Censored Quantile Regression Forest

Alexander Hanbo Li [1]
Amazon Research

Jelena Bradic
University of California San Diego

Abstract

Random forests are powerful non-parametric regression method but are severely limited in their usage in the presence of randomly censored observations, and naively applied can exhibit poor predictive performance due to the incurred biases. Based on a local adaptive representation of random forests, we develop its regression adjustment for randomly censored regression quantile models. Regression adjustment is based on a new estimating equation that adapts to censoring and leads to quantile score whenever the data do not exhibit censoring. The proposed procedure named *censored quantile regression forest*, allows us to estimate quantiles of time-to-event without any parametric modeling assumption. We establish its consistency under mild model specifications. Numerical studies showcase a clear advantage of the proposed procedure.

1 Introduction

Censored data exists in many different areas. In economics, policies such as minimum wage and minimum transaction fee result in left-censored data. In biomedical study, researchers cannot always observe the time until the occurrence of an event of interest because the time span of the study is limited or the patient withdraws from the experiment, resulting in right-censored data.

Classical statistical approaches always assume an underlying model like accelerated failure time model (Koul et al., 1981; Robins and Tsiatis, 1992; Robins, 1992; Wei, 1992; Zeng and Lin, 2007; Huang et al., 2007). These methods perform well when the model is correctly specified, but quickly break down when the model assumption is wrong or the error distribution is heteroscedastic. Other non-parametric methods (Louis, 1981; Hoover et al., 1998) and rank-based methods (Jin et al., 2003) strive to achieve assumption-lean modeling of the mean. Forest algorithms (Breiman, 2001; Geurts et al., 2006; Meinshausen, 2006; Athey et al., 2019) are non-parametric and allow for flexible modeling of covariate interactions. However, it is non-trivial to adapt forest algorithms to censored data. Random survival forests (Ishwaran et al., 2008) or bagging survival trees (Hothorn et al., 2004, 2005) rely on building survival trees using survival function as the splitting criterion, and are only applicable to right-censored data. Moreover, any technique developed for uncensored data cannot be easily extended to censoring scenario. For example, generalized random forest can effectively deal with heteroscedastic data, but applying the same technique to heteroscedastic and censored data is non-trivial.

In this paper, we propose a novel method that connects the quantile forest algorithms to the censored data problem. This is done by a carefully designed estimating equation. In this way, any technique developed for quantile forests on uncensored data can be seamlessly applied to censored data. One of the promising applications of the introduced method is in the estimation of heterogeneous treatment effects when the response variable is censored. We will show in the experiments that the introduced methods do achieve the best performance on both simulated and real datasets. Especially on one heteroscedastic data, the proposed method is the only working solution among many other forest algorithms, including random survival forest. The proposed method, *censored quantile regression forest*, is motivated by the observation that random forests actually define a local similarity metric (Lin and Jeon, 2006; Li and Martin, 2017; Athey et al., 2019) which is essentially a data-driven kernel. Using this kernel, random forests can be rephrased as locally weighted regressions. We will review the regression adjustments for forests in Section 2.

[1] Work mostly done at University of California San Diego.

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

1.1 Related Work

In the case of right censoring, most non-parametric recursive partitioning algorithms rely on survival tree or its ensembles. Ishwaran et al. (2008) proposed random survival forest (RSF) algorithm in which each tree is built by maximizing the between-node log-rank statistic. However, it is not directly estimating the conditional quantiles but instead estimating the cumulative hazard. Zhu and Kosorok (2012) proposed the recursively imputed survival trees (RIST) algorithm with the same splitting criterion for each individual tree but different ensemble scheme. Other similar methods relying on different kinds of survival trees were proposed in Gordon and Olshen (1985), Segal (1988), Davis and Anderson (1989), LeBlanc and Crowley (1992) and LeBlanc and Crowley (1993). All these methods as mentioned above use splitting rules specifically designed for the right censored data, and they all rely on the proportional hazard assumption and cannot reduce to a loss-based method that might ordinarily be used in the situation with no censoring. Molinaro et al. (2004) proposed a tree method based on the inverse probability censoring (Robins et al., 1994) weighted (IPCW) loss function which reduces to the full data loss function used by CART in the absence of censoring. Hothorn et al. (2005) then extended the IPCW idea and proposed a forest-type method in which each tree is trained on resampled observations according to inverse probability censoring weights. However, the censored data always get weights zero and hence only uncensored observations will be resampled. As pointed out by Robins et al. (1994), the inverse probability weighted estimators are inefficient because of their failure to utilize all the information available on observations with missing or partially missing data.

2 Regression Adjustments for Forests

We will briefly review random forest and generalized forest in this section and show that they can be written as weighted regression problems. We also introduce *"forest weights"* that is an essential concept in this paper.

Random Forest Let θ denote the random parameter determining how a tree is grown, and $\{(X_i, Y_i) : i = 1, \ldots, n\} \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^p \times \mathbb{R}$ denote the training data. For each tree $T(\theta)$, let R_l denotes its l-th terminal leaf. We let the index of the leaf that contains x to be $l(x; \theta)$.

As shown in Meinshausen (2006), for any single tree $T(\theta)$, the prediction on x can be written as $\sum_{i=1}^{n} w(X_i, x; \theta) Y_i$ where $w(X_i, x; \theta) = \mathbb{1}_{\{X_i \in R_{l(x;\theta)}\}} / \#\{j : X_j \in R_{l(x;\theta)}\}$. Then a random forest containing m trees formulates a prediction of $\mathbb{E}[Y|X = x]$ as $\sum_{i=1}^{n} w(X_i, x) Y_i$ where

$$w(X_i, x) = \frac{1}{m} \sum_{t=1}^{m} w(X_i, x; \theta_t). \tag{1}$$

From now on, we call the weight $w(X_i, x)$ in equation 1 as *random forest weight*. The above representation of the random forest prediction of the mean can be equivalently obtained as a solution to the least-squares optimization problem $\min_{\lambda \in \mathbb{R}} \sum_{i=1}^{n} w(X_i, x)(Y_i - \lambda)^2$. Therefore, a least-squares regression adjustment, as the above, is equivalent to Breiman (2001) representation of random forests. However, when we move to estimation quantities that are not the mean, the latter representation is very powerful. Namely, a quantile random forest of Meinshausen (2006) can be seen as a quantile regression adjustment (Li and Martin, 2017), i.e., as a solution to the following optimization problem

$$\min_{\lambda \in \mathbb{R}} \sum_{i=1}^{n} w(X_i, x) \rho_\tau (Y_i - \lambda),$$

where ρ_τ is the τ-th quantile loss function, defined as $\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0))$. Local linear regression adjustment was also recently utilized in Athey et al. (2019) to obtain a smoother and more poweful generalized forest algorithm.

Generalized Random Forests Athey et al. (2019) proposed to generalize random forest using a more sophisticated splitting criterion which is model-free. The new criterion aims to maximize the in-sample heterogeneity, formally defined as

$$\tilde{\Delta}(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{|\{i : X_i \in C_j\}|} \left(\sum_{\{i : X_i \in C_j\}} \rho_i \right)^2$$

where ρ_i is a pseudo-response defined similarly as in Gradient Boosting (Friedman, 2001). Note that in the original random forest, we simply have $\rho_i = Y_i - \bar{Y}_P$ where $\bar{Y}_P$ is the mean response in the parent node. The generalized random forest, while applied to quantile regression problem, can deal with *heteroscedasticity* because the splitting rule directly targets changes in the quantiles of the Y-distribution.

Just like the random forest algorithm, the generalized random forest is also an ensemble of trees and hence defines a weight or similarity between two samples using equation 1. The main difference hence lies in how they split the samples into different terminal regions. Therefore, in the following sections, whenever we refer to *forest weight*, it can be calculated from either random forest or generalized random forest. In the experiment section, we will distinguish them by *RF-weights* and *GRF-weights*.

3 Censored Quantile Regression Forest

The forest regressions cannot be directly applied to censored data $\{(X_i, Y_i)\}$ because the conditional quantile of Y is different from the quantiles of the latent variable T due to the censoring. Moreover, there is no explicitly defined quantile loss function for randomly censored data. In this section, we design a new approach to achieve both tasks. We will motivate and derive our method using sections 3.1 to 3.3.

3.1 No Censoring and Locally Invariate T_i

Temporary Assumptions: To simplify the exposition of the proposed method, we make two temporary assumptions: (**1**) there is no censoring on the data, and (**2**) the latent variable T_i has the same conditional probability in a neighborhood R_x of x. These assumptions will all be released in section 3.3.

Following the regression adjustment reasoning, we could estimate the τ-th quantile of T_i at x as

$$q_{\tau,x} = \arg\min_{q\in\mathbb{R}} \sum_{i=1}^{n} w(X_i, x)\rho_\tau(T_i - q).$$

The above optimization problem has the following estimating equation

$$U_n(q; x) = (1-\tau) - \sum_{i=1}^{n} w(X_i, x)\mathbb{1}(T_i > q) \approx 0. \quad (2)$$

Now out of the n data points, assume $\{X_1, \cdots, X_k\} \subset R_x$ and $w(X_i, x) = k^{-1}\mathbb{1}\{X_i \in R_x\}$. Because of the assumption 1, the estimating equation becomes

$$U_k(q) \quad = (1-\tau) - \tfrac{1}{k}\textstyle\sum_{i=1}^{k} \mathbb{1}(T_i > q) \approx 0. \qquad (3)$$

Now conditional on $\{x\}\cup\{X_i\}_{i=1}^k$, we have the expected estimating equation

$$\mathbb{E}\left[U_k(q)|x, X_i, i = 1, \ldots, k\right] = (1-\tau) - \mathbb{P}(T > q|x)$$

which will be zero at q^* where $\mathbb{P}(T > q^*|x) = 1-\tau$, that is, at the true τth quantile at x.

3.2 With Censoring and Locally Invariate T_i

Let's now consider the case that we could only observe $Y_i = \min\{T_i, C_i\}$ and the censoring indicator $\delta_i = \mathbb{1}(T_i \le C_i)$. Note that the following analysis extends straightforwardly to left censoring. In order to estimate $q_{\tau,x}$, we cannot simply replace T_i with Y_i in equation (3) as the τ-th quantile of T_i is no longer the τ-th quantile of Y_i. However, because C_i and T_i are conditionally independent, we have the following relation:

$$\begin{aligned}\mathbb{P}(Y_i > q_{\tau,x}|x) &= \mathbb{P}(T_i > q_{\tau,x}|x)\mathbb{P}(C_i > q_{\tau,x}|x)\\ &= (1-\tau)G(q_{\tau,x}|x)\end{aligned}$$

where $G(u|x)$ is the conditional survival function of C_i at x. That is to say, the τ-th quantile of T_i is actually the $1-(1-\tau)G(q_{\tau,x}|x)$-th quantile of Y_i at x, leading to a new estimating equation that closely resembles equation (3):

$$S_k^o(q; x) = (1-\tau)G(q|x) - \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}(Y_i > q) \approx 0, \quad (4)$$

and we still have $\mathbb{E}[S_k^o(q_{\tau,x})|x] = 0$. An intuitive explanation for using (4) is that because the τ-th quantile of T_i is just the $1-(1-\tau)G(q_{\tau,x}|x)$-th quantile of Y_i at x, instead of estimating the former which is not available because of the censoring, we could just estimate the later one.

The survival function $G(\cdot|x)$ can be estimated by any consistent estimator, for example, the Kaplan-Meier estimator $\hat{G}(\cdot|x)$ using $\{Y_i\}_{i=1}^k$ and $\{\delta_i\}_{i=1}^k$, and we can then solve for

$$S_k(q; x) = (1-\tau)\hat{G}(q|x) - \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}(Y_i > q) \approx 0. \quad (5)$$

3.3 Full Model

In the previous section, we assume that $\mathbb{P}(T|X) = \mathbb{P}(T|x)$ for all $X \in R_x$. But in reality, this assumption is not always true, and that is why $w(X_i, x)$ plays an important rule in our final estimator, as it "corrects" the empirical probability of each T_i at x. Intuitively, if X_i is more similar to x than X_j, then Y_i should play a more important role than Y_j on estimating the quantile at x. Now let $w(X_i, x)$ denote a similarity measure between X_i and x. In order for $\sum_{i=1}^n w(X_i, x)\mathbb{1}(T_i \le q)$ to be a proper estimation of $\mathbb{P}(T \le q|x)$, it needs to satisfy two conditions:

(**1**) $\sum_{i=1}^n w(X_i, x) = 1$; (**2**) $\sum_{i=1}^n w(X_i, x)\mathbb{1}(T_i \le q) \xrightarrow{p} \mathbb{P}(T \le q|x)$ for all q.

One may think that any fixed Kernel weights, $K(X_i, x)$, could be a suitable choice, but in fact they would not be able to satisfy the second condition for every distribution $\mathbb{P}(T|x)$. Fortunately, as shown in Meinshausen (2006) and Athey et al. (2019), the data-adaptive (generalized) random forest weight $w(X_i, x)$ perfectly satisfies both conditions. Therefore if we define

$$U_n(q_{\tau,x}) = (1-\tau) - \sum_{i=1}^{n} w(X_i, x)\mathbb{1}(T_i > q_{\tau,x}), \quad (6)$$

we have $U_n(q_{\tau,x}) \xrightarrow{p} 0$ asymptotically. Then following the same logic of how we get equation 5, a general case estimating equation for censoring data will be

$$S_n(q;x) = (1-\tau)\hat{G}(q|x) - \sum_{i=1}^{n} w(X_i,x)\mathbb{1}(Y_i > q) \approx 0. \tag{7}$$

3.4 Estimators for $G(q|x)$

Many consistent estimators for the conditional survival functions exist. For example, the nonparametric estimator (Beran, 1981)

$$\tilde{G}(q|x) = \prod_{Y_i \le q} \left\{ 1 - \frac{W_i(x,a_n)}{\sum_{j=1}^{n} \mathbb{1}(Y_j \ge Y_i) W_j(x,a_n)} \right\}^{1-\delta_i} \tag{8}$$

is shown to be consistent (Beran, 1981; Dabrowska, 1987, 1989; Gonzalez-Manteiga and Cadarso-Suarez, 1994; Akritas, 1994; Li and Doss, 1995; Van Keilegom and Veraverbeke, 1996). Here, $W_i(x,a_n)$ is the Nadaraya-Watson weight. However, since we already have an adaptive version of kernel – the forest weights $w(X_i,x)$, we propose the following two new estimators for $G(q|x)$:

Kaplan-Meier using nearest neighbors. We first find the k nearest neighbors of x according to the magnitude of the weights $w(X_i,x)$, and denote these points as a set N_x. Then we define the Kaplan-Meier estimator on N_x as

$$\prod_{i:X_i \in N_x, Y_i \le q} \left(1 - \frac{1}{\sum_{j=1}^{n} \mathbb{1}(Y_j \ge Y_i)\mathbb{1}(X_j \in N_x)} \right)^{1-\delta_i}. \tag{9}$$

Here, the number of nearest neighbors k will be a tuning hyperparameter.

Beran estimator with forest weights. We replace the Nadaraya-Watson weights in equation 8 with the forest weights $w(X_i,x)$:

$$\hat{G}(q|x) = \prod_{Y_i \le q} \left\{ 1 - \frac{w(X_i,x)}{\sum_{j=1}^{n} \mathbb{1}(Y_j \ge Y_i) w(X_j,x)} \right\}^{1-\delta_i}. \tag{10}$$

In fact equation 9 is a special case of equation 10 when the weight $w(X_i,x) = 1/k$ for $X_i \in R_x$ and 0 otherwise.

3.5 Algorithm

We summarize our algorithm in Algorithm 1. The details for choosing the candidate set $\mathcal{C}$ is in Section A.1.

Algorithm 1 Censored quantile regression forest

Input: number of trees: B, minimum node size: m, the number of nearest neighbors: k (if using equation 9), test set $\mathcal{A}$, training set $\mathcal{D} = \{(X_i,Y_i,\delta_i)\}_{i=1}^{n}$, quantile τ

Step 0: Train a forest on $(\mathcal{D})$ with B trees and minimum node size m.

for $x \in \mathcal{A}$ **do**

Step 1: Calculate forest weights $w(x,X_i)$.

Step 2: Calculate the survival function estimate $\hat{G}(q|x)$.

Step 3: Get the quantile estimation:

$$\hat{q}(x) \leftarrow \arg\min_{q \in \mathcal{C}} |S_n(q;x)|$$

$\{\mathcal{C}$ is a candidate set as discussed in Section A.1$\}$

$\{S_n$ is defined in equation 7$\}$

end for

4 Theoretical Develoments

In this section, we will show the consistency of the proposed quantile estimator. The time complexity analysis is in the Appendix.

4.1 Consistency

The consistency of random forest has been extensively studied (Arlot and Genuer, 2014; Athey et al., 2019; Biau et al., 2008; Biau and Devroye, 2010; Biau, 2012; Denil et al., 2014; Lin and Jeon, 2006; Scornet et al., 2015; Wager and Walther, 2015; Wager and Athey, 2018). Following the common settings, we also assume the covariate space $\mathcal{X} = [0,1]^p$ and the parameter $q \in \mathcal{B} \subset \mathbb{R}$ where $\mathcal{B}$ is a compact subset of $\mathbb{R}$. In our case, since q stands for the quantile, the assumption means that there exists some $r > 0$ such that $q \in [-r,r] = \mathcal{B}$. We also make another standard assumption that the density of X is bounded away from 0 and ∞. Note that since $\mathcal{X}$ is a compact support, the density condition holds true for Gaussian distribution and more broadly any symmetric and continuous distribution with unbounded support.

Condition 1 (Lipschitz in x)**.** *Denote $F(y|x) = \mathbb{P}(Y \le y|x)$. There exists a constant L such that $F(y|x)$ is Lipschitz continuous with parameter L, that is, for all $x, x' \in \mathcal{X}$,*

$$\sup_y |F(y|x) - F(y|x')| \le L\|x - x'\|_1.$$

This Condition 1 appears in all existing work related to quantile regression and inference thereafter.

Condition 2 (Identification)**.** *For any fixed x, the latent variable T and the censoring variable C are con-*

ditionally independent, and the conditional distribution $\mathbb{P}(T \leq q|x)$ *and* $\mathbb{P}(C \leq q|x)$ *are both strictly increasing in* q.

Conditional independence of T and C is a very standard assumption and can be traced back to Robins and Tsiatis (1992) among other works.

Condition 3 (Tree splitting)**.** *For each tree splitting, the probability that each variable is chosen for the split point is bounded from below by a positive constant, and every child node contains at least* γ *proportion of the data in the parent node, for some* $\gamma \in (0, 0.5]$. *[Quantile forest (Meinshausen, 2006)] The terminal node size* $m \to \infty$ *and* $m/n \to 0$ *as* $n \to \infty$. *[Generalized forest (Athey et al., 2019)] The forest is honest and built via subsampling with subsample size* s *satisfying* $s/n \to 0$ *and* $s \to \infty$.

The first two requirements of Condition 3 are shared in Meinshausen (2006) and Athey et al. (2019). For quantile random forest (Meinshausen, 2006), they require that the leaf node size of each tree should increase with the sample size n, but at a slower rate. Our experiments also justify that the required leaf node size of Meinshausen's quantile forest is larger than the node size of the generalized forest. In general, using the generalized forest weights give us more stable estimations because the trees are honest and regular (Wager and Athey, 2018).

Condition 4 (Censoring variable)**.** *For any* $x \in \mathcal{X}$, $\hat{G}(q|x)$ *is a uniformly consistent estimator of the true conditional survival function* $G(q|x)$ *for* $q \in \mathcal{B}$.

Condition 4 is satisfied, for example, by the Kaplan-Meier estimator equation 8 (Dabrowska, 1989). Please take a look at Figure 5 where we compare finite sample properties of the newly introduced estimators equation 9 and equation 10. We observe that the new distributional estimators are more adaptive and yet seemingly inherit consistency to that of the traditional KM estimator.

We proceed to showcase asymptotic properties of the proposed estimating equations. We begin by illustrating a concentration of measure phenomenon for the introduced score equations.

Theorem 1. *Define*

$$S(q;\tau) = (1-\tau)G(q|x) - \mathbb{P}(Y > q). \tag{11}$$

Under Conditions 1 – 4, for any $x \in \mathcal{X}$, $r > 0$ *and* $\tau \in (0,1)$, *we have*

$$\sup_{q \in [-r,r]} |S_n(q;\tau) - S(q;\tau)| = o_p(1).$$

Next, we present our main result that illustrates an asymptotic consistency of the proposed conditional quantile estimator. The proof is given in Appendix.

Theorem 2. *Under Conditions 1 – 4, for fixed* $\tau \in (0,1)$ *and* $x \in \mathcal{X}$, *define* q^* *to be the root of* $S(q;\tau) = 0$, *and* $r > 0$ *to be some constant so that* $q^* \in [-r, r]$. *Also define* q_n *to be* $\arg\min_{q \in [-r,r]} |S_n(q;\tau)|$. *Then* $\mathbb{P}(T \leq q^*|x) = \tau$, *and* $q_n \xrightarrow{p} q^*$ *as* $n \to \infty$.

5 Experiments

In this section, we will compare the proposed model, censored regression forest (*crf*), with generalized random forest (*grf*) (Athey et al., 2018), quantile random forest (*qrf*) (Meinshausen, 2006) and random survival forest (*rsf*) (Hothorn et al., 2005) on various simulated and real datasets. On the simulated datasets, we report both censored and oracle results for *qrf* and *grf*. To obtain the censored result, we directly apply generalized random forest and quantile random forest to the censored data, and denote the results by *grf* and *qrf* respectively. For oracle result, we instead train the models using the oracle responses without censoring (i.e. T_i's), and call the results *grf-oracle* and *qrf-oracle*. [1]

5.1 Simulation Study

In this section, we denote censored regression forest with generalized forest weights as *crf-generalized* and the one with (quantile) random forest weights as *crf-quantile*. We first define the evaluation metric used in this section – **quantile loss**. The τ-th quantile loss is defined as follows. Let $\hat{q}_i^\tau$ be the estimated τ-th quantile at X_i, then

$$L_{quantile}(\hat{q}_1^\tau, \ldots, \hat{q}_n^\tau) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(T_i - \hat{q}_i^\tau). \tag{12}$$

We could use this metric because we know the latent responses T_i's in simulations.

5.1.1 Accelerated Failure Time Data

In this section, we generate data from a accelerated failure time (AFT) model. We sample $n = 1000$ independent and identically distributed examples where X_i is uniformly distributed over $[0,2]^p$ with $p = 20$, and T_i is conditional on $(X_i)_1$ and $\log(T_i|X_i) = (X_i)_1 + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 0.3^2)$. The censoring variable $C_i \sim \text{Exp}(\lambda = 0.08)$ and $Y_i = \min(T_i, C_i)$, resulting in about 23% censoring level. For more comprehensive study on different censoring levels, please see the Appendix B.3. The other 19 covariates are noise. We estimate the quantiles at $\tau = 0.1$ and 0.9, and draw the predicted quantiles in Figure 1 for node size 40.

[1] Our codes are available at `https://github.com/AlexanderYogurt/censored_ExtremelyRandomForest`.

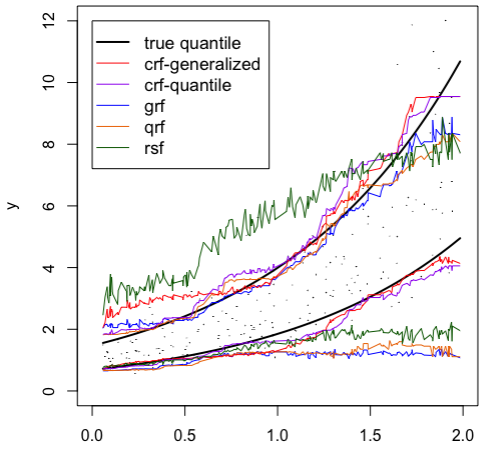


Figure 1: Quantile estimates on AFT model when $\tau = 0.1$ and 0.9. The minimum node size is 40 and each forest contains 1000 trees.

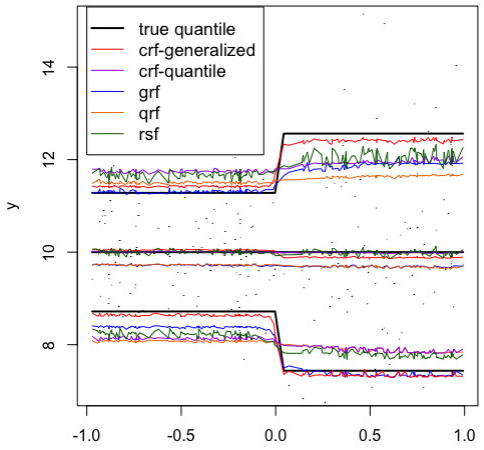


Figure 2: Quantile estimates on the censored heteroscedastic data for $\tau = 0.1$, 0.5 and 0.9. The minimum node size for all trees is 150 and each forest contains 2000 trees.

From the results in Figure 1, both *grf* and *qrf* are severely biased downwards because of the right censoring. The proposed methods *crf-generalized* and *crf-quantile* both provide consistent quantile estimation that is almost identical to the true quantiles. Random survival forest is very unstable on predicting the quantiles and is not able to correct the censoring bias. We increase the node size to from 20 to 80, all the methods except for *rsf* become more biased and less variant, but *rsf* is still volatile.

We then repeat the above experiment ten times for different node sizes ranging from 10 to 80, and report the average and standard deviation of the quantile losses in Figure 3. We observe that the performances of both *crf-generalized* and *crf-quantile* are close to their corresponding oracles, and are much better than *grf* or *qrf* on censored data. *crf-quantile* (*qrf*) performs slightly better than *crf-generalized* (*grf*), implying that original quantile random forest can be more effective when the data is homoscedastic. The random survival forest *rsf* behaves only slightly better than the biased *grf* and *qrf*, but is much worse than the proposed methods.

5.1.2 Heteroscedastic Data

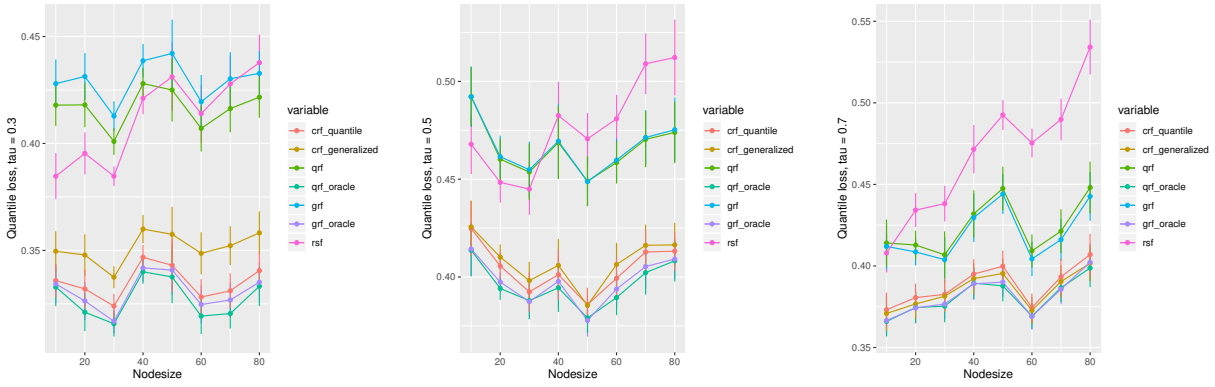
We test the proposed method on a heteroscedastic dataset. The dataset is taken from Athey et al. (2019). We sample $n = 2000$ independent and identically distributed examples where X_i is uniformly distributed over $[-1, 1]^p$ with $p = 40$, and T_i is Gaussian conditionally on $(X_i)_1$ and $T_i|X_i \sim \mathcal{N}(10, (1 + \mathbb{1}\{(X_i)_1 > 0\})^2)$. The censoring variable $C_i \sim 8 + \text{Exp}(\lambda = 0.10)$ and $Y_i = \min(T_i, C_i)$. The other 39 covariates are noise. The censoring ratio is about 20% in this example. We shift the mean of $T_i|X_i$ to 10 because the random survival forest only allows positive responses. We estimate the quantiles at $\tau = 0.1$, 0.5, 0.9. The results are in Figure 2.

When the data is heteroscedastic, using generalized forest weights (*crf-generalized*) provides much more accurate quantile estimation than all the other methods. The predicted quantiles by *crf-generalized* are almost identical to the truths. Our results are inline with Athey et al. (2019) that generalized random forest is very effective at dealing with heteroscedasticity. Note that the random survival forest (*rsf*) also fails to recognize the variance shift. This experiment indicates that our method coupled with generalized forest weights is the most, arguably the only effective method when dealing with heteroscedastic data.

We also repeat the above experiment ten times for different node sizes and report the quantile losses in Figure 4. We again observe that *crf-generalized* achieves almost the same performance of *grf-oracle*. *rsf* and *crf-quantile* have similar performance on this dataset, and are both slightly worse than even *grf* when $\tau = 0.1$. This shows that the splitting rule of random survival forest or quantile forest do not work well on heteroscedastic data.

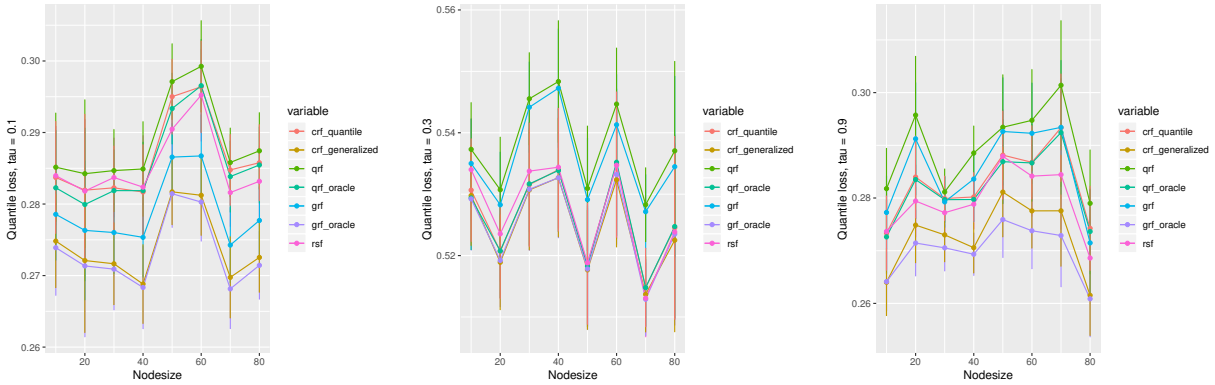
5.2 Conditional Survival Functions

In this section, we compare the two proposed conditional survival function estimators equation 9 and equation 10. We generate examples from the AFT model, and then choose four test points $\{x_1 = 0.4, x_2 = 0.8, x_3 = 1.2, x_4 = 1.6\}$ to plot the conditional survival function estimations on them. The results are shown



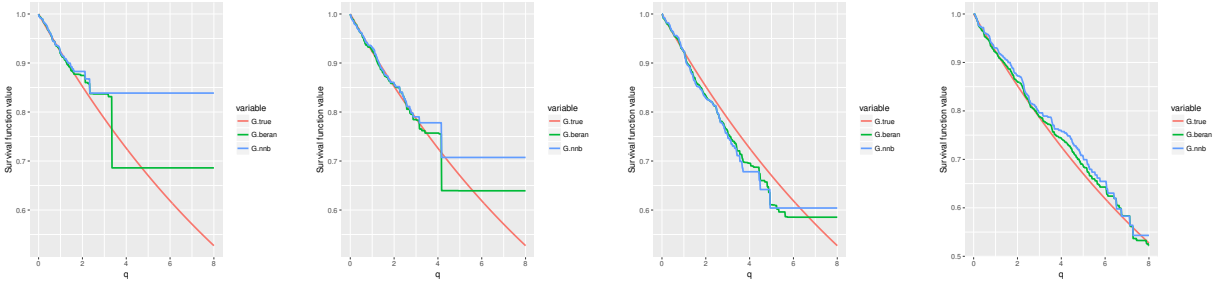
(a) $\tau = 0.3$ (b) $\tau = 0.5$ (c) $\tau = 0.7$

Figure 3: Quantile losses on multi-dimensional AFT data with different node sizes.



(a) $\tau = 0.1$ (b) $\tau = 0.3$ (c) $\tau = 0.9$

Figure 4: Quantile losses on heteroscedastic examples.



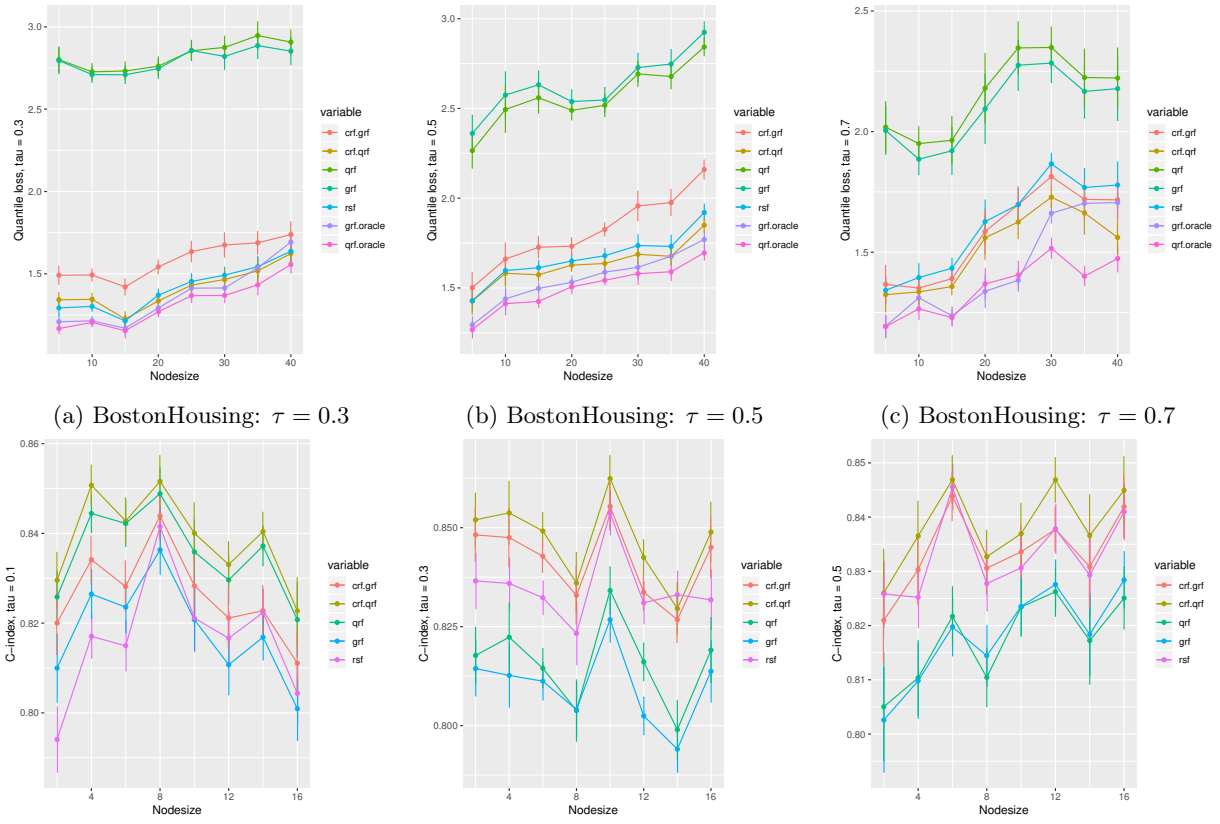
(a) $x = 0.4$ (b) $x = 0.8$ (c) $x = 1.2$ (d) $x = 1.6$

Figure 5: Comparison of the two conditional survival estimators on the AFT data. The sample size is 5000. For the nearest neighbor estimator equation 9, we set the number of neighbors to be 10% of the sample size.

in Figure 5.

We observe that when n increases, two curves become closer and are both good approximations of the true survival curve. But the first method equation 9 has an extra tuning parameter k – the number of nearest neighbors. Therefore, in the experiments, we always choose to use the second estimator equation 10 which is parameter free.

Note that the estimated survival function will degenerate at the tail of the distribution when the test point x is small. This is a common phenomenon even for the regular KM estimator because there is no censored observations beyond some time point. In the AFT model, when x is small, the conditional mean of T is also small, and hence we could not observe most of the censoring values, leading to degenerated survival curves.



(a) BostonHousing: $\tau = 0.3$ (b) BostonHousing: $\tau = 0.5$ (c) BostonHousing: $\tau = 0.7$

(d) pbc: $\tau = 0.1$ (e) pbc: $\tau = 0.3$ (f) pbc: $\tau = 0.5$

Figure 6: The results on real datasets. On BostonHousing, we report the quantile losses (the lower the better), and on pbc dataset, we report the C-index (the higher the better).

5.3 Real Data

In this section, we compare the proposed method with other forest algorithms on two real datasets, BostonHousing (Dua and Graff, 2017) and Primary Biliary Cirrhosis (PBC) Data (Fleming and Harrington, 2011). On the BostonHousing data, we manually generate censoring variables from $\text{Exp}(\lambda = 1/2\bar{y})$ where $\bar{y}$ is the sample mean of the house prices. The censoring level is about 40%. We evaluate the models using quantile loss because we know the true responses in this case. The PBC data is already right-censored, and the censoring rate is about 60%. On this dataset, we cannot use quantile loss for evaluation because we do not know the true responses of the censored data in the test set. (Note that we can sample uncensored data to form a test set, but these data will be biased.) Instead, we use the Harrell's concordance index (C-index) (Harrell Jr et al., 1982). We repeat the experiment for each node size for 50 times and report the mean and standard deviation of the C-index. For each experiment, we randomly sample 80% of the data for training and the rest for testing. All the forests contain 1000 trees. The results are in Figure 6.

Overall, the proposed method with random forest weights *crf-quantile* has the best performance. It agrees with our observation in the simulations that *crf-quantile* works better than the other methods if there is no clear heteroscedasticity in the data.

6 Discussion

In this article, we introduced *censored quantile regression forest*, a novel non-parametric method for quantile regression problems that is integrated with the censored nature of the observations. While preserving information carried by the censored observations, the novel estimating equation maintains the flexibility of general forest approaches. One of the promising applications of the introduced method is in the estimation of heterogeneous treatment effects when the response variable is censored. Treatment discovery with right-censored observations is an important and yet poorly understood research area. Equipping this literature with the proposed fully non-parametric approach would lead to a significant broadening of the now more known parametric approaches. We also observe that our estimating equations can be easily replaced with another kind that targets treatment effects directly.

References

Michael G Akritas. Nearest neighbor estimation of a bivariate distribution under random censoring. *The Annals of Statistics*, pages 1299–1327, 1994.

Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Forthcoming in the Annals of Statistics*, 2018.

Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Rudolf Beran. Nonparametric regression with randomly censored survival data. Technical report, Technical Report, Univ. California, Berkeley, 1981.

GÃŠrard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095, 2012.

GÃŠrard Biau, Luc Devroye, and GÃĄbor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9 (Sep):2015–2033, 2008.

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(10):2499–2518, 2010.

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

Dorota M Dabrowska. Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics*, pages 181–197, 1987.

Dorota M Dabrowska. Uniform consistency of the kernel conditional kaplan-meier estimate. *The Annals of Statistics*, pages 1157–1167, 1989.

Roger B Davis and James R Anderson. Exponential survival trees. *Statistics in Medicine*, 8(8):947–961, 1989.

Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *International conference on machine learning*, pages 665–673, 2014.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Thomas R Fleming and David P Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63 (1):3–42, 2006.

W Gonzalez-Manteiga and C Cadarso-Suarez. Asymptotic properties of a generalized kaplan-meier estimator with some applications. *Communications in Statistics-Theory and Methods*, 4(1):65–78, 1994.

Louis Gordon and Richard A Olshen. Tree-structured survival analysis. *Cancer treatment reports*, 69(10): 1065–1069, 1985.

Frank E Harrell Jr, Robert M Califf, David B Pryor, Kerry L Lee, Robert A Rosati, et al. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

Donald R. Hoover, John A. Rice, Colin O. Wu, and Li-Ping Yang. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822, 1998.

Torsten Hothorn, Berthold Lausen, Axel Benner, and Martin Radespiel-Tröger. Bagging survival trees. *Statistics in medicine*, 23(1):77–91, 2004.

Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, 2005.

Jian Huang, Shuangge Ma, and Huiliang Xie. Least absolute deviations estimation for the accelerated failure time model. *Statistica Sinica*, pages 1533–1548, 2007.

Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

Zhezhen Jin, DY Lin, LJ Wei, and Zhiliang Ying. Rank-based inference for the accelerated failure time model. *Biometrika*, 90(2):341–353, 2003.

H Koul, V v Susarla, J Van Ryzin, et al. Regression analysis with randomly right-censored data. *The Annals of statistics*, 9(6):1276–1288, 1981.

Michael LeBlanc and John Crowley. Relative risk trees for censored survival data. *Biometrics*, pages 411–425, 1992.

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

Alexander Hanbo Li and Andrew Martin. Forest-type regression with general losses and robust forest. In *International Conference on Machine Learning*, pages 2091–2100, 2017.

Gang Li and Hani Doss. An approach to nonparametric regression for life history data using local linear fitting. *The Annals of Statistics*, pages 787–823, 1995.

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

Thomas A. Louis. Nonparametric analysis of an accelerated failure time model. *Biometrika*, 68(2):381–390, 1981.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

Annette M Molinaro, Sandrine Dudoit, and Mark J Van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1):154–177, 2004.

James Robins. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79(2):321–334, 1992.

James Robins and Anastasios A Tsiatis. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*, 79(2): 311–319, 1992.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

Erwan Scornet, Gérard Biau, Jean-Philippe Vert, et al. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

Mark Robert Segal. Regression trees for censored data. *Biometrics*, pages 35–47, 1988.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Ingrid Van Keilegom and Noël Veraverbeke. Uniform strong convergence results for the conditional kaplan-meier estimator and its quantiles. *Communications in Statistics–Theory and Methods*, 25(10):2251–2265, 1996.

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.

Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

Lee-Jen Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15): 1871–1879, 1992.

Donglin Zeng and DY Lin. Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, 102(480):1387–1396, 2007.

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