size 10^4 in each run. The simulation setup is described in Appendix C.

Table 2.1: Comparison of IPWE, tree and random forest estimates of optimal treatment regime in complete data case

	# Noise terms	IPWE	Tree	Random Forest	True value
$\mathbb{E}[Y^*(g^{opt})]$	0	12.444(2.988)	15.84(0.86)	17.06 (0.46)	17.2
Percentage	0	0.525(0.1337)	0.515(0.051)	0.7(0.0278)	1
Weighted Error	0	4.73(2.99)	1.337(0.756)	0.116(0.0315)	0
$\mathbb{E}[Y^*(g^{opt})]$	8	11.62(1.23)	16.125(0.92)	17(0.52)	17.2
Percentage	8	0.53(0.07)	0.536(0.044)	0.597(0.008)	1
Weighted Error	8	5.64(1.187)	1.13(0.725)	0.25(0.009)	0
$\mathbb{E}[Y^*(g^{opt})]$	18	11.24(0.69)	16.32(0.7)	17.04(0.45)	17.2
Percentage	18	0.515 (0.047)	0.549(0.048)	0.59(0.005)	1
Weighted Error	18	6.05(0.64)	0.97(0.53)	0.26(0.006)	0

The value function that is the mean outcome from the true optimal regime is 17.2. The mean outcome under any regime can calculated using simulation since we know the true generating distributions. We can see that the mean outcome if the optimal treatment regime is estimated using a random forest is closest to 17.2.

Percentage of times the optimal regime estimated matches the true one is also better using a random forest $(60 - 70\%)$ as compared to 50% in case of IPWE. Instead an prediction error we consider the weighted error as the sum of outcome

values when regime does not match the true regime, this sums up the outcome value lost due to wrong treatment assignment since we aim at maximizing mean outcome. The weighted error for tree and random forest is also much lower than IPWE. The variance of the tree method is more than that of random forest. As the dimension of the covariate space/noise increases the performance of the IPWE method deteriorates but the both the tree and the random forest method give consistent estimates.

Summarizing the simulation results, we can see that a misspecified parametric model can be outperformed by non-parametric methods such as tree or random forest, the latter being preferred in terms of variance and stable predictions.

2.3 Background and Notations

We use the standard setting and assumptions used in survival analysis, the observed data $(X_i, Y_i, \delta_i)_{1 \leq i \leq n}$ are i.i.d. samples of covariates (X), observed survival time/time to event (Y) and the censoring indicator (δ). Denoting T_i as the true survival time (T) and C_i to be the censoring time (C), the observed survival time $Y_i = \min(T_i, C_i)$, and $\delta_i = 1(T_i \leq C_i)$ for $1 \leq i \leq n$. Further, we assume that $S(t|X)$ is the conditional survival function, $S(t|X) = P(t > t|X)$, $f(.|X)$ is the conditional density of $T|X$, $\lambda(.|X)$ the conditional hazard function, and $\Lambda(.|X)$ the conditional cumulative hazard function.

In a single stage decision setting we consider a finite discrete set of K possible treatments, $\mathcal{A} = \{0, \ldots, K-1\}$. We consider the potential survival time to be $T^*(a)$ and potential censoring time to be $C^*(a)$ under any treatment $a \in \mathcal{A}$. The usual assumptions of causal inference now in survival analysis setting are postulated. The assumption of consistency translates to $T = \sum_{a \in \mathcal{A}} \mathbb{I}(A = a) T^*(a)$ and $C = \sum_{a \in \mathcal{A}} \mathbb{I}(A = a) C^*(a)$. The assumption of

unmeasured confounders in this case is that $\{T^*(a) : a \in \mathcal{A}\} \perp A|X$. We also consider non-informative censoring such that $\{C^*(a) : a \in \mathcal{A}\} \perp \{T^*(a) : a \in \mathcal{A}\}|X, A$.

The restricted survival time is the survival time up to a fixed follow up time ($\boldsymbol{\tau}$). Denoting restricted survival time as $\widetilde{\mathbf{T}} = \min(T, \boldsymbol{\tau})$, we get the restricted mean survival time to be $\mathbb{E}[\widetilde{\mathbf{T}}] = \mathbb{E}[\min(T, \boldsymbol{\tau})]$. The restricted mean survival depends on the treatment regime since survival depends on the treatment assigned. We propose to use the restricted mean survival time as a metric to assess the effectiveness of a treatment regime.

Given covariates X and possible treatments, $A \in \mathcal{A}$, treatment regime, $g : X \to \mathcal{A}$ is treatment assignment on the basis of covariates. We then consider $\widetilde{\mathbf{T}}^*(g)$ to be the potential restricted survival time when treatment regime g is assigned, hence $\widetilde{\mathbf{T}}^*(g) = \sum_{a \in \mathcal{A}} \mathbb{I}(g(X) = a)\widetilde{\mathbf{T}}^*(a)g(X)$ which follows from the consistency assumption. We aim to select a treatment regime which maximizes the restricted mean survival. We denote $\mathbf{V}(g)= \mathbb{E}[\widetilde{\mathbf{T}}^*(g)]$ to be the restricted mean survival time if patients are assigned treatments according to regime g based on covariates X. We will omit the $\boldsymbol{\tau}$ from all notations, but assume that it is fixed throughout.

2.3.1 Optimal Treatment Regime

For right-censored survival data, we aim at maximizing restricted survival which is unobserved. In an observable outcome case in a binary treatment setting, as in Zhang et al. [2012], the optimal treatment regime ($\mathbb{I}[\mathbb{E}[Y_1|X] > \mathbb{E}[Y_0|X]]$) results in a linear form if the conditional outcome regression model is linear. But a right censored survival setting would mostly not result in a treatment regime which has a linear form such as $\{\boldsymbol{\beta}X > 0\}$. In fact a parametric survival function may also not provide an explicit solution to the optimal treatment regime. Moreover, any

form of the treatment regime would depend on the parametric or semi-parametric form of the survival model used which would restrict the regime class.

The methodology we introduce has two levels of estimation: estimating the survival curve using censored data and prognostic covariates and then estimating the marginal restricted survival mean. We estimate the survival function using survival random forests, so that regimes are not restricted to a class.

2.3.1.1 Preliminaries

In a censored setting we fail to observe the outcome and have to estimate or impute to find an optimal treatment regime. Since we aim to maximize the restricted mean survival, we will utilize the conditional distribution of survival time to realize an equivalent form which we can estimate with available tools.

Lemma 2.3.1 *Assuming that $T|X$ has a continuous density $f(.|X)$,*
$\mathbb{E}[\min(T, \boldsymbol{\tau})] = \mathbb{E}_{\mathbf{X}}[\int_0^{\boldsymbol{\tau}} S(t|x)dt]$ *where $S(.|X)$ is the conditional survival function.*

Corollary 2.3.2 *Assuming that $T|X$ has a continuous density $f(.|X)$,*

$$\mathbb{E}[\min(T, \boldsymbol{\tau})|X = x] = \int_0^{\boldsymbol{\tau}} S(t|x)dt.$$

Suppose $\hat{S}(.|X)$ is the estimator of the survival function $S(.|X)$ at the unique death times $0 \leq t_1 \leq t_2 \leq \ldots \leq t_N$ as observed in the training set. Since we are interested in the integral up to $\boldsymbol{\tau}$, we would consider $N_{\boldsymbol{\tau}}$ time points such that $N_{\boldsymbol{\tau}} = \max\{0 \leq i \leq N : t_i \leq \boldsymbol{\tau}\}$ where $t_0 = 0$. We use the plug-in estimator of $\mathbb{E}[\min(T, \boldsymbol{\tau})|X = x]$ derived in Wey, Vock, Connett, and Rudser (2016)Wey et al. [2016],

$$\hat{\mathbb{E}}[\min(T, \boldsymbol{\tau})|X = x] = [\sum_{j=0}^{N_{\boldsymbol{\tau}}} (t_{j+1} - t_j)\hat{S}(j+1|x)]. \tag{2.3.1}$$

Given the data, (Y, δ, A, X), we would like to maximize $\mathbf{V}(g) = E[\widetilde{\mathbf{T}}^*(g)]$ over the entire function space of g, where $\widetilde{\mathbf{T}}^*(g)$ is the potential restricted survival time $\widetilde{\mathbf{T}} = \min(T, \boldsymbol{\tau})$ under treatment regime g.

NOTE: The restriction $\boldsymbol{\tau}$ is just used for the estimation of $E[\widetilde{\mathbf{T}}^*(g)]$ and not for data generation or survival time model.

Proposition 2.3.3 *Optimal treatment regime in case of K treatments,*
$\mathcal{A} = \{0, \ldots, K-1\}$,

$$g^*(X) = \arg \max_{0 \le i \le K-1} \mathbb{E}[\widetilde{\mathbf{T}}^*(i)|X].$$

The proof of proposition 2.3.3 has been delegated to Appendix D.1. Denoting $\boldsymbol{\mu}_a(X) = E[\widetilde{\mathbf{T}}^*(a)|X]$, we get $g^*(X) = \arg \max_{0 \le a \le K-1} \boldsymbol{\mu}_a(X)$ from Proposition 2.3.3. Hence, the plug-in estimator of optimal treatment regime is

$$\hat{g}^{\text{opt}}(X) = \arg \max_{0 \le a \le K-1} \hat{\boldsymbol{\mu}}_a(x). \tag{2.3.2}$$

where $\hat{\boldsymbol{\mu}}_a(x) = [\sum_{j=0}^{N_{\boldsymbol{\tau}^a}} (t^a_{j+1} - t^a_j) \hat{S}_a(j+1|x)]$ as suggested in 2.3.1. Here $t^a_0 = 0$ and t^a_j's are the time points of the estimated survival function $\hat{S}_a(.|X)$, under treatment a.

2.3.2 Methodology

The plug-in estimator in 2.3.2 provides us with an optimal regime which requires us to estimate the conditional survival function under each possible treatment group and then plug it in to get the optimal treatment regime at $x \in \mathcal{X}$. We would be using random forests for the conditional survival function estimation motivated from the complete case situation where a misspecified parametric model performed poorly.

Algorithm 1 Optimal Treatment Regime Estimation

procedure
  INPUT: Data (Y, δ, A, X), Restriction τ
  $\hat{S}_a(.|x) \leftarrow$ Estimate from **random survival forest** for the subgroup of patients with treatment a.
  For each $a \in \mathcal{A}$, uniquely order death times $t_1^a < t_2^a < \ldots$
  $N_{\tau_a} \leftarrow$ index of maximum death/failure time $< \tau$
  $\hat{\mathbb{E}}[\tilde{T}(a)|x] = \hat{\boldsymbol{\mu}}_a(X) \leftarrow \sum_{j=1}^{N_{\tau_a}} (t_{j+1}^a - t_j^a)\hat{S}_a(j+1|x)$
  OUTPUT : $\hat{g}^{\text{opt}}(X) \leftarrow \arg\max_{a \in \mathcal{A}} \hat{\boldsymbol{\mu}}_a(x)$

$S_a(.|x)$, the survival function under treatment a and covariates x

We would establish a few properties of this optimal treatment regime in a general k treatments single stage setting. We would then move on to establish the consistency when survival random forests are used to estimate the conditional survival function in the case of binary treatments.

2.3.3 Properties of the Optimal Regime

2.3.3.1 Bayes Risk

It can be shown as in Qian and Murphy [2011] that the value function or the restricted survival mean under a treatment regime g, $\mathbf{V}(g) = \mathbb{E}[\frac{\widetilde{\mathbf{T}}}{\boldsymbol{\pi}(A,X)}\mathbb{I}(A = g(X))]$ where $\boldsymbol{\pi}(a, x) = \mathbb{P}[A = a|X = x]$. The maximization of $\mathbf{V}(g)$ is equivalent to minimization of $\mathbb{E}[\frac{\widetilde{\mathbf{T}}}{\boldsymbol{\pi}(A,X)}\mathbb{I}(A \neq g(X))]$. The problem of maximization can be then thought of as a weighted classification problem where a treatment regime(g) is a classifier, the misclassification is $A \neq g(X)$, weights are $\frac{\widetilde{\mathbf{T}}}{\boldsymbol{\pi}(A,X)}$. We define the loss function for any classifier d as $\mathbb{L}(a, x, \widetilde{\mathbf{T}}, d) = \frac{\widetilde{\mathbf{T}}}{\boldsymbol{\pi}(a,x)}\mathbb{I}(a \neq d(x))$.

For K classes, we can define $g^*(X) = \arg\max_a \boldsymbol{\mu}_a(X)$ where

$$\boldsymbol{\mu}_a(X) = \mathbb{E}[\widetilde{\mathbf{T}}^*(a)|X], a \in \{0, 1, \ldots K-1\}$$

.

Theorem 2.3.4 *In the single stage decision setting with $\mathcal{A} = \{0, 1, \ldots, K-1\}$, the optimal treatment regime, g^* minimizes Bayes risk.*

The proof of this theorem is in Appendix D.2.

Hence, we have established that an optimal treatment regime of the form as in Proposition 2.3.3 minimizes Bayes risk. This performance of this estimator depends on the modelling of the conditional survival function. We suggest using a random forest so as to not restrict the regime class as in parametric methods. Therefore, to establish consistency we will now consider a survival random forest setting.

2.4 Asymptotic Properties of the Optimal Regime Estimate

Some asymptotic properties of the estimated cumulative hazard function from a survival tree and survival random forests are established in Ishwaran et al. [2008] and Cui et al. [2017]. Ishwaran et al. [2008] assume discrete feature space and suggest approaching continuous covariates by factorizing them. Cui et al. [2017] extends it without such restrictions on the feature space, establishing consistency of cumulative hazard function under various splitting rules and conditions on the dimension of the covariate space. In an optimal treatment regime setting we would wish to establish the consistency of the value function, i.e., the restricted mean survival under the estimated optimal treatment regime. To establish consistency we now introduce a few notations, definitions and assumptions for a survival random forest [Cui et al., 2017][Ishwaran and Kogalur, 2010]. Random survival trees and forests are random variables the distribution of which depends

on the splitting rule at each node. We assume that $\mathcal{X}$ has fixed and finite dimension $d < \infty$, the training sample size n.

For this section we will consider the setting $\mathcal{A} = \{0, 1\}$. $\hat{S}(t|x)$ the survival forest KM estimator of the survival function $S(t|x)$, $S_a(t|x)$ the survival function under the treatment $A = a$. Assuming that the survival random forest was built from B bootstrap samples and hence has B trees, the KM estimator of the b^{th} tree survival function $S^{(b)}(t|x)$ is denoted as $\hat{S}^{(b)}(t|x)$ for $1 \leq b \leq B$.

2.4.1 Notations and Assumptions

Definition 2.4.1 ($\{\alpha, k\}$ valid) *We use the setup used in Cui et al. [2017], which defines $\{\alpha, k\}$ valid tree partition as all such trees such that during splitting, each child node contains at least a fraction $\alpha \in (0, 0.5)$ of the training samples of the parent node. It also requires at least k training samples to be contained in each of the terminal nodes.*

We state this in our first assumption as follows

Assumption 2.4.2 *The random forest is built from trees from each of the B bootstrap samples, each tree being $\{\alpha, k\}$ valid.*

Common Assumption in Survival Analysis,

Assumption 2.4.3 *There exists fixed positive constants $\boldsymbol{\tau}_0 < \infty$ and $M_0 \in (0, 1)$, such that $Pr[Y_i \geq \boldsymbol{\tau}_0|X] \geq M_0$, uniformly for all $X \in \mathcal{X}$.*

To allow dependency among covariates,

Assumption 2.4.4 *Covariates $X \in [0, 1]^d$ are distributed according to a density $p(.)$ satisfying $1/\boldsymbol{\zeta} \leq p(x) \leq \boldsymbol{\zeta}$ for all x and some $\boldsymbol{\zeta} \geq 1$.*

Note that $\boldsymbol{\zeta} = 1$ is the case of uniform independent covariates.

They also set a restriction on the tuning parameter, the minimum terminal node size k such that k grows with the training size n and dimension d at the rate,

Assumption 2.4.5 *Assume that k is bounded below so that*

$$\lim_{n\to\infty} \frac{\log(n)\max\{\log(d), \log\log(n)\}}{k} = 0. \tag{2.4.1}$$

They also consider a smoothness assumption on the hazard function,

Assumption 2.4.6 *For any fixed time point t, the cumulative hazard function $\Lambda(t|x)$ is L_1-Lipschitz continuous in terms of x, and the hazard function $\lambda(t|x)$ is L_2-Lipschitz continuous in terms of x, i.e., $|\Lambda(t|x_1) - \Lambda(t|x_2)| \leq L_1||x_1 - x_2||$ and $|\lambda(t|x_1) - \lambda(t|x_2)| \leq L_2||x_1 - x_2||$, respectively, where $||.||$ is the Euclidean norm.*

To ensure that treatment effects are distinguishable.

Assumption 2.4.7 *There exists $c > 0, \boldsymbol{\gamma} \geq 0$ such that*

$$\mathbb{P}[|\boldsymbol{\mu}_1(X) - \boldsymbol{\mu}_0(X)| \leq \epsilon] \leq c\epsilon^{\boldsymbol{\gamma}}. \tag{2.4.2}$$

With these assumptions stated, we proceed to prove consistency starting from a survival tree. The existing literature considers cumulative hazard functions, but our methodology uses the estimate of the survival function from the survival random forest for estimating the causal effects. Hence, we prove the uniform consistency of survival function estimates of a tree and random forest, before considering the optimal treatment regime setup.

2.4.2 Consistency of Survival Function Estimate

Proposition 2.4.8 (Consistency of a survival tree) *For this proposition, we denote $\hat{S}(t|x)$ to be the tree KM estimator of the survival function $S(t|x)$. Under Assumptions 1-5 and random splitting, the survival tree KM estimator of survival function is consistent, i.e., for each $x \in [0,1]^d$,*

$$\sup_{0\le t\le \boldsymbol{\tau}} \mathbb{E}_{\mathbf{X}}[|\hat{S}(t|x) - S(t|x)|] = O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}}\Big) \quad (2.4.3)$$

with probability greater than $(1 - w_n)$ which approaches 1 as $n \to \infty (w_n \to 0)$.

where $w_n = \frac{2}{\sqrt{n}} + d\exp^{-\frac{c_2^2 \log_{1/\alpha}(n/k)}{2d}} + d\exp^{-\frac{(1-c_2)c_3c_4^2\log_{1/\alpha}(n/k)}{2d}}$, c_1, c_2, c_3, c_4 are constants such that $c_2, c_4 \in (0,1)$, $c_3 = (1-2\alpha)/8$ and $c_1 = \frac{c_3(1-c_2)(1-c_4)}{\log_{1-\alpha}(\alpha)}$.

The proof of this proposition is relegated to Appendix E.1.

Random survival forest is an ensemble of survival trees for which the uniform consistency of the survival function has been shown in Proposition 2.4.8. We will now show the uniform consistency of a forest estimator.

Theorem 2.4.9 (Consistency of survival random forest) *Under Assumptions 1-5 and random splitting, the survival forest estimator of survival function is consistent,*

$$\begin{aligned} &\lim_{B\to\infty} \sup_{0\le t\le \boldsymbol{\tau}} \mathbb{E}_X[|(\hat{S}(t|x) - S(t|x))|] \\ &= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}} + 2w_n\Big), \end{aligned} \quad (2.4.4)$$

where w_n is as in 2.4.3 in Proposition 2.4.8.

The proof of this theorem is in Appendix E.2.

Our main aim is to show that the restricted survival mean under the estimated optimal regime (say $\hat{g}$) is consistent. The optimal regime is estimated using a survival random forest, hence we would use Theorem 2.4.9 in our next step. For this part we are back to the causal setting and would be requiring an additional assumption (Assumption 6). We denote the optimal regime as g^*.

2.4.3 Consistency of Optimal Regime

Theorem 2.4.10 (Consistency of restricted mean survival) *Under Assumptions 1-6, random splitting and fixed covariate space dimension $d < \infty$, for any $\delta > 0$,*

$$\begin{aligned} &|\mathbf{V}(\hat{g}) - \mathbf{V}(g^*)| \\ &\leq \frac{\boldsymbol{\tau}}{2}\sqrt{c\delta^{\boldsymbol{\gamma}+\frac{\boldsymbol{\tau}}{\delta}}O\left(\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}}+(\frac{k}{n})^{\frac{c_1}{d}}+2w_n\right)} \end{aligned} \tag{2.4.5}$$

where w_n is as in 2.4.3 in Proposition 2.4.8.

The proof of this theorem is relegated to Appendix E.3.

Hence, the mean expected restricted survival time from estimated regime $\hat{g}$ approaches that from optimal regime g^*. Therefore, we get the convergence rate for $\mathbf{V}(\hat{g}) \to^p \mathbf{V}(g^*)$. Since the final aim is to maximize restricted mean survival time, it is preferable that $\mathbf{V}(\hat{g})$ approaches the true quantity.

2.5 Simulations

We compare the proposed methodology using random forests to the Cox model in a single stage decision problem in a binary treatment setting. In higher dimensions we compare survival random forest(randomForestSRC package in R)

to lasso penalized Cox model(glmnet package in R). The tuning parameter of the Cox model is chosen using cross validation. The algorithm for both procedures is the same, except at the stage of estimation of the survival function we compare different models(semi-parametric and non-parametric). For comparison, we use few different measures, such as the restricted mean survival time if the treatment regime is applied across the entire population. The main aim is to maximize RMST and hence the treatment regime from the preferred method would have a higher RMST. In certain cases RMST may be maximized but less proportion of patients may be getting the right treatment. Hence another measure compared is the percentage of times the correct treatment was suggested by the treatment regime. Lastly we compare the bias in the survival function estimation.

2.5.1 Simulation Setting

The failure time is generated from a Weibull distribution for both treatment groups and hence is not a proportional hazards model. The censoring variable has the same uniform distribution for both treatment groups. We compare the survival random forest with Cox proportional hazards model in lower dimensions and the lasso penalized Cox model in higher dimensions.

The predictors, $X_1, X_2, \cdots, X_p; A$ are generated in two steps,
$(\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_p)^T$ such that $X_1, X_2 \sim$ iid Unif$(0,1)$; and $X_3, \ldots, X_p \sim$ iid $N(0,1)$. The treatments are randomized, $A|X_1, X_2, \cdots, X_p \sim$ Binomial$(\frac{1}{2})$.

The survival times are generated for each of the 2 treatment groups (since binary treatment setting) as $T_0 \sim$ 2 Weibull$(0.4, 50)$ and $T_1 \sim (2.5 - X_2)$ Weibull$(0.1 + 2X_1, 50 + 10X_1)$. Survival time is $T = T_0(1-A) + T_1 A$. The censoring time is generated as $C \sim$ Unif$(0, 300)$.

We compare the performance for dimensions 2, 12 and 102, the training size

being 500. We consider the test data size to be 10000 and run the simulations 200 times. The measures are averaged over the 200 simulation runs.

The treatment regimes from various methods are compared using

$$RMST = E[\tilde{T}^*(g)]$$

$$percentage = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(g^{\text{opt}}(X_i) = \hat{g}^{\text{opt}}(X_i))$$

and

$$bias = \frac{1}{n}\sum_{i=1}^{n}(\mu_1(X_i) - \mu_0(X_i)) - (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$$

The RMST, percentage and bias for any regime g is estimated by averaging over all test sample means from 100 Monte Carlo simulations. (training size 500 , test size 10000)

2.5.2 Simulation Results

We take the sample size, n to be 500, test size to be 10000 and explore across the covariate dimensions $p = 2, 12, 102$. The true generating model uses 2 covariates, hence the number of noise terms compared are 0, 10, 100. The results are tabulated from 200 simulation runs. In brackets are the standard errors.

Table 2.2: Simulation Results

Noise	τ	$E[\tilde{T}^*(g^{\text{opt}})]$	Method	$E[\tilde{T}^*(\hat{g}^{opt})]$	percentage	bias
0	20	17.65	RF	17.49(0.014)	0.91(0.005)	-0.01(0.081)
0	20		Cox	17.01(0.142)	0.78(0.028)	-2.97(0.057)
10	20		RF	17.41(0.011)	0.87(0.003)	0.002(0.099)
10	20		Cox	15.16(0.241)	0.41(0.048)	-3.16(0.067)
100	20		RF	17.33(0.004)	0.85(0.001)	0.22(0.106)
100	20		Cox	14.49(0.194)	0.28(0.039)	-3.25(0.073)
0	50	39.48	RF	38.93(0.043)	0.87(0.006)	-0.07(0.278)
0	50		Cox	37.77(0.434)	0.76(0.032)	-7.62(0.218)
10	50		RF	38.96(0.024)	0.86(0.003)	-0.27(0.259)
10	50		Cox	33.07(0.650)	0.41(0.048)	-8.30(0.222)
100	50		RF	38.89(0.011)	0.85(0.001)	0.20(0.279)
100	50		Cox	31.27(0.523)	0.28(0.039)	-8.64(0.236)
0	70	51.37	RF	50.48(0.068)	0.85(0.007)	-0.15(0.408)
0	70		Cox	49.21(0.570)	0.76(0.032)	-9.60(0.333)
10	70		RF	50.62(0.040)	0.85(0.003)	-0.50(0.363)
10	70		Cox-l_1	43.05(0.852)	0.41(0.048)	-10.63(0.341)
100	70		RF	50.63(0.020)	0.84(0.001)	0.11(0.394)
100	70		Cox-l_1	40.69(0.686)	0.28(0.03897)	-11.17(0.359)

The estimated optimal regime using random forest is consistent with increasing noise variables in moderately high dimensions. The performance of Cox model deteriorates with increasing dimensions, especially for higher restriction value for τ. In the case of no noise terms, the restricted mean survival time from Cox and survival forest are comparable, the latter being higher. The percentage of patients assigned the correct optimal treatment shows a clear demarcation between random forest and Cox even in lower dimensions.

2.6 Real data analysis

2.6.1 Estimates for Data Analysis

Given data$((Y_i, \delta_i, X_i, A_i)1 \leq i \leq n)$, we compare the estimated optimal treatment regime g^* from our algorithm, to the treatment (A) assigned during the experiment (since we do not know the true treatment regime).

To get an estimate of the performance of various regimes, we use random forest to find the estimate of the survival function, $\hat{S}_a(.|x)$ given x and under treatment a.

$$\hat{\mu}_a(x) = [\sum_{j=1}^{N_\tau}(t_{j+1} - t_j)\hat{S}_a(j+1|x)]. \tag{2.6.1}$$

Next, we estimate the restricted survival mean under any treatment regime (g) as

$$\hat{\mathbb{E}}[\tilde{T}(g)] = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}}\mathbb{I}(a = g(X_i))\hat{\mu}_a(X_i). \tag{2.6.2}$$

2.6.1.1 Notations for Data Analysis

We denote g^* to be the estimated optimal regime from the survival random forest method and A to be the regime or treatment assignment during the trial/experiment. We denote $\hat{\mu}$ to be the estimated restricted mean survival if treatments are assigned as given in dataset, $\hat{\mu}^{\text{rf}}$ to be the estimated restricted mean survival if treatment regime g^* is followed. $(\#A = g^* = \text{treatment } a)$ denotes the number of patients who were assigned the treatment a under both the suggested optimal treatment using random forest as well as in the given dataset. $(\#g^* = \text{treatment } a)$ denotes the number of patients who were assigned

the treatment a under the suggested optimal treatment regime using random forest.

2.6.2 Lung Cancer Data Analysis

For the first data analysis, we consider a lung cancer microarray dataset, a study of which has been done by Zhu et al. [2010]. Zhu et al. [2010] derives a stage independent prognostic gene expression signature which separated patients into low and high risk groups. They also show that the signature gene expression is predictive of survival benefit from ACT.

The data has 2 possible post surgical treatments as the treatment groups, observation and ACT (Adjuvant Chemotherapy). There are various kinds of survival data available but we consider the disease specific survival for our analysis. We consider this dataset to illustrate the performance in a high dimensional case since we consider gene expression data with 22283 different gene expressions along with 5 other clinical covariates such as stage of cancer, age, gender, histology type, DCC type. We consider only those gene expressions with complete data, hence finally we have 22220 covariates. There are 2 kinds of treatments, Obs (placebo), ACT. The data has total 133 patients (complete cases). The overall censoring rate is 35.5%, with censoring rate of patients in Obs group being 36.17% and that of patients on ACT being 34.88%.

The survival time attains a maximum of 9.2 years with the median being 5.97 years. If we just restrict to the patients in the observation group then patients survived till 0.6-9 years with half of them surviving more than 6 years. Among the patients on ACT, 50% survive more than 5.8 years, survival time varying from few days to 9.3 years. Hence we make survival random forest model for the survival in each of the 2 groups and then integrate to estimate the restricted mean survival

for each group. Given any patient's covariates, the restricted mean is calculated for each group and the group with maximum mean is estimated to be the optimal treatment for that patient.

After removing columns of gene expression data with missing values,

Table 2.3: Lung cancer data analysis results

τ	$\hat{\mu}$	$\hat{\mu}^{\text{rf}}$	$\#A = g^*(X) = ACT$	$\#A = g^*(X) = OBS$	trt
2	1.663283	1.764571	53/71	49/62	66
3	2.326970	2.526097	57/71	36/62	83
4	2.874764	3.181439	57/71	32/62	87
5	3.262828	3.774773	57/71	8/62	111
6	4.011463	4.605347	57/71	23/62	96

Here trt = Number of patients who are assigned treatment ACT according to the optimal regime estimated using survival random forest.

As the restriction increases, the number of patients suggested to be assigned to ACT also increases. The 5-year and 6-year survival is maximized by assigning similar number of patients to adjuvant chemotherapy.

2.6.3 Colon Cancer Data Analysis

This data is from one of the first successful trials of adjuvant chemotherapy for colon cancer, the first study being originally described in Laurie et al. [1989]. The dataset has evolved and analyzed in many papers, we use the dataset for chemotherapy for stage B/C colon cancer in the survival package in R. Levamisole is a low-toxicity compound which was used to treat worm infestations in animals and 5-FU is a moderately toxic chemotherapy agent. The records of patients with event as death is considered (not recurrence). This is a randomized trial. We

consider this dataset to illustrate the performance of our proposed method in case of more than 2 treatments.

There are 3 kinds of treatments, Obs (placebo), Lev, Lev+5FU and 11 clinical covariates. We have data for total 888 patients (complete cases), 305 on placebo, 294 given Lev, 289 given Lev+FU. The overall censoring rate is 51.57%, with censoring rate of patients in Obs group being 46.22%, censoring rate of patients on Levamisole being 49.31%, and that of patients on Levamisole + 5-FU being 59.51%.

The survival time for the entire dataset is in the range 23-3329 days with the median being 1983 days. If we just restrict to the patients in the observation group then patients survived till 113-3214 days with half of them surviving less than 1852 days. Among the patients on Levamisole, 50% survive more than 1936 days, survival time varying from 24 to 3329 days. The survival time of patients on a combination Levamisole + 5-FU ranges from 23 to 3309 days, with 50% patients surviving more than 2099 days which is more than any of the other treatment groups. But their are other factors to balance when compare survival times, or restricted survival times. Hence we make survival random forest model for the survival in each of the groups and then integrate to estimate the restricted mean survival for each group. Given any patient's covariates, the restricted mean is calculated for each group and the group with maximum mean is estimated to be the optimal treatment for that patient.

Table 2.4: Colon cancer data analysis results

τ	$\hat{\mu}$	$\hat{\mu}^{\text{rf}}$	$\#g^* = Obs$	$\#g^* = Lev$	$\#g^* = Lev + 5FU$
1200	983.723	1009.341	495	387	6
1500	1154.129	1188.258	509	296	83
1800	1336.118	1372.218	451	414	23
2000	1432.764	1475.304	425	406	57
2500	1680.391	1731.723	337	533	18

We can see from the table that even though the median survival of the Lev+5FU group was higher, as the restriction increases, the number of patients assigned to Lev+5FU to optimize survival is not increasing much. It seems that the initial effect over 2000 days might be good but the toxicity of 5FU starts to set in after that.

2.7 Discussion

A correctly specified Cox model will be faster and more stable than survival random forest. The results show that survival random forest does better than a misspecified Cox proportional hazards model. The differences are more when the restrictions are higher and noise is more. In higher dimensions we may use a penalized Cox model but it will be difficult to account for all the interaction terms. Survival random forest might be a better choice in that case. Moreover, giving the same treatment to everyone might be cost effective in certain cases but in the simulation example provided, a personalized approach to assign treatment is more effective for patients to maximize their restricted survival. These differences are statistically significant but the improvement may not be practically significant. Considering this in mind, we can also change the objective

function to include cost of the treatment assignment which might result in assigning treatment only if the treatment does substantially better in terms of survival.

A correctly specified parametric model performs best but is difficult to achieve. In high dimensions we may do a regularized semiparametric form as L_1 penalized Cox regression or non-parametric methods such as survival random forests. When the dimension of the feature/covariates space is larger than the sample size, penalized regression do not provide unique solutions. Hence for such cases we might want to use non-parametric methods.

A possible future work consideration might be to change the objective function to include cost of the treatment assignment. Also this case only considers discrete and finite set of treatments $\mathcal{A}$, sometimes we might be interested in a further complicated setup of continuous treatments.

Chapter 3

Quantile Random Forest

3.1 Introduction

Decision trees are a class of machine learning algorithms which divide the covariate space into numerous partitions and then fit a simple model for each of the partitions resulting in a conditional prediction model. Trees were first introduced as "automatic interaction detection" (AID) by Morgan and Sonquist [1963]. Further various other algoirthms for splitting data into partitions and estimating arise. THAID was the first classification tree developed by Messenger and Mandell [1972] and then came CART(classification and regression trees) by Breiman et al. [1983]. These were then extended to censored data , longitudinal data, multivariate responses etc. However a single tree can be highly variable and hence to stabilize the predictions from a single tree Breiman [1996] suggested bootstrap aggregating(bagging) which is a way to make an ensemble of weak predictors to improve their accuracy. Random forests are ensembles of decision trees, the ensembles can differ by how the samples are randomized for each tree and how the trees are built.

Random forest first introduced by Breiman [2001], has been extended to quantile regression framework by Meinshausen [2006] and to the context of survival analysis by Ishwaran et al. [2008]. Random forest is an ensemble method constructed by building multiple decision trees and aggregating the predictions from the trees(bagging introduced by Breiman [1996]). Numerous algorithms have been proposed to improve upon the accuracy and computation. Motivated from applications in bioinformatics such as analysing DNA microarray data, Amaratunga et al. [2008] introduce enriched random forests which is obtained by modifying the random forest algorithm to use weighted random sampling instead of simple random sampling to favor informative features. Geurts et al. [2006] introduce extremely randomized trees, another tree based ensemble method

which randomizes both features and cut points at tree-node splits to achieve better computational efficiency. Mentch and Hooker [2016] propose a subbagging based random forest such that it can be visualized as a U-statistics estimator for inference. Sub-bagging consists of proper subsamples of training set instead of complete bootstrap samples whereas bagging [Breiman, 1996] consists of growing trees on bootstrap samples of the training data.

Random forest is widely used for regression, however inference using random forest is limited. It is cumbersome to establish mathematical and asymptotic properties and hence derive confidence intervals(bootstrap intervals being the computationally intensive option out.) Breiman [2001] derives an upper bound on the generalization error of forests. Lin and Jeon [2006] show that random forests can be viewed as adaptively weighted k-PNN method, they further introduce a splitting scheme for desirable adaptivity. Lin and Jeon [2006] derive a lower bound for the rate of convergence of the MSE of random forests with nonadaptive splitting schemes. Biau and Devroye [2010] further prove the consistency of the bagged nearest neighbor method in context of regression and classification. In the context of classification, Biau et al. [2008] derive the consistency of various versions of random forest classifiers and other randomized ensemble classifiers. Wager et al. [2014] derive confidence intervals of bagged predictors using infinitesimal Jackknife procedure and study its sampling distribution. Zhu et al. [2015] derive consistency for reinforcement learning trees, they also introduce a variable muting feature which prevents noise variables from being considered for splitting. Ishwaran et al. [2014] derive consistency of survival random forests. Mentch and Hooker [2016] develop inference procedures by defining a variant of bagging and random forests estimator as an incomplete U-statistic. We extend this framework of using U-statistics for quantile random forest(Meinshausen [2006])

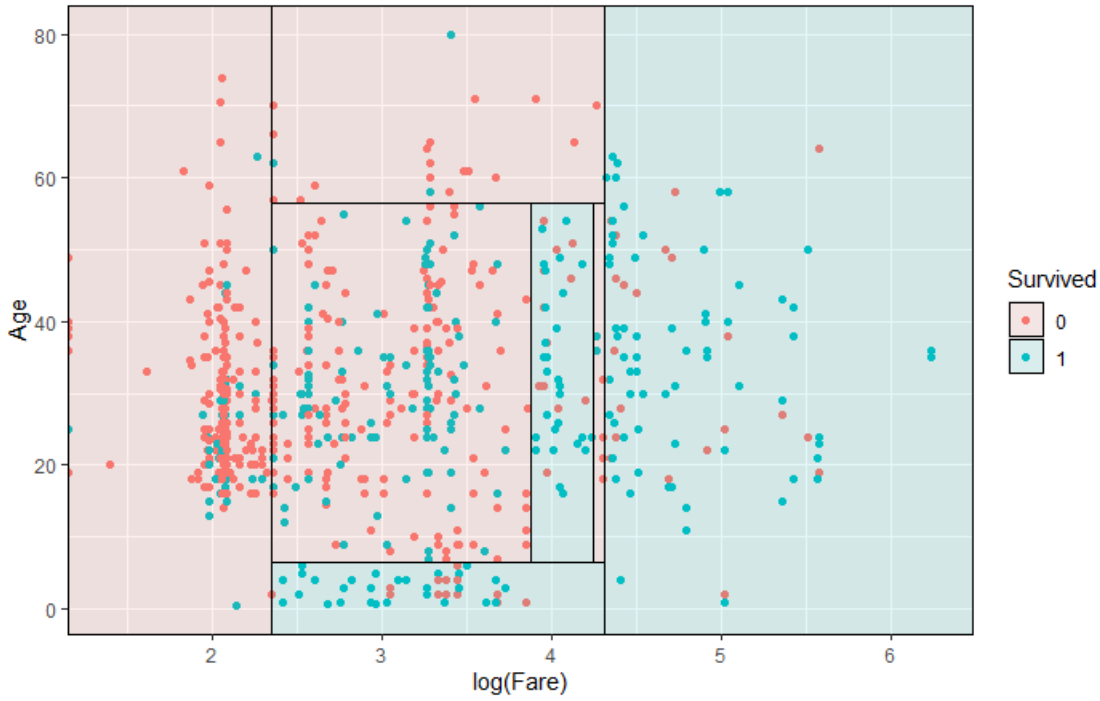
3.2 Review: U-statistic and Random Forest

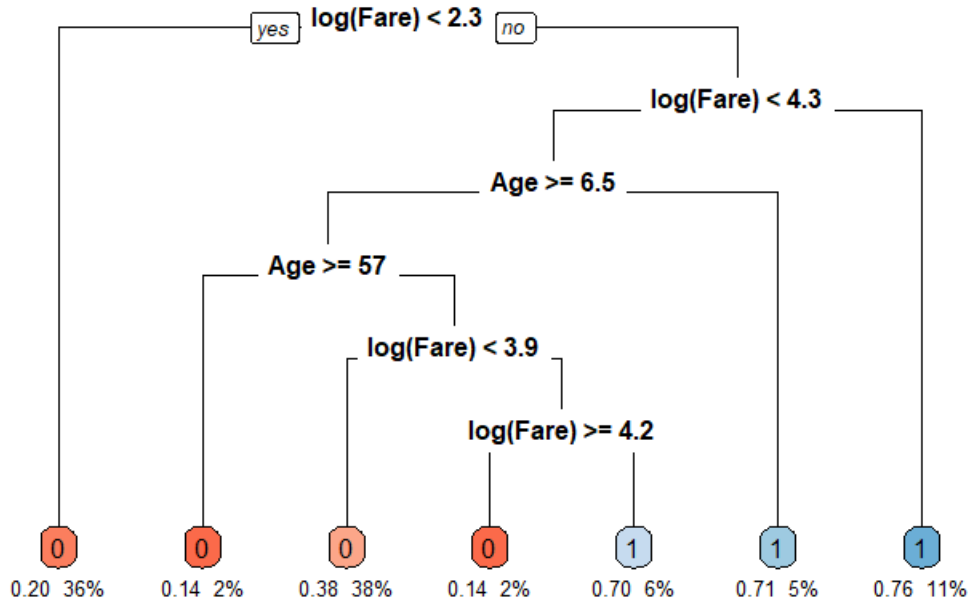
3.2.1 Decision Trees

CART introduced by Breiman et al. [1983], begins with all the observations and then recursively splits the data into "nodes" based on a criteria. For a binary tree, at the first step a feature say X is chosen at random for splitting, all splits of the feature of the form "$X \leq c$" are considered, which splits the feature space into two. Now how the split is chosen depends on the final aim, for regression response variance would be minimized over all partitions whereas for classification Gini index is considered. This creates two partitions called nodes each of which can be further partitioned in the same manner till a stopping rule is reached such as minimum number of observations at a node or further partitioning does not gain much. The nodes at this final step are referred to as leaves. We will denote a leaf of a tree by l. For prediction at say the covariate vector $\tilde{x}$, it is dropped down the tree meaning that the partition or the leaf of the tree which contains it is found(note that the partitions are mutually exclusive and exhaustive so there will exist one leaf and only one leaf). The observations at that leaf are then used to generate a prediction, such as the leaf response mean for regression and the leaf majority vote for classification.

To demonstrate the partitions created by rpart package in R(Therneau and Atkinson [2019]), which is based on the CART algorithm(Breiman et al. [1983]) we utilize the Titanic survival dataset in the titanic package in R(Hendricks [2015]). We use attributes age and log(Fare) of passengers to predict if they survived or not(1=Survived, 0 if not). The classification rpart tree creates the partitions of the 2 dimensional feature space as shown in figure 3.1 (created using Milborrow [2020] and McDermott [2021]).

Figure 3.1: Decision Tree partitions for Titanic Survival Data





We see that higher fare and lower age are mostly associated with better survival. This is a simple yet good representation of how the feature space is split into 7 leaves or seven partitions. Both the geometric and algorithmic representation have been provided for the overall picture.

3.2.2 Random Forest

Random forests estimate the conditional mean response $\mathbb{E}[Y|X = x]$ using weighted mean of the observed response values. For conditional mean at given $X = x$, we need to drop x down the trees in the random forest. We use the notations of Breiman [2001] and Meinshausen [2006]. Suppose ω denotes the random parameter for a tree(random parameter determines how the tree is grown). Let $R_l(x, \omega)$ denote the X-values of the observations in the leaf of tree ω which is the partition consisting of x. Then the weights at observation i for each tree,

$$w_i(x, \omega) = \frac{\mathbb{I}(X_i \in R_l(x, \omega))}{\#(j : X_j \in R_l(x, \omega))} \tag{3.2.1}$$

The estimate from each tree, ω is $\hat{\mu}_\omega(x) = \sum_{i=1}^{n} w_i(x, \omega) Y_i$. Hence estimate of $\mu(x) = \mathbb{E}[Y|X = x]$ from a forest with m trees is as follows,

$$\begin{aligned} \hat{\mu}(x) &= \frac{1}{m} \sum_{t=1}^{m} \hat{\mu}_{\omega_t}(x) \\ &= \frac{1}{m_n} \sum_{t=1}^{m} [\sum_{i=1}^{n} w_i(x, \omega_t) Y_i] \\ &= \sum_{i=1}^{n} [\frac{1}{m} \sum_{t=1}^{m} w_i(x, \omega_t) Y_i] \end{aligned} \tag{3.2.2}$$

The random forest conditional mean estimate in equation 3.2.2 is derived using the weights in Meinshausen [2006] and the sub-bagging procedure as per Mentch and Hooker [2016] so that it can be visualized in the U-statistic framework.

3.2.3 Quantile Random Forest

The entire conditional cumulative distribution $F_Y(.|X = x)$ of the response Y informs much more than just the conditional average $\mathbb{E}[Y|X = x]$ of a response variable given covariates X. As per Meinshausen [2006], the quantile random forest algorithm estimates the conditional distribution function as follows using the same weights w as for a random forest,

Algorithm 2 Quantile Random Forest

1: **procedure**
2: INPUT: Data (Y, X), point of interest (y, x)
3: Grow m trees as in a random forest. Store all the observations in every leaf of every tree.
4: Given $X = x$, drop x down each tree in the forest.
5: For each tree, compute weight $w_i(x, \theta)$ where θ is the random parameter representing the tree for all observations $1 \leq i \leq n$.
6: Compute weight $w_i(x) = \frac{1}{m} \sum_{j=1}^{m} w_i(x, \theta_j)$.
7: OUTPUT : $\hat{F}(y|X = x) = \sum_{i=1}^{n} w_i(x) \mathbb{I}_{\{Y_i \leq y\}}$

3.2.4 U-Statistic

Suppose θ is the parameter of interest and the training data is $Z_1, \ldots, Z_n \sim \text{iid} F_Z$ where F_Z is the cumulative distribution function. We consider θ is a functional of F_Z of the form $\theta(F) = \int \cdots \int h(x_1, \ldots, x_k) dF(x_1) \ldots dF(x_k)$. Hence h is an unbiased estimator of θ such that,

$$\theta = \mathbb{E}[h(Z_1, \ldots, Z_k)] \tag{3.2.3}$$

Here the function h has k arguments, hence we would need $n \geq k$ observations to estimate θ. There are $\binom{n}{k}$ ways of estimating θ from a training sample of size n. The function h can be assumed to be symmetric without loss of generality($h(X_1, \ldots, X_k)$ can be redefined as an average of h evaluated at each permuation of $(X_1, \ldots, X_k)$ making it symmetric). We can now define a U-statistic estimator of the parameter θ as follows,

$$U_n = \frac{1}{\binom{n}{k}} \sum_i h(Z_{i_1}, \ldots, Z_{i_k}) \tag{3.2.4}$$

The sum is over all possible subsamples of the training data which are of size k. This is a U-statistic of rank k and kernel h. Halmos [1946] shows that among all unbiased estimators of the parameter of interest θ, the U-statistics has the minimum variance. Hoeffding [1948] further derive asymptotic normality for U-statistics.

Hoeffding [1948] shows that the asymptotic variance of the U-statistic is $\frac{k^2}{n}\zeta_{1,k}$, where,

$$\zeta_{1,k} = \text{Var}(\mathbb{E}[h(Z_1, \ldots, Z_k)|Z_1 = z]) \tag{3.2.5}$$

3.2.5 Subbagging

Suppose the training data is $Z_i : (X_i, Y_i) \sim \text{iid}F_{X,Y}$ for all $i = 1, \ldots, n$ where Y is the observed response and $X \in \mathcal{X} \subseteq \mathbb{R}^p$ denotes a p-dimensional set of covariates. Suppose we consider that a tree is built using k observations. Given covariates $x \in \mathcal{X}$, the prediction from a tree can be written as a function $T_x : (\mathcal{X} \times \mathbb{R})^k \to \mathbb{R}$. There are $\binom{n}{k}$ ways of building trees with k observations out of total n observations in the training data. Hence aggregating over all such $\binom{n}{k}$ trees we get subbagged

prediction(Mentch and Hooker [2016]) at x as follows,

$$b_n(x) = \frac{1}{\binom{n}{k}} \sum_i T_x((X_{i_1}, Y_{i_1}), \ldots, (X_{i_k}, Y_{i_k})) \tag{3.2.6}$$

Building $\binom{n}{k}$ trees is computationally infeasible hence Mentch and Hooker [2016] suggest building m_n trees ($m_n < \binom{n}{k}$), it seems reasonable to have number of trees dependent on sample size. The m_n samples of size k from training data is selected uniformly at random with replacement from a total of $\binom{n}{k}$. However k should also grow with sample size n hence we consider tree size to be k_n(dependent on n). Hence the final random forest estimate is an average obtained from the tree estimates,

$$b_{n,k_n,m_n}(x) = \frac{1}{m_n} \sum_i T_x((X_{i_1}, Y_{i_1}), \ldots, (X_{i_k}, Y_{i_k})) \tag{3.2.7}$$

3.3 Theoretical Framework

A random forest is an ensemble of trees and as mentioned in the introduction, there can be various algorithms to build a tree. The procedure of choosing features for splitting at every node of tree and then selecting a cutoff for split, introduces a randomization parameter for each tree, and hence the entire ensemble. We propose the estimate of the conditional distribution function of response Y as given in Meinshausen [2006]. The estimate of the cumulative distribution function F from tree built from randomization parameter ω, with the weights same as before,

$$\hat{F}_\omega(y^*|X = x) = \sum_{i=1}^{n} w_i(x, \omega)\mathbb{I}(Y_i \leq y^*) = \frac{\#(j : X_j \in R_l(x, \omega), Y_i \leq y^*)}{\#(j : X_j \in R_l(x, \omega))} \tag{3.3.1}$$

If we create an ensemble of m trees, where t^{th} tree is built from randomization

parameter ω_t for $1 \leq t \leq m$, then the estimate from the random forest with m trees is as follows,

$$\begin{aligned}\hat{F}(y^*|X=x) &= \frac{1}{m}\sum_{t=1}^{m}\hat{F}_{\omega_t}(y^*|X=x) \\ &= \frac{1}{m}\sum_{t=1}^{m}[\sum_{i=1}^{n} w_i(x,\omega_t)\mathbb{I}(Y_i \leq y^*)] \\ &= \sum_{i=1}^{n}[\frac{1}{m}\sum_{t=1}^{m} w_i(x,\omega_t)\mathbb{I}(Y_i \leq y^*)]\end{aligned} \tag{3.3.2}$$

This is the estimator proposed in Meinshausen [2006]. We have introduced the necessary background on which our proposed method is established. We now move on to introduce a method which is inherited from ideas both in Meinshausen [2006] and Mentch and Hooker [2016] so that we can derive a quantile random forest in U-statistic framework in order to derive its asymptotic properties.

3.3.1 Estimate of Conditional Cumulative Distribution Function

Now we introduce the main procedure, we propose the sub-bagging procedure of Mentch and Hooker [2016] to build the quantile random forest similar to Meinshausen [2006]. The main difference in comparison to Meinshausen [2006] is that the trees are built on sub-samples of the training data (as in Mentch and Hooker [2016]) in comparison to complete bootstrap samples.

For a tree, built from training set $\tilde{Z} = (Z_1, \ldots, Z_{k_n})$ of size, k_n ($k_n \leq n$), we denote the weight at observation i as $w_i(x, \omega)$; ω referring to the randomization parameter used to build the tree. The estimate of the distribution function from

a forest with m_n trees,

$$\hat{F}(y^*|X = x^*) = \frac{1}{m_n}\sum_{j=1}^{m_n}[\sum_{j=1}^{n} w_i(x^*, \omega_j)\mathbb{I}(Y_i \leq y^*)] \tag{3.3.3}$$

where

$$w_i(x^*, \omega) = \frac{\mathbb{I}(X_i \in R_l(x^*, \omega))}{\#(j : X_j \in R_l(x^*, \omega))} \tag{3.3.4}$$

and $R_l(x^*, \omega)$ is the set of x observations contained in the leaf l of the tree which reached when x^* is dropped down the tree.

Algorithm 3 Proposed Random Forest

1: **procedure**
2: INPUT: Data (Y, X), point of interest (y^*, x^*)
3: **for** `j in 1:`m_n **do**
4: Subsample $(Z_{j_1}, \ldots, Z_{j_{k_n}})$ with replacement from $(Z_1, \ldots, Z_n)$.
5: Drop x down and note all observations in the leaf l containing x.
6: Compute weight $w_i(x^*, \omega_j)$ for all observations $1 \leq i \leq n$.
7: Compute ensemble weight $w_i(x^*) = \frac{1}{m_n}\sum_{j=1}^{m_n} w_i(x^*, \omega_j)$.
8: OUTPUT : $\hat{F}(y^*|X = x^*) = \sum_{i=1}^{n} w_i(x^*)\mathbb{I}_{\{Y_i \leq y\}}$

3.3.2 Incomplete U-Statistic

Denoting $Z = (X, Y)$, the prediction from a tree, ω can be written as

$$T_{x^*,k,y^*}(Z_1, \ldots, Z_k) = \sum_{i=1}^{n} w_i(x, \omega)\mathbb{I}(Y_i \leq y^*)$$

We can write the estimate of the distribution function at y^* using a random forest with m_n trees with each tree being built from k_n observations as a incomplete

U statistic as follows,

$$U_{n,k_n,m_n}(y^*) = \frac{1}{m_n} \sum_{(i)} T_{x^*,k_n,y^*}(Z_{i_1}, \ldots, Z_{i_{k_n}}) \tag{3.3.5}$$

with T_{x^*,k_n,y^*} as the kernel function.

3.4 Asymptotic Properties

3.4.1 Asymptotic Normality

We will begin with the general incomplete U-statistic notation and denote $T_{x,k_n,y}(Z_1, \ldots, Z_{k_n})$ as the estimate from a tree at y given x, which is built from k_n observations $(Z_1, \ldots, Z_{k_n})$. Note that we will not be referring to any specific method of building trees until much later, for now we ignore the randomization parameter ω.

$$\begin{aligned}
&h^*_{1,k_n,y}(Z) = \mathbb{E}[T_{x,k_n,y}(Z, Z_2, \ldots, Z_{k_n})|Z] \\
&\theta_{k_n}(y) = \mathbb{E}[T_{x,k_n,y}(Z_1, \ldots, Z_{k_n})] \\
&\zeta_{1,k_n}(y) = \mathbb{V}\text{ar}[h^*_{1,k_n,y}(Z)] = \mathbb{C}\text{ov}(T_{x,k_n,y}(Z_1, Z_2, \ldots, Z_{k_n}), T_{x,k_n,y}(Z_1, Z'_2, \ldots, Z'_{k_n}))
\end{aligned} \tag{3.4.1}$$

Theorem 3.4.1

$$\frac{\sqrt{n}(U_{n,k_n,m_n}(y) - \theta_{k_n}(y))}{k_n \sqrt{\zeta^*_{1,k_n}(y)}} \to^{\mathcal{D}} \mathbb{N}(0,1) \tag{3.4.2}$$

if

1. $\lim_{n\to\infty} \frac{n}{m_n} = 0$

2. $\lim_{n\to\infty} \frac{k_n}{\sqrt{n}} = 0$

3. $\lim_{n\to\infty} \zeta_{1,k_n} \neq 0$

4. $\lim_{n\to\infty} \min_{1\leq l\leq L} \zeta_{1,k_n}(y_l) \neq 0$

Conditions 1,2,3, same as that from Mentch and Hooker [2016].

Proof: To show that $U_{n,k_n,m_n}(y)$ converges to a Gaussian process in y we need to show finite dimensional convergence and tightness.

To show finite dimensional convergence, enough to show that any $1 \times r$ random vector, $(U_{n,k_n,m_n}(y_1), \ldots, U_{n,k_n,m_n}(y_L))$ converges weakly to a multivariate normal for any $L \in \mathbb{N}$. To show that, it is enough to show that a linear combination $\sum_{l=1}^{L} \alpha_l U_{n,k_n,m_n}(y_l)$ converges weakly to a univariate normal for any $\alpha_l \in \mathbb{R}$.

Notations:

$$
\begin{aligned}
U^*_{n,k_n,m_n} &= \sum_{l=1}^{L} \frac{U_{n,k_n,m_n}(y_l) - \theta_{k_n}(y_l)}{\zeta_{1,k_n}(y_l)} = \frac{1}{m_n} \sum_{i=1}^{m_n} \sum_{l=1}^{L} \alpha_l \frac{T_{x,k_n,y_l}(Z_{i_1}, \ldots, Z_{i_{k_n}}) - \theta_{k_n}(y_l)}{\zeta_{1,k_n}(y_l)} \\
h^*_{k_n}(Z_1, \ldots, Z_{k_n}) &= \sum_{l=1}^{L} \alpha_l \frac{T_{x,k_n,y_l}(Z_{i_1}, \ldots, Z_{i_{k_n}}) - \theta_{k_n}(y_l)}{\zeta_{1,k_n}(y_l)} \\
\theta^*_{k_n} &= \mathbb{E}[h^*_{k_n}(Z_1, \ldots, Z_{k_n})] = 0 \\
h^*_{1,k_n}(z) &= \mathbb{E}[h^*_{k_n}(z, Z_2, \ldots, Z_{k_n}) - \theta^*_{k_n}] = \mathbb{E}[h^*_{k_n}(z, Z_2, \ldots, Z_{k_n})] \\
\zeta^*_{1,k_n} &= \mathrm{Cov}(h^*_{k_n}(Z_1, Z_2, \ldots, Z_{k_n}), h^*_{k_n}(Z_1, Z'_2, \ldots, Z'_{k_n}))
\end{aligned}
\tag{3.4.3}
$$

$$
\begin{aligned}
&\mathrm{Var}[E[h^*_{1,k_n}(Z)]] \\
&= \mathrm{Var}_Z[E[h^*_{k_n}(Z, Z_2, \ldots, Z_{k_n})|Z]] \\
&= \mathrm{Cov}(h^*_{k_n}(Z_1, Z_2, \ldots, Z_{k_n}), h^*_{k_n}(Z_1, Z'_2, \ldots, Z'_{k_n})) \\
&= \zeta^*_{1,k_n}
\end{aligned}
\tag{3.4.4}
$$

Lindeberg condition : For $\delta > 0$, $\lim_{n\to\infty} \frac{1}{\zeta^*_{1,k_n}} \int_{|h^*_{1,k_n}(Z_1)|\geq\delta\sqrt{n\zeta^*_{1,k_n}}} h^{*2}_{1,k_n}(Z_1)dP = 0$

$$\begin{aligned}
&\frac{1}{\zeta^*_{1,k_n}} \int_{|h^*_{1,k_n}(Z_1)|\geq\delta\sqrt{n\zeta^*_{1,k_n}}} h^{*2}_{1,k_n}(Z_1)dP \\
&\leq \frac{4}{\zeta^*_{1,k_n}} \Big\{ \sum_{l=1}^{L} \frac{|\alpha_l|}{\zeta_{1,k_n}(y_l)} \Big\}^2 \mathbb{P}[|h^*_{1,k_n}(Z_1)| \geq \delta\sqrt{n\zeta^*_{1,k_n}}] \\
&\leq \frac{4(\sum_{j=1}^{r} |\alpha_j|)^2}{\zeta^*_{1,k_n} \min\limits_{1\leq l\leq L} \zeta^*_{1,k_n}(y_l)} \frac{\mathrm{Var}_{Z_1}[h^*_{1,k_n}(Z_1)]}{\delta^2 n\zeta^*_{1,k_n}} \\
&= \frac{4(\sum_{j=1}^{r} |\alpha_j|)^2}{\zeta^*_{1,k_n} \delta^2 n \min\limits_{1\leq l\leq L} \zeta^*_{1,k_n}(y_l)}
\end{aligned} \tag{3.4.5}$$

The rest of the proof is the same as Mentch and Hooker [2016]. Hence for given r, α_i's, $\delta > 0$, Lindeberg's condition is satisfied if

1. $\lim\limits_{n\to\infty} \frac{n}{m_n} = 0$
2. $\lim\limits_{n\to\infty} \frac{k_n}{\sqrt{n}} = 0$
3. $\lim\limits_{n\to\infty} \zeta^*_{1,k_n} \neq 0$
4. $\lim\limits_{n\to\infty} \min\limits_{1\leq l\leq L} \zeta^*_{1,k_n}(y_l) \neq 0$

Hence,

$$\frac{\sqrt{n}(U^*_{n,k_n,m_n} - \theta^*_{k_n})}{k_n\sqrt{\zeta^*_{1,k_n}}} \to^{\mathcal{D}} \mathbb{N}(0,1) \tag{3.4.6}$$

$$\frac{\sqrt{n}(\sum_{l=1}^{L} \frac{U_{n,k_n,m_n}(y_l) - \theta_{k_n}(y_l)}{\zeta_{1,k_n}(y_l)})}{k_n\sqrt{\zeta^*_{1,k_n}}} \to^{\mathcal{D}} \mathbb{N}(0,1) \tag{3.4.7}$$

□

However a random forest consists of trees which are not only built from random sub-samples but there is also a randomization parameter associated with the trees which refers to how it is built. The sub-bagging method does not limit how each tree is built. As stated in the introduction of this chapter, the literature mentions numerous algorithms for building trees for a random forest. For

example choosing a subset of attributes to split at each node and then selecting a feature at random to split on bring randomness to the tree. We accommodate this randomness by denoting the randomization parameter by ω. We assume that each tree is built according to the same randomization procedure. The result derived in theorem 3.4.1 is applicable for fixed kernel U-statistic whereas a random forest estimator would be random kernel U-statistic. We will first define the random forest estimator and then show that the random kernel U-statistic is a consistent estimator of the fixed kernel U-statistic defined before. We can then extend the asymptotic result to the random forest estimator.

As in Mentch and Hooker [2016], we define the randomized kernel U-statistic as,

$$U_{\omega;n,k_n,m_n} = \frac{1}{m_n} \sum_i h_{k_n}^{\omega_i}(Z_{i_1}, \ldots, Z_{i_{k_n}}) \tag{3.4.8}$$

where ω_i is the randomness parameter associated with i^{th} tree. We assume that $\omega_1, \ldots, \omega_{m_n} \sim \text{iid} F_\omega$ i.e. each tree is built according to same randomization procedure, however F_ω is open to choice. The estimate from i^{th} tree if denoted by $T^{(\omega_i)}_{x,k_n,y}(Z_{i_1}, \ldots, Z_{i_{k_n}})$. Then we can define the random forest estimator as follows,

$$r_{n,k_n,m_n}(y|x) = \frac{1}{m_n} \sum_i T^{(\omega_i)}_{x,k_n,y}(Z_{i_1}, \ldots, Z_{i_{k_n}}) \tag{3.4.9}$$

where,

$$\begin{aligned} T^{(\omega)}_{x,k_n,y}(Z_1, \ldots, Z_{k_n}) &= \sum_{i=1}^{k_n} w_i(x,\omega)\mathbb{I}(Y_i \leq y) \\ h^{(\omega)*}_{1,k_n,y}(Z) &= \mathbb{E}[T^{(\omega)}_{x,k_n,y}(Z, Z_2, \ldots, Z_{k_n})|Z] \end{aligned} \tag{3.4.10}$$

Now by defining $\tilde{U}_{\omega;n,k_n,m_n} = \mathbb{E}_\omega[U_{\omega;n,k_n,m_n}] = \mathbb{E}_\omega[\frac{1}{m_n}\sum_i h_{k_n}^{\omega_i}(Z_{i_1}, \ldots, Z_{i_{k_n}})]$, we take an expectation over the randomization parameter ω and fix the kernel such

that we can apply theorem 3.4.1 to $\tilde{U}_{\omega;n,k_n,m_n}$. Now we need to show the asymptotic normality of $\tilde{U}_{\omega;n,k_n,m_n}$. The following lemma follows from theorem 2 of Mentch and Hooker [2016].

Lemma 3.4.2 *Suppose* $\lim \frac{k_n}{\sqrt{n}} = 0$, $\lim \frac{n}{m_n} = \alpha$, $\lim \zeta_{1,k_n} \neq 0$ *and*

$$\lim_{n\to\infty} \mathbb{E}[h^{\omega_i}_{k_n}(Z_{\beta_1}, \ldots, Z_{\beta_{k_n}}) - \mathbb{E}(h^{\omega_i}_{k_n}(Z_{\beta_1}, \ldots, Z_{\beta_{k_n}}))] \neq 0 \tag{3.4.11}$$

then

$$\frac{\sqrt{n}(U_{\omega;n,k_n,m_n} - \tilde{U}_{\omega;n,k_n,m_n})}{\sqrt{k_n^2 \zeta_{1,k_n}(y)}} \to^p 0 \tag{3.4.12}$$

Hence by Slutsky, asymptotic normality holds for $\tilde{U}_{\omega;n,k_n,m_n}$ i.e. $r_{n,k_n,m_n}(y|x)$ in this case.

$$\frac{\sqrt{n}(r_{n,k_n,m_n}(y|x) - \theta_{k_n}(y))}{\sqrt{k_n^2 \zeta_{1,k_n}(y)}} \to^{\mathcal{D}} \mathbb{N}(0,1) \tag{3.4.13}$$

However a point to note is that the limiting distribution would be for the random forest estimate for $\theta_{k_n}(y)$ and not the true value $\theta(y)$ of the underlying distribution function. This next step would be dependent upon if the tree building algorithm is consistent such that $\theta_{k_n}(y) \to^p \theta(y)$ as $n \to \infty$. In this case $\theta(y) = F(y|X = x)$ and

$\theta_{k_n}(y) = \mathbb{E}[T_{x,k_n,y}(Z_1, \ldots, Z_{k_n})] = \mathbb{E}[\sum_{i=1}^{k_n} w_i(x,\theta)\mathbb{I}(Y_i \leq y)]$

We will introduce the assumptions as in Meinshausen [2006], to derive the asymptotic normality for the random forest as an estimator of the true value at $\theta(y)$.

An assumption on the distribution of covariate space.

Assumption 3.4.3 $\mathcal{X} = [0,1]^p$ *and* $X \sim$ *Unif*$[0,1]^p$

Next is an assumption on the process how the trees are built. Meinshausen [2006] have attempted to have minimal possible restrictions.

Denoting the node size at the leaf l of a tree constructed using randomization parameter ω as $\tilde{k}_\omega(l) = \#\{i : X_i \in R_{l(x,\theta)}\}$,

Assumption 3.4.4 *The proportions of observations at a node of each tree diminishes for large n,* $\max_{l,\omega} \tilde{k}_\omega(l) = o(n)$. *The minimal number of observations in a node grows with n, that is* $1/\min_{l,\omega} \tilde{k}_\omega(l) = o(1)$

Assumption 3.4.5 *The probability of choosing an attribute for a split at any node is bounded below by a positive constant. Further a split is chosen only if each of the daughter nodes contain at least a proportion* γ *of the observations at the original node,* $0 < \gamma \leq 0.5$.

The last 2 assumptions are on the true distribution function.

Assumption 3.4.6 *The underlying distribution function* $F(y|X = x)$ *is Lipschitz continuous with parameter L, for all* $x, x' \in \mathcal{X}$,

$$\sup_y |F(y|X = x) - F(y|X = x')| \leq L||x - x'||_1 \tag{3.4.14}$$

Assumption 3.4.7 *The true conditional distribution function* $F(y|X = x)$ *is strictly monotonous in y for all* $x \in \mathcal{X}$.

Lemma 3.4.8 *Under above stated assumptions 1-5, at a fixed* $y \in \mathbb{R}$

$$\theta_{k_n}(y) \to^p \theta(y) \text{ as } n \to \infty$$

Proof: Defining $\widehat{\theta}_{k_n}(y) = \frac{1}{m_n} \sum_{i=1}^{m_n} T_{x,k_n,y}(Z_{i_1}, \ldots, Z_{i_{k_n}})$

$$\theta_{k_n}(y) - \theta(y) = \theta_{k_n}(y) - \widehat{\theta}_{k_n}(y) + \widehat{\theta}_{k_n}(y) - \theta(y) \tag{3.4.15}$$

Since random sub-samples for each tree are selected and built via the same randomization mechanism, $T_{x,k_n,y}(Z_{i_1},\ldots,Z_{i_{k_n}})$ are iid and hence $\theta_{k_n}(y) - \widehat{\theta}_{k_n}(y) \to^p 0$. Further the assumptions stated are from Meinshausen [2006], where it is derived that $\widehat{\theta}_{k_n}(y) - \theta(y) \to^p 0$ under these assumptions. Hence the lemma holds true. □

However note that we would need the consistency rate to tie it with the asymptotic normality result. We will discuss this further in the discussion section.

3.4.2 Consistent Estimator of Variance ζ_{1,k_n}

The asymptotic distribution of the random forest estimator depends on the unknown variance parameter ζ_{1,k_n} and unknown mean θ_{k_n}. Consistent estimates of ζ_{1,k_n} are required for further inference.

We wish to estimate ζ_{1,k_n} where,

$$\zeta_{1,k_n}(y) = \mathrm{Var}_z(\mathbb{E}[h(Z_1,\ldots,Z_{k_n})|Z_1 = z]) \tag{3.4.16}$$

Mentch and Hooker [2016] derive the intuitive estimate by initially selecting and fixing $\tilde{z}$ from the training set $(Z_1,\ldots,Z_n)$ where $Z_i = (X_i, Y_i)$. N subsamples of size k_n are then randomly selected from the training set with the restriction that each of them includes $\tilde{z}$. Each of these subsamples are then used to build a tree and the average of these are stored as an ensemble prediction. This process is repeated $n_{\tilde{z}}$ times with different values of $\tilde{z}$ randomly selected each time. Finally the variance of these ensemble predictions are calculated as an estimate of ζ_{1,k_n}.

$$\widehat{\zeta}_{1,k_n}(y) = \mathrm{Var}(\frac{1}{N}\sum_{j=1}^{N} T_{x,k_n,y}^{\omega_{1,j}}(S_{1,j}),\ldots,\frac{1}{N}\sum_{j=1}^{N} T_{x,k_n,y}^{\omega_{\tilde{z},j}}(S_{n_{\tilde{z}},j})) \tag{3.4.17}$$

$\frac{1}{N}\sum_{j=1}^{N} T^{\omega_{i,j}}_{x,k_n,y}(S_{i,j})$ are iid since the training data $(Z_1,\ldots,Z_n)$ are iid from $F_{X,Y}$ and also ω are iid from F_ω. The sample variance is a U-statistic and hence the sample variance here is a consistent estimator of ζ_{1,k_n}. The algorithm to calculate $\widehat{\theta}_{k_n}(y)$ is as follows,

Algorithm 4 Variance Estimation

Require: $y, x, k_n, N, Z_i = (X_i, Y_i)$ for $1 \leq i \leq n$

procedure
  for `i in 1 to` $n_{\tilde{z}}$ **do**
    Choose $\tilde{z}_i$ at random from $(Z_1,\ldots,Z_n)$
    for `j in 1 to` N **do**
      Select subsample $\mathcal{S}$ of size $(k_n - 1)$ from training data excluding $\tilde{z}$.
      $\mathcal{S} = \mathcal{S} \bigcup \tilde{z}$
      Build a tree using $\mathcal{S}$ and use it to predict at y given x, store as $\hat{r}_{ij}$.
    $r_i = \frac{1}{N}\sum_j^N r_{ij}$
  OUTPUT1 : $\widehat{\zeta}_{1,k_n}(y) = \mathrm{Var}(r_1,\ldots,r_{n_{\tilde{z}}})$
  OUTPUT2 : $\widehat{\theta}_{k_n}(y) = \frac{1}{n_{\tilde{z}}}\sum_i^{n_{\tilde{z}}} r_i$

This procedure calculates the ensemble estimate as well as the variance estimator at the same time saving computation time. This is referred to as internal estimation of variance by Mentch and Hooker [2016]. The variance estimation can also be done outside the ensemble estimation, referred to as external estimation.

Lastly we state the limiting distribution which can be utilized for inference.

Theorem 3.4.9 *Under the stated 5 assumptions, fixed* $y \in \mathbb{R}$ *and conditions as in 3.4.1,* $\lim_{n\to\infty} \frac{n}{m_n} = 0$, $\lim_{n\to\infty} \frac{k_n}{\sqrt{n}} = 0$, $\lim_{n\to\infty} \zeta_{1,k_n} \neq 0$ *and* $\lim_{n\to\infty} \min_{1\leq l\leq L} \zeta_{1,k_n}(y_l) \neq 0$,

$$\frac{\sqrt{n}(r_{n,k_n,m_n}(y|x) - \theta_{k_n}(y))}{\sqrt{k_n^2\widehat{\zeta}_{1,k_n}(y)}} \to^{\mathcal{D}} \mathbb{N}(0,1) \tag{3.4.18}$$

The proof of this follows from application of Slutsky theorem. In fact this point-

wise convergence can be extended to finite dimensional convergence similar to the proof of 3.4.1.

3.5 Simulations

We use a short simulation study to demonstrate the limiting gaussian distribution in equation 3.4.18 of the proposed random forest estimate of the distribution function.

3.5.1 Asymptotic Normality

The data is generated using covariates,

$$X_1 \sim N(5, 10), X_2 \sim N(-3, 5), X_3, \ldots, X_7 \sim \text{ iid Unif}(-1, 1)$$

. The response is generated conditionally as $Y = -1 + 2.5X_1X_2 + 5X_2^3 + \epsilon$, where $\epsilon \sim N(0, 1)$.We look at the limiting distribution of $\hat{F}$ at 2 different values y_1 and y_2 given $x = [4.5, -3, 0, 0, 0, 0, 0]$.

We consider two values of Y, $y = 50\%$ and 90% quantile of Y given x=[4.5,-3,0,0,0,0,0]. We take n=1000, $m_n = 1996$, $k_n = 22$, $\alpha = \lim_{n\to\infty} \frac{n}{m_n} = 0$.

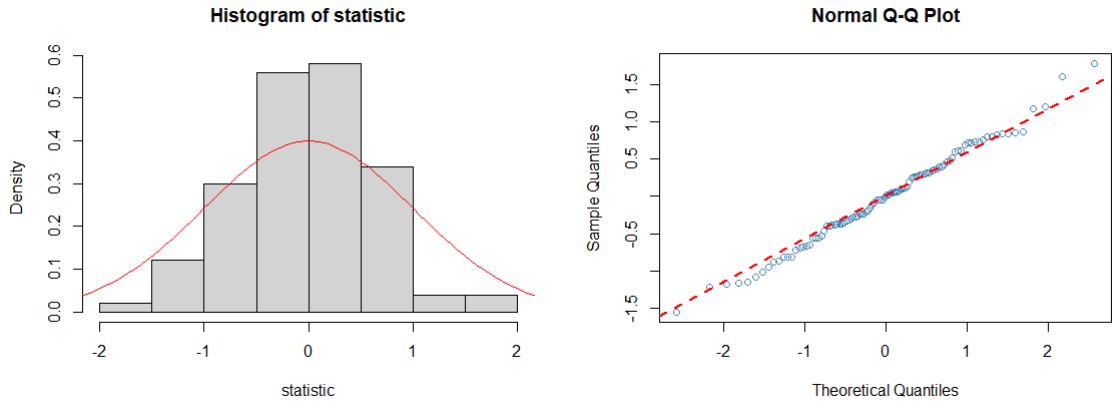


Figure 3.2: Quantile Random Forest : 50% Quantile

Shapiro-Wilk normality test

data: statistic W = 0.99246, p-value = 0.8532

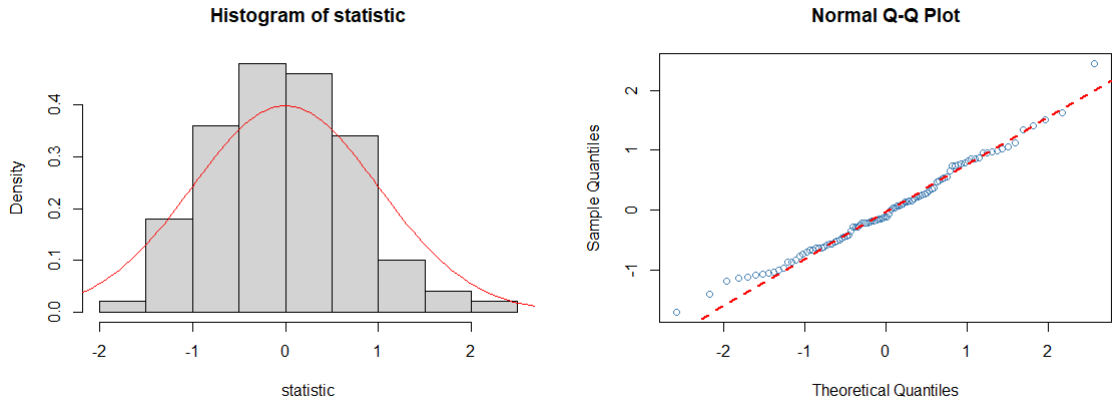


Figure 3.3: Quantile Random Forest : 90% Quantile

Shapiro-Wilk normality test

data: statistic W = 0.98555, p-value = 0.3479

3.6 Discussion

3.6.1 Alternate Method : Honest Trees

In the context of causal inference in a heterogenous treatment effect setting, Wager and Athey [2018] introduce the concept of "honesty" , Athey and Imbens [2016] having first introduced honest trees and causal forests. They define "honesty" to be the property of a tree such that each training sample outcome Y_i is either used to estimate within leaf treatment effect or to split at a node, but not both. They then discuss double sample trees and propensity trees, 2 causal forest algorithms which are built using trees following the honesty property. The double sample tree algorithm splits the subsample used to build a tree into 2, $\mathcal{I}$ and $\mathcal{J}$. The tree is split at each node using observations in $\mathcal{J}$ and covariates and treatment assignments in $\mathcal{I}$(not the outcome Y). Finally at the leaf estimates are derived only using the outcome observations in set $\mathcal{I}$.

Wager and Athey [2018] show that adaptive forests have bias that exceeds sampling variation and hence centered confidence intervals cannot be derived. Hence the consistency rate if derived for the sub-sampled quantile random forest would converge slow. We would need an alternate tree building algorithm for the asymptotic normality to the the true value $\theta(y)$.

Honest forests are unbiased regardless of n, and their mean squared error also decreases with sample size. A disadvantage of double sample method is that for a high dimensional covariate space and small samples, one may require a 70-30 split of $\mathcal{I}, \mathcal{J}$ for better estimation of the random forest(no longer enough information to choose high-quality splits). Wager and Athey [2018] compare adaptive forests and honest causal forests in treatment effect setting. Honest causal forest outperforms in terms of bias and RMSE. Wager and Athey [2018] also explain that for large

n, adaptive forests push outliers in the corners of feature space in causal forest, Wager et al. [2014] having observed a similar phenomenon for regression forests. Honest trees however do not have this issue because different samples are used for node-splitting and estimation. However adaptive forests might be pointwise biased in corners of the covariate space.

Keeping the same properties and bias phenomenon in mind, we suggest random forest with double sample honest trees as suggested by Wager and Athey [2018] but which use node splits as in Meinshausen [2006]. We ran a short simulation to compare the perfomance of quantile random forests(bagged estimates of Meinshausen [2006]), generalized random forest (Athey et al. [2018]) and the proposed random forests with honest trees.

We study 2 simulation settings, both with covariate dimension $p = 40$ one with iid gaussian error and another with heteroskedastic error.

$X_i \sim \text{Unif}(-1, 1) \forall 1 \leq i \leq 40$

Model 1 : $Y|X = 0.8\mathbb{I}(X_1 > 0) + \epsilon, \epsilon \sim N(0, 1)$

Model 2 : $Y|X = (1 + \mathbb{I}(X_1 > 0))\epsilon, \epsilon \sim N(0, 1)$

Let $\hat{q}^{\text{grf}}, \hat{q}^{\text{qrf}}, \hat{q}^{\text{rf}}$ be the estimates of the conditional quantile of $Y|X$ from generalized random forest, quantile random forest(Meinshausen (2006), not honest) and using our method(forest with honest trees which use regression splits). Each method is used to estimate q for 100 Monte Carlo iterations. For every run, training sample size n, training and test samples(test size 1000) are generated. The estimtes of the error and loss from these runs are then calculated as follows.

Suppose $\hat{q}^{(j)}(\tau|X = \tilde{x_i})$ is the estimate of τ^{th} quantile from the j^{th} simulation run for the i^{th} test data $\tilde{x}_i$. We use the following metrics on test data for comparison,

$$\widehat{\text{Error}} = \frac{1}{n_{MC}} \sum_{j=1}^{n_{MC}} \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left[|\hat{q}^{(j)}(\tau|X = \tilde{x_i}) - q(\tau|X = \tilde{x_i})| \right] \tag{3.6.1}$$

$$\widehat{\text{Loss}} = \frac{1}{n_{MC}} \sum_{j=1}^{n_{MC}} \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left[\hat{q}^{(j)}(\tau|X = \tilde{x_i})(\tau - \mathbb{I}(\hat{q}^{(j)}(\tau|X = \tilde{x_i}) < 0)) \right] \tag{3.6.2}$$

Table 3.1: Model 1 : Comparison of GRF, QuantregForest, Honest forest

	$Q_{0.5}$		$Q_{0.7}$	
Model	Error	Loss	Error	Loss
grf	0.0634(0.0021)	0.2044(0.0012)	0.0657(0.0018)	0.6564(0.0020)
qrf	0.1212(0.0010)	0.2299(0.0011)	0.1244(0.0010)	0.6607(0.0020)
rf	0.0514(0.0015)	0.2072(0.0012)	0.0536(0.0013)	0.6532(0.0020)

Table 3.2: Model 2 : Comparison of GRF, QuantregForest, Honest forest

	$Q_{0.5}$		$Q_{0.7}$	
Model	Error	Loss	Error	Loss
grf	0.0579(0.0182)	0.0290(0.0091)	0.5696(0.0428)	0.5429(0.0300)
qrf	0.1407(0.0134)	0.0703(0.0067)	0.6806(0.0374)	0.5379(0.0273)
rf	0.0457(0.0159)	0.0228(0.0079)	0.7133(0.0278)	0.5064(0.0256)

We see that both test set L_1 prediction error and quantile loss increases when we move farther away from 0.5 quantile towards 0 or 1. Generalized random forest and the proposed random forest with honest trees perform better on the given metrics, however the standard error of quantile random forest appears to be slightly lesser than the other 2. This needs to be investigated further and the theoretical framework studied for the proposed random forest.

3.6.2 Confidence Intervals and Tests of Significance

In the previous sections we have derived the point-wise and finite dimensional limiting distribution of the estimate of the distribution function. We can now derive a confidence interval for the random forest estimates and conduct tests of significance. However we would not have centered confidence intervals for $\theta(y)$ but instead for $\theta_{k_n}(y)$. We could still conduct hypothesis tests as suggested in Mentch and Hooker [2016].

However if we were to prove the asymptotic normality using honest trees, we could then derive centered confidence intervals and conduct desired hypothesis tests.

The finite dimensional limiting distribution at a vector of points $(y_1, \ldots, y_L)$ can be used to get confidence intervals for linear combinations of $(\theta(y_1), \ldots, \theta(y_L))$.

If we were to establish weak asymptotic convergence of the random forest process $r_{n,k_n,m_n}(.|x)$ to a Gaussian process, we could then extend this asymptotic normality to quantiles using delta method. We denote the quantile estimate from the random forest to be

$$q_{n,k_n,m_n}(\tau|x) = \inf\{y : r_{n,k_n,m_n}(y|x) \geq \tau\}. \tag{3.6.3}$$

The quantile of the of the true underlying distribution is,

$$q(\tau|x) = \inf\{y : F(y|X = x) \geq \tau\}. \tag{3.6.4}$$

Applying delta method, the quantile estimates q_n would converge at the same rate as for r_n. The limiting Gaussian distribution would then have a derivative term involving the density $f(y|x)$ of $Y|X$ exaluated at the required quantile. We could also derive the confidence interval for quantile estimates. However we would

need pointwise consistent estimates for the true density function f such as kernel density estimate say $\hat{f}$. By continuous mapping theorem and under a few regularity conditions, $\hat{f}(q_{n,k_n,m_n}(\tau|x))$ would be consistent for $f(F^{-1}(\tau|x))$. The quantiles of the limiting distruibution can be used to generate confidence intervals as well as test for significance for the hypothesis $H_0 : q(\tau|X = x) = c$.

Chapter 4

Censored Quantile Regression

4.1 Introduction

There exist various survival metrics to consider from a survival curve such as mean survival time, median survival time, restricted mean survival time, survival probabilities instead of just visual inspection of survival curves. However quantiles are known to assess the distributional impact of covariates better than an average especially for the cases of unknown and heteroskedastic error distributions. Quantiles in the context of regression had been first introduced in Koenker and Bassett [1978]. Powell [1984] proposed least absolute deviations estimation of parameters, replacing the least square estimator in the context of regression median when the dependent variable is censored. Powell [1986] further extended this to more general quantiles in a censored setting with fixed and known censoring. Portnoy [2003] relaxed the fixed censoring assumption to random censoring with conditional independence of survival and censoring times for censored quantile regression. Portnoy [2003] suggest a recursive weighting idea to generalize the Kaplan-Meier estimator.

We consider the setting with right censored survival times and possible high dimensional covariates. Censored quantile regression is an alternative to the proportional hazards constraint when modeling heterogeneity in data. Ying et al. [1995] derive a semiparametric method for median censored regression for survival data. Koenker and Geling [2001] further talk about general quantile regression in survival analysis however considering that the survival times are observed. Lindgren [1997] construct a quantile regression estimate in censored case using generalized L_1 minimization however they do not establish its theoretical validity. Peng and Huang [2008] render the quantile estimating equation as a stochastic integration, utilizing the martingale property associated with censored data. This stochastic setting helps prove uniform consistency, weak

convergence and also derive a closed form of the limit process of the estimated regression quantiles. However both Portnoy [2003] and Peng and Huang [2008] consider a stringent condition where for estimation of quantile τ, the lower quantiles are all linear. Wang and Wang [2009] refer to this consition as "global linearity" and derives an estimating procedure where linearity is assumed only at the quantile of interest say τ. They propose a weighted quantile regression where the weights are derived such that to accomodate for the censoring (using the redistribution of mass idea of Efron [1967]) However the weights needs to be estimated, being a function of the true conditional cumulative distribution function $F_0(.|x)$. Wang and Wang [2009] propose to use the local Kaplan-Meier estimator of $F_0(.|x)$. Gannoun et al. [2003] also propose a local linear estimator for quantile regression. However a local method such as this is computation intensive in high dimensions and also suffers from curse of dimensionality. Moreover, it fails to accommodate categorical attributes. We propose estimating using random survival forest instead to overcome the stated concerns.

A widely prevalent problem in many applications ranging from finance to healthcare, is to measure the policy or treatment effects to compare impacts of different policy or treatment regimes. Most existing research focus on ATE or average treatment effects. Conditional treatment effects(conditional on covariates) deal with informative data heterogenity. However when the outcome is skewed, the average treatment effect might not result in a good policy or regime. Abadie et al. [2002] studies the effects of job training program, JTPA on salary. It might be the case that 2 job trainings perform similar on an average but differently for people at different salary ranges(salary being the outcome variable being measured). A training program might do well for employees in the higher salary range or lower range, both implying very different policy situations. This example showcases the distributional impacts of the treatment effect and

how that might change a policy or regime. Doksum [1974] derive a complete treatment effect function using 2 marginal outcome distributions as an alternative to constant unconditional mean treatment effect. Doksum [1974] and Lehmann and D'Abrera [1998] introduce the idea of quantile treatment effect as the horizontal distance between 2 marginal cumulative outcome distributions. Quantile treatment effect(QTE) at various quantiles help explore beyond just the mean (ATE). One should note that differences in quantiles is one of the possible discretized metrics derived to address treatment heterogeneity. Firpo [2007] defines QTE(overall unconditional quantile treatment effect) and QTT(quantile treatment effect among the treated) and estimates by minimizing a propensity weighted quantile loss function (under unconfoundedness assumption). Abadie et al. [2002] and Chernozhukov and Hansen [2005] study quantile treatment effects in the case of instrumental variables. Imbens and Wooldridge [2009] discusses conditional quantile treatment effect(CQTE) among a wide range of approaches for empirical researchers for program or policy interventions.

Similarly in the case of survival times if we consider patients likely to survive longer and the impact of treatments in that group vs impact of a treatment in weaker patients with shorter lifespan, we might devise different treatment regimes. Hence in our setting of survival times, QTE would be apt metric to be considered when comparing treatments to capture the heterogeneity of treatment effects.

4.1.1 Background and Notations

We use the standard setting and assumptions used in survival analysis, the observed data $(X_i, Y_i, \delta_i)_{1 \leq i \leq n}$ are i.i.d. samples of covariates (X), observed survival time/time to event (Y) and the censoring indicator (δ). Denoting T_i as the true survival time (T) and C_i to be the censoring time (C), the observed

survival time $Y_i = \min(T_i, C_i)$, and $\delta_i = 1(T_i \leq C_i)$ for $1 \leq i \leq n$. Further, we assume that $S(t|X)$ is the conditional survival function, $S(t|X) = P(t > t|X)$, $f(.|X)$ is the conditional density of $T|X$, $\lambda(.|X)$ the conditional hazard function, and $\Lambda(.|X)$ the conditional cumulative hazard function. Assuming conditional independence of time to event and the censoring time $T \perp C|X$, for a specific quantile $\tau \in (0, 1)$ we estimate the coefficients β in the latent regression model as follows,

$$T = X^T \beta_0(\tau) + e(\tau) \tag{4.1.1}$$

where $e(\tau)$ is the error term such that $Q_e(\tau|X) = \inf\{t : P(e \leq t|X) \geq \tau\} = 0$. This problem of quantile regression in a censored setting is referred to as censored quantile regression.

An advantage of quantiles is that they are robust to monotonic transformations. This expands the space of models we can explore. We can extend equation 4.1.1 to,

$$\tilde{T} = h(T) = X^T \beta_0(\tau) + e(\tau) \tag{4.1.2}$$

such that $h(Q_T(\tau|X)) = Q_{h(T)}(\tau|X)$ where h is a monotonic function. Without loss of generality h can be assumed to be a monotonically increasing function(consider $-h$). This represents the familiar parametric class of survival models called accelerated failure times. When h is exponential, and the error follows a Gaussian distribution we get the log-normal family. Similarly if the error follows a generalized extreme value distribution, parametric families such as Weibull and exponential are also included since Weibull and exponential distributions are special cases of EVD distributions. Since survival time is non-negative, considering wider class of models such as equation 4.1.2 is more reasonable in the context of estimating quantiles of survival time.

4.2 Methodology

We propose a two step procedure , first step estimating weights to adjust for censoring and next step estimating the coefficients in the regression model.

The latent regression model is equivalent to ,

$$Q_{T_i}(\tau|X_i = x_i) = \inf\{t : F_0(t|x_i) \geq \tau\} = x_i^T \beta_0(\tau) \tag{4.2.1}$$

In an uncensored setting we observe T_i, and we can minimize the objective function,

$$S_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - x_i\beta) \tag{4.2.2}$$

where $\rho_\tau(z) = z\{\tau - \mathbb{I}(z < 0)\}$ is the quantile loss function. However in a censored setting if $F_0(.|X)$ is known, we consider minimizing the following weighted objective function,

$$Q_n(\beta, F_0) = \frac{1}{n}\sum_{i=1}^{n}\left[w_i(F_0)\rho_\tau(Y_i - x_i\beta) + (1 - w_i(F_0))\rho_\tau(Y^{+\infty} - x_i\beta)\right] \tag{4.2.3}$$

Similar to Wang and Wang [2009], we use Efron's redistribution of mass idea that redistributes the mass of each censored observation to the uncensored ones to the right.

$$w_i(F) = \begin{cases} 1 & \delta_i = 1 \text{ or } F(C_i|X_i) > \tau \\ \frac{\tau - F(C_i|X_i)}{1 - F(C_i|X_i)} & \delta_i = 0 \text{ and } F(C_i|X_i) < \tau \end{cases} \tag{4.2.4}$$

We propose to estimate $F_0(.|X_i)$, the true conditional distribution function of $T|X$ using survival random forest. We will explain the intuition behind the redistribution of mass in weights, in the more general setting of $\tilde{T} = h(T)$. We will denote $\tilde{C} = h(C)$ and $\tilde{Y} = h(Y)$.

In the monotonic transformation setting, i.e. 4.1.2 we need to solve the following latent regression model,

$$Q_{\tilde{T}_i}(\tau|X_i = x_i) = \inf\{t : \tilde{F}(t|x_i) \geq \tau\} = x_i^T \beta_0(\tau) \tag{4.2.5}$$

where $\tilde{F}$ is the cumulative distribution function of $\tilde{T}$. In an uncensored setting we observe $Y_i = T_i$, and we can minimize the objective function,

$$S_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(\tilde{T}_i - x_i\beta) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(\tilde{Y}_i - x_i\beta) \tag{4.2.6}$$

where ρ_τ is the quantile loss function. Here we denote $\tilde{F}$ as the cumulative distribution of $\tilde{T}$. In a censored setting if $\tilde{F}(.|X)$ is known, we consider minimizing the following weighted objective function,

$$Q_n(\beta, \tilde{F}) = \frac{1}{n}\sum_{i=1}^{n} \left[\tilde{w}_i(\tilde{F})\rho_\tau(\tilde{Y}_i - x_i\beta) + (1 - \tilde{w}_i(\tilde{F}))\rho_\tau(\tilde{Y}^{+\infty} - x_i\beta) \right] \tag{4.2.7}$$

where now the new weights $\tilde{w}$ for F are defined as follows,

$$\tilde{w}_i(F) = \begin{cases} 1 & \delta_i = 1 \text{ or } F(\tilde{C}_i|X_i) > \tau \\ \frac{\tau - F(\tilde{C}_i|X_i)}{1 - F(\tilde{C}_i|X_i)} & \delta_i = 0 \text{ and } F(\tilde{C}_i|X_i) < \tau \end{cases} \tag{4.2.8}$$

Suppose $\tilde{F}$ is the cumulative distribution function of $\tilde{T}$ and F_0 is the corresponding cumulative distribution of T. Given h is a monotonically increasing function,

$$\tilde{F}(\tilde{C}|X) = P(\tilde{T} \leq \tilde{C}|X) = P(T \leq C|X) = F_0(C|X) \tag{4.2.9}$$

Hence we see that $\tilde{w}_i(\tilde{F}) = w_i(F_0)$.

Alternate explanation and intuition of weights : To minimize the objective function in equation 4.2.6, we need to solve the negative subgradient condition. The subgradient involves the gradient of the quantile loss, which only depends on the sign of the residual $(\tilde{T}_i - x_i^T\beta_0(\tau))$(the minimization is wrt parameter β and β is only involved in the sign of this residual). In a censored setting, $\delta = \mathbb{I}(T \leq C) = \mathbb{I}(\tilde{T} \leq \tilde{C})$ since h is monotonically increasing. For uncensored observations, $\tilde{Y}_i) = \tilde{T}$ and hence the sign of the residual is $\mathbb{I}(\tilde{T} - x_i^T\beta_0(\tau) < 0)$ is observed. For censored observations, if $\tilde{T} \geq \tilde{Y}_i = \tilde{C}_i > x_i^T\beta_0(\tau)$ then $\mathbb{I}(\tilde{T}_i - x_i^T\beta_0(\tau) < 0) = 0$. The remaining case is $\tilde{T} \geq \tilde{Y}_i = \tilde{C}_i$ but $\tilde{C}_i < x_i^T\beta_0(\tau)$, for which given (C_i, x_i)

$$\begin{aligned}
&\mathbb{E}[\mathbb{I}(\tilde{T} - x_i^T\beta_0(\tau) < 0)|T_i > C_i] \\
&= \frac{P(\tilde{C}_i < \tilde{T}_i < x_i^T\beta_0(\tau))}{P(\tilde{C}_i > \tilde{T}_i)} \\
&= \frac{\tau - \tilde{F}(\tilde{C}_i|x_i)}{1 - \tilde{F}(\tilde{C}_i|x_i)} \\
&= \frac{\tau - F_0(C_i|x_i)}{1 - F_0(C_i|x_i)}
\end{aligned} \tag{4.2.10}$$

$\tilde{F}$ is the cumulative distribution function of $\tilde{T}$ and F_0 is the cumulative distribution of T. The weights assigned are 1 for uncensored observations and when $\tilde{T}_i \geq \tilde{Y}_i = \tilde{C}_i > x_i^T\beta_0(\tau)$. For the last situation, a weight of $\tilde{w}_i(F) = \frac{\tau - F(C_i|x_i)}{1 - F(C_i|x_i)}$ is assigned to the pseudo-observations Y_i and the rest of the weight $(1 - \tilde{w}_i(F))$ is redistributed to any point above (x_i, C_i) for example $(x_i, Y^{+\infty})$. The situation $\tilde{T} \geq \tilde{Y}_i = \tilde{C}_i > x_i^T\beta_0(\tau)$ can be rewritten as $\tilde{F}(\tilde{C}_i) > \tau$ and equivalently $F_0(C_i) > \tau$. Hence $\tilde{w}_i(\tilde{F}) = w_i(F_0)$

Algorithm 5 Censored quantile regression

1: **procedure**
2: INPUT: Data (Y, δ, X), Quantile τ
3: $\hat{S}(.|x) \leftarrow$ Estimate of survival function of T from **random survival forest**.
4: $\hat{F}(.|x) \leftarrow 1 - \hat{S}(.|x)$
5: For each observation (Y_i, δ_i, X_i), find the corresponding weight, $w_i(\hat{F})$ at quantile τ.
6: Run weighted quantile regression(at specified quantile τ) by regressing $\tilde{Y}$ on X using weights $w_i; 1 \leq i \leq n$.
7: OUTPUT : Coefficients $\hat{\beta}$ from weighted quantile regression.

$S(.|x)$, the survival function covariates x

4.3 Asymptotic Properties

The asymptotic properties of the estimated cumulative hazard function from a survival tree and survival random forests are established in Ishwaran et al. [2008] and Cui et al. [2017]. Ishwaran et al. [2008] assume discrete feature space and suggest approaching continuous covariates by factorizing them. Cui et al. [2017] extends it without such restrictions on the feature space, establishing consistency of cumulative hazard function under various splitting rules and conditions on the dimension of the covariate space.

To establish consistency we now introduce a few notations, definitions and assumptions for a survival random forest [Cui et al., 2017][Ishwaran and Kogalur, 2010]. Random survival trees and forests are random variables, the distribution of which depends on the splitting rule at each node. We assume that $\mathcal{X}$ has fixed and finite dimension $d < \infty$, the training sample size n.

For this section we will consider $\hat{S}(t|x)$ the survival forest KM estimator of the survival function $S(t|x)$. Assuming that the survival random forest was built from B bootstrap samples and hence has B trees, the KM estimator of the b^{th} tree survival function $S^{(b)}(t|x)$ is denoted as $\hat{S}^{(b)}(t|x)$ for $1 \leq b \leq B$.

4.3.1 Consistency

4.3.1.1 Notations and Assumptions

Definition 4.3.1 ($\{\alpha, k\}$ valid) *$\{\alpha, k\}$ valid tree partition are all such trees such that during splitting, each child node contains at least a fraction $\alpha \in (0, 0.5)$ of the training samples of the parent node. It also requires at least k training samples to be contained in each of the terminal nodes.*

We state this in our first assumption as follows

Assumption 4.3.2 *The random forest is built from trees from each of the B bootstrap samples, each tree being $\{\alpha, k\}$ valid.*

Common Assumption in Survival Analysis,

Assumption 4.3.3 *There exists fixed positive constants $\boldsymbol{\tau}_0 < \infty$ and $M_0 \in (0, 1)$, such that $Pr[Y_i \geq \boldsymbol{\tau}_0 | X] \geq M_0$, uniformly for all $X \in \mathcal{X}$.*

To allow dependency among covariates,

Assumption 4.3.4 *Covariates $X \in [0, 1]^d$ are distributed according to a density $p(.)$ satisfying $1/\boldsymbol{\zeta} \leq p(x) \leq \boldsymbol{\zeta}$ for all x and some $\boldsymbol{\zeta} \geq 1$; and $\mathbb{E}(xx^T)$ is a positive definite $p \times p$ matrix.*

Note that $\boldsymbol{\zeta} = 1$ is the case of uniform independent covariates.

They also set a restriction on the tuning parameter, the minimum terminal node size k such that k grows with the training size n and dimension d at the rate,

Assumption 4.3.5 *Assume that k is bounded below so that*

$$\lim_{n \to \infty} \frac{\log(n) \max\{\log(d), \log\log(n)\}}{k} = 0. \tag{4.3.1}$$

They also consider a smoothness assumption on the hazard function,

Assumption 4.3.6 *For any fixed time point t, the cumulative hazard function $\Lambda(t|x)$ is L_1-Lipschitz continuous in terms of x, and the hazard function $\lambda(t|x)$ is L_2-Lipschitz continuous in terms of x, i.e., $|\Lambda(t|x_1) - \Lambda(t|x_2)| \leq L_1||x_1 - x_2||$ and $|\lambda(t|x_1) - \lambda(t|x_2)| \leq L_2||x_1 - x_2||$, respectively, where $||.||$ is the Euclidean norm.*

Assumption 4.3.7 *The functions $F_0(t|x)$ and $G(t|x)$ have first derivatives w.r.t. t, denoted as $f_0(t|x)$ and $g(t|x)$ which are uniformly bounded away from infinity. In addition, $F(t|x)$ and $G(t|x)$ have bounded (uniformly in t) second-order partial derivatives w.r.t. x.*

Assumption 4.3.8 *For β in the neighborhood of $\beta_0(\tau)$,*
$\mathbb{E}[xx^T f_0(x^T\beta|x)\{1 - G(x^T\beta|x)\}]$ is positive definite.

With these assumptions stated, we proceed to prove consistency starting from a survival tree. The existing literature considers cumulative hazard functions, but our methodology uses the estimate of the survival function from the survival random forest for estimating the causal effects. Hence, we prove the uniform consistency of survival function estimates of a tree and random forest. We then proceed to prove consistency of the censored regression estimator.

Lemma 4.3.9 *Under assumptions 1-5,*

$$||\hat{F} - F||_{\mathcal{H}} = O_p\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}}\Big)$$

Proof: From Theorem 4.3 in Cui et al.(2019), for a survival tree,

$$\sup_{0\leq t\leq\tau_0} |(\hat{\Lambda}(t|x) - \Lambda(t|x))| = O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}}\Big), \quad (4.3.2)$$

with probability greater than $(1 - w_n)$ which approaches 1 as $n \to \infty (w_n \to 0)$.

For $0 \leq t \leq \tau_0$ and the b^{th} tree estimator $\hat{S}$ of the true survival function S,

$$\begin{aligned} & |\hat{S}^b(t|x) - S(t|x)| \\ & \leq [2 + M] \sup_{0 \leq t \leq \tau} |(\hat{\Lambda}(t|x) - \Lambda(t|x))| \\ & = O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}}\Big), \end{aligned} \tag{4.3.3}$$

with probability greater than $(1 - w_n)$ which approaches 1 as $n \to \infty (w_n \to 0)$.

For b^{th} tree for $1 \leq b \leq B$, over the set, say A_b the probability of which is greater than $1 - w_n$, $|(\hat{S}^{(b)}(t|x) - S(t|x))| = O\Big(\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k \log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}}\Big)$ for any $0 \leq t \leq \tau_0$.

For a forest with B trees, we require the intersection of the sets $A_1, \ldots, A_B$. $\mathbb{P}[A_1 \cap \cdots \cap A_B] \geq \sum_{b=1}^{B} \mathbb{P}[A_b] - (B - 1) \geq B(1 - w_n) - (B - 1)$ from Frechet inequalities.

The forest estimator being $\hat{S}(t) = \frac{1}{B} \sum_{b=1}^{B} \hat{S}^b(t)$, for any $0 \leq t \leq \tau_0$.

$$\begin{aligned} & |\hat{S}(t|x) - S(t|x)| \\ & = |\frac{1}{B} \sum_{b=1}^{B} \hat{S}^b(t|x) - S(t|x)| \\ & \leq \frac{1}{B} \sum_{b=1}^{B} |\hat{S}^b(t|x) - S(t|x)| \\ & = O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}}\Big), \end{aligned} \tag{4.3.4}$$

with probability greater than $B(1 - w_n) - (B - 1)$ which approaches 1 as $n \to \infty (w_n \to 0)$.

Therefore

$\sup_{0\le t\le\tau_0}\sup_x|\hat{S}(t|x)-S(t|x)| = O_p\Big(\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}}+(\frac{k}{n})^{\frac{c_1}{d}}\Big)$

Finally,

$$\begin{aligned}
&||\hat{F}-F||_{\mathcal{H}}\\
&=\sup_{0\le t\le\tau_0}\sup_x|\hat{F}(t|x)-F(t|x)|\\
&=\sup_{0\le t\le\tau_0}\sup_x|\hat{S}(t|x)-S(t|x)|\\
&=O_p\Big(\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}}+(\frac{k}{n})^{\frac{c_1}{d}}\Big)
\end{aligned} \tag{4.3.5}$$

□

Theorem 4.3.10 *Consistency*

Under assumptions 1-7, for iid (Y_i, X_i, δ_i) *and conditional independence,* $T \perp C|x$,

$$\hat{\beta}(\tau)\to\beta_0(\tau) \text{ in probability as } n\to\infty \tag{4.3.6}$$

Proof: Note that we intend to prove consistency of the estimator of β for the more general framework $Q_{\tilde{T}}(\tau|X=x)=\beta^T x$. The first step consists of estimating $\hat{F}_0$ where F_0 is the true distribution function of T and then we define $\hat{\tilde{F}}(t) = \hat{F}_0(h^{-1}(t))$. We have shown that the weights for the next step are $\tilde{w}(\tilde{F}) = w(F_0)$. The way the estimators are defined, we also get that $\tilde{w}(\hat{\tilde{F}}) = w(\hat{F}_0)$. Then we proceed to weighted quantile regression on $\tilde{Y}$ with weights $w(\hat{F}_0)$.

In essence we can also consider this as estimating $\hat{\tilde{F}}$ in the first step and then using the weights $\tilde{w}(\hat{\tilde{F}})$ for weighted quantile regression on $\tilde{Y}$. This is to eliminate the confusion when using both T and $\tilde{T}$. We will now just use $\tilde{T}$ in the consistency proof.

Negative subgradient of the quantile estimating equation gives us,

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i, \tilde{Y}_i, \delta_i, \hat{\tilde{F}}) = 0 \tag{4.3.7}$$

where

$$m(X_i, \tilde{Y}_i, \delta_i, \hat{\tilde{F}}) = X_i(\tau - \tilde{w}_i(\hat{\tilde{F}}))\mathbb{I}(\tilde{Y}_i < \beta^T X)) \tag{4.3.8}$$

We will use the setting and theorem 1 of Chen et al. [2003] to prove consistency similar to Wang and Wang [2009]. As shown in Wang and Wang [2009], $M(\beta_0, \tilde{F}) = 0$ because the change proposed is for the estimator and not the setting or space. We further need to verify conditions 1.1-1.4 and 1.5' in Chen et al. [2003].

Conditions 1.2, 1.3 do not need the form of $\hat{\tilde{F}}$ hence the same proof as that of Wang and Wang [2009] can be applied. In Wang and Wang [2009], for their proof of Theorem 1, conditions 1.1 and 1.5' do not require the form of $\hat{\tilde{F}}$ hence the same proof works.

Condition 1.4 follows from lemma 4.3.9. We have defined $\widehat{\tilde{F}}$, the estimate of $\tilde{F}$ as $\widehat{\tilde{F}}(t) = \hat{F}_0(h^{-1}(t))$.

$$\tilde{F}(t) = P(\tilde{T} \leq t) = P(\tilde{T} \leq t) = P(T \leq h^{-1}(t)) = F_0(h^{-1}(t))$$

We consider the space of distribution functions on $\tilde{T}$ to be $F \in \tilde{\mathcal{H}}$. Noting that F would have 2 parameters t and x since $F(t|x)$ is the cumulative distribution function st t given x. Defining the norm $||f(t,x)||_{\tilde{\mathcal{H}}} = \sup_{h(0)\leq t\leq h(\tau_0)} \sup_x |f(t,x)|$

$$
\begin{aligned}
&||\widehat{\tilde{F}} - F||_{\tilde{\mathcal{H}}} \\
&= \sup_{h(0)\leq t\leq h(\tau_0)} \sup_x |\widehat{\tilde{F}}(t|x) - \tilde{F}(t|x)| \\
&= \sup_{h(0)\leq t\leq h(\tau_0)} \sup_x |\hat{F}_0(h^{-1}(t)|x) - F_0(h^{-1}(t)|x)| \\
&= \sup_{h^{-1}(h(0))\leq h^{-1}(t)\leq h^{-1}(h(\tau_0))} \sup_x |\hat{F}_0(t|x) - F_0(t|x)| \\
&= \sup_{0\leq t\leq \tau_0} \sup_x |\hat{F}_0(t|x) - F_0(t|x)| \\
&\leq ||\hat{F}_0 - F_0||_{\mathcal{H}} \text{ (for big enough } \tau_0)
\end{aligned}
\tag{4.3.9}
$$

Hence $||\widehat{\tilde{F}} - F||_{\tilde{\mathcal{H}}} = o_p(1)$ follows from lemma 4.3.9. Hence conditions 1.1-1.4 and 1.5' of Chen et al. [2003] hold true.

□

4.4 Quantile Treatment Effects under Censoring

Further in a causal setting, treatment effect heterogeneity necessitates the need to consider conditional quantile treatment effect instead of comparing population level survival metrics. We focus on estimation of conditional quantile treatment effect which essentially measures the difference in conditional counterfactual survival times at a specific quantile.

4.4.1 Preliminaries

In a single stage decision setting we consider a finite discrete set of 2 possible treatments, $\mathcal{A} = \{0, 1\}$. We consider the potential survival time to be $\tilde{T}^*(a)$ and potential censoring time to be $\tilde{C}^*(a)$ under any treatment $a \in \mathcal{A}$. The usual

assumptions of causal inference now in survival analysis setting are postulated. The assumption of consistency translates to $\tilde{T} = \sum_{a\in\mathcal{A}} \mathbb{I}(A = a)\tilde{T}^*(a)$ and $\tilde{C} = \sum_{a\in\mathcal{A}} \mathbb{I}(A = a)\tilde{C}^*(a)$. The assumption of unmeasured confounders in this case is that $\{\tilde{T}^*(a) : a \in \mathcal{A}\} \perp A|X$. We also consider non-informative censoring such that $\{\tilde{C}^*(a) : a \in \mathcal{A}\} \perp \{\tilde{T}^*(a) : a \in \mathcal{A}\}|X, A$.

Let $\tilde{T}^*(a)$ denote the survival time of a randomly selected individual been given treatment $A = a$, $a \in \{0, 1\}$. $\tilde{T}^*(a)$ is referred to as the potential survival time. The survival time $\tilde{T}$ is : $\tilde{T} = A\tilde{T}^*(1) + (1 - A)\tilde{T}^*(0)$. Assume,

$$\tilde{T}^*(a) = \beta_0 + \alpha_0 a + \beta^T X + a\gamma^T X + \epsilon \tag{4.4.1}$$

where $\mathbb{P}(\epsilon \leq 0|X) = \tau$, $0 \leq \tau \leq 1$ and X is a high dimensional vector of covariates.

The τ-th conditional quantile of $\tilde{T}^*(a)$ given X is $Q_\tau(\tilde{T}^*(a)) = \beta_0 + \alpha_0 a + \beta^T X + a\gamma^T X$. The conditional quantile treatment effect is ,

$$\text{QTE} = Q_\tau(\tilde{T}^*(1)) - Q_\tau(\tilde{T}^*(0)) = \beta_0 + \alpha_0 + \beta^T X + \gamma^T X - \beta_0 - \beta^T X = \alpha_0 + \gamma^T X \tag{4.4.2}$$

Due to random censoring, we may not observe $\tilde{T}$. The observed outcome is denoted by $\tilde{Y}$, $\tilde{Y} = \min(\tilde{T}, \tilde{C})$ where $\tilde{C}$ is the censoring time. We assume $\tilde{T} \perp \tilde{C}|X$ that is $\tilde{T}$ and $\tilde{C}$ are independent given covariates X. $\delta = \mathbb{I}(T \leq C)$.

The goal of our work is to estimate the conditional quantile treatment effect when,

(1) the survival time $\tilde{T}$ is subject to random censoring.

(2) the covariate dimension is large.

(3) the data are from an observational study rather than a randomized clinical trial.

Let $\pi(X) = \mathbb{P}(A = 1|X)$ be the propensity score of the treatment. Let $F_0(.|x, a)$

be the conditional distribution of T given $X = x, A = a$. We assume that the treatment assignment only depends on X that is $A \perp (\tilde{T}, \tilde{C})|X$.

4.4.2 Methodology

We will estimate the parameter set $(\alpha_0, \beta^T, \gamma^T)$ by estimating the conditional quantile of $\tilde{T}|X, A$ i.e., solving the following equation,

$$(\beta_0, \alpha_0, \beta^T, \gamma^T) = \arg\min \rho_\tau(\tilde{T} - \beta_0 - \alpha_0 a - \beta^T X - a\gamma^T X) \tag{4.4.3}$$

We do not observe $\tilde{T}$, hence we need weights to adjust for censoring similar to previous setting with no treatments. The weights in this case would be ,

$$w_i(F) = \begin{cases} 1 & \delta_i = 1 \text{ or } F(C_i|X_i, A_i) > \tau \\ \frac{\tau - F(C_i|X_i, A_i)}{1 - F(C_i|X_i, A_i)} & \delta_i = 0 \text{ and } F(C_i|X_i, A_i) < \tau \end{cases} \tag{4.4.4}$$

We estimate $F_0(.|X_i, A_i)$, the true conditional distribution function of $T|X, A$ using survival random forest for each treatment arm. Denoting the estimates of parameters to be $(\hat{\beta}_0, \hat{\alpha}_0, \hat{\beta}^T, \hat{\gamma}^T)$, the estimate of conditional quantile effect then is ,

$$\widehat{\text{QTE}} = \hat{\alpha}_0 + \hat{\gamma}^T X \tag{4.4.5}$$

Algorithm 6 Quantile Treatment Effect

1: **procedure**
2: INPUT: Data (Y, δ, X, A), Quantile τ
3: $\hat{S}(.|a, x) \leftarrow$ Estimate of survival function of T from **random survival forest** inputting both A and X as covariates to estimate the conditional survival function.
4: $\hat{F}(.|a, x) \leftarrow 1 - \hat{S}(.|a, x)$
5: For each observation (Y_i, δ_i, X_i), find the corresponding weight, $w_i(\hat{F})$ at quantile τ.
6: Run weighted quantile regression(at specified quantile τ) by regressing $\tilde{Y}$ on (X, A, AX) and the weights $w_i; 1 \leq i \leq n$.
7: OUTPUT : Coefficients $(\hat{\beta}_0, \hat{\alpha}_0, \hat{\beta}^T, \hat{\gamma}^T)$ from weighted quantile regression.

4.4.3 Consistency

Theorem 4.4.1 *Consistency*

Under model assumptions 1-7, for iid $(Y_i, X_i, \delta_i, A_i)$ *and conditional independence,* $T \perp C|x$,

$$(\hat{\beta}_0(\tau), \hat{\alpha_0}(\tau), \hat{\beta^T}(\tau), \hat{\gamma^T}(\tau)) \rightarrow (\beta_0(\tau), \alpha_0(\tau), \beta^T(\tau), \gamma^T(\tau)) \text{ in probability as } n \rightarrow \infty \tag{4.4.6}$$

This proof is very similar to that of theorem 4.3.10. The proof has been done in details in Appendix F.1.

Theorem 4.4.2 *Consistency QTE*

Under model assumptions 1-7, for iid $(Y_i, X_i, \delta_i, A_i)$ *and conditional independence,* $T \perp C|x$,

$$(\hat{\alpha_0}(\tau), \hat{\gamma^T}(\tau)) \rightarrow (\alpha_0(\tau), \gamma^T(\tau)) \text{ in probability as } n \rightarrow \infty \tag{4.4.7}$$

Proof: The proof follows from a simple application of the Slutsky theorem to equation 4.3.6 of Theorem 4.3.10 when we replace estimating coefficients of

covariates X by estimating the effect of interactions of covariates X and treatment assignment A on the quantile of survival time.

Taking $\tilde{c} = (1, \mathbf{0}_{d+1}, \mathbf{1}_d)$, applying Slutsky to 4.3.10,

$$\tilde{c}(\hat{\beta}_0(\tau), \hat{\alpha_0}(\tau), \hat{\beta^T}(\tau), \hat{\gamma^T}(\tau)) \to \tilde{c}(\beta_0(\tau), \alpha_0(\tau), \beta^T(\tau), \gamma^T(\tau)) \tag{4.4.8}$$

in probability as $n \to \infty$.

Hence,

$$(\hat{\alpha_0}(\tau), \hat{\gamma^T}(\tau)) \to (\alpha_0(\tau), \gamma^T(\tau)) \text{ in probability as } n \to \infty \tag{4.4.9}$$

□

4.5 Simulations

4.5.1 Simulations for Censored Quantile Regression

We compare our proposed algorithm with the oracle(known true time to event and features), Portnoy's method(Portnoy [2003] crq), local method of Wang and Wang [2009] (lcrq), and 2 naive methods. Naive method 1 ignores censoring and naive method 2 models using only uncensored observations in the sample for usual quantile regression.

As per the equation 4.1.2, we choose the function $h = \log()$ such that the survival time is positive and the error to be distributed according to generalized extreme value distribution. If $X \sim \text{GEV}(\mu, \sigma, 0)$ then $\sigma \exp(-\frac{X-\mu}{\mu\sigma}) \sim \text{Weibull}(\sigma, \mu)$ (Wikipedia contributors [2021]). If $X \sim \text{Exponential}(1)$, (Exponential distribution) then $\mu - \sigma \log X \sim \text{GEV}(\mu, \sigma, 0)$ (Wikipedia contributors [2021]). Hence, this setting

encompasses a few familiar parametric survival models.

We study simulation settings in small to moderate dimensions(high dimensiona can be implemented using regularization sucha as LASSO to weighted quantile regression). We explore survival settings which involve correlated covariates, heteroskedastic errors and sparse models.

The main aim is to estimate the survival model coefficients, so the table reports the bias and standard errors in the estimation of the coefficients. We also use the L_1 norm of bias and MSE as metrics to compare different methods for simulations. $Bias = \mathbb{E}(\hat{\beta}) - \beta$. Suppose $\beta_j^{(k)}$ is the estimate of the j^{th} coefficient β_j from the k^{th} simulation run, then the estimates aggregated from N simulation runs are defined as follows,

$$\widehat{Bias}_j = \frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_j^{(k)} - \beta_j) \tag{4.5.1}$$

We aggregate the bias and squared errors in coefficient estimation by considering the metrics, L_1 prediction error as follows,

$$||\text{Bias}||_1 = \sum_{j=0}^{d}|\frac{1}{N}\sum_{i=1}^{N}(\hat{\beta}_j^{(k)} - \beta_j)| \tag{4.5.2}$$

4.5.1.1 Simulation Setting 1

The covariates are generated as $X_1, X_2, \ldots, X_{10} \sim \mathrm{N}(0_{10}, \Sigma)$ where Σ is an **AR(1)** covariance structure with $\sigma = 1$ and $\rho = 0.8$.
The true time to event is generated as
$\log(T) = 1+X_1+X_3+X_5+(0.3+2(X_2-0.5)^2)(Z-\phi^{-1}(\tau))$ where $Z \sim \mathrm{EVD}(0,1,0)$ and ϕ is the extreme value distribution.

The censoring time is generated as $\log(C) \sim 1.5 + X_2 + X_4 + 0.5(Z - \phi^{-1}(\tau))$. The observed time is $Y_i = \min(T_i, C_i)$. This results in $40-45\%$ censoring. We also consider $\log(C) \sim 3 + X_2 + X_4 + 0.5(Z - \phi^{-1}(\tau))$ which results in $23-30\%$ censoring. The results in the next 3 tables are from 100 simulation runs. We check multiple settings with $\tau \in \{0.5, 0.7\}$, and $n \in \{500, 1000\}$.

Note that the true survival time is not associated with features X_2 and X_4 but these features are used to generate the censoring time. The bias of the local method(Wang and Wang) and both the naive methods are high for both τ and across sample sizes low or high. The bias in the local method can be attributed to the curse of dimensions. Portnoy's method and the proposed method have comparable bias but the bias of Portnoy's method increases with sample size whereas the proposed method is consistent. For lower censoring we again see that Portnoy's method and the proposed RFRCRQ method perform similar. As for the individual coefficients we see that the proposed estimator is unbiased for all coefficients, whereas CRQ(Portnoy's method) has high bias for β_2. This can be attributed to the heteroscedasticity of error whcih involves X_2.

Table 4.1: Setting 1 : Bias with $\tau = 0.5$ and $n = 1000$

	β_0	β_1	β_2	β_3	β_4	β_5
Omni bias	-0.0019	0.0068	0	0.0069	0	-0.0033
Omni se	(0.0068)	(0.0076)	(0)	(0.0083)	(0)	(0.006)
crq bias	0.0474	0.0132	-0.1442	0.01	8e-04	7e-04
crq se	(0.0106)	(0.009)	(0.0197)	(0.0116)	(0.0114)	(0.0121)
lcrq bias	-0.1549	-0.1947	0.2211	-0.198	0.2052	-0.2011
lcrq se	(0.0067)	(0.0073)	(0.0124)	(0.0094)	(0.009)	(0.0101)
rfcrq bias	-0.0064	0.0102	-0.0023	0.0073	-0.0103	0.0055
rfcrq se	(0.0095)	(0.0092)	(0.0187)	(0.0132)	(0.0116)	(0.012)
naive1 bias	-0.2895	-0.4725	0.4796	-0.4733	0.4831	-0.4822
naive1 se	(0.0056)	(0.0069)	(0.01)	(0.0081)	(0.0076)	(0.009)
naive2 bias	-0.8891	-0.3153	1.2434	-0.3276	0.3261	-0.3132
naive2 se	(0.0094)	(0.0108)	(0.0196)	(0.0146)	(0.0131)	(0.0124)

	β_6	β_7	β_8	β_9	β_{10}
Omni bias	0	0	0	0	0
Omni se	(0)	(0)	(0)	(0)	(0)
crq bias	0.0086	-0.0034	-0.0034	0.0149	-0.0086
crq se	(0.0114)	(0.0105)	(0.011)	(0.0096)	(0.0078)
lcrq bias	0.0022	0.0019	-0.0027	0.0081	-0.0085
lcrq se	(0.0087)	(0.0083)	(0.0089)	(0.0081)	(0.0066)
rfcrq bias	0.0053	-0.0046	-0.0045	0.0118	-0.0083
rfcrq se	(0.0115)	(0.0105)	(0.0121)	(0.0112)	(0.0084)
naive1 bias	0.0071	-0.0012	-0.005	0.0031	-0.0052
naive1 se	(0.007)	(0.0071)	(0.0079)	(0.0067)	(0.0056)
naive2 bias	-0.0196	0.0062	0.0053	-0.0081	0.0064
naive2 se	(0.0104)	(0.0102)	(0.0118)	(0.0111)	(0.0091)

Table 4.2: Setting 1 : Bias with $\tau = 0.7$ and $n = 1000$

	β_0	β_1	β_2	β_3	β_4	β_5
Omni bias	0.0017	0.0033	0	0.0116	0	-1e-04
Omni se	(0.0081)	(0.0093)	(0)	(0.0109)	(0)	(0.0076)
crq bias	0.0829	0.0162	-0.2107	0.0171	-0.0224	0.0178
crq se	(0.0128)	(0.012)	(0.0227)	(0.0139)	(0.0147)	(0.0157)
lcrq bias	-0.2605	-0.2888	0.3544	-0.2854	0.2821	-0.2886
lcrq se	(0.0068)	(0.0087)	(0.0129)	(0.0094)	(0.0098)	(0.0111)
rfcrq bias	-0.0684	-0.0269	0.0873	-0.0365	0.0207	-0.0384
rfcrq se	(0.0099)	(0.0117)	(0.0216)	(0.0137)	(0.0136)	(0.0155)
naive1 bias	-0.3635	-0.5048	0.5537	-0.5071	0.5107	-0.514
naive1 se	(0.0057)	(0.0067)	(0.0091)	(0.008)	(0.0079)	(0.0092)
naive2 bias	-0.9478	-0.3403	1.3818	-0.3435	0.3443	-0.3422
naive2 se	(0.0075)	(0.0096)	(0.0196)	(0.0126)	(0.0118)	(0.0138)

	β_6	β_7	β_8	β_9	β_{10}
Omni bias	0	0	0	0	0
Omni se	(0)	(0)	(0)	(0)	(0)
crq bias	-0.0012	0.0121	-0.0047	0.0048	-0.0066
crq se	(0.0163)	(0.0135)	(0.0148)	(0.0123)	(0.0087)
lcrq bias	0.0038	0.0067	-0.0026	0.0077	-0.0117
lcrq se	(0.0098)	(0.009)	(0.01)	(0.0087)	(0.0068)
rfcrq bias	-0.0125	0.0016	0.002	0.0104	-0.0194
rfcrq se	(0.016)	(0.0125)	(0.0139)	(0.0127)	(0.0089)
naive1 bias	8e-04	0.0038	8e-04	-6e-04	-0.0067
naive1 se	(0.0082)	(0.008)	(0.0088)	(0.0071)	(0.0059)
naive2 bias	-0.0037	0.0025	-0.0037	0.0111	-0.0052
naive2 se	(0.011)	(0.0096)	(0.0108)	(0.011)	(0.0084)

Table 4.3: Censored Quantile Regression : Setting 1
L_1 Prediction Error

	23-30% Censoring				40-45% Censoring			
	n=500		n=1000		n=500		n=1000	
	$Q_{0.5}$	$Q_{0.7}$	$Q_{0.5}$	$Q_{0.7}$	$Q_{0.5}$	$Q_{0.7}$	$Q_{0.5}$	$Q_{0.7}$
Omni	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
CRQ	0.2	0.26	0.08	0.11	0.34	0.51	0.26	0.4
LCRQ	0.27	0.44	0.24	0.44	1.19	1.82	1.2	1.79
RFCRQ	0.24	0.15	0.16	0.07	0.17	0.25	0.08	0.32
Naive1	0.55	0.69	0.52	0.71	2.68	2.95	2.7	2.97
Naive2	0.55	0.69	0.52	0.71	2.68	2.95	2.7	2.97

4.5.2 Simulations for Quantile Treatment effect

In the treatment assignment setting, we will compare both the quantile regression coefficients as well the quantile treatment effect coefficients. We explore 2 settings, the first one a randomized trial and the second being an observational setup. We have already stated that our method is applicable in both settings. We will compare both the individual coefficient bias and standard error as well that the L_1 prediction error of the coefficient vector.

4.5.2.1 Simulation Setting 2

The covariates are generated as $X_1, X_2, \ldots, X_5 \sim \mathrm{N}(0_5, \Sigma)$ where Σ is an **AR(1)** covariance structure with $\sigma = 1$ and $\rho = 0.8$. $A \sim \mathrm{binom}(0.5)$

The true time to event is generated as

$\log(T) = 1 + X_1 + 0.5X_3 + 1.5X_5 + 0.5A(X_3 - X_5) + (0.3 + 2(X_3 - 0.5)^2)(Z - \phi^{-1}(\tau))$

where $Z \sim \mathrm{EVD}(0, 1, 0)$ and ϕ is the extreme value distribution.

The censoring time is generated as $\log(C) \sim 1.5 + X_2 + X_4 + 0.5(Z - \phi^{-1}(\tau))$.

The observed time is $Y_i = \min(T_i, C_i)$.

The results in next 3 tables are from 100 simulation runs.

Table 4.4: Setting 2 : Bias with $\tau = 0.5$ and $n = 1000$

	β_0	β_1	β_2	β_3	β_4	β_5
Omni bias	0.0036	-4e-04	0	-3e-04	0	-0.0032
Omni se	(0.0075)	(0.005)	(0)	(0.0193)	(0)	(0.0072)
crq bias	0.0725	0.0107	2e-04	-0.1581	-0.007	0.0053
crq se	(0.0121)	(0.0118)	(0.017)	(0.0227)	(0.0161)	(0.016)
lcrq bias	-0.1605	-0.1942	0.2024	-0.0835	0.2036	-0.3025
lcrq se	(0.0081)	(0.0114)	(0.0147)	(0.0177)	(0.012)	(0.0119)
rfcrq bias	0.0054	-0.0028	0.0129	-0.0289	-0.004	-0.0038
rfcrq se	(0.012)	(0.0149)	(0.0187)	(0.0224)	(0.0176)	(0.0179)
naive1 bias	-0.3113	-0.475	0.4907	-0.2466	0.4965	-0.7357
naive1 se	(0.0077)	(0.0093)	(0.0133)	(0.0141)	(0.0112)	(0.0114)
naive2 bias	-0.9268	-0.3313	0.348	0.7312	0.3767	-0.5041
naive2 se	(0.0122)	(0.0148)	(0.0186)	(0.0287)	(0.0218)	(0.0185)

	α_0	γ_1	γ_2	γ_3	γ_4	γ_5
Omni bias	0	0	0	0.0184	0	-0.0082
Omni se	(0)	(0)	(0)	(0.0213)	(0)	(0.0106)
crq bias	-0.0191	-0.021	0.0143	-0.0158	-0.0043	-0.0029
crq se	(0.019)	(0.0151)	(0.0237)	(0.0328)	(0.022)	(0.0209)
lcrq bias	0.0359	-0.0128	0.0018	-0.1183	-0.0072	0.0949
lcrq se	(0.0118)	(0.0137)	(0.0185)	(0.0269)	(0.0184)	(0.0158)
rfcrq bias	-0.005	-0.0189	0.0014	-3e-04	-0.008	-0.0061
rfcrq se	(0.0174)	(0.0182)	(0.0246)	(0.0331)	(0.0246)	(0.0229)
naive1 bias	0.0647	-0.0231	0.0094	-0.2509	-0.0108	0.2468
naive1 se	(0.0103)	(0.0133)	(0.0171)	(0.0197)	(0.0156)	(0.014)
naive2 bias	0.0602	-0.0104	-0.0121	-0.0888	-0.0417	0.1702
naive2 se	(0.0166)	(0.0202)	(0.023)	(0.0434)	(0.0303)	(0.0234)

Table 4.5: Setting 2 : Bias with $\tau = 0.7$ and $n = 1000$

	β_0	β_1	β_2	β_3	β_4	β_5
Omni bias	0.0105	0.0034	0	-0.0028	0	-0.0036
Omni se	(0.0099)	(0.0058)	(0)	(0.0228)	(0)	(0.0102)
crq bias	0.1224	0.0454	-0.0246	-0.2406	-0.0161	0.0291
crq se	(0.0154)	(0.0168)	(0.0228)	(0.0271)	(0.0199)	(0.018)
lcrq bias	-0.2579	-0.2735	0.2865	-0.0986	0.289	-0.4321
lcrq se	(0.0089)	(0.0118)	(0.0155)	(0.0178)	(0.0138)	(0.0143)
rfcrq bias	-0.032	-0.0318	0.0471	-0.0767	0.065	-0.0621
rfcrq se	(0.014)	(0.0159)	(0.0228)	(0.027)	(0.0188)	(0.0178)
naive1 bias	-0.377	-0.5109	0.5185	-0.2169	0.5137	-0.7699
naive1 se	(0.0079)	(0.0098)	(0.0125)	(0.014)	(0.0117)	(0.0126)
naive2 bias	-0.9741	-0.3474	0.3456	0.8595	0.3747	-0.5565
naive2 se	(0.01)	(0.0137)	(0.0193)	(0.0284)	(0.0186)	(0.0173)

	α_0	γ_1	γ_2	γ_3	γ_4	γ_5
Omni bias	0	0	0	0.0277	0	-0.0036
Omni se	(0)	(0)	(0)	(0.025)	(0)	(0.0141)
crq bias	-0.022	-0.0476	0.0312	-0.0078	-0.0051	-0.0276
crq se	(0.0251)	(0.0248)	(0.0317)	(0.0381)	(0.0291)	(0.0236)
lcrq bias	0.0386	-0.0243	0.0151	-0.1563	-0.0091	0.1323
lcrq se	(0.0136)	(0.016)	(0.0201)	(0.0258)	(0.0176)	(0.0174)
rfcrq bias	-0.0193	-0.0303	0.0304	0.0602	-0.0369	-0.0332
rfcrq se	(0.0209)	(0.0238)	(0.033)	(0.0415)	(0.0281)	(0.0239)
naive1 bias	0.0523	-0.0189	0.0116	-0.2668	-0.0082	0.2477
naive1 se	(0.0114)	(0.0149)	(0.0179)	(0.02)	(0.0147)	(0.0142)
naive2 bias	0.0525	-0.0142	0.0196	-0.1399	-0.0197	0.1961
naive2 se	(0.0156)	(0.019)	(0.0263)	(0.0453)	(0.026)	(0.0237)

Table 4.6: Censored Quantile Regression : Setting 2
L_1 Prediction Error

	Complete Quantile Model		Quantile Treatment Effect	
	$Q_{0.5}$	$Q_{0.7}$	$Q_{0.5}$	$Q_{0.7}$
Omni	0.03	0.05	0.03	0.03
CRQ	0.33	0.62	0.08	0.14
LCRQ	1.42	2.01	0.27	0.38
RFCRQ	0.1	0.53	0.04	0.21
Naive1	3.36	3.51	0.61	0.61
Naive2	3.36	3.51	0.61	0.61

4.5.2.2 Simulation Setting 3

We next explore a treatment assignment setting which depends on covariates. We consider the conditional logistic model which is such that higher the quantile treatment effect higher the probability of assigning treatment $A = 1$. This is one of the most natural form of confounding which may occur where the propensity score is proportional to the quantile treatment effect (in binary treatment case).

The covariates are generated as $X_1, X_2, \ldots, X_5 \sim \mathrm{N}(0_5, \Sigma)$ where Σ is an **AR(1)** covariance structure with $\sigma = 1$ and $\rho = 0.8$. $A \sim \text{logistic}(X_3 - X_5)$
The true time to event is generated as
$\log(T) = 1 + X_1 + 0.5X_3 + 1.5X_5 + 0.5A(X_3 - X_5) + (0.3 + 2(X_3 - 0.5)^2)(Z - \phi^{-1}(\tau))$
where $Z \sim \mathrm{EVD}(0, 1, 0)$ and ϕ is the extreme value distribution.
The censoring time is generated as $\log(C) \sim 1.5 + X_2 + X_4 + 0.5(Z - \phi^{-1}(\tau))$.
The observed time is $Y_i = \min(T_i, C_i)$. Here the quantile treatment effect is $(X_3 - X_5)$ and the propensity score is also proportional to $(X_3 - X_5)$. The results

in the next 3 tables are from 100 simulation runs.

Table 4.7: Setting 3 : Bias with $\tau = 0.5$ and $n = 1000$

	β_0	β_1	β_2	β_3	β_4	β_5
Omni bias	0.0047	-0.0014	0	-0.0028	0	-0.0108
Omni se	(0.0076)	(0.005)	(0)	(0.0184)	(0)	(0.0089)
crq bias	0.075	0.0035	-0.0022	-0.162	-0.0101	0.0089
crq se	(0.014)	(0.0131)	(0.0178)	(0.0245)	(0.0176)	(0.0171)
lcrq bias	-0.1578	-0.2215	0.2236	-0.0575	0.2186	-0.3368
lcrq se	(0.008)	(0.012)	(0.015)	(0.0186)	(0.0137)	(0.0134)
rfcrq bias	0.0087	-0.0109	0.0137	-0.0146	1e-04	-2e-04
rfcrq se	(0.0121)	(0.0158)	(0.0187)	(0.0231)	(0.0189)	(0.0195)
naive1 bias	-0.288	-0.5282	0.533	-0.2189	0.5233	-0.8053
naive1 se	(0.0079)	(0.01)	(0.0124)	(0.0139)	(0.0114)	(0.0111)
naive2 bias	-0.9462	-0.3485	0.3575	0.9421	0.3584	-0.5259
naive2 se	(0.0114)	(0.0183)	(0.0195)	(0.0268)	(0.0226)	(0.0207)

	α_0	γ_1	γ_2	γ_3	γ_4	γ_5
Omni bias	0	0	0	0.0239	0	0.0085
Omni se	(0)	(0)	(0)	(0.0233)	(0)	(0.0129)
crq bias	-0.011	-0.0015	0.0165	-0.0323	0.0085	-0.0119
crq se	(0.0199)	(0.0194)	(0.0259)	(0.0389)	(0.0227)	(0.0224)
lcrq bias	0.0386	0.0237	-0.0234	-0.1791	-0.0206	0.1366
lcrq se	(0.0127)	(0.0163)	(0.0215)	(0.0294)	(0.0186)	(0.0173)
rfcrq bias	-0.0055	-0.006	0.004	-0.0588	0.0031	-0.0164
rfcrq se	(0.0186)	(0.021)	(0.027)	(0.0375)	(0.0243)	(0.0248)
naive1 bias	0.0398	0.0442	-0.0477	-0.2966	-0.0416	0.3301
naive1 se	(0.0116)	(0.0135)	(0.0175)	(0.021)	(0.0153)	(0.0135)
naive2 bias	0.1017	-0.0153	0.0041	-0.4885	0.0054	0.1798
naive2 se	(0.0172)	(0.0245)	(0.027)	(0.0418)	(0.0298)	(0.0251)

Table 4.8: Setting 3 : Bias with $\tau = 0.7$ and $n = 1000$

	β_0	β_1	β_2	β_3	β_4	β_5
Omni bias	0.0108	0.0016	0	0.0018	0	-0.0126
Omni se	(0.01)	(0.006)	(0)	(0.0242)	(0)	(0.0118)
crq bias	0.1232	0.0219	-0.0011	-0.2249	-0.0458	0.0429
crq se	(0.0153)	(0.0174)	(0.0228)	(0.0304)	(0.025)	(0.0257)
lcrq bias	-0.2615	-0.3157	0.313	-0.0437	0.3128	-0.4793
lcrq se	(0.0082)	(0.013)	(0.0165)	(0.0219)	(0.0161)	(0.0158)
rfcrq bias	-0.0458	-0.0682	0.0749	-0.0181	0.049	-0.061
rfcrq se	(0.0117)	(0.0177)	(0.0225)	(0.0312)	(0.0255)	(0.0241)
naive1 bias	-0.3633	-0.5552	0.5517	-0.18	0.5384	-0.8331
naive1 se	(0.0075)	(0.0104)	(0.0125)	(0.0162)	(0.0125)	(0.0122)
naive2 bias	-0.9961	-0.3708	0.3791	1.0761	0.3682	-0.5554
naive2 se	(0.0096)	(0.0147)	(0.0188)	(0.0297)	(0.0163)	(0.0174)

	α_0	γ_1	γ_2	γ_3	γ_4	γ_5
Omni bias	0	0	0	0.0254	0	0.016
Omni se	(0)	(0)	(0)	(0.0304)	(0)	(0.0159)
crq bias	-0.008	0.011	0.006	-0.0632	0.0282	-0.0244
crq se	(0.021)	(0.0258)	(0.0322)	(0.0462)	(0.0305)	(0.032)
lcrq bias	0.0489	0.0366	-0.0171	-0.2706	-0.0278	0.1932
lcrq se	(0.0119)	(0.0186)	(0.0243)	(0.0333)	(0.0204)	(0.019)
rfcrq bias	0.0032	0.0456	-0.0139	-0.0799	0.0098	-0.0345
rfcrq se	(0.0181)	(0.0275)	(0.0328)	(0.048)	(0.0324)	(0.0305)
naive1 bias	0.0438	0.0466	-0.0322	-0.3465	-0.0368	0.3327
naive1 se	(0.011)	(0.0147)	(0.0186)	(0.0243)	(0.0177)	(0.0155)
naive2 bias	0.1175	0.0072	-0.0161	-0.627	0.022	0.1757
naive2 se	(0.0143)	(0.0204)	(0.0256)	(0.0421)	(0.0237)	(0.0217)

Table 4.9: Censored Quantile Regression : Setting 3
L_1 Prediction Error

	Complete Quantile Model		Quantile Treatment Effect	
	$Q_{0.5}$	$Q_{0.7}$	$Q_{0.5}$	$Q_{0.7}$
Omni	0.05	0.07	0.03	0.04
CRQ	0.34	0.6	0.08	0.14
LCRQ	1.64	2.32	0.42	0.59
RFCRQ	0.14	0.5	0.09	0.19
Naive1	3.7	3.86	0.8	0.84
Naive2	3.7	3.86	0.8	0.84

For both the randomized trial and observational case we note that the naive methods and the local method of Wang and Wang [2009] have high bias whereas Portnoy's method is unbiased for most coefficients(except β_0 and β_3). The proposed method however is unbiased for all coefficients. When comparing the L_1 aggregated error in β estimation we notice that the proposed method performs similar to Portnoy's method when comparing only the quantile treatment effect coefficients. This is because Portnoy's method has very low bias for those coefficients.

However when interpreting a model all the coefficients contribute to the interpretation as we will see in the next section. The quantile treatment effect will help make treatment recommendations whereas the individual coefficients will help understand the effect of all the covariates at each quantile. Hence even in simulation settings we focus on the individual coefficients in complete model as well as for the treatment effect. As we move away from the median, the bias increases hence we would explore multiple quantile models to have a more complete picture.

4.6 Real Data Analysis

We explore the colon cancer dataset(Laurie et al. [1989] in survival package in R Therneau [2021]) to demonstrate our proposed method. For real data analysis, our proposed method needs to be adjusted to the context of the data while choosing the terms for regression. We suggest screening variables for the random forest if dimensions are greater than 20, via methods such as marginal screening using marginal p-values or random forest variable importance. For moderately low dimensional data random forest would pick up the signals but for the second step we would need to choose the terms when fitting the parametric model. Although both survival time and any increasing monotonic transformation of it can be represented as survival time framework but in this case the parametric model is sensitive to model specification. Both the issues stated, could be addressed by a combination of graphical representation of associations within the data and goodness of fit using measures such as model AIC/BIC/SIC and quantile loss. We use goodness of fit to choose the transformation of the response and association plots to choose the terms for regression, further we use step-wise model selection to reach the final quantile regression model. Further baseline attributes with less explanatory power i.e. larger p-value were dropped from the model sequentially starting from the one with highest p-value till all p-values are < 0.1.

4.6.1 Colon Cancer Data Analysis

This data is from one of the first successful trials of adjuvant chemotherapy for colon cancer, the first study being originally described in Laurie et al. [1989]. The dataset has evolved and analyzed in many papers, we use the dataset for chemotherapy for stage B/C colon cancer in the survival package in R. Levamisole

is a low-toxicity compound which was used to treat worm infestations in animals and 5-FU is a moderately toxic chemotherapy agent. The records of patients with event as death is considered (not recurrence). This is a randomized trial. We consider this dataset to illustrate the performance of our proposed method in case of more than 2 treatments.

There are 3 kinds of treatments, Obs (placebo), Lev, Lev+5FU and 11 clinical covariates. We have data for total 888 patients (complete cases), 305 on placebo, 294 given Lev, 289 given Lev+FU. The overall censoring rate is 51.57%, with censoring rate of patients in Obs group being 46.22%, censoring rate of patients on Levamisole being 49.31%, and that of patients on Levamisole + 5-FU being 59.51%.

Figure 4.1: Colon Cancer Survival : Treatment Groups

	Obs (N=305)	Lev (N=294)	Lev+5FU (N=289)	Overall (N=888)
time				
Mean (SD)	1600 (856)	1630 (894)	1800 (858)	1670 (873)
Median [Min, Max]	1850 [113, 3210]	1940 [24.0, 3330]	2100 [23.0, 3310]	1980 [23.0, 3330]
status				
0	141 (46.2%)	145 (49.3%)	172 (59.5%)	458 (51.6%)
1	164 (53.8%)	149 (50.7%)	117 (40.5%)	430 (48.4%)

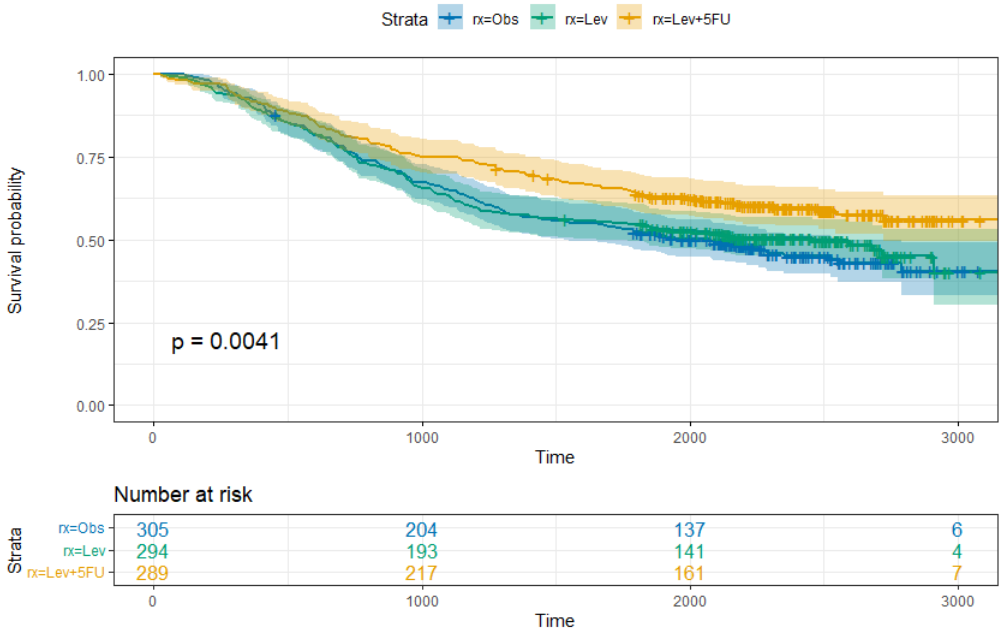


Figure 4.2: Colon Cancer Survival curves

We first look at the coefficients at each quantile and then proceed to inference on treatment effects.

Table 4.10: Colon Cancer : Censored Quantile Regression

	log(Time)		
	$Q_{0.25}$	$Q_{0.5}$	$Q_{0.75}$
sex1	−0.023	−0.107	0.015
	(0.116)	(0.106)	(0.047)
obstruct1	−0.450***		
	(0.079)		
age		0.000	−0.0003
		(0.003)	(0.002)
perfor1			−0.375*
			(0.218)
adhere1		−0.227	−0.001
		(0.157)	(0.078)
nodes	−0.106***	−0.113***	−0.075***
	(0.009)	(0.011)	(0.018)
surg1	−0.145	−0.114	−0.054
	(0.092)	(0.078)	(0.055)
differ11	−0.470***	−0.304**	−0.347**
	(0.170)	(0.154)	(0.172)
extent11	−0.546***		−0.062
	(0.118)		(0.053)
aa2	−0.029	0.004	−0.166*
	(0.126)	(0.166)	(0.085)
aa3	−0.133	0.006	−0.085
	(0.148)	(0.159)	(0.089)
sex1:aa2	−0.180	0.060	
	(0.184)	(0.173)	
sex1:aa3	0.465**	0.251	
	(0.205)	(0.161)	
perfor1:aa2			0.316
			(0.368)
perfor1:aa3			0.493**
			(0.234)
nodes:aa2		−0.005	0.045**
		(0.038)	(0.022)
nodes:aa3		0.047	0.046
		(0.034)	(0.029)
differ11:aa2			0.303
			(0.204)
differ11:aa3			0.376*
			(0.215)
Constant	7.869***	8.056***	8.255***
	(0.129)	(0.177)	(0.135)

Note: *p<0.1; **p<0.05; ***p<0.01

Denoting Q_τ as the τ^{th} quantile of survival time $\log(T)$ given baseline covariates,

$$
\begin{aligned}
&\mathrm{Q}_{0.25} = -0.023\mathbb{I}(\text{Gender=Male}) - 0.45\text{Obstruct} - 0.106\text{Nodes} \\
&- 0.145\text{Surg} - 0.47\text{Differ} - 0.546\text{Extent} \\
&+ \mathbb{I}(\text{trt} = \text{Lev})(-0.029 - 0.18\mathbb{I}(\text{Gender=Male})) \\
&+ \mathbb{I}(\text{trt} = \text{Lev+5FU})(-0.133 + 0.465\mathbb{I}(\text{Gender=Male})) \\
&\mathrm{Q}_{0.5} = -0.107\mathbb{I}(\text{Gender=Male}) - 0.227\text{Adhere} - 0.113\text{Nodes} \\
&- 0.114\text{Surg} - 0.304\text{Differ} \\
&+ \mathbb{I}(\text{trt} = \text{Lev})(0.004 + 0.06\mathbb{I}(\text{Gender=Male}) - 0.005\text{Nodes} \\
&\mathbb{I}(\text{trt} = \text{Lev+5FU})(0.006 + 0.251\mathbb{I}(\text{Gender=Male}) + 0.047\text{Nodes}) \\
&\mathrm{Q}_{0.75} = 0.015\mathbb{I}(\text{Gender=Male}) - 0.375\text{Perfor} - 0.001\text{Adhere} - 0.075\text{Nodes} \\
&- 0.0003\text{Age} - 0.054\text{Surg} - 0.347\text{Differ} - 0.062\text{Extent} \\
&+ \mathbb{I}(\text{trt} = \text{Lev})(-0.166 + 0.316\text{Perfor} + 0.045\text{Nodes} + 0.303\text{Differ}) \\
&+ \mathbb{I}(\text{trt} = \text{Lev+5FU})(-0.085 + 0.493\text{Perfor} + 0.046\text{Nodes} + 0.376\text{Differ})
\end{aligned}
\tag{4.6.1}
$$

Attributes such as obstruction of colon by tumor, number of nodes, time from surgery, differentiation of tumor and extent of local spread have negative correlation with the 25% quantile. The 25% quantile of survival times represent weaker patients among which we see that Levamisole is more effective than Levamisole+5FU (although this effect is not significant). Males have higher survival in comparison to females when treated with Levamisole+5FU (interaction of gender and treatment is significant at 0.05 level of significance).

At higher quantiles 50% and 75%, perforation of colon and adherence to nearby organs also have negative association with survival time. In comparison to higher quantiles which represent patients who may live longer, weaker patients do not

seem to have much benefit from any of the treatments. However at higher quantiles we find significant interaction of treatments and disease related attributes. We now focus on quantile treatment effects, where we compare each treatment with the control(no treatment). Follwoing are the quantile treatment effects derived from the complete conditional models,

$$
\begin{aligned}
&\mathrm{QTE}^{0.25}_{\mathrm{Lev}} \\
&= -0.029 - 0.18\mathbb{I}(\mathrm{Gender{=}Male}) \\
&\mathrm{QTE}^{0.25}_{\mathrm{Lev{+}5FU}} \\
&= -0.133 + 0.465\mathbb{I}(\mathrm{Gender{=}Male}) \\
&\mathrm{QTE}^{0.5}_{\mathrm{Lev}} \\
&= 0.004 + 0.06\mathbb{I}(\mathrm{Gender{=}Male}) - 0.005\mathrm{Nodes} \\
&\mathrm{QTE}^{0.5}_{\mathrm{Lev{+}5FU}} \\
&= 0.006 + 0.251\mathbb{I}(\mathrm{Gender{=}Male}) + 0.047\mathrm{Nodes} \\
&\mathrm{QTE}^{0.75}_{\mathrm{Lev}} \\
&= -0.166 + 0.316\mathrm{Perfor} + 0.045\mathrm{Nodes} + 0.303\mathrm{Differ} \\
&\mathrm{QTE}^{0.75}_{\mathrm{Lev{+}5FU}} \\
&= -0.085 + 0.493\mathrm{Perfor} + 0.046\mathrm{Nodes} + 0.376\mathrm{Differ}
\end{aligned}
\tag{4.6.2}
$$

For higher quantiles 50% and 75% we see that Lev+5FU dampens the impact of number of lymph nodes with detectable cancer. At 75% both Levamisole and Lev+5FU have less negative effect of perforation of colon and poor differentiation of tumour. The conditional quantile treatment effects also imply that for patients with less perforation of colon, less nodes impacted with cancer and well differentiated cells, the treatments Levamisole and Lev+5FU are not suggested.

4.7 Discussion

Comparing censored quantile regression methods, we observe that the local method of Wang and Wang [2009] is not preferable for dimensions higher than 2-3 due to curse of dimensionality. This method is also not appropriate for categorical covariates. In lower dimensions, random forest performs better when heterogeneity is present. Portnoy [2003] assumes global linearity, hence is impacted by heterogeneity. Computation time of the censored quantile regression algorithm by Portnoy [2003] higher in comparison to other methods.

We observe that bias more for quantiles away from 0.5. Random forest is consistent in comparison to other methods. (we compare n=500 vs 1000 for simulation setting 2) We also consider different levels of censoring. In the case of estimation of quantile treatment effects, we explore both randomized treatment assignments and observational data with different levels of confounding. We further study sensitivity of the proposed method to missing confounders and misspecified models.

The proposed method is valid for observational data as well. We would need regularization such as LASSO for second step in very high dimensions. This study can be further extended to accommodate dimensions than 20-25. This method can also be extended to multiple treatments by defining quantile treatment effects as difference between treatments and control, one treatment at a time.

Further as per our setup, quantile method is robust to monotonic transformations and less dependent on error distributions than ATE and hence applicable in a few common settings such as AFT with various distributions. When implementing this method for real data, we would first need to screen for terms for random forest, this can be done using marginal p-values or Random Forest variable importance. Further we would need to look at plots for quantile

regression terms. The final model interaction terms can then be used to suggest personalized optimal treatments at specific quantiles. Looking at different quantiles might yield different regimes. One would then have to make sense of the meaning of quantiles in context of the data to make recommendations.

Chapter 5

Discussion and Areas for Future Work

5.1 Conclusions

When modeling conditional survival times, there are various models to consider depending on the situation. Cox model, accelerated failure time, survival trees, survival forests etc are a few such models. Semiparametric Cox model has the proportional hazards assumption and accelerated failure time model is a paramteric survival model which also needs a error model specified. The results in chapter 2 show that survival random forest does better than a misspecified Cox proportional hazards model. The differences are more when the restrictions are higher and noise is more. A correctly specified parametric model performs best but is difficult to achieve. In high dimensions we may do a regularized semiparametric form as L_1 penalized Cox regression or non-parametric methods such as survival random forests. When the dimension of the feature/covariates space is larger than the sample size, penalized regression do not provide unique solutions. Hence for such cases we might want to use non-parametric methods. Other than data assumptions, non parametric methods also do not restrict the class of treatment regimes since we wish to optimize treatment regimes.

More than often it is not the case of low dimensional data where it is possible to explore the effect of covariates and treatments on the outcome graphically and build an efficient parametric model. Machine learning algorithms such as random forests help those not familiar with the complications of statistical theory and model assumptions or when data assumptions such as Gaussian error do not hold. However there are various other such possible models which can be considered in different situations and compared to arrive at a conclusion when dealing with real datasets.

5.2 Future Work

A possible future work consideration might be to have a setting where treatment assignments can be made in multiple stages and also have time dependent covariates such that the optimal treatment regime adjusts at every stage. Also this case only considers discrete and finite set of treatments $\mathcal{A}$, whereas we might be interested in a further complicated setup of continuous treatments. The assumption of unmeasured confounders can be very restrictive in complex confounding structures such as reverse causation, in which case the estimators proposed do not suffice.

Chapter 3 can be extended by changing the random forest algorithm such that it is built from honest trees. Asymptotic weak convergence of the distribution function to a gaussian process once derived, one would then need to invert the distribution function to show the assymptotic normality of the quantile process estimate from quantile random forest.

The 2 step method proposed in chapter 4 estimate the survival function using random forest but the next step is a parametric quantile regression setup. In case of high dimensions, regularization might be necessary. This method can be extended to a L_1-regularized censored quantile regression setting in such cases. In real data cases it is often difficult to estimate the exact parametric terms where additive quantile regression(Fasiolo et al. [2017]) or splines might be a good option to consider for modeling.

References

Alberto Abadie, Joshua Angrist, and Guido Imbens. Instrumental variable estimates of the effect of subsidized training on the quantile of trainee earnings. *Econometrica*, 70:91–117, 02 2002. doi: 10.2139/ssrn.195733.

Dhammika Amaratunga, Javier Cabrera, and Yung-Seop Lee. Enriched random forests. *Bioinformatics (Oxford, England)*, 24:2010–4, 09 2008. doi: 10.1093/bioinformatics/btn356.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. doi: 10.1080/01621459.1996.10476902. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1996.10476902`.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. ISSN 0027-8424. doi: 10.1073/pnas.1510489113. URL `https://www.pnas.org/content/113/27/7353`.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests, 2018.

Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(66): 2015–2033, 2008. URL http://jmlr.org/papers/v9/biau08a.html.

Gérard Biau and Luc Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101:2499–2518, 11 2010. doi: 10.1016/j.jmva.2010.06.019.

Gérard Biau and Erwan Scornet. A random forest guided tour. *TEST*, 25, 11 2015. doi: 10.1007/s11749-016-0481-7.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

L. Breiman, J. Friedman, R. Olshen, and C. J. Stone. Classification and regression trees. 1983.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Brantly Callaway. Quantile treatment effects in r: The qte package. 2019.

Xiaohong Chen, Oliver Linton, and Ingrid Van Keilegom. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica*, 71(5):1591–1608, 2003. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1555514.

Victor Chernozhukov and Christian Hansen. An iv model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/3598944.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972. ISSN 00359246. URL http://www.jstor.org/stable/2985181.

Yifan Cui, Ruoqing Zhu, Mai Zhou, and Michael Kosorok. Some asymptotic results of survival tree and forest models. 07 2017.

Kjell Doksum. Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, 2(2):267–277, 1974. ISSN 00905364. URL http://www.jstor.org/stable/2958036.

Bradley Efron. The two-sample problem with censored data. *Proceedings Fifth Berkeley Symposium in Mathematical Statistics*, IV:831–853, 1967.

Matteo Fasiolo, Yannig Goude, Raphael Nedellec, and Simon N. Wood. *Fast calibrated additive quantile regression.*, 2017. URL https://arxiv.org/abs/1707.03307.

Sergio Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276, 2007. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/4123114.

Jerome H. Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010. ISSN 1548-7660. doi: 10.18637/jss.v033.i01. URL https://www.jstatsoft.org/v033/i01.

Ali Gannoun, Jérôme Saracco, and Keming Yu. Nonparametric prediction by conditional median and quantiles. *Journal of Statistical Planning and Inference*, 117:207–223, 12 2003. doi: 10.1016/S0378-3758(02)00384-1.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63:3–42, 04 2006. doi: 10.1007/s10994-006-6226-1.

Yair Goldberg and Michael R. Kosorok. Q-learning with censored data. *Ann. Statist.*, 40(1):529–560, 02 2012. doi: 10.1214/12-AOS968. URL `https://doi.org/10.1214/12-AOS968`.

Paul R. Halmos. The Theory of Unbiased Estimation. *The Annals of Mathematical Statistics*, 17(1):34 – 43, 1946. doi: 10.1214/aoms/1177731020. URL `https://doi.org/10.1214/aoms/1177731020`.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Paul Hendricks. *titanic: Titanic Passenger Survival Data Set*, 2015. URL `https://CRAN.R-project.org/package=titanic`. R package version 0.1.0.

Wassily Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293 – 325, 1948. doi: 10.1214/aoms/1177730196. URL `https://doi.org/10.1214/aoms/1177730196`.

Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986. ISSN 01621459. URL `http://www.jstor.org/stable/2289064`.

Guido Imbens and Joshua Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–75, 1994. URL `https://EconPapers.repec.org/RePEc:ecm:emetrp:v:62:y:1994:i:2:p:467-75`.

Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):

5–86, March 2009. doi: 10.1257/jel.47.1.5. URL https://www.aeaweb.org/articles?id=10.1257/jel.47.1.5.

H. Ishwaran and U.B. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2021. URL https://cran.r-project.org/package=randomForestSRC. R package version 2.11.0.

Hemant Ishwaran and Udaya B. Kogalur. Consistency of random survival forests. *Statistics and Probability Letters*, 80(13-14):1056–1064, 7 2010. ISSN 0167-7152. doi: 10.1016/j.spl.2010.02.020.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *Ann. Appl. Stat.*, 2(3):841–860, 09 2008. doi: 10.1214/08-AOAS169. URL https://doi.org/10.1214/08-AOAS169.

Hemant Ishwaran, Thomas A. Gerds, Udaya B. Kogalur, Richard D. Moore, Stephen J. Gange, and Bryan M. Lau. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 04 2014. ISSN 1465-4644. doi: 10.1093/biostatistics/kxu010. URL https://doi.org/10.1093/biostatistics/kxu010.

E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. ISSN 01621459. URL http://www.jstor.org/stable/2281868.

Roger Koenker. *quantreg: Quantile Regression*. URL https://CRAN.R-project.org/package=quantreg. R package version 5.86.

Roger Koenker. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. doi: 10.1017/CBO9780511754098.

Roger Koenker. Censored quantile regression redux. *Journal of Statistical Software, Articles*, 27(6):1–25, 2008. ISSN 1548-7660. doi: 10.18637/jss.v027.i06. URL https://www.jstatsoft.org/v027/i06.

Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1): 33–50, 1978. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1913643.

Roger Koenker and Gilbert Bassett. Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 50(1):43–61, 1982. URL https://EconPapers.repec.org/RePEc:ecm:emetrp:v:50:y:1982:i:1:p:43-61.

Roger Koenker and Olga Geling. Reappraising medfly longevity: A quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454):458–468, 2001. ISSN 01621459. URL http://www.jstor.org/stable/2670284.

Max Kuhn and Davis Vaughan. *parsnip: A Common API to Modeling and Analysis Functions*, 2021. URL https://CRAN.R-project.org/package=parsnip. R package version 0.1.6.

J A Laurie, C G Moertel, T R Fleming, H S Wieand, J E Leigh, J Rubin, G W McCormack, J B Gerstner, J E Krook, and J Malliard. Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. the north central cancer treatment group and the mayo clinic. *Journal of Clinical Oncology*, 7(10):1447–1456, 1989. doi: 10.1200/JCO.1989.7.10.1447. URL https://doi.org/10.1200/JCO.1989.7.10.1447. PMID: 2778478.

E.L. Lehmann and H.J.M. D'Abrera. *Nonparametrics: Statistical Methods Based*

on Ranks. Prentice Hall, 1998. ISBN 9780139977350. URL https://books.google.com/books?id=zNNFAQAAIAAJ.

Chenlei Leng and Xingwei Tong. A quantile regression estimator for censored data. *Bernoulli*, 19(1):344 – 361, 2013. doi: 10.3150/11-BEJ388. URL https://doi.org/10.3150/11-BEJ388.

Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006. ISSN 01621459. URL http://www.jstor.org/stable/27590719.

Anna Lindgren. Quantile regression with censored data using generalized l1 minimization. *Computational Statistics Data Analysis*, 23(4):509–524, 1997. ISSN 0167-9473. doi: https://doi.org/10.1016/S0167-9473(96)00048-5. URL https://www.sciencedirect.com/science/article/pii/S0167947396000485.

Grant McDermott. *parttree: Simple functions for plotting 2D decision tree partition plots*, 2021. URL https://github.com/grantmcdermott/parttree.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35):983–999, 2006. URL http://jmlr.org/papers/v7/meinshausen06a.html.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17(26):1–41, 2016. URL http://jmlr.org/papers/v17/14-168.html.

Robert Messenger and Lewis Mandell. A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association*, 67(340):768–772, 1972. doi: 10.1080/01621459.1972.10481290. URL https://doi.org/10.1080/01621459.1972.10481290.

Stephen Milborrow. *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*, 2020. URL https://CRAN.R-project.org/package=rpart.plot. R package version 3.0.9.

J. Morgan and J. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415–434, 1963.

Tereza Neocleous, Karlien Vanden Branden, and Stephen Portnoy. Correction to censored regression quantiles by s. portnoy, 98 (2003), 1001–1012. *Journal of the American Statistical Association*, 101(474):860–861, 2006. doi: 10.1198/016214506000000087. URL https://doi.org/10.1198/016214506000000087.

PCAST (President's Council of Advisors on Science and Technology). Priorities for personalized medicine. president's council of advisors on science and technology. Sep 2008.

National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. National Academies Press (US), 2011. ISBN 978-0-309-22222-8. doi: 10.17226/13284.

Limin Peng and Yijian Huang. Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 103(482):637–649, 2008. ISSN 01621459. URL http://www.jstor.org/stable/27640086.

Stephen Portnoy. Censored regression quantiles. *Journal of the American Statistical Association*, 98(464):1001–1012, 2003. ISSN 01621459. URL http://www.jstor.org/stable/30045346.

David Powell. *A New Framework for Estimation of Quantile Treatment Effects: Nonseparable Disturbance in the Presence of Covariates*. RAND Corporation, Santa Monica, CA, 2013. doi: 10.7249/WR824-1.

James L. Powell. Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25(3):303–325, 1984. URL https://EconPapers.repec.org/RePEc:eee:econom:v:25:y:1984:i:3:p:303-325.

James L. Powell. Censored regression quantiles. *Journal of Econometrics*, 32(1):143–155, 1986. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(86)90016-3. URL https://www.sciencedirect.com/science/article/pii/0304407686900163.

Min Qian and Susan A. Murphy. Performance guarantees for individualized treatment rules. *Ann. Statist.*, 39(2):1180–1210, 04 2011. doi: 10.1214/10-AOS864. URL https://doi.org/10.1214/10-AOS864.

D. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

Donald B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26, 1977. doi: 10.3102/10769986002001001. URL https://doi.org/10.3102/10769986002001001.

A. G. Stephenson. evd: Extreme value distributions. *R News*, 2(2), June 2002. URL https://CRAN.R-project.org/doc/Rnews/.

Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2019. URL https://CRAN.R-project.org/package=rpart. R package version 4.1-15.

Terry M Therneau. *A Package for Survival Analysis in R*, 2021. URL https://CRAN.R-project.org/package=survival. R package version 3.2-11.

Stefan Wager. Asymptotic theory for random forests, 2016.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839. URL https://doi.org/10.1080/01621459.2017.1319839.

Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests, 2016.

Stefan Wager, Trevor Hastie, and Bradley Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(48):1625–1651, 2014. URL http://jmlr.org/papers/v15/wager14a.html.

Huixia Judy Wang and Lan Wang. Locally weighted censored quantile regression. *Journal of the American Statistical Association*, 104(487):1117–1128, 2009. doi: 10.1198/jasa.2009.tm08230. URL https://doi.org/10.1198/jasa.2009.tm08230.

Andrew Wey, David M. Vock, John Connett, and Kyle Rudser. Estimating restricted mean treatment effects with stacked survival models. *Statistics in Medicine*, 35(19):3319–3332, 2016. doi: 10.1002/sim.6929. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6929.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Wikipedia contributors. Generalized extreme value distribution — Wikipedia, the free encyclopedia, 2021. URL https://en.wikipedia.org/w/index.php?title=Generalized_extreme_value_distribution&oldid=1016905218. [Online; accessed 28-May-2021].

Z. Ying, S. H. Jung, and L. J. Wei. Survival analysis with median regression models. *Journal of the American Statistical Association*, 90(429):178–184, 1995. ISSN 01621459. URL http://www.jstor.org/stable/2291141.

Donglin Zeng. Estimating marginal survival function by adjusting for dependent censoring using many covariates. *Ann. Statist.*, 32(4):1533–1555, 08 2004. doi: 10.1214/009053604000000508. URL https://doi.org/10.1214/009053604000000508.

Baqun Zhang, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012. doi: 10.1111/j.1541-0420.2012.01763.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2012.01763.x.

Y. Q. Zhao, D. Zeng, E. B. Laber, R. Song, M. Yuan, and M. R. Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168, 12 2014. ISSN 0006-3444. doi: 10.1093/biomet/asu050. URL https://doi.org/10.1093/biomet/asu050.

Yingqi Zhao, Donglin Zeng, A John Rush, and Michael R Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the*

American Statistical Association, 107(449):1106—1118, September 2012. ISSN 0162-1459. doi: 10.1080/01621459.2012.695674. URL http://europepmc.org/articles/PMC3636816.

Yufan Zhao, Donglin Zeng, Mark A. Socinski, and Michael R. Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011. doi: 10.1111/j.1541-0420.2011.01572.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2011.01572.x.

Chang-Qi Zhu, Keyue Ding, Dan Strumpf, Barbara A. Weir, Matthew Meyerson, Nathan Pennell, Roman K. Thomas, Katsuhiko Naoki, Christine Ladd-Acosta, Ni Liu, Melania Pintilie, Sandy Der, Lesley Seymour, Igor Jurisica, Frances A. Shepherd, and Ming-Sound Tsao. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non–small-cell lung cancer. *Journal of Clinical Oncology*, 28(29):4417–4424, 2010. doi: 10.1200/JCO.2009.26.4325. URL https://doi.org/10.1200/JCO.2009.26.4325. PMID: 20823422.

Ruoqing Zhu, Donglin Zeng, and Michael R. Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015. doi: 10.1080/01621459.2015.1036994. URL https://doi.org/10.1080/01621459.2015.1036994. PMID: 26903687.

Appendix A

Notations for Chapters 2 and 4

Symbol	Description
n	Training size
a	Treatment
$\mathcal{A}$	Treatment set $\{0, \ldots, K-1\}$
X	Covariates
$\mathcal{X}$	Space of covariates
d	Dimension of covariate space
g	Treatment regime $g : \mathcal{X} \to \mathcal{A}$
Y	Observed outcome
$Y^*(a)$	Potential outcome under treatment a
$Y^*(g)$	Potential outcome under treatment regime g
T	Time to event
$T^*(a)$	Potential time to event under treatment a
$T^*(g)$	Potential time to event under treatment regime g
C	Censoring time
$C^*(a)$	Potential censoring time under treatment a

$C^*(g)$	Potential censoring time under treatment regime g
$S_a(.\|x)$	Survival function under treatment a and covariates x
$\lambda_a(.\|x)$	Hazard function under treatment a and covariates x
$\Lambda_a(.\|x)$	Cumulative hazard function under treatment a and covariates x
$\boldsymbol{\tau}$	Restriction for Chapter 2
$\boldsymbol{\tau}$	Quantile for Chapter 4
$\widetilde{\mathbf{T}}$	Restricted survival time $\min(T, \boldsymbol{\tau})$ for Chapter 2
$\widetilde{\mathbf{T}}^*(a)$	Potential restricted survival time under a for Chapter 2
$\widetilde{\mathbf{T}}^*(g)$	Potential restricted survival time under treatment regime g for Chapter 2
$\widetilde{\mathbf{T}}$	$h(T)$ for Chapter 4
$\widetilde{\mathbf{T}}^*(a)$	Potential $\widetilde{\mathbf{T}}$ under treatment a for Chapter 4
$\widetilde{\mathbf{T}}^*(g)$	Potential $\widetilde{\mathbf{T}}$ under treatment regime g for Chapter 4
$\mathbf{V}(g)$	Restricted mean survival under treatment regime g
$\boldsymbol{\mu}_a(X)$	Restricted mean survival under treatment a and covariates
α	minimum proportion of training samples of the parent node in the child node
k	minimum number of samples in terminal nodes
$\boldsymbol{\tau}, M_0, L_1, \zeta,$ $c, \boldsymbol{\gamma}, c_1, c_2, c_3, c_4$	constants

Appendix B

Notations for Chapter 3

Symbol	Description
n	Training size
X	Covariates
$\mathcal{X}$	Space of covariates
p	Dimension of covariate space
Y	Observed outcome $\in \mathbb{R}$
k_n	Training size used to build a tree
m_n	Number of trees in the forest
$F(y\|X = x)$	Conditional cumulative function of Y
$q(\tau\|X = x)$	Conditional τ^{th} quantile of Y
l	leaf of tree
ω	Randomization parameter denoting how tree was constructed
$T_{x,k,y}$	Estimate of $F(y\|x)$ from a tree build from training size k at y given x.
$r_{n,k_n,m_n}(y\|x)$	Estimate of $F(y\|x)$ from random forest at y given x.
$q_{n,k_n,m_n}(y\|x)$	Estimate of $Q(y\|x)$ from random forest at y given x.

Appendix C

Simulation Setting : Chapter 2

Predictors $X_1, X_2, \cdots, X_p; A$ are generated in two steps.

We first generate $(\tilde{X}_1, \tilde{X}_2, \cdots, \tilde{X}_p)^T$ from the multivariate uniform distribution such that $\tilde{X}_i \sim$ Unifrom$[-1.5, 1.5]$ for all $1 \leq i \leq k$ with $\text{Cov}(\tilde{X}_j, \tilde{X}_k) = 0$, $1 \leq j, k \leq p$.

The next step is to generate $A \in \{0, 1\}$ given $(X_1, X_2, \cdots, X_p)$ as $A|X_1, X_2, \cdots, X_p \sim$ Binomial$(\frac{1}{2})$. The scalar response is generated according to the model

$$Y = 2 - 1.5X_1^2 - 1.5X_2^2 + 3X_1X_2 + a(-0.1 - X_1 + X_2 + 2X_1X_2) + \epsilon,$$

where $\epsilon \sim N(0, 1)$ is independent of the covariates. The global optimal regime is $I(-0.1 - x_1 + x_2 + 2x_1x_2 > 0)$.

For each treatment we fit a tree/random forest model for the response Y based on the predictor X variables, model0 for $a = 0$ and model1 for $a = 1$. The predictions $\tilde{Y}_0$ and $\tilde{Y}_1$ from each of the models are then compared to estimate $\max_g \mathbb{E}_{\mathbf{X}}[Y^*(g)|X]$.

Appendix D

D.1 Proof of Proposition 2.3.3

Proof: Using the consistency assumption, we have

$$\tilde{T}^*(g) = \sum_{a=0}^{K-1} \mathbb{I}(g(X) = a)\tilde{T}^*(a)$$

$$\begin{aligned} \mathbf{V}(g) &= \mathbb{E}[\tilde{T}^*(g)] \\ &= \mathbb{E}[\sum_{a=0}^{K-1} \mathbb{I}(g(X) = a)\tilde{T}^*(a)] \\ &= \mathbb{E}_{\mathbf{X}}\mathbb{E}[\sum_{a=0}^{K-1} \mathbb{I}(g(X) = a)\tilde{T}^*(a)|X] \\ &= \mathbb{E}_{\mathbf{X}}[\sum_{a=0}^{K-1} \mathbb{I}(g(X) = a)\mathbb{E}[\tilde{T}^*(a)|X]]. \end{aligned} \tag{D.1.1}$$

where $\tilde{T}^*(a)$ is the survival time if treatment a is assigned.

Denote $g^*(X) = \arg\max_{0\le i\le K-1} \mathbb{E}[\widetilde{\mathbf{T}}^*(i)|X]$, then

$$\sum_{a=0}^{K-1} \mathbb{I}(g^*(X)=a)\mathbb{E}[\tilde{T}^*(a)|X] = \max_{0\le i\le K-1} \mathbb{E}[\widetilde{\mathbf{T}}^*(i)|X].$$

For any treatment regime g,

$$\begin{aligned}\mathbf{V}(g) &= \mathbb{E}_{\mathbf{X}}[\sum_{a=0}^{K-1} \mathbb{I}(g(X)=a)\tilde{T}^*(a)] \\ &\le \mathbb{E}_{\mathbf{X}}[\max_{0\le i\le K-1} \mathbb{E}[\widetilde{\mathbf{T}}^*(i)|X]] \\ &= \mathbf{V}(g^*)\end{aligned} \tag{D.1.2}$$

We can see that, $g^* = \arg\max_g \mathbf{V}(g)$.

Hence, the **optimal treatment regime** is

$$g^*(X) = \arg\max_{0\le i\le K-1} \mathbb{E}[\widetilde{\mathbf{T}}^*(i)|X].$$

□

D.2 Proof of Theorem 2.3.4

Proof: The risk R of any classifier d is as follows,

$$R(d) = \mathbb{E}[\frac{\tilde{T}}{\pi(A,X)}\mathbb{I}(A\neq d(X))]. \tag{D.2.1}$$

Since there are k treatments, let $\mathcal{A} = \{1, 2, \ldots, K\}$,

$\pi(a,x) = \sum_{a=1}^{K} \mathbb{I}(A=a)\mathbb{P}(A=a|x)$ and we have consistency assumption in

this case as $\tilde{T} = \sum_{j=1}^{K} \mathbb{I}(A = j)\tilde{T}^*(j)$.

$$\begin{aligned} &\mathbb{E}[\frac{\tilde{T}}{\pi(A,X)}\mathbb{I}(A \neq d(X))] \\ &= \mathbb{E}[\frac{\tilde{T}}{\pi(A,X)}|X] - \mathbb{E}[\frac{\tilde{T}}{\pi(A,X)}\mathbb{I}(A = d(X))]. \end{aligned} \tag{D.2.2}$$

The assumption of no unmeasured confounders implies, $Y^*(a)$ is independent of A given X for $a = \{1, 2, \ldots, K\}$. Moreover, $\tilde{T}\mathbb{I}(A = a) = \tilde{T}^*(a)\mathbb{I}(A = a)$ using consistency assumption. Hence, second term in the Bayes risk equation is

$$\begin{aligned} &\mathbb{E}[\frac{\tilde{T}}{\pi(A,X)}\mathbb{I}(A = d(X))|X] \\ &= \mathbb{E}[\frac{\tilde{T}}{P(A = a|X)}\mathbb{I}(A = a)|X] \quad \text{if } d(X) = a \\ &= \mathbb{E}[\frac{\tilde{T}}{P(A = a|X)}\mathbb{I}(A = a)|X] \quad \text{if } d(X) = a \\ &= \mathbb{E}[\frac{\tilde{T}^*(a)}{P(A = a|X)}\mathbb{I}(A = a)|X] \quad \text{if } d(X) = a \\ &= \mathbb{E}[\tilde{T}^*(a)|X]\mathbb{E}[\frac{\mathbb{I}(A = a)}{P(A = a|X)}|X] \quad \text{if } d(X) = a \\ &= \mathbb{E}[\tilde{T}^*(a)|X] \quad \text{if } d(X) = a. \end{aligned} \tag{D.2.3}$$

Hence, Bayes risk is

$$\begin{aligned} &\mathbb{E}[\frac{\tilde{T}}{\pi(A,X)}\mathbb{I}(A \neq d(X))|X] \\ &= \mathbb{E}[\frac{\tilde{T}}{\pi(A,X)}|X] - \mathbb{E}[\tilde{T}^*(a)|X] \quad \text{if } d(X) = a. \end{aligned} \tag{D.2.4}$$

The first term is constant across any treatment assignment a. Hence, to minimize Bayes posterior risk over a we need to minimize the second term that is maximize $\mathbb{E}[\tilde{T}^*(a)|X]$.

The classifier d minimizes Bayes risk if $d(X) = \arg\max_{a \in \mathcal{A}} \mathbb{E}[\tilde{T}^*(a)|X]$. Hence, our classifier g^* is a Bayes rule. $\square$

Appendix E

E.1 Proof of Proposition 2.4.8

Proof: For this proposition, we denote $\hat{S}(t|x)$ to be the tree KM estimator of the survival function $S(t|x)$ and $\hat{\Lambda}$ to be the tree NA estimator of the cumulative hazard function Λ. $\hat{S}(t|x) = \Pi_{[0,t]}(1 - \hat{\Lambda}(ds))$ and $S(t|x) = \Pi_{[0,t]}(1 - \Lambda(ds))$.

Using Duhamel equationZeng [2004] in step 3, for $t \in [0, \infty)$,

$$
\begin{aligned}
&|\hat{S}(t|x) - S(t|x)| \\
&= |\Pi_{[0,t]}(1 - \hat{\Lambda}(ds)) - \Pi_{[0,t]}(1 - \Lambda(ds))| \\
&= |-\int_0^t \Pi_0^{v-}(1 - \hat{\Lambda}(du))(\hat{\Lambda}(dv) - \Lambda(dv))\Pi_v^t(1 - \Lambda(du))| \\
&= |-\int_0^t \hat{S}(v-)(\hat{\Lambda}(dv) - \Lambda(dv))\frac{S(t)}{S(v)}| \\
&= |-S(t)[\int_0^t \frac{\hat{S}(v-)}{S(v)}(\hat{\Lambda}(dv) - \Lambda(dv))]| \\
&= |S(t)[\int_0^t \frac{\hat{S}(v-)}{S(v)}(\hat{\Lambda}(dv) - \Lambda(dv))]| \\
&= |S(t)[\int_0^t \frac{\hat{S}(v-)}{S(v)}(\hat{\Lambda}(dv) - \Lambda(dv))]| \\
&= |S(t)[\frac{\hat{S}(t-)}{S(t)}(\hat{\Lambda}(t) - \Lambda(t)) - \\
&\quad \int_0^t (\hat{\Lambda}(v) - \Lambda(v))\frac{S(v)d(\hat{S}(v-)) - \hat{S}(v-)d(S(v))}{(S(v))^2}]| \\
&\leq \sup_{0 \leq s \leq t} |(\hat{\Lambda}(s) - \Lambda(s))|[|S(t)\frac{\hat{S}(t-)}{S(t)}| + |-S(t)\int_0^t \frac{d(\hat{S}(v-))}{S(v)}| + \\
&\quad |S(t)\int_0^t \frac{\hat{S}(v-)d(S(v))}{(S(v))^2}|] \\
&= \sup_{0 \leq s \leq t} |(\hat{\Lambda}(s) - \Lambda(s))|[|\hat{S}(t-)| + |S(t)\int_0^t \frac{d(\hat{S}(v-))}{S(v)}| + \\
&\quad |S(t)\int_0^t \frac{\hat{S}(v-)d(S(v))}{(S(v))^2}|] \\
&\leq \sup_{0 \leq s \leq t} |(\hat{\Lambda}(s) - \Lambda(s))|[1 + S(t)\int_0^t |\frac{d(\hat{S}(v-))}{S(v)}| + \\
&\quad S(t)\int_0^t |\frac{\hat{S}(v-)d(S(v))}{(S(v))^2}|]
\end{aligned}
\tag{E.1.1}
$$

$$
\begin{aligned}
&\leq \sup_{0\leq s\leq t} |(\hat{\Lambda}(s) - \Lambda(s))|[1 + S(t)\int_0^t \frac{|d(\hat{S}(v-))|}{S(t)} + S(t)\int_0^t \frac{|d(S(v))|}{(S(t))^2}] \\
&\leq \sup_{0\leq s\leq t} |(\hat{\Lambda}(s) - \Lambda(s))|[1 + \int_0^t |d(\hat{S}(v-))| + \int_0^t \frac{|d(S(v))|}{S(t)}] \\
&\leq \sup_{0\leq s\leq t} |(\hat{\Lambda}(s) - \Lambda(s))|[1 + \int_0^t \hat{f}(v-)dv + \int_0^t \frac{f(v)dv}{S(t)}] \\
&\leq \sup_{0\leq s\leq t} |(\hat{\Lambda}(s|x) - \Lambda(s|x))|[2 + \frac{1}{S(t|x)}].
\end{aligned}
$$

Assuming $\sup_{x\in\mathcal{X}} \frac{1}{S(t|x)} < M$ for all $t > 0$. Hence, for $t \leq \tau$,

$$
|\hat{S}(t|x) - S(t|x)| \leq [2 + M] \sup_{0\leq t\leq\tau} |(\hat{\Lambda}(t|x) - \Lambda(t|x))|, \tag{E.1.2}
$$

with probability 1.

From Theorem 4.3 in Cui et al.(2019)Cui et al. [2017],

$$
\begin{aligned}
&\sup_{0\leq t\leq\tau} |(\hat{\Lambda}(t|x) - \Lambda(t|x))| \\
&= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}}\Big),
\end{aligned} \tag{E.1.3}
$$

with probability greater than $(1 - w_n)$ which approaches 1 as $n \to \infty (w_n \to 0)$.

Hence,

$$
\begin{aligned}
&\sup_{0\leq t\leq\tau} |\hat{S}(t|x) - S(t|x)| \\
&\leq [2 + M] \sup_{0\leq t\leq\tau} |(\hat{\Lambda}(t|x) - \Lambda(t|x))| \\
&= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}}\Big),
\end{aligned} \tag{E.1.4}
$$

with probability greater than $(1 - w_n)$ which approaches 1 as $n \to \infty (w_n \to 0)$.

For b^{th} tree for $1 \leq b \leq B$, over the set, say A_b, where with probability less

than w_n,

$$|(\hat{S}^{(b)}(t|x) - S(t|x))| \leq |(\hat{S}^{(b)}(t|x)| + |S(t|x))| \leq 2 \text{ for any } 0 \leq t \leq \tau.$$

Hence, for $t \in [0, \tau]$,

$$\begin{aligned} &\sup_{0 \leq t \leq \tau} \mathbb{E}_X[|(\hat{S}^{(b)}(t|x) - S(t|x))|] \\ &= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}} + 2w_n\Big). \end{aligned} \tag{E.1.5}$$

□

E.2 Proof of Theorem 2.4.9

Proof: For b^{th} tree for $1 \leq b \leq B$,

$$\begin{aligned}
&\sup_{0\leq t\leq \boldsymbol{\tau}} \mathbb{E}_X[|(\hat{S}^{(b)}(t|x) - S(t|x))|] \\
&= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}} + 2w_n\Big).
\end{aligned} \tag{E.2.1}$$

Next, for the random survival forest estimator $\hat{S_B}$ of S,

$$\begin{aligned}
&|(\hat{S}(t|x) - S(t|x))| \\
&= |\frac{1}{B}\sum_b (\hat{S}^{(b)}(t|x) - S(t|x))| \\
&\leq \frac{1}{B}\sum_b |(\hat{S}^{(b)}(t|x) - S(t|x))|.
\end{aligned} \tag{E.2.2}$$

Hence,

$$\begin{aligned}
&\sup_{0\leq t\leq \boldsymbol{\tau}} \mathbb{E}_X[|(\hat{S}(t|x) - S(t|x))|] \\
&\leq \frac{1}{B}\sum_b \sup_{0\leq t\leq \boldsymbol{\tau}} \mathbb{E}_X[|(\hat{S}^{(b)}(t|x) - S(t|x))|] \\
&= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}} + 2w_n\Big).
\end{aligned} \tag{E.2.3}$$

Hence,

$$\begin{aligned}
&\lim_{B\to\infty} \sup_{0\leq t\leq \boldsymbol{\tau}} \mathbb{E}_X[|(\hat{S}(t|x) - S(t|x))|] \\
&= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}} + 2w_n\Big).
\end{aligned} \tag{E.2.4}$$

□

E.3 Proof of Theorem 2.4.10

Proof: $\mathbf{V}(g) = \mathbb{E}[\tilde{T}(g)] = \mathbb{E}_{\mathbf{X}}[\mathbb{E}[\tilde{T}(g)|X]]$

From the assumption of consistency, we have

$$\begin{aligned} &\mathbb{E}[\tilde{T}(g)|X] \\ &= g(X)\mathbb{E}[\tilde{T}^*(1)|X] + (1-g(X))E[\tilde{T}^*(0)|X] \\ &= g(X)\mu_0(X) + (1-g(X))\mu_1(X). \end{aligned} \tag{E.3.1}$$

Hence,

$$\begin{aligned} &|\mathbf{V}(\hat{g})(X) - \mathbf{V}(g^*)(X)| \\ &= |\mathbb{E}_{\mathbf{X}}[\mathbb{E}[\tilde{T}(\hat{g})|X]] - \mathbb{E}_{\mathbf{X}}[\mathbb{E}[\tilde{T}(g^*)|X]]| \\ &\leq \mathbb{E}_{\mathbf{X}}[|\mathbb{E}[\tilde{T}(\hat{g})|X] - \mathbb{E}[\tilde{T}(g^*)|X]|]. \end{aligned} \tag{E.3.2}$$

Since we have 2 possible treatments 0 and 1,

$$|\mathbb{E}[\tilde{T}(\hat{g})|X] - \mathbb{E}[\tilde{T}(g^*)|X]| = \begin{cases} |\mu_1(X) - \mu_0(X)| & \text{if } \hat{g}(X) \neq g^*(X), \\ 0 & \text{otherwise.} \end{cases} \tag{E.3.3}$$

Hence,

$$|\mathbb{E}[\tilde{T}(\hat{g})|X] - \mathbb{E}[\tilde{T}(g^*)|X]| = |(\hat{g}(X) - g^*(X))(\mu_1(X) - \mu_0(X))|.$$

Defining $h(X) = \mu_1(X) - \mu_0(X)$, we get that $g^*(X) = \mathbb{I}(h(X) > 0)$. Similarly,

$\hat{h}(X) = \hat{\mu}_1(X) - \hat{\mu}_0(X)$, $\hat{g}(X) = \mathbb{I}(\hat{h}(X) > 0)$. Hence,

$$\begin{aligned}
&|\mathbf{V}(\hat{g}) - \mathbf{V}(g^*)| \\
&\leq \mathbb{E}|(\hat{g}(X) - g^*(X))(\mu_1(X) - \mu_0(X))| \\
&= \mathbb{E}|(\mathbb{I}(\hat{h}(X) > 0) - \mathbb{I}(h(X) > 0))(\mu_1(X) - \mu_0(X))| \qquad \text{(E.3.4)} \\
&\quad \text{using Cauchy-Schwarz inequality, we have} \\
&\leq \frac{1}{2}[\mathbb{E}(\mathbb{I}(\hat{h}(X) > 0) - \mathbb{I}(h(X) > 0))^2]^{\frac{1}{2}}[\mathbb{E}(\mu_1(X) - \mu_0(X))^2]^{\frac{1}{2}}.
\end{aligned}$$

Let $S_a(.|X)$ be the survival function under treatment $a \in \{0, 1\}$. $\mu_a(X) = \int_0^\tau S_a(t|X)dt$. Hence, $0 \leq \mu_a(X) \leq \tau$ for $a \in \{0, 1\}$ and $|\mu_1(X) - \mu_0(X)| \leq \tau$, which implies $[\mathbb{E}(\mu_1(X) - \mu_0(X))^2]^{\frac{1}{2}} \leq \tau$.

$$\begin{aligned}
&\mathbb{E}(\mathbb{I}(\hat{h}(X) > 0) - \mathbb{I}(h(X) > 0))^2 \\
&= \mathbb{P}(\hat{h}(X) \neq h(X)) \qquad \text{(E.3.5)} \\
&= \mathbb{P}(\hat{h}(X) < 0 < h(X)) + \mathbb{P}(h(X) < 0 < \hat{h}(X)).
\end{aligned}$$

For any $\delta > 0$,

$$\mathbb{P}(\hat{h}(X) < 0 < h(X)) \leq \mathbb{P}(\delta \geq h(X) > 0) + \mathbb{P}(|\hat{h}(X) - h(X)| > \delta). \qquad \text{(E.3.6)}$$

$$\begin{aligned}
&\mathbb{P}(h(X) < 0 < \hat{h}(X)) \\
&= \mathbb{P}(-\hat{h}(X) < 0 < -h(X)) \qquad \text{(E.3.7)} \\
&\leq \mathbb{P}(-\delta \leq h(X) < 0) + \mathbb{P}(|\hat{h}(X) - h(X)| > \delta).
\end{aligned}$$

Combining E.3.6 and E.3.7, we have

$$\mathbb{E}(\mathbb{I}(\hat{h}(X) > 0) - \mathbb{I}(h(X) > 0))^2 \leq \mathbb{P}(|h(X)| < \delta) + 2\mathbb{P}(|\hat{h}(X) - h(X)| > \delta) \qquad \text{(E.3.8)}$$

For the first term in E.3.8, from Assumption 6 we have that

$$\mathbb{P}(|h(X)| < \delta) = \mathbb{P}[|\mu_1(X) - \mu_0(X)| \leq \delta] \leq c\delta^{\gamma}. \tag{E.3.9}$$

For the second term in E.3.8,

$$\begin{aligned}
&\mathbb{P}(|\hat{h}(X) - h(X)| > \delta) \\
&\leq \frac{\mathbb{E}_X[|\hat{h}(X) - h(X)|]}{\delta} \\
&= \frac{1}{\delta}\mathbb{E}_X|(\hat{\mu}_1(X) - \mu_1(X)) - (\hat{\mu}_0(X) - \mu_0(X))| \\
&= \frac{1}{\delta}[\mathbb{E}_X|(\hat{\mu}_1(X) - \mu_1(X))| + \mathbb{E}_X|(\hat{\mu}_0(X) - \mu_0(X))|] \\
&= \frac{1}{\delta}[\mathbb{E}_X|\int_0^{\tau} \hat{S}_1(t|X)dt - \int_0^{\tau} S_1(t|X)dt| \\
&\quad + \mathbb{E}_X|\int_0^{\tau} \hat{S}_0(t|X)dt - \int_0^{\tau} S_0(t|X)dt|] \\
&\leq \frac{1}{\delta}[\mathbb{E}_X \int_0^{\tau} |\hat{S}_1(t|X)dt - S_1(t|X)|dt \\
&\quad + \mathbb{E}_X \int_0^{\tau} |\hat{S}_0(t|X)dt - S_0(t|X)|dt] \\
&\leq \frac{\tau}{\delta}[\sup_{0\leq t\leq\tau} \int_{\mathcal{X}} |\hat{S}_1(t|X)dt - S_1(t|X)|d\mu(X) \\
&\quad + \sup_{0\leq t\leq\tau} \int_{\mathcal{X}} |\hat{S}_0(t|X)dt - S_0(t|X)|d\mu(X)].
\end{aligned} \tag{E.3.10}$$

To get the convergence rate, we need the convergence rate of the terms in E.3.10. From 2.4.4, for each $a \in \{0, 1\}$,

$$\begin{aligned}
&\lim_{B\to\infty} \sup_{0\leq t\leq\tau} \mathbb{E}_X[|(\hat{S}_a(t|x) - S_a(t|x))|] \\
&= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}} + 2w_n\Big).
\end{aligned} \tag{E.3.11}$$

Combining the inequalities E.3.9 and E.3.10, we get that for any $\delta > 0$,

$$\begin{aligned} &|\mathbf{V}(\hat{g}) - \mathbf{V}(g^*)| \\ &\leq \frac{\tau}{2}\sqrt{c\delta^{\gamma} + \frac{\tau}{\delta}O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + (\frac{k}{n})^{\frac{c_1}{d}} + 2w_n\Big)}. \end{aligned} \tag{E.3.12}$$

□

Appendix F

F.1 Proof of Theorem 4.4.1

Proof: We intend to prove consistency of the estimator of $(\alpha_0, \beta, \gamma)$ for the framework $Q_{\tilde{T}}(\tau|X = x, A = a) = \alpha_0 a + \beta^T x + a\gamma^T x$. Denoting $Z = (A, X, AX)$ and $\beta_z = (\alpha_0, \beta, \gamma)$, we get that $Q_{\tilde{T}}(\tau|Z = z) = \beta_z^T z$. The

The first step consists of estimating $\hat{F}_0(.|A, X)$ where $F_0(.|A, X)$ is the true distribution function of T given covariates A, X and then we define $\hat{\tilde{F}}(t|A, X) = \hat{F}_0(h^{-1}(t)|A, X)$. Note that the information contained in Z is the same as that in (A, X), hence $F_0(.|Z) \equiv F_0(.|A, X)$. Random forests consider interaction terms, we do not need to explicitly give it the interaction term AX as we need to for the parametric step. Hence we can write $\hat{F}(.|Z)$ instead of $\hat{F}(.|A, X)$. We now consider the entire setup in terms of Z and can draw parallel with the censored quantile regression consistency proof from the previous section. The quantile estimating equation is as follows,

$$Q_n(\beta_z, \tilde{F}) = \frac{1}{n}\sum_{i=1}^{n}\left[\tilde{w}_i(\tilde{F})\rho_\tau(\tilde{Y}_i - z_i\beta_z) + (1 - \tilde{w}_i(\tilde{F}))\rho_\tau(\tilde{Y}^{+\infty} - z_i\beta_z)\right] \quad \text{(F.1.1)}$$

Similar to previous section, the weights ($\tilde{w}(\tilde{F})$) for the next step can be shown to be $\tilde{w}(\tilde{F}) = w(F_0)$. The way the estimators are defined, we also get that $\tilde{w}(\hat{\tilde{F}}) = w(\hat{F}_0)$. Then we proceed to weighted quantile regression on $\tilde{Y}$ on Z with weights $w(\hat{F}_0)$.

In essence we can also consider this as estimating $\hat{\tilde{F}}$ in the first step and then using the weights $\tilde{w}(\hat{\tilde{F}})$ for weighted quantile regression on $\tilde{Y}$. This is to eliminate the confusion when using both T and $\tilde{T}$. We will now just use $\tilde{T}$ in the consistency proof.

The negative subgradient of the quantile estimating equation gives us,

$$\frac{1}{n}\sum_{i=1}^{n} m(Z_i, \tilde{Y}_i, \delta_i, \hat{\tilde{F}}) = 0 \tag{F.1.2}$$

where

$$m(Z_i, \tilde{Y}_i, \delta_i, \hat{\tilde{F}}) = Z_i(\tau - \tilde{w}_i(\hat{\tilde{F}})\mathbb{I}(\tilde{Y}_i < \beta_z^T Z)) \tag{F.1.3}$$

We will use the setting and theorem 1 of Chen et al. [2003] to prove consistency similar to Wang and Wang [2009]. As shown in Wang and Wang [2009], $M(\beta_z, \tilde{F}) = 0$ because the change proposed is for the estimator and not the setting or space. We further need to verify conditions 1.1-1.4 and 1.5' in Chen et al. [2003].

Conditions 1.2, 1.3 do not need the form of $\hat{\tilde{F}}$ hence the same proof as that of Wang and Wang [2009] can be applied. In Wang and Wang [2009], for their proof of Theorem 1, conditions 1.1 and 1.5' do not require the form of $\hat{\tilde{F}}$ hence the same proof works.

Condition 1.4 follows from lemma 4.3.9. We have defined $\widehat{\tilde{F}}$, the estimate of $\tilde{F}$ as $\widehat{\tilde{F}}(t) = \hat{F}_0(h^{-1}(t))$.

$$\tilde{F}(t) = P(\tilde{T} \leq t) = P(\tilde{T} \leq t) = P(T \leq h^{-1}(t)) = F_0(h^{-1}(t))$$

We consider the space of distribution functions on $\tilde{T}$ to be $F \in \tilde{\mathcal{H}}$. Noting that F would have 2 parameters t and z since $F(t|z)$ is the cumulative distribution function st t given z. Defining the norm $||f(t,z)||_{\tilde{\mathcal{H}}} = \sup_{h(0)\leq t\leq h(\tau_0)} \sup_x |f(t,z)|$

$$
\begin{aligned}
&||\widehat{\tilde{F}} - F||_{\tilde{\mathcal{H}}} \\
&= \sup_{h(0)\leq t\leq h(\tau_0)} \sup_z |\widehat{\tilde{F}}(t|z) - \tilde{F}(t|z)| \\
&= \sup_{h(0)\leq t\leq h(\tau_0)} \sup_z |\hat{F}_0(h^{-1}(t)|z) - F_0(h^{-1}(t)|z)| \\
&= \sup_{h^{-1}(h(0))\leq h^{-1}(t)\leq h^{-1}(h(\tau_0))} \sup_z |\hat{F}_0(t|z) - F_0(t|z)| \\
&= \sup_{0\leq t\leq \tau_0} \sup_z |\hat{F}_0(t|z) - F_0(t|z)| \\
&\leq ||\hat{F}_0 - F_0||_{\mathcal{H}} \text{ (for big enough } \tau_0)
\end{aligned}
\tag{F.1.4}
$$

Hence $||\widehat{\tilde{F}} - F||_{\tilde{\mathcal{H}}} = o_p(1)$ follows from lemma 4.3.9. Hence conditions 1.1-1.4 and 1.5' of Chen et al. [2003] hold true.

□