Title

Censored Quantile Regression Forest

Authors

Li, Alexander Hanbo
Bradic, Jelena

Publication Date

2020

Peer reviewed

Censored Quantile Regression Forests

Alexander Hanbo Li and Jelena Bradic

Department of Mathematics
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 new estimating equations that adapt to censoring and lead 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.

Keywords: Random Forest; Censored quantile regression; Nonparametric regression; Kaplan-Meier estimation.

1 Introduction

In many applications, we want to predict and estimate the effect of a covariate on survival time of interests. Examples include treatment, surgical procedure, or immunization on survival time of patients, who for example, could be individuals who have metastatic breast cancer, military casualties suffering from various injuries, or survival time of infectious diseases. Classically, most datasets have been too small to meaningfully examine the heterogeneity of the data beyond dividing them into a few sub-populations. In the past few years, however, there has been an explosion of experimental settings where it is potentially feasible to explore heterogeneity to its full extent.

An impediment to exploring heterogeneous effects is the fear that scientists with two opposite agendas could hypothetically string together two opposite but coherent results by searching through many different possible models and then reporting only the very extreme ones – highlighting solely spurious results (Olken, 2015). Thus, protocols for clinical trials must specify in advance the pre-analysis plans and then learn from the data. However, such restrictions can make it challenging to discover unexpected effects due to heterogeneity. Here, we aim to address this challenge by developing a robust, non-parametric method for estimation in regression settings with censored response variable, which yields consistent estimator that adapts to the heterogeneity of the data and hence can be broadly applied to many different models and achieves further improvement. One example is the accelerated failure time model.

Classical approaches to accelerated failure time model include non-parametric maximum likelihood method, semi-parametric approaches, as well as traditional parametric approaches; see e.g., Zeng and Lin (2007), Robins and Tsiatis (1992), Robins (1992), as well as Koul et al. (1981), Wei (1992), Huang et al. (2007) etc. These methods perform well in settings where the model error is correctly specified, but quickly break down when the distribution of the error has any heterogeneous, asymmetric, or outlier structure. Other non-parametric methods, like Louis (1981), Hoover et al. (1998), and rank-based methods, like Jin et al. (2003), strive to achieve certain level of assumption-lean modeling of the mean in the accelerated failure time model. In this paper, we explore the use of ideas

from the machine learning literature to improve the performance of these classical methods in a non-parametric fashion that is adaptive to heterogeneous or heavy-tailed error distributions.

Random forest algorithms introduced by Breiman (2001) allow for flexible modeling of covariate interactions, and are related to kernels and nearest-neighbor methods in the sense that they make predictions using a weighted average of "nearby" observations. However, random forests depart from the above principles in that they have a data-driven way to determine which nearby observations receive more weight, something that is especially important in environments with many covariates or complex interactions among covariates.

Despite their wide-spread success at estimation and prediction, application of random forests to censored regression models is far from being easily understood. Not all response variables being observed makes it difficult to understand how to evaluate the prediction arising from simple tree structures. Namely, a simple average is only enough for conditional mean estimation without censoring. Median estimates, and quantiles more generally, are not easy to construct. In particular, quantile random forests (Meinshausen, 2006) have mainly been undeveloped for censored observations.

This paper addresses these limitations, developing a forest-based method for estimation of quantiles for randomly censored observations (right or left censored). More formally, let T be a real-valued latent variable (e.g., survival time) and X be a (possibly high-dimensional) predictor variable. Let C denote the censoring variable, which prevents us from observing all of the information regarding T. In left-censored data, we only observe $Y_i = \max(T_i, C_i)$, whereas in right-censored data, we only observe $Y_i = \min(T_i, C_i)$. Our goal is to estimate the τ-th quantile of T_i, non-parametrically, using observations (Y_i, X_i, δ_i) where $\delta_i = \mathbb{1}\{T_i \leq C_i\}$.

We take on the perspective of random forests as that of an nonparametric, adaptive kernel smoothing method. This interpretation follows work by Athey et al. (2018), Bloniarz et al. (2016), Hothorn et al. (2004), Li and Martin (2017), and Meinshausen (2006), and supplements the customary view of forests as an ensemble method (i.e., an average of predictions made by separate trees). In this view, random forest predictions, of the mean

for example, evaluated at a new test point x, can be represented by

$$\sum_{i=1}^{n} w_i(X_i, x) Y_i, \qquad \sum_{i=1}^{n} w_i(X_i, x) = 1, \qquad w_i(X_i, x) \geq 0$$

where the weights encode the similarity between the new test point x and the observed covariates X_i. It is worth pointing that for the conditional mean, the averaging and weighting views of random forests are equivalent; however, once we move to more general settings, the weighting-based perspective proves substantially more powerful. The goal is then to utilize these local neighborhood weights and quantile regression adjustments to design a new non-parametric quantile estimate of

$$Q_{T|x}(\tau|x), \qquad \tau \in (0, 1), \tag{1}$$

based on observations $(Y_i, X_i, \delta_i)_{i=1}^n$. In their simplest form, censored quantile forests just take the forest weights, $w_i(X_i, x)$, and use them for quantile regression:

$$\hat{q}_\tau(x) = \arg\min_q \left\{ \left\| \sum_{i=1}^{n} w_i(X_i, x) \mathcal{S}_\tau(X_i, Y_i, q, x) \right\|_l \right\},$$

where $\|\cdot\|_l$ denotes any l-norm with $l \geq 1$. Here, we define a new score function, $\mathcal{S}_\tau$, that is censoring and quantile adaptive:

$$\mathcal{S}_\tau(X_i, Y_i, q, x) = (1 - \tau)\hat{G}(q|x) - \mathbb{1}(Y_i > q)$$

with $\hat{G}$ denoting an estimate of $G(u) = \mathbb{P}(C \geq u | X = x)$, the survival function of the censoring time C given X. In the rest of the document we focus on the case of $l = 1$; however, results remain true for general cases of l as well.

Formally, we study the performance of the above non-parametric quantile estimator $\hat{q}_\tau(x)$ of (1) while considering right-censored observations. All of our results can be trivially extended to the left-censored case. Most of the theoretical work focuses on establishing Theorem 3 that ensures the consistency of the censored quantile forest estimator $\hat{q}_\tau(x)$ at a given test point x while allowing minimal assumptions on the regression model as well as the distribution of the model error. This result also allows us to construct prediction intervals regarding (1) that adapt to the model structure of the data generating process. Censored

quantile forest estimator improves on prediction error compared with many state-of-the-art censored regression methods, and yet it retains the flexibility of random forest methods designed for non-censored observations. Finally, we note that our method can also be seen as an improvement over classical non-parametric approaches for censored observations. While the latter only perform well in low-dimensional problems, ours performs well even with a large number of covariates. One important reason is that random forest weights adapt reasonably well to the dimensionality increase, whereas kernel methods suffer from the curse of dimensionality.

1.1 Related Work

There has been a long-time understanding that proportional hazard models, and cox's model, in particular, are especially powerful for right-censored observations and regression problems. However, they do not adapt to possibly left-censored observations; besides, they heavily rely on the proportionality assumption which can sometimes be inappropriate, necessitating stratification of the baseline hazard or some other weakening of the proportional hazards condition (Koenker et al., 2008).

A more flexible approach for random censoring problem is to model the conditional quantiles of the response variable directly. This approach offers greater flexibility as it does not restrict the structure of the hazard function (Koenker et al., 2008) and merely is more intuitive. To estimate the conditional quantiles, Portnoy (2003) proposed a recursive method which estimates a sequence of linear conditional quantile functions recursively. It can be treated as a generalization to regression of the Kaplan Meier estimator. Another closely related quantile regression model proposed by Peng and Huang (2008) instead makes the linkage to the Nelson-Aalen estimator of the cumulative hazard function, upon which they developed a complete asymptotic theory. The closest work to ours is that of Backer et al. (2018) where authors propose a new censored loss function. However, they only discuss the properties of quantile estimates in the case of linear quantile model.

The parametric methods, including those mentioned above, always rely on the linearity

assumption on the conditional quantiles, that is,

$$Q_{\log(T)|x}(\tau|x) = x^\top \beta. \tag{2}$$

Here, the log transformation is arbitrary but popular in survival analysis and can be replaced by any monotone function. This linearity assumption is too restrictive in many cases, especially when data lie on a complex manifold. Therefore, non-parametric methods are necessary and play an important role in modeling data heterogeneity.

In the case of right censoring, most non-parametric recursive partitioning algorithms in the existing literature 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 to deal with the right censored data. Despite different splitting strategies, 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.

This work aims to build a non-parametric conditional quantile estimator for randomly

censored data that reduces to the ordinal quantile forest estimators when full data are observed, and efficiently utilizes all available information on both uncensored and censored observations. Furthermore, it does not require the specific modification of ordinal regression tree (e.g., CART) to survival tree, and hence works on both left and right censored problems.

Fundamentally different from the aforementioned forests methods, in which the censoring information is considered directly in the tree constructing process, our method avoids this complexity and only requires building ordinal regression trees (e.g., CART) for the first step, treating all observations equally. The censoring effects are then taken care of in the second step by solving a locally weighted estimating equation. These local weights can be directly calculated from the random forest constructed in the first step; weights derived from the fraction of trees in which an observation appears in the same leaf as the target value of the covariate vector. This locally weighted view of random forests was previously advocated by Hothorn et al. (2005) and Bloniarz et al. (2016); original random forest algorithm (Breiman, 2001) utilized ensemble learning literature. Differently from kernel weights, typically employed in local maximum likelihood method, for example, and whose performance suffers greatly whenever the dimensionality of the covariate space is more than two or three, our random-forests weights adapt well to moderate dimensionality increases.

Additional challenges arise due to the random censoring nature of the observations. For fixed censoring, one observes all the censoring values and hence can straightforwardly modify the objective used in the general framework of Athey et al. (2018), for example. However, it is unclear how to develop a non-parametric estimator that adapts to unknown censoring in the observations.

1.2 Organization

The paper is organized as follows. In Section 2.1 we provide local adaptive nature of random forests and regression adjustment utilizing those weights. In Sections 2.2 and 2.3, we showcase the development of a new loss function and its power in predicting any

conditional quantile of the latent variable T by solving an ingenious estimating equation, which is designed to correct the censoring effect. The Algorithm is then described in details in Section 2.4. In Section 3, we analyze the time complexity of our algorithm and prove consistency of the proposed estimator. Section 4 contains extensive numerical studies where we compare our algorithm with other forest algorithms on simulated and real censored data sets. Proofs are collected in the Appendix.

2 Censored Quantile Regression Forest

The quantile random forest cannot be directly applied to censored data $\{(X_i, Y_i)\}$ because the conditional quantile of Y is different from that 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.

2.1 Regression adjustment for random forests

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. Since the space $\mathcal{X}$ is split into disjoint leaves by $T(\theta)$, we know for any $x \in \mathcal{X}$, there is exactly one leaf containing x. We let the index of the leaf be $l(x;\theta)$ and we say $x \in R_{l(x;\theta)}$.

Then for any single tree $T(\theta)$, the prediction on any data point $x \in \mathcal{X}$ is $\sum_{i=1}^n w(X_i, x; \theta) Y_i$ where

$$w(X_i, x; \theta) = \frac{\mathbb{1}_{\{X_i \in R_{l(x;\theta)}\}}}{\#\{j : X_j \in R_{l(x;\theta)}\}}. \tag{3}$$

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

$$w(X_i, x) = \frac{1}{m} \sum_{t=1}^m w(X_i, x; \theta_t). \tag{4}$$

From now on, we call the weight $w(X_i, x)$ in (4) *random forest weight*. One can easily show that $\sum_{i=1}^{n} w(X_i, x) = 1$. The above representation of the random forest prediction of the mean can be equivalently obtained as a solution to the following 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. (2018) to obtain a smoother and more poweful random forest algorithm.

2.2 Motivation

Let us consider the case of no censored observations. Full data serve as a motivation for developing suitable estimating equations. Following the regression adjustment reasoning, for the case of fully observed data, we could estimate the τ-th quantile of T_i at x, denoted as $q_{\tau,x}$, as a solution to

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

Equivalently, such estimate would solve the following estimating equations

$$\begin{aligned} U_n(q) &= \sum_{i=1}^{n} w(X_i, x)\Big\{(1-\tau) - \mathbb{1}(T_i > q)\Big\} \\ &= (1-\tau) - \sum_{i=1}^{n} w(X_i, x)\,\mathbb{1}(T_i > q) \approx 0, \end{aligned} \tag{5}$$

where the second equality is true because $\sum_{i=1}^{n} w(X_i, x) = 1$.

For simplicity and better illustration of the idea, we first assume the latent variable T_i has the same conditional probability in a neighborhood R_x of x. Out of the n data points, assume $\{X_1, \cdots, X_k\} \subset R_x$ and $w(X_i, x) = 1/k$ when $X_i \in R_x$ and 0 otherwise. Now the estimating equation becomes

$$\begin{aligned} U_k(q) &= \frac{1}{k}\sum_{i=1}^{k}\Big\{(1-\tau) - \mathbb{1}(T_i > q)\Big\} \\ &= (1-\tau) - \frac{1}{k}\sum_{i=1}^{k}\mathbb{1}(T_i > q). \end{aligned} \tag{6}$$

Now conditional on $\{x\} \cup \{X_i\}_{i=1}^k$,

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

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

Let's now consider the case of right-censored setting, where we further have the censoring variable C_i, which is independent of T_i conditional on X_i, and we could only observe $Y_i = \min\{T_i, C_i\}$ and censoring indicator $\delta_i = \mathbb{1}(T_i \leq C_i)$. In order to estimate $q_{\tau,x}$, we cannot simply replace T_i with Y_i in (6) as the τ-th quantile of T_i is no longer the τ-th quantile of Y_i because of the censoring. However, we can observe and utilize the following relationship

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

where $G(u|x)$ is the 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. Now, we define a new estimating equation that resembles (6) as follows

$$S_k^o(q) = \frac{1}{k}\sum_{i=1}^{k}\Big\{(1-\tau)G(q|x) - \mathbb{1}(Y_i > q)\Big\} \approx 0. \tag{7}$$

If we substitute $G(q|x)$ with $G(q_{\tau,x}|x)$, an intuitive explanation for (7) is that as the τ-th quantile of T_i happens to be 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 turn to estimate

the later one. Namely, the conditional expectation, $\mathbb{E}[S_k^o(q)]$, will still be zero at the same root q^* for (6).

The survival function $G(\cdot|x)$ can be estimated by a consistent estimate, 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 define

$$S_k(q) = \frac{1}{k}\sum_{i=1}^{k}\Big\{(1-\tau)\hat{G}(q|x) - \mathbb{1}(Y_i > q)\Big\} \approx 0. \tag{8}$$

2.3 Full model

In the previous subsection, we made an assumption that $\mathbb{P}(T|X) = \mathbb{P}(T|x)$ for all $X \in R_x$, where R_x is a neighborhood of 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.

For example, say we have n data points $\{(X_i, T_i)\}_{i=1}^n$ and have two cases: (1) at all X_i's we have the same conditional probability of T, i.e. $\mathbb{P}(T|X_i) = \mathbb{P}(T|X_j)$ for all i, j; (2) T has different conditional probabilities at different locations. In the setting (1), X_i's become irrelevant and the point mass on each T_i is $1/n$. We share the mass uniformly to the n points T_i's as they are equally important. When $n \to \infty$, it is known that for any q,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}(T_i \le q) \to \mathbb{P}(T \le q|x). \tag{9}$$

However, in the case (2), the convergence (9) is no longer valid. We cannot simply put a mass $1/n$ on each T_i because the probability of T_i showing up at X_i could be severely different than the probability it shows up at x. An extreme example is when $\mathbb{P}(T|x) = \text{Unif}(x-1, x+1)$. Then if $|X_i - x| > 1$, any T_i showing up at X_i should not even be counted when estimating $\mathbb{P}(T|x)$ because $\mathbb{P}(T_i|x) = 0$. In another word, we should give T_i mass 0 instead of $1/n$.

Therefore, a measure of "similarity" between points X_i and x needs to come into play, because we can no longer uniformly distribute the mass; observe that some T_i's are more important than others for estimating $\mathbb{P}(T|x)$. For instance, if $X_i = x + 0.01$ and $X_j = x + 2$ in the previous example, then T_i should be assigned much more weight than T_j.

Now let $w(X_i, x)$ denote the weight (mass) we assign to T_i when we are estimating $\mathbb{P}(T|x)$. In the setting (1), we just have $w(X_i, x) = 1/n$ uniformly. But in the setting (2), we should have $w(X_i, x) > w(X_j, x)$ when X_i is more similar to x than X_j in some sense. Therefore, the estimator for $\mathbb{P}(T \leq q|x)$ is then

$$\sum_{i=1}^{n} w(X_i, x)\, \mathbb{1}(T_i \leq q)$$

and it becomes clear that a proper weight $w(X_i, x)$ needs to satisfy:

$$(1)\ \sum_{i=1}^{n} w(X_i, x) = 1; \quad (2) \sum_{i=1}^{n} w(X_i, x)\, \mathbb{1}(T_i \leq q) \xrightarrow{p} \mathbb{P}(T \leq q|x)\ \forall q. \tag{10}$$

One may naively think that any fixed Kernel weights, $K(X_i, x)$, could be a suitable choice. However, they would not be able to satisfy the second condition in (10) for any distribution $\mathbb{P}(T|x)$. Fortunately, as shown in Meinshausen (2006), the data-adaptive random forest weight $w(X_i, x)$ introduced in Section 2.1 perfectly satisfy both conditions in (10). And therefore going back to (5), we now consider estimating equations,

$$U_n(q_{\tau,x}) = (1 - \tau) - \sum_{i=1}^{n} w(X_i, x)\, \mathbb{1}(T_i > q_{\tau,x}) \xrightarrow{p} 0 \tag{11}$$

when $n \to \infty$. Then following the same logic of how we get (8), a heuristic extension of (5) to censoring case will be

$$\begin{aligned} S_n(q; \tau) &= \sum_{i=1}^{n} w(X_i, x)\Big\{(1 - \tau)\hat{G}(q|x) - \mathbb{1}(Y_i > q)\Big\} \\ &= (1 - \tau)\hat{G}(q|x) - \sum_{i=1}^{n} w(X_i, x)\, \mathbb{1}(Y_i > q). \end{aligned} \tag{12}$$

2.4 Forest Algorithm

In the simplified example in Section 2.2, we assume that Y has the same conditional probability $\mathbb{P}(Y|X)$ in a neighborhood R_x of x, and hence, we can estimate $G(q|x)$ by

Kaplan-Meier estimator (Kaplan and Meier, 1958) (assuming no tied events)

$$\begin{aligned}\hat{G}(q|x) &= \prod_{i:X_{(i)}\in R_x, Y_{(i)}\leq q}\left(1-\frac{1}{k-i+1}\right)^{1-\delta_{(i)}} \\ &= \prod_{i:X_i\in R_x, Y_i\leq q}\left(1-\frac{1}{\sum_{j=1}^n \mathbb{1}(Y_j\geq Y_i)\,\mathbb{1}(X_j\in R_x)}\right)^{1-\delta_i} \end{aligned} \tag{13}$$

where $k = |R_x|$. In the more complex case like in Section 2.3, many consistent estimators for the conditional survival functions exist. For example, the nonparametric estimator proposed by Beran (1981)

$$\tilde{G}(q|x) = \prod_{Y_i\leq q}\left\{1-\frac{W_i(x,a_n)}{\sum_{j=1}^n \mathbb{1}(Y_j\geq Y_i)W_j(x,a_n)}\right\}^{1-\delta_i} \tag{14}$$

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, $W_i(x, a_n)$ are the Nadaraya-Watson weights

$$W_i(x,a_n) = \frac{K((x-X_i)/a_n)}{\sum_{j=1}^n K((x-X_j)/a_n)},$$

$K(\cdot)$ is a known kernel and $\{a_n\}$ is a bandwidth sequence tending to zero as n tends to infinity. We can then simply use $\tilde{G}(q|x)$ as $\hat{G}(q|x)$ in (12).

However, since we already have an adaptive version of kernel – the random forest weights $w(X_i, x)$, we will propose the following two new estimators for $G(q|x)$.

Kaplan-Meier using nearest neighbors. The first estimator is resembles that of (13). We first find the k nearest neighbors of x according to the weights $w(X_i, x)$. Denoting these points as a set N_x, then we can simply use the Kaplan-Meier estimator on N_x

$$\hat{G}(q|x) = \prod_{i:X_i\in N_x, Y_i\leq q}\left(1-\frac{1}{\sum_{j=1}^n \mathbb{1}(Y_j\geq Y_i)\,\mathbb{1}(X_j\in N_x)}\right)^{1-\delta_i}. \tag{15}$$

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

Beran estimator with random forest weights. In the second proposal, we will replace the Nadaraya-Watson weights in (14) with random forest weights and get

$$\hat{G}(q|x) = \prod_{Y_i \le q} \left\{ 1 - \frac{w(X_i, x)}{\sum_{j=1}^{n} \mathbb{1}(Y_j \ge Y_i) w(X_j, x)} \right\}^{1-\delta_i}. \tag{16}$$

One could observe that (15) is a special case of (16) when the weight $w(X_i, x) = 1/k$ for $X_i \in R_x$ and 0 otherwise.

Finally, we summarize our main algorithm in Algorithm 1. The details for choosing the candidate set $\mathcal{C}$ is in Section 3.1. The choice to minimize the absolute value of $S_n(q;\tau)$ is arbitrary. The goal is to find the approximate root of $S_n(q;\tau) = 0$.

Algorithm 1 Censored quantile regression forest

All tuning parameters, including the number of trees, B, in the forest as well as the minimum size of the leaves, m, in the individual trees are pre-determined. If using the KM-estimator (15), the number of nearest neighbors, k, is also given.

procedure FOREST-CQR(test point x, training set $\mathcal{D} = \{(X_i, Y_i, \delta_i)\}_{i=1}^n$, quantile τ)

tree $\mathcal{T} \leftarrow$ REGRESSION FOREST $(\mathcal{D})$ with B trees and leaf sizes m

random forest weights $w(x, X_i) \leftarrow$ (4), for all i

survival function $\hat{G}(q|x) \leftarrow$ (15) or (16)

quantile estimator $\hat{q}$ such that

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

▷ $\mathcal{C}$ is a candidate set as discussed in Section 3.1 ▷ S_n is (12)

return $\hat{q}$ ▷ The τ-th quantile of survival time T at x

The function RANDOM FOREST creates a regression tree and can be implemented to be as any of the standard forest algorithms.

3 Theoretical Develoments

In this section, we will assume the random forest has terminal node size m, feature vector $X_i \in \mathbb{R}^p$, sample size is n, and k nearest neighbors are chosen if using (15).

3.1 Time complexity

The step 6 in Algorithm 1 involves of finding the q^* in a candidate set $\mathcal{C}$ that sets the estimating equation $S_n(q;\tau)$ closest to zero. We simply evaluate the function $S_n(q;\tau)$ for all possible q in $\mathcal{C}$ and find the minimum point. Note that for any fixed τ, $S_n(q;\tau)$ is a step function in q with jumps at Y_i's because the discontinuities only happen at Y_i's for $\hat{G}(q|x)$ (both (15) and (16)) and $\sum_{i=1}^n w(X_i,x)\,\mathbb{1}(Y_i > q)$. Therefore, the candidate set $\mathcal{C} \subset \{Y_i\}_{i=1}^n$, and $|\mathcal{C}| = n$ in the worst case.

But in fact, for any fixed x, only Y_i's with the corresponding feature vector $X_i \in R_x$ (15) or with $w(X_i,x) > 0$ (16) will be jump points, and hence, we can refine $\mathcal{C} = \{Y_i : X_i \in R_x\}$ for (15) or $\mathcal{C} = \{Y_i : w(X_i,x) > 0\}$ for (16). We then have the following theorem. The proof is given in the Appendix 6.

Theorem 1. *For a fixed test point x, depending on whether $G(q|X)$ is estimated by (15) or (16), the time complexity for Algorithm 1 is $O(n\max\{k,\log(n)\})$ or $O(nm\log(n)^{p-1})$, respectively.*

3.2 Consistency

In this section, we will show that for any fixed $\tau \in (0,1)$, $S_n(q;\tau)$ in (12) will converge in probability to $(1-\tau)G(q|x) - \mathbb{P}(Y_i > q)$ uniformly for q.

Condition 1. *The density of X is positive and bounded from above and below by positive constants on the support $\mathcal{X}$.*

We note that Condition 1 is a very primitive condition on the distribution of the covariates. It is satisfied for example for Gaussian distribution and more broadly for most symmetric, continuous distributions with unbounded support. The case of bounded or discrete covariates is beyond the scope of the current work.

Condition 2. *The terminal node size $m \to \infty$ and $m/n \to 0$ as $n \to \infty$. Furthermore, 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 $\gamma \in (0, 0.5]$.*

The two requirements of Condition 2 are also required in Meinshausen (2006) (see Assumptions 2 and 3 therein). This condition states that the leaf node size of each tree should increase with the sample size n, but at a slower rate. Intuitively, first, the trees that we are using need to be shallow (i.e., with large leaves) in order to estimate a more complex model, reliably. Secondly, there can not be leaves with no samples, i.e., each leaf must be large enough to capture the local estimating equations more adequately. Our experiments also justify the necessity of Condition 2, as the performance of our model, will deteriorate if we keep a small leaf node size but increase the sample size. We will talk about this in detail in Section 4.2.6.

Condition 3. *Denote $F(y|x) = \mathbb{P}(Y \leq y|x)$. There exists a constant L such that $F(y|x)$ is Lipschitz continuous with parameter L, that is, for all $x, x^{'} \in \mathcal{X}$,*

$$\sup_{y} |F(y|x) - F(y|x^{'})| \leq L\|x - x^{'}\|_1.$$

We note that Condition 3 appears in all existing work related to quantile regression and inference thereafter.

Condition 4. *The response variable T and the censoring variable C are independent conditional on X, and the conditional distribution $\mathbb{P}(T \leq q|x)$ and $\mathbb{P}(C \leq q|x)$ are both positive and strictly increasing in q for all $x \in \mathcal{X}$.*

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 5. *For any $x \in \mathcal{X}$, the estimator $\hat{G}(q|x)$ converges pointwisely to the true conditional survival function $G(q|x)$.*

Condition 5 is satisfied, for example, by the Kaplan-Meier estimator (14) (Dabrowska, 1989). Please take a look at Figure 4 and Figure 5 where we compare finite sample properties of the newly introduced estimators (15) and (16). 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 2. *Define*

$$S(q;\tau) = (1-\tau)G(q|x) - \mathbb{P}(Y > q). \tag{17}$$

Under Conditions 1 – 5, for any $x \in \mathcal{X}$, $r > 0$, $\tau \in (0,1)$,

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

Next, we present our main result that illustrates an asymptotic consistency of the proposed conditional quantile estimator. The proof is given in Appendix 6.

Theorem 3. *Under Conditions 1 – 5, 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*

$$\mathbb{P}(T \le q^*|x) = \tau,$$

and as $n \to \infty$,

$$q_n \xrightarrow{p} q^*.$$

4 Experiments

In this section, we will compare our model, censored forest regression (*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 simulated and real data sets.

On the simulated data sets, we will apply *qrf* and *grf* to the censored data directly, and get biased models which we denote by *qrf* and *grf*, respectively. We also apply *qrf* and *grf* to the data with uncensored responses, and call the resulted models *qrf-oracle* and *grf-oracle*.

Throughout this section, we fix the number of trees for each forest to be 1000. The only tuning parameter we have is the node size of each tree. All other parameters are kept as default.

4.1 Illustrative example

In this section, we generate the true response, i.e., latent, variables $T_i \sim \text{Unif}(0,1)$, and consider the censoring variables $C_i \sim \mathcal{N}(0.8, 0.2^2)$. The censored responses, i.e. observed responses, are then taken as $Y_i = \min(T_i, C_i)$. We compare the estimating equation on the latent variables T_i

$$U_1(q) = (1-\tau) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(T_i > q)$$

to the estimating equation of our proposed algorithm

$$U_2(q) = (1-\tau)\hat{G}(q) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(Y_i > q),$$

where $\hat{G}(q)$ is the one-dimensional Kaplan-Meier estimator for the survival function of censoring variable C. The results are shown in Figure 1. We consider $\tau = 0.5$ but the results persist for many other choices of τ.

In Figure 1 we present the two estimating equations as functions of q and illustrate that the solutions to $U_1(q) = 0$ and $U_2(q) = 0$ are closer and closer together when the sample size grows. The solution for $U_1(q) = 0$ can be treated as an oracle solution where the oracle observes "uncensored" (true) response variable. Figure 1 therefore indicated that the root of our method's estimating equation is very close to the oracle root and that we are therefore finding a good approximation to the unknown parameter of interest.

4.2 Simulation results

In this section, we will compare different forest algorithms on simulated data sets including accelerated failure time model (AFT) and many non-parametric censored regression models.

4.2.1 One-dimensional AFT model

We simulate data from an one-dimensional AFT model

$$\log(T) = X + \epsilon$$

where $X \sim \text{Unif}(0,2)$ and $\epsilon \sim \mathcal{N}(0, 0.3^2)$. Then the censoring variable $C \sim \text{Exp}(\lambda = 0.08)$, and the observed response $Y = \min(T, C)$ and the censoring indicator $\delta = \mathbb{1}(T \leq C)$. The

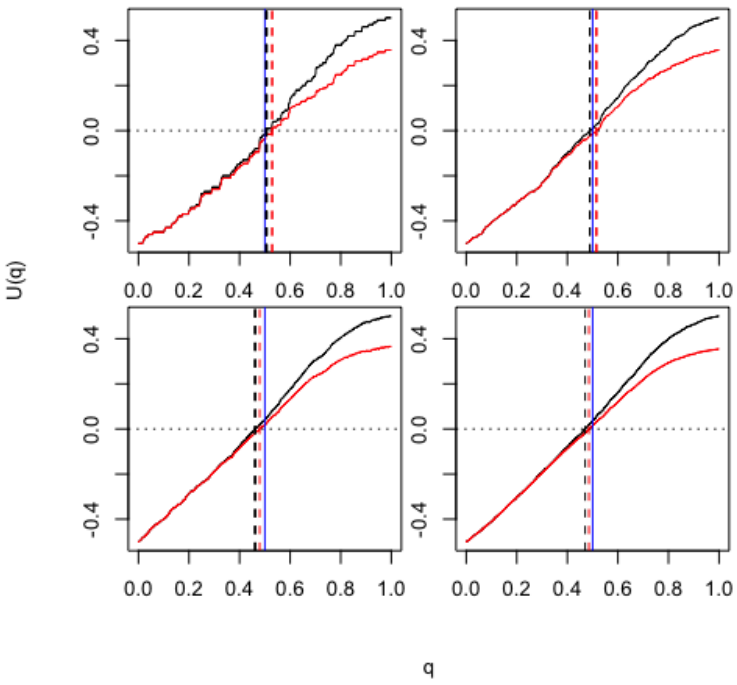
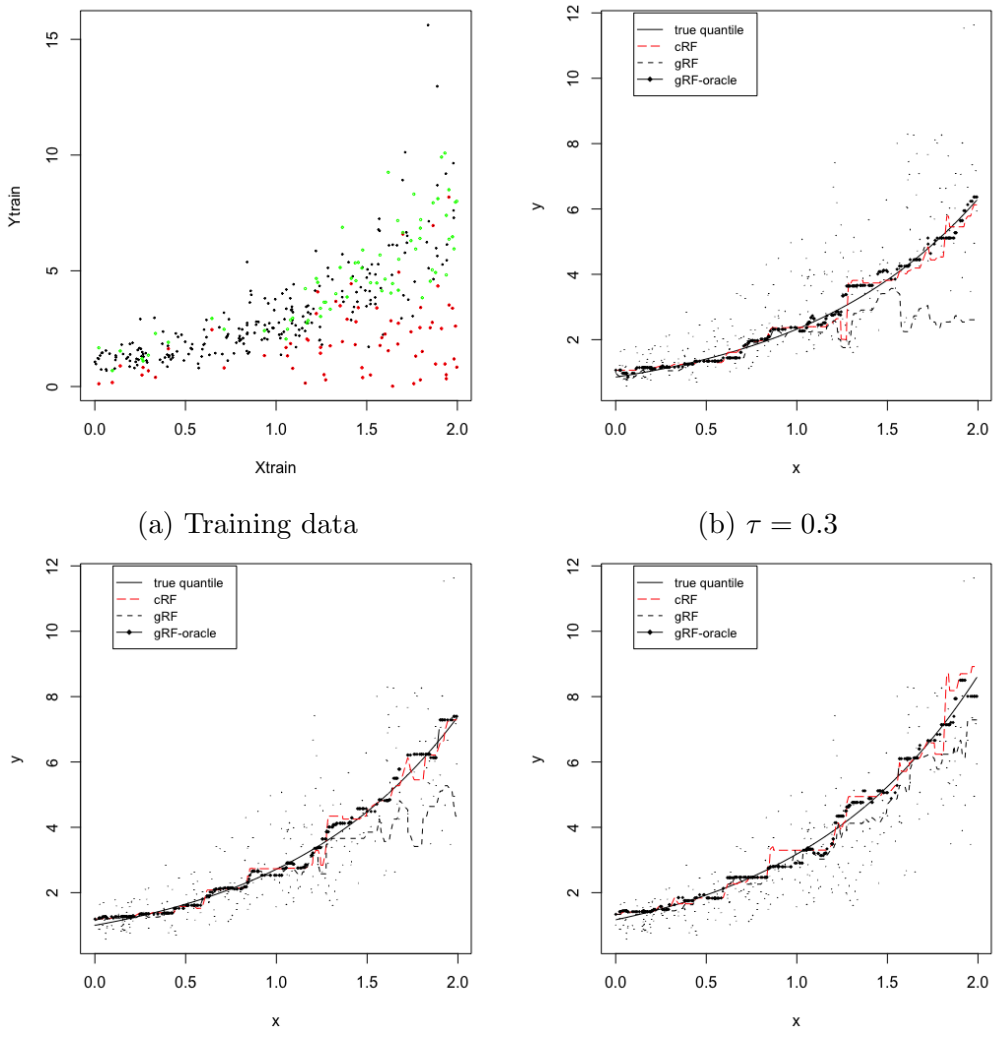


Figure 1: Sample loss plot for different sample size when $\tau = 0.5$. For upper left figure, we have sample size 100, for upper right, $n = 500$, for lower left, $n = 1000$, and for lower right, $n = 5000$. The black curve is the value of $U_1(q)$, the red cureve is the value of $U_2(q)$, the black dotted vertical line is the root of $U_1(q)$, the red dotted vertical line is the root of $U_2(q)$, and the blue vertical line is $q = \tau$.

average censoring rate is about 20%. The number of training data, validation data and test data are all 300. All the forests consist of 1000 trees. The node size of each forest is determined by cross–validation. We plot out one set of training data and the corresponding quantile predictions for $\tau = 0.3, 0.5, 0.7$ on a set of test data in Figure 2. We only show the results of *crf*, *grf*, and *grf-oracle* because in one dimension, *qrf*'s performance is visually indistinguishable from *grf*. There we observe a consistency of our method as well as superior behavior to the competing methods. Namely, the generalized random forest that ignores the censoring component of the data, incurs large bias; due to the right censoring, bias is larger for lower values of the quantiles. We observe that the proposed *crf* follows closely the oracle estimator and is extremely close to the true quantile regardless of the τ in the

study.



(a) Training data

(b) $\tau = 0.3$

(c) $\tau = 0.5$

(d) $\tau = 0.7$

Figure 2: One-dimensional AFT model results with $n = 300$ and $B = 1000$. In (a), black points stand for observations that are not censored; red points are observations that are censored, i.e. the truly observed data points, and the green points are the counterpart of the red points, that is, they are the latent values of those red points if they were not censored.

Moreover, we proceed further and for a set of values $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, we repeat the process 40 times, and for each time, we calculate the MSE and MAD between

the estimated quantiles and the true quantiles, and the τ-th quantile loss. To be more specific, let T_i be the response in test set (all uncensored), Q_i^τ be the true τ-th quantile, and $\hat{Q}_i^\tau$ be the estimated quantile, then

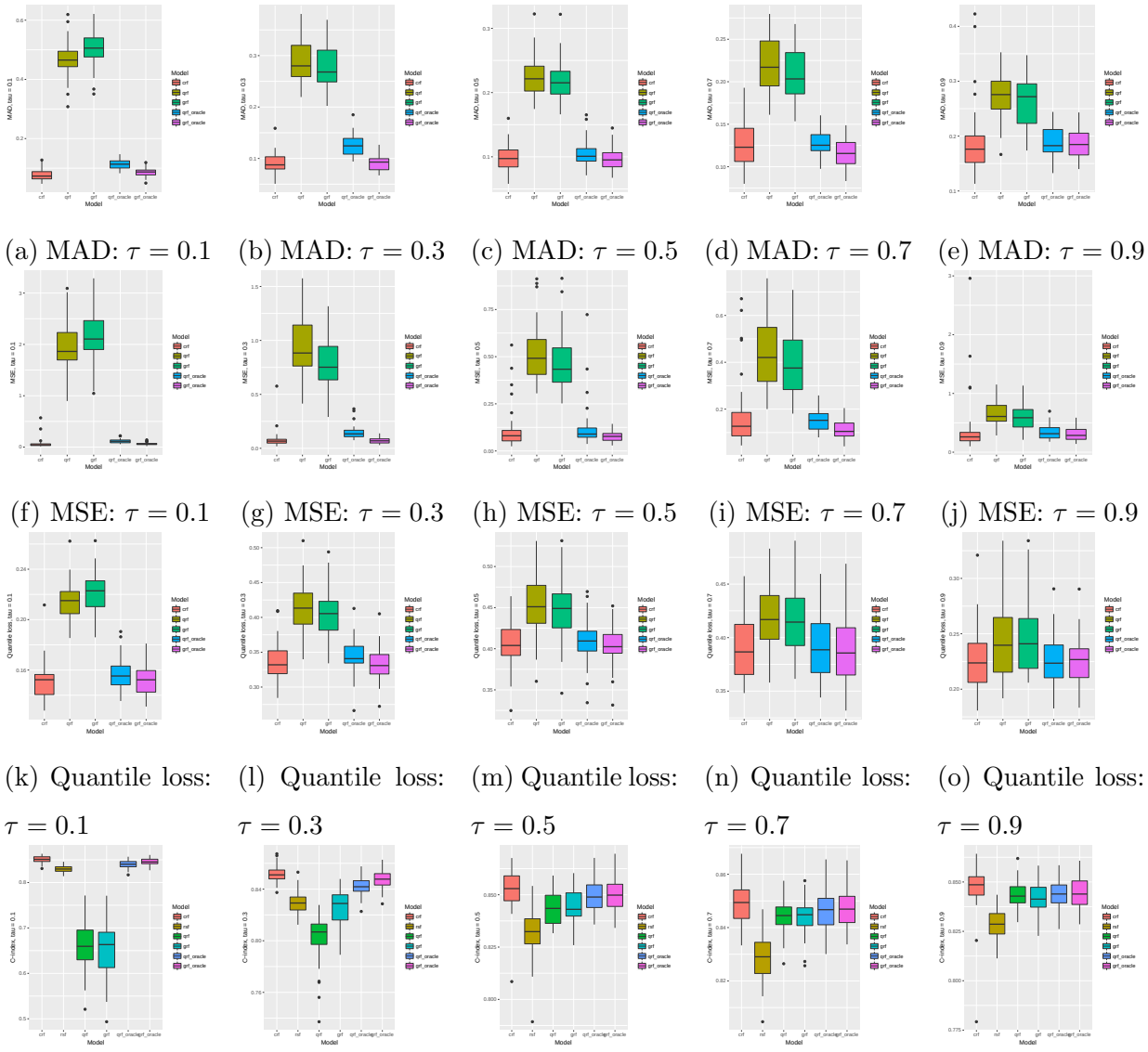
$$L_{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Q}_i^\tau - Q_i^\tau)^2, \tag{18}$$

$$L_{MAD} = \frac{1}{n}\sum_{i=1}^{n}|\hat{Q}_i^\tau - Q_i^\tau|, \tag{19}$$

$$L_{quantile} = \frac{1}{n}\sum_{i=1}^{n}\rho_\tau(T_i - \hat{Q}_i^\tau). \tag{20}$$

The reason we use $L_{quantile}$ to measure the quality of quantile predictions is that, by Meinshausen (2006), the τ-th quantile of T at x equals to $\arg\min_{q\in\mathbb{R}}\mathbb{E}[\rho_\tau(T-q)|X=x]$. The results are illustrated in Figure 3 where besides the abose three measures we compare the concordance index (C-index) (Harrell Jr et al., 1982), which is related to the area under the ROC curve (Heagerty and Zheng, 2005). It estimates the probability that, in a randomly selected pair of cases, the case that fails first had a worse predicted outcome. In Ishwaran et al. (2008), they use the ensemble mortality as the predictive outcome for their random survival forest, and the predicted survival time for random forest regression. For our method *crf* and the other two methods, *qrf* and *grf*, we will use the τ-th conditional quantile as the predicted outcome. Since the outcomes will be different for different τ, we report the results for all $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

In Figure 3 we observe an oracle like behavior of the proposed *crf* method in terms all four measures of the quality of estimation and/or prediction. Namely, we observe that MAD, MSE and quantile losses are extremely small whereas C-index is high and all are close to the corresponding oracle estimators (colored purple and blue). Moreover, we observe that the proposed *crf* method, although not primarily build for the hazard rates, is even better than survival random forest: see for example discrepancies between red and brown boxplots in the last row of Figure 3 where the larger the C-index is the better the method is.



(a) MAD: $\tau = 0.1$ (b) MAD: $\tau = 0.3$ (c) MAD: $\tau = 0.5$ (d) MAD: $\tau = 0.7$ (e) MAD: $\tau = 0.9$

(f) MSE: $\tau = 0.1$ (g) MSE: $\tau = 0.3$ (h) MSE: $\tau = 0.5$ (i) MSE: $\tau = 0.7$ (j) MSE: $\tau = 0.9$

(k) Quantile loss: $\tau = 0.1$ (l) Quantile loss: $\tau = 0.3$ (m) Quantile loss: $\tau = 0.5$ (n) Quantile loss: $\tau = 0.7$ (o) Quantile loss: $\tau = 0.9$

(p) C-index: $\tau = 0.1$ (q) C-index: $\tau = 0.3$ (r) C-index: $\tau = 0.5$ (s) C-index: $\tau = 0.7$ (t) C-index: $\tau = 0.9$

Figure 3: One-dimensional AFT model box plots with $n = 300$ and $B = 1000$. For the metrics MSE (18), MAD (19) and quantile loss (20), the smaller the value is the better. For C-index, the larger it is the better.

4.2.2 Comparison of different conditional survival estimators

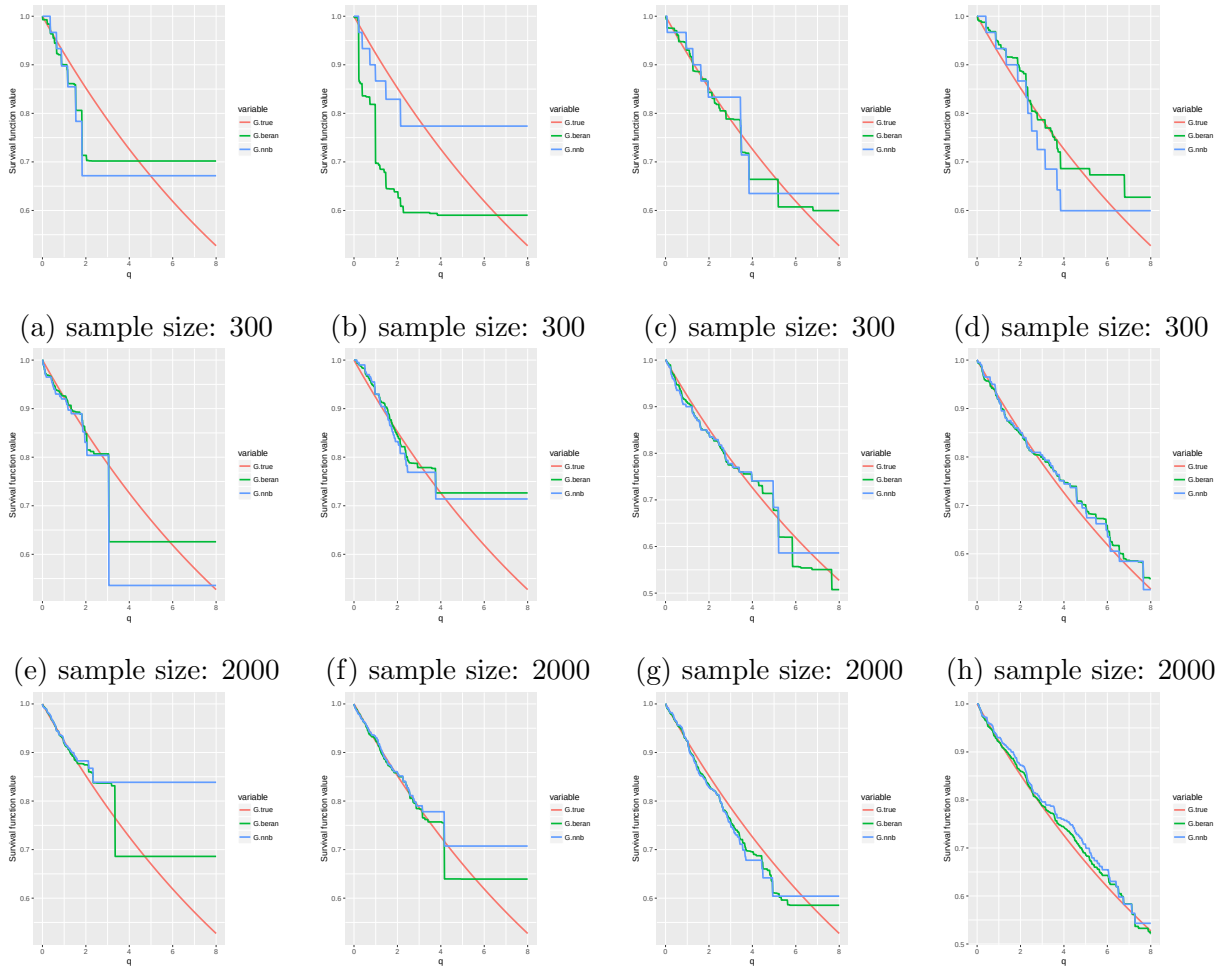
In this section, we will compare the two different conditional survival function estimators (15) and (16). We generate training data and test data from the one-dimensional AFT model defined in the previous section, but with two different censoring rate:

- $C \sim \text{Exp}(\lambda = 0.08)$, in this case, the censoring rate is about 20%.
- $C \sim \text{Exp}(\lambda = 0.20)$, in this case, the censoring rate is about 50%.

We then choose four test points $\{x_1 = 0.4, x_2 = 0.8, x_3 = 1.2, x_4 = 1.6\}$, and then plot out the conditional survival function estimators $\hat{G}(q|x_i)$ by the two different methods (15) and (16) on these four points. The results are shown in Figure 4 and 5 for three different training sample sizes $n \in \{300, 2000, 5000\}$. For the nearest neighbor estimator (15), we set the number of neighbors to be $n/10$, which is also the node size we choose.

We can observe that when n increases, two curves become closer and are both good approximations of the true survival curve. But the first method (15) does have an extra tuning parameter k – the number of nearest neighbors, so in the experiments, we always choose to use the second estimator (16), which is more adaptive and parameter free.

Note that the estimated survival function will degenerate at the tail of the distribution when the test point x is small; see the first two columns in Figure 4 and 5. 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, the conditional distribution of the latent variable depends on the location x. When x is small, the conditional mean of T is also small, and we could not observe most of the censoring values where $C_i > T_i$, leading to degenerated survival curves. However, if we continue increasing the sample size n, we should be able to recover the entire curve even for smaller x. In fact, when we increase the censoring level from 20% (Figure 4) to 50% (Figure 5), we find that both estimators give better performance because we can observe more censored values.



(a) sample size: 300 (b) sample size: 300 (c) sample size: 300 (d) sample size: 300

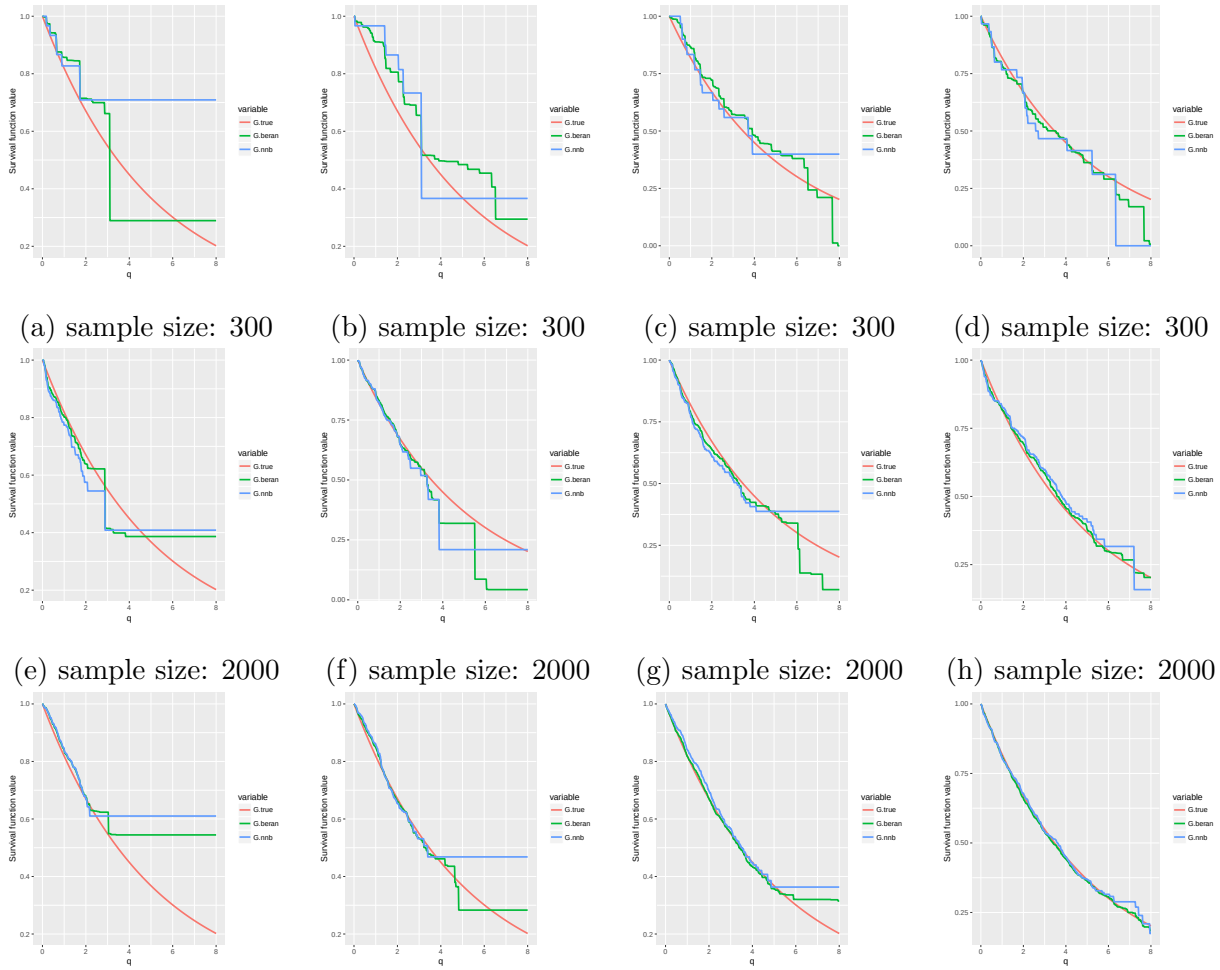
(e) sample size: 2000 (f) sample size: 2000 (g) sample size: 2000 (h) sample size: 2000

(i) sample size: 5000 (j) sample size: 5000 (k) sample size: 5000 (l) sample size: 5000

Figure 4: Comparison of different conditional survival estimators on the one-dimensional AFT model. In this case, the censoring variable $C \sim \text{Exp}(\lambda = 0.08)$, and the average censoring rate is around 20%. From left-most column to right-most column, we plot the conditional survival estimators for four test points, $x = 0.4, 0.8, 1.2, 1.6$.

4.2.3 One-dimensional censored sine model

Since our forest regression method *crf* is nonparametric and does not rely on any parametric assumption between response and explanatory variables, it can be used to estimate quantiles for any general model $T = f(X) + \epsilon$. In this section, we let $f(x) = \sin(x)$ and



(a) sample size: 300 (b) sample size: 300 (c) sample size: 300 (d) sample size: 300

(e) sample size: 2000 (f) sample size: 2000 (g) sample size: 2000 (h) sample size: 2000

(i) sample size: 5000 (j) sample size: 5000 (k) sample size: 5000 (l) sample size: 5000

Figure 5: Comparison of different conditional survival estimators on the one-dimensional AFT model. In this case, the censoring variable $C \sim \text{Exp}(\lambda = 0.20)$, and the average censoring rate is around 50%. From left-most column to right-most column, we plot the conditional survival estimators for four test points, $x = 0.4, 0.8, 1.2, 1.6$.

have the model

$$T = 2.5 + \sin(X) + \epsilon$$

where $X \sim \text{Unif}(0, 2\pi)$ and $\epsilon \sim \mathcal{N}(0, 0.3^2)$. Then the censoring variable $C \sim 1 + \sin(X) + \text{Exp}(\lambda = 0.2)$, and the responses $Y = \min(T, C)$. All the other settings are the same as in Section 4.2.1. We plot out the training data and the quantile predictions for $\tau =$

$0.3, 0.5, 0.7$ in Figure 6. The censoring level is about 25%. We observe that for all $\tau \in \{0.3, 0.5, 0.7\}$, *crf* can produce comparable quantile predictions to *grf-oracle*. Especially when $\tau = 0.3$, the quantile prediction by grf (blue dotted curve) severely deviates from the true quantile, while our method *crf* can still predict the correct quantile and performs as good as *grf-oracle*. We want to emphasize that *grf-oracle* uses the latent responses T_i while our method only uses the observed responses Y_i and censoring indicators δ_i. We then repeat the experiments for 40 times and report the box plots in Figure 7. Again we can see that for all quantiles, our method *crf* behaves almost as good as *qrf-oracle* and *grf-oracle*, and consistently better than *qrf* and *grf*. For example the order of magnitude of our error is twice and sometimes more than two times smaller than that of quantile or generalized random forest.

4.2.4 Multi-dimensional AFT model results

In this section, we test our algorithm on a multi-dimensional AFT model

$$\log(T) = X^\top \beta + \epsilon,$$

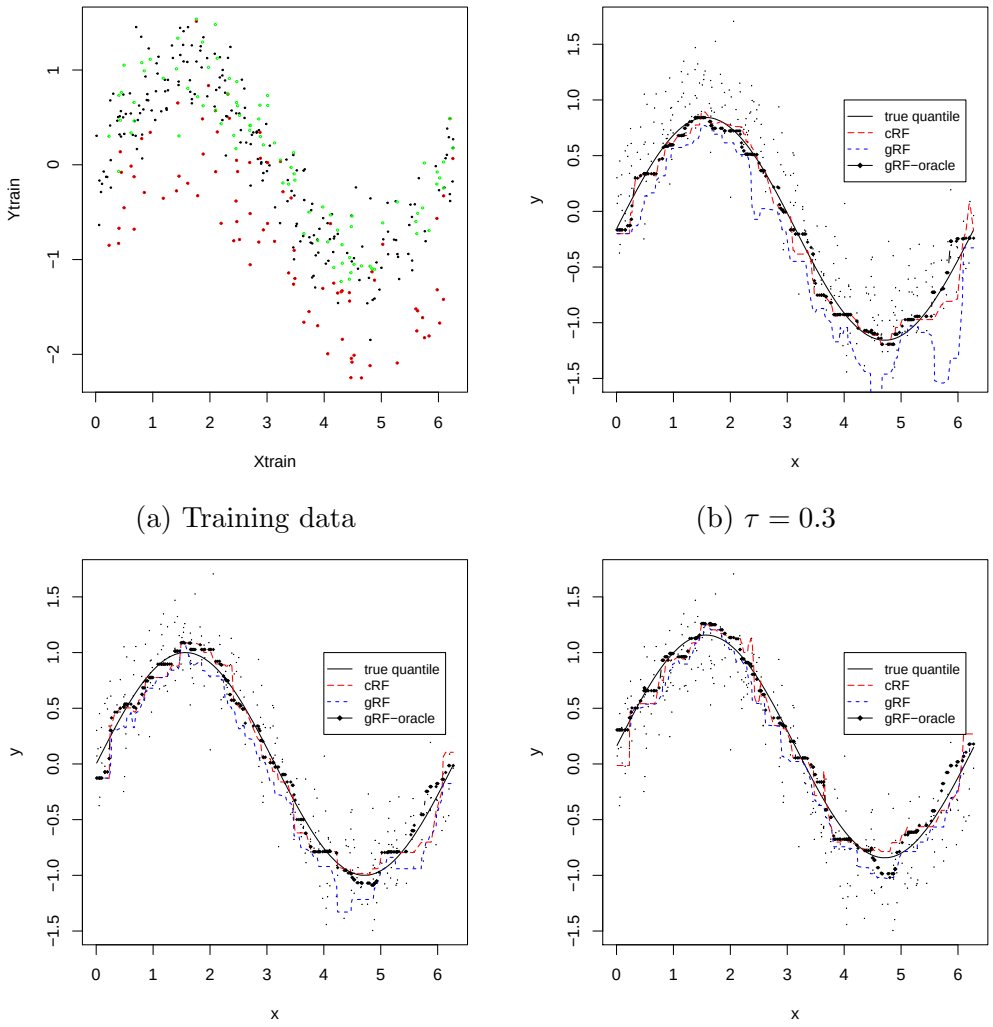
where $\boldsymbol{\beta} = (0.1, 0.2, 0.3, 0.4, 0.5)$, $X_{\cdot,j} \sim \text{Unif}(0, 2)$, and $\epsilon \sim \mathcal{N}(0, 0.3^2)$. The censoring variable $C \sim \text{Exp}(\lambda = 0.05)$, and $Y = \min(T, C)$. The censoring level is about 22%. The number of training data is 500 and the number of test points is 300. All the forests consist of 1000 trees. The result is in Figure 8. Our model *crf* still outperforms *qrf* and *grf* significantly, and is comparable to *qrf-oracle* and *grf-oracle*.

4.2.5 Multi-dimensional censored complex manifold

In this section, we construct a complex model

$$T = 5 + \frac{1}{5}\left(\sin(X_{\cdot,1}) + \cos(X_{\cdot,2}) + X_{\cdot,3}^2 + \exp(X_{\cdot,4}) + X_{\cdot,5}\right) + \epsilon,$$

where $X_{\cdot,j}$ stands for j-th dimension of $\boldsymbol{X} \in \mathbb{R}^5$, and $\epsilon \sim \mathcal{N}(0, 0.3^2)$. Then we consider a censoring variable independent of $\boldsymbol{X}$ and T: $C \sim \text{Exp}(\lambda = 0.015)$. The result is summarized in Figure 9. Although the model is highly non-linear our method is able to capture that



(a) Training data

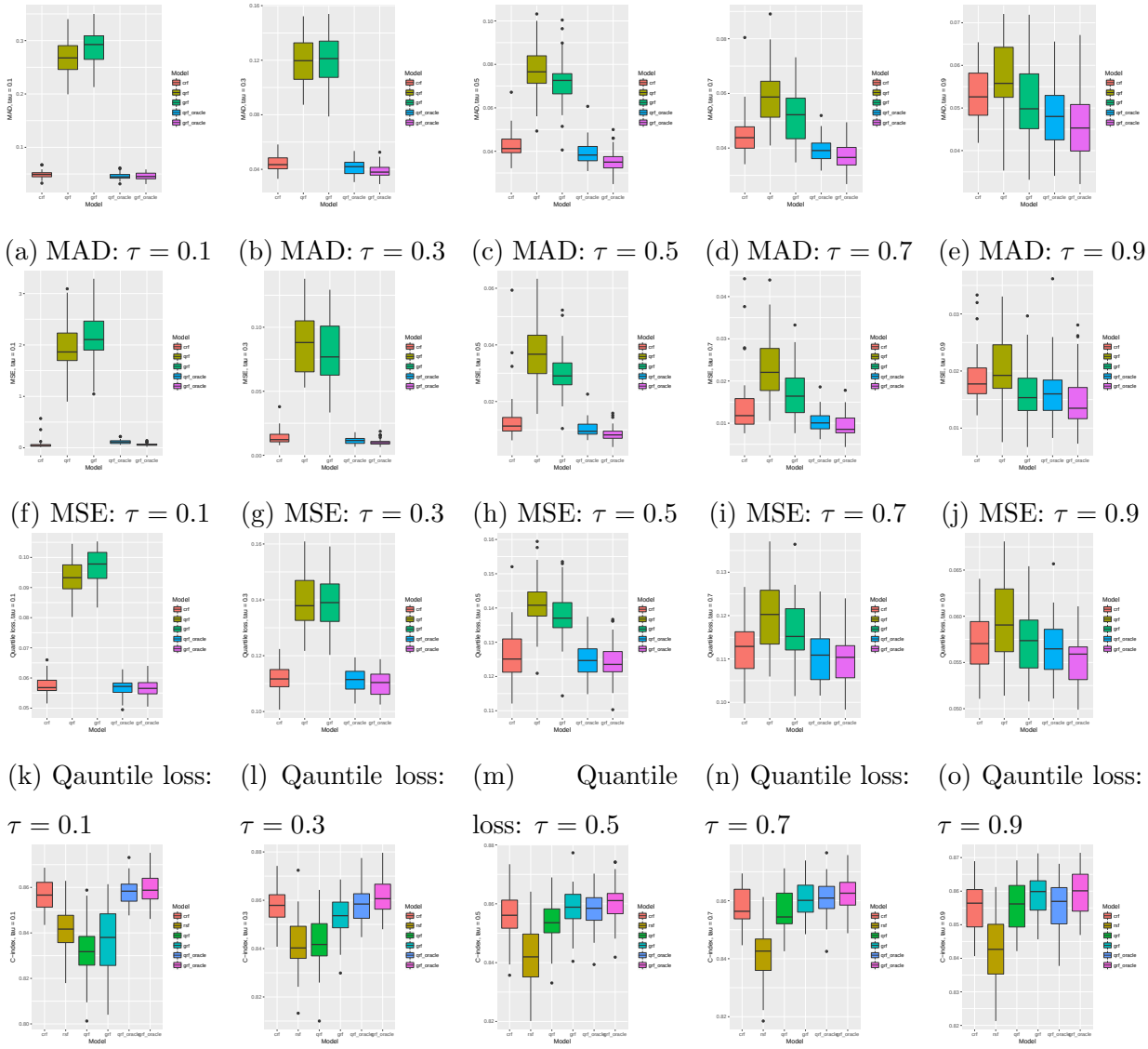
(b) $\tau = 0.3$

(c) $\tau = 0.5$

(d) $\tau = 0.7$

Figure 6: One-dimensional Sine model. In (a), black points stand for observation that are not censored; red points are observations that are censored, and the green points are the counterpart of the red points, that is, they are the latent values of those red points if they were not censored.

and estimate the quantile of interest consistently. However, other methods are unable to be consistent. Behavior matches that of linear models closely. In this case we notice that the oracular generalized random forests have much better performance; note that since we are estimating equations, our algorithm can be easily tweaked to match *grf-oracle* –



(a) MAD: $\tau = 0.1$ (b) MAD: $\tau = 0.3$ (c) MAD: $\tau = 0.5$ (d) MAD: $\tau = 0.7$ (e) MAD: $\tau = 0.9$

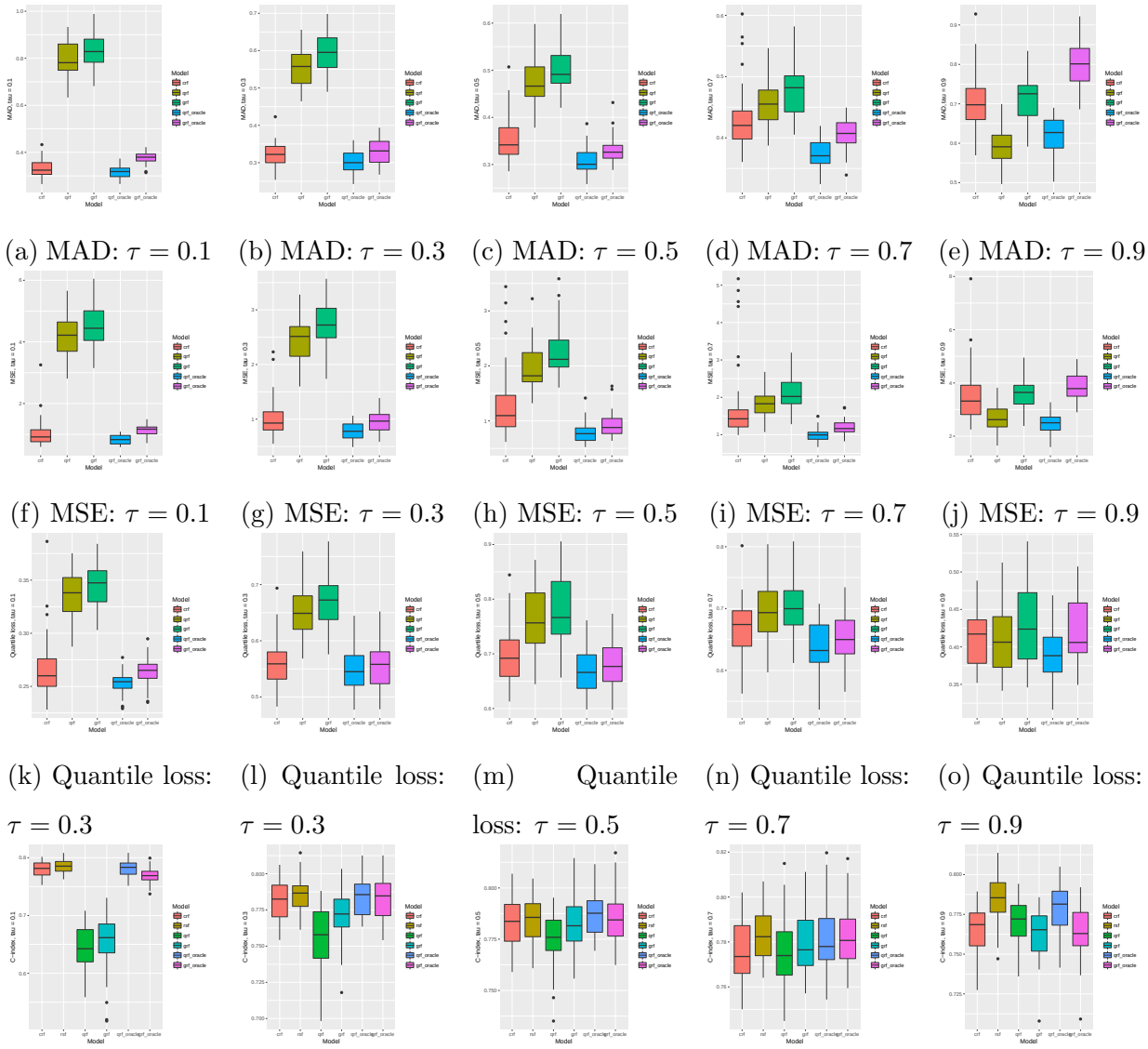
(f) MSE: $\tau = 0.1$ (g) MSE: $\tau = 0.3$ (h) MSE: $\tau = 0.5$ (i) MSE: $\tau = 0.7$ (j) MSE: $\tau = 0.9$

(k) Qauntile loss: $\tau = 0.1$ (l) Qauntile loss: $\tau = 0.3$ (m) Quantile loss: $\tau = 0.5$ (n) Quantile loss: $\tau = 0.7$ (o) Qauntile loss: $\tau = 0.9$

(p) C-index: $\tau = 0.1$ (q) C-index: $\tau = 0.3$ (r) C-index: $\tau = 0.5$ (s) C-index: $\tau = 0.7$ (t) C-index: $\tau = 0.9$

Figure 7: One-dimensional censored sine model box plots with $n = 300$ and $B = 1000$. For the metrics MSE (18), MAD (19) and quantile loss (20), the smaller the value is the better. For C-index, the larger it is the better.

the only change would be to design the splits in the initial tree construction to match our estimating equations, use subsampling and save a separate sample for minimizing the equations themselves.



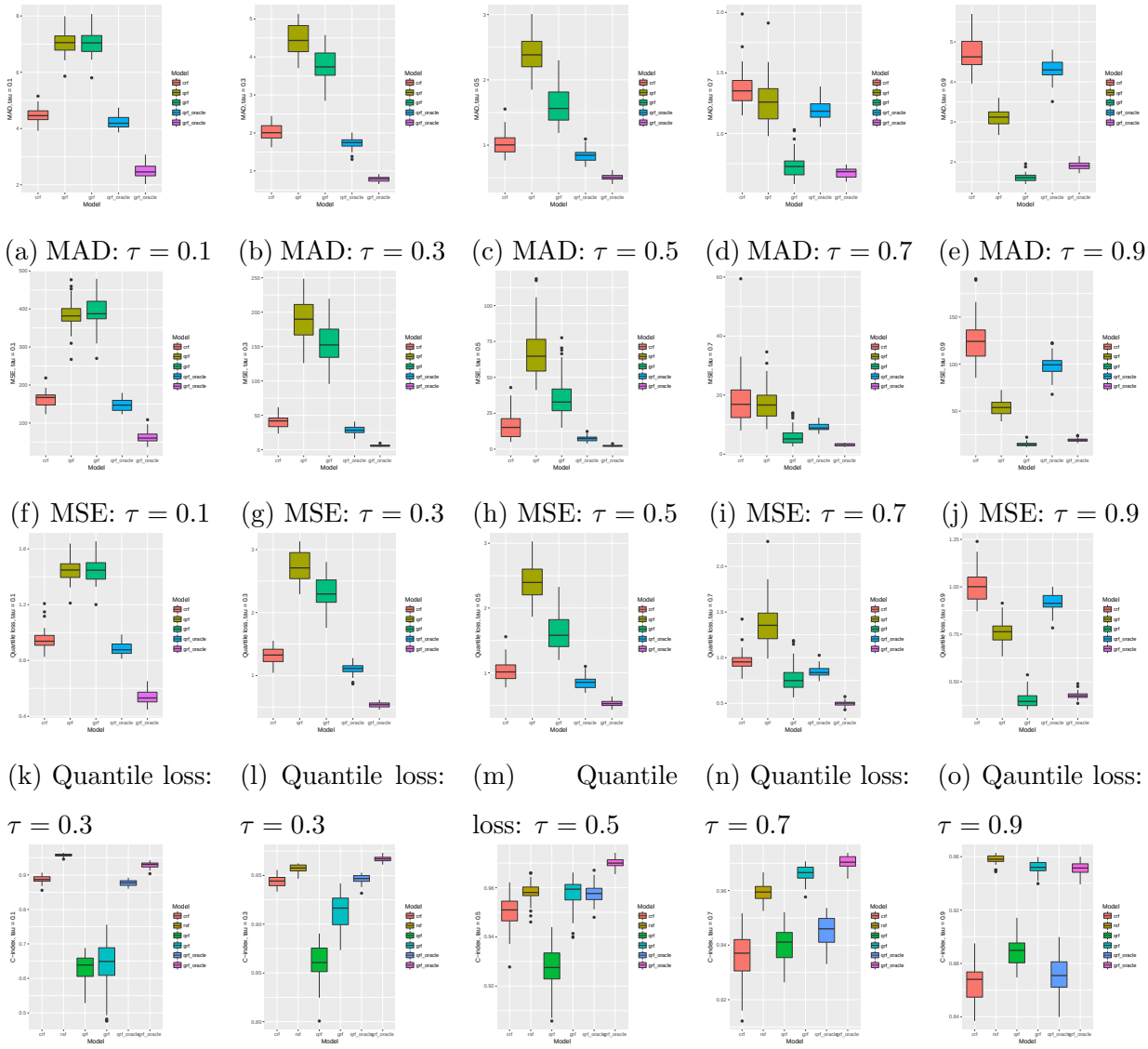
(a) MAD: $\tau = 0.1$ (b) MAD: $\tau = 0.3$ (c) MAD: $\tau = 0.5$ (d) MAD: $\tau = 0.7$ (e) MAD: $\tau = 0.9$

(f) MSE: $\tau = 0.1$ (g) MSE: $\tau = 0.3$ (h) MSE: $\tau = 0.5$ (i) MSE: $\tau = 0.7$ (j) MSE: $\tau = 0.9$

(k) Quantile loss: $\tau = 0.3$ (l) Quantile loss: $\tau = 0.3$ (m) Quantile loss: $\tau = 0.5$ (n) Quantile loss: $\tau = 0.7$ (o) Qauntile loss: $\tau = 0.9$

(p) C-index: $\tau = 0.3$ (q) C-index: $\tau = 0.3$ (r) C-index: $\tau = 0.5$ (s) C-index: $\tau = 0.7$ (t) C-index: $\tau = 0.9$

Figure 8: Multi-dimensional AFT model box plots with dimension five and $n = 800$ and $B = 1000$: 500 training and 300 testing size. For the metrics MSE (18), MAD (19) and quantile loss (20), the smaller the value is the better. For C-index, the larger it is the better.



(a) MAD: $\tau = 0.1$ (b) MAD: $\tau = 0.3$ (c) MAD: $\tau = 0.5$ (d) MAD: $\tau = 0.7$ (e) MAD: $\tau = 0.9$

(f) MSE: $\tau = 0.1$ (g) MSE: $\tau = 0.3$ (h) MSE: $\tau = 0.5$ (i) MSE: $\tau = 0.7$ (j) MSE: $\tau = 0.9$

(k) Quantile loss: $\tau = 0.3$ (l) Quantile loss: $\tau = 0.3$ (m) Quantile loss: $\tau = 0.5$ (n) Quantile loss: $\tau = 0.7$ (o) Qauntile loss: $\tau = 0.9$

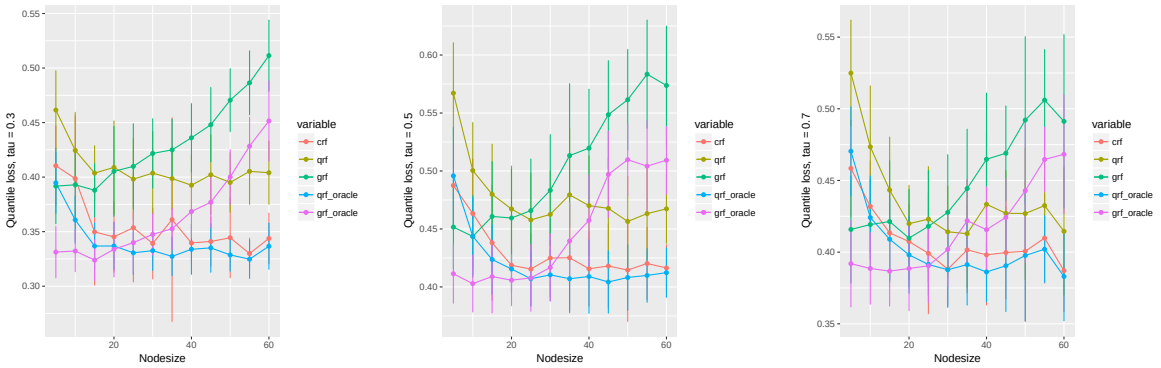
(p) C-index: $\tau = 0.3$ (q) C-index: $\tau = 0.3$ (r) C-index: $\tau = 0.5$ (s) C-index: $\tau = 0.7$ (t) C-index: $\tau = 0.9$

Figure 9: Censored regression model with five-dimensional complex manifold structure: box plots with $n = 300$ and $B = 1000$. For the metrics MSE (18), MAD (19) and quantile loss (20), the smaller the value is the better. For C-index, the larger it is the better.

4.2.6 Node size

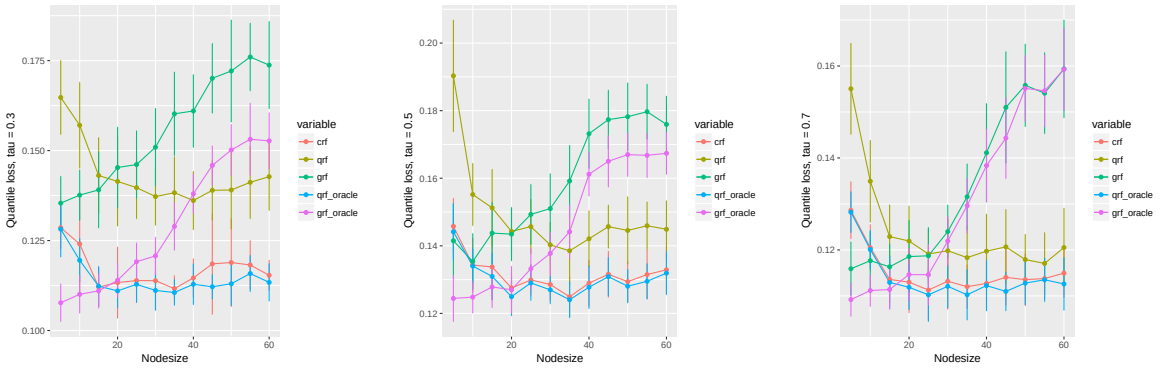
In this section, we investigate the impact of the choice of the node size on different methods. The data we use will be generated from the one-dimensional and multi-dimensional AFT

and Sine models as defined in the previous sections. We increase the node size from 5 to 60 with step size of 5.



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

Figure 10: Quantile losses of 1D AFT model with different node sizes.

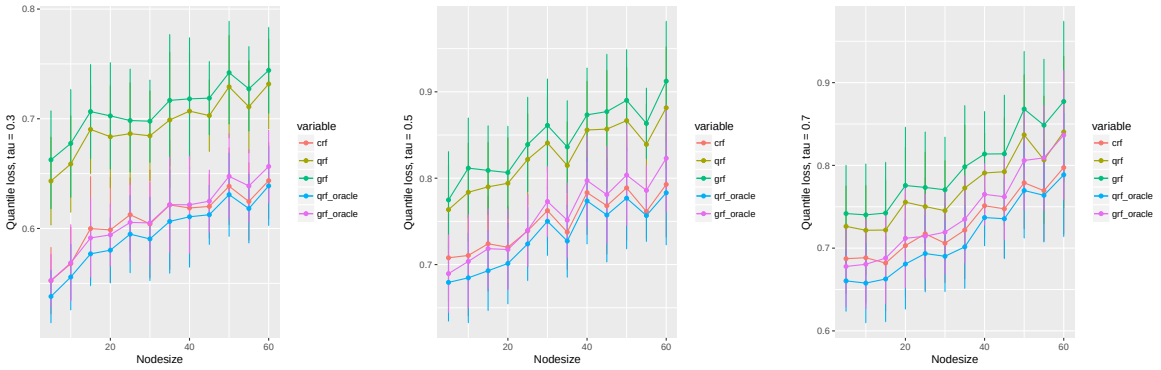


(a) $\tau = 0.3$ (b) $\tau = 0.5$ (c) $\tau = 0.7$

Figure 11: Quantile losses of Sine model with different node sizes.

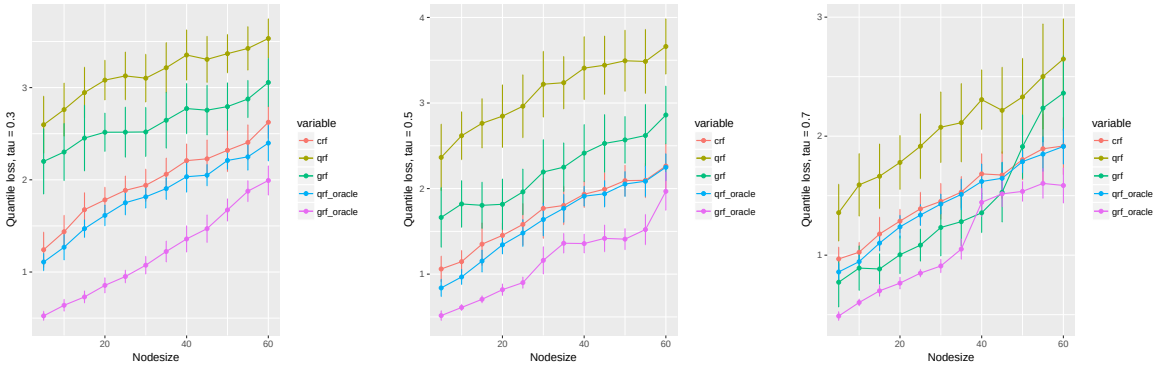
One-dimensional AFT and Sine models The result of sine model is summarized in Figure 11. One can see that for both *qrf* and our model, *crf*, the quantile loss will first decrease when node size increases. It attains minimum around node size of 30. However, for *grf*, its quantile loss is almost monotonically increasing, and attains minimum at node size of 5. But both *qrf-oracle* and *grf-oracle* can attain the best quantile loss of about 0.125. And one impressive observation is that our model, *crf*, almost performs the same as

qrf-oracle for all node sizes. Similar conclusion can be made from the AFT result which is in Figure 10.



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

Figure 12: Quantile losses of multi-dimensional AFT model with different node sizes.



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

Figure 13: Quantile losses of multi-dimensional complex model with different node sizes.

Multi-dimensional AFT model From Figure 12, we observe that *qrf*, *qrf-oracle* and *grf-oracle* all give similar results. The performance of our model *crf* is only slightly worse than *qrf-oracle*, but is even better than *grf-oracle*.

Multi-dimensional complex model The result is summarized in Figure 13. The censoring level is about 25%. From the figure, we observe that the behavior of *crf* is still only

slightly worse than *qrf-oracle*. In this experiment, *grf-oracle* behaves the best. All of *crf*, *qrf-oracle* and *grf-oracle* are significantly better than the biased models *qrf* and *grf*. When $\tau = 0.7$, *grf* behaves slightly better than *qrf-oracle* when node size is small. The reason is that the conditional quantiles of Y and T are closer when τ is larger, and *grf* is more stable and smooth on the data in this experiment. But we still observe that the performance of *crf* and *qrf-oracle* are very close.

4.3 Real Data

In this section, we compare our censored forest (crf) with quantile random forest (qrf) (Meinshausen, 2006) and generalized forest (grf) (Athey et al., 2018) on real datasets. In order to evaluate the performances unbiasedly, we manually add censoring to the data. In addition, we apply qrf and grf to the data without censoring and we call the resulted models qrf-oracle and grf-oracle, respectively.

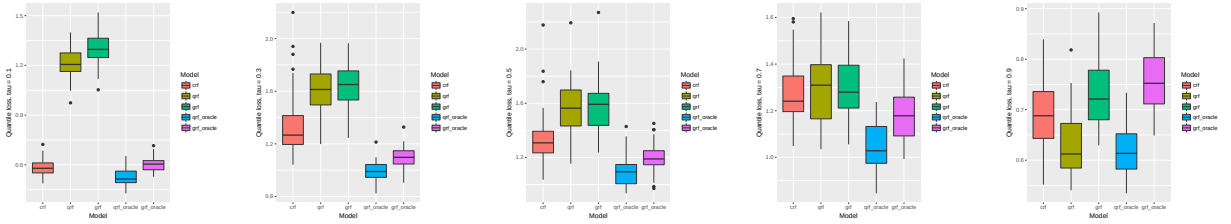
For all these methods, bagged versions of the training data are used for each of the 1000 trees. We use 5-fold cross validation to select the best node sizes for different methods. For all the other parameters, we keep the default settings.

Datasets We use datasets *BostonHousing*, *Ozone* from the R packages *mlbench* and *alr3*. For all the datasets, we sample censoring variables from Exponential distributions with λ set so that the censoring level is roughly 20%. For *BostonHousing* dataset, we set $\lambda = 0.01$. For *Ozone*, $\lambda = 0.025$. For *Abalone* dataset, we random sample 1000 observations and take the log-transformation of the response variable *rings*. We then set $\lambda = 0.10$.

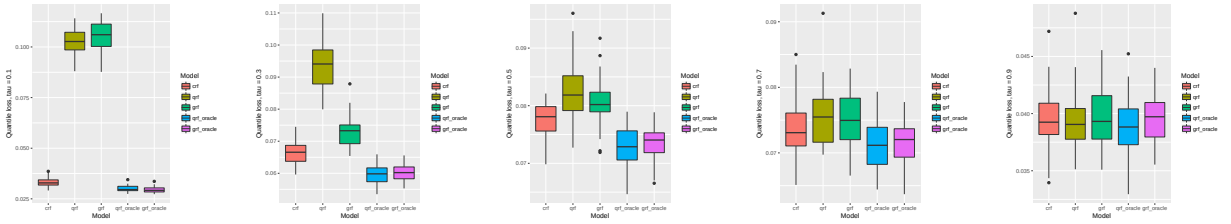
Evaluation For each dataset, we train our model on bootstrapped version of the data, and test the performance on out-of-bag observations. This process is repeated for 40 times, and we calculate the mean and standard deviation of the prediction errors. In our context, the prediction error is measured by the τ-th quantile loss for τ-th quantile estimation. The results are illustrated in Figure 14.

On all data sets, our proposed method behaves better than quantile forest and generalized forest in terms of quantile losses. Especially when $\tau = 0.1$, 0.3 or 0.5, the performance

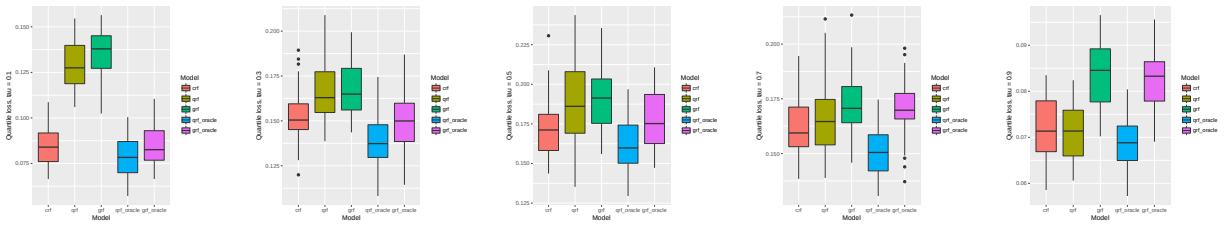
of our method is significantly better than *qrf* and *grf*, and is even comparable to that of oracle *qrf* and *grf*. It agrees with our observation in the one-dimensional example (Figure 2 and 3). While estimating larger quantiles, the true τ-th quantile of T_i and Y_i are close, and hence the performance of all five models are similar. But when τ is small, the τ-th quantile of T_i and Y_i are different because of the censoring, and in this case, our model has superior advantage and find the true quantiles of T_i almost as good as the oracle methods.



(a) BostonHousing: $\tau = 0.1$ (b) BostonHousing: $\tau = 0.3$ (c) BostonHousing: $\tau = 0.5$ (d) BostonHousing: $\tau = 0.7$ (e) BostonHousing: $\tau = 0.9$

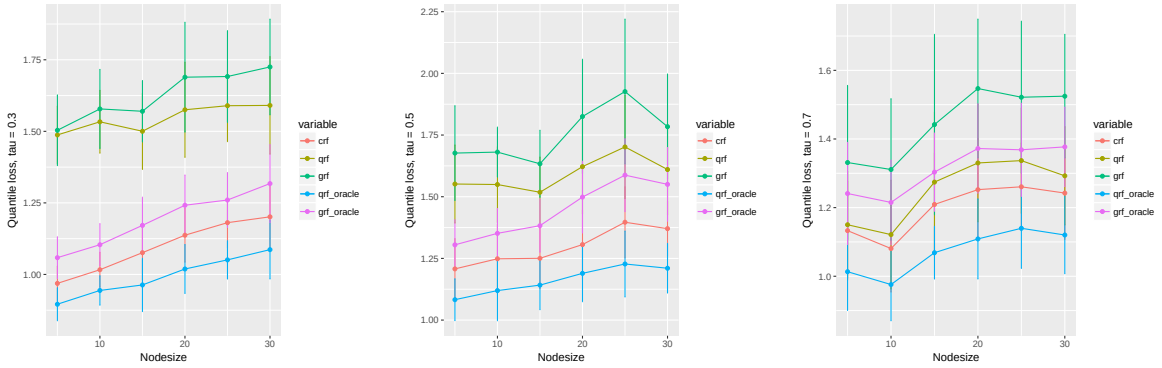


(f) Abalone: $\tau = 0.1$ (g) Abalone: $\tau = 0.3$ (h) Abalone: $\tau = 0.5$ (i) Abalone: $\tau = 0.7$ (j) Abalone: $\tau = 0.9$

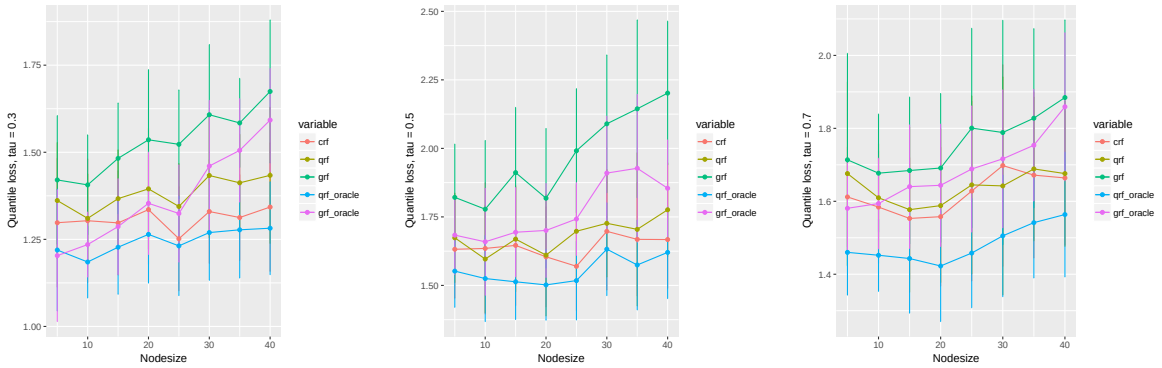


(k) Ozone: $\tau = 0.1$ (l) Ozone: $\tau = 0.3$ (m) Ozone: $\tau = 0.5$ (n) Ozone: $\tau = 0.7$ (o) Ozone: $\tau = 0.9$

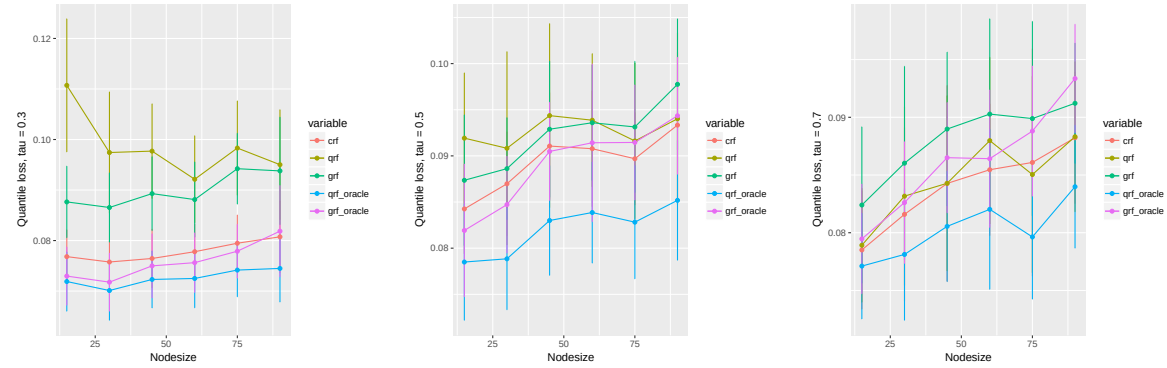
Figure 14: Quantile losses evaluated on a hold-out set of the three real data sets: each one presented in a corresponding row. Censoring level is approximately 20%. We compare our method with *grf* (green) and *qrf* (olive) and their corresponding oracle versions, *grf-oracle* (purple) and *qrf-oracle* (blue).



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



(d) Ozone: $\tau = 0.3$ (e) Ozone: $\tau = 0.5$ (f) Ozone: $\tau = 0.7$



(g) Abalone: $\tau = 0.3$ (h) Abalone: $\tau = 0.5$ (i) Abalone: $\tau = 0.7$

Figure 15: Hold-out Quantile losses with different node sizes on the three real data sets: each one represented in one row.

Nodesize For each dataset, we train different models using different nodesizes and compare the performance. For each node size, we bootstrap the data and repeat the experiments for 20 times, and we calculate the mean and standard deviation of the quantile predictions

for quantiles $\tau = 0.3$, 0.5, and 0.7. The result is in Figure 15. We observe that our method, *crf*, is uniformly better than *qrf* and *grf*, proving that *crf* is able to correct the bias introduced by censoring. Moreover, the quantile loss of *crf* is always competitive to that of *qrf-oracle* and *grf-oracle*, and is remarkably always better than *grf-oracle*, only slightly worse than *qrf-oracle*.

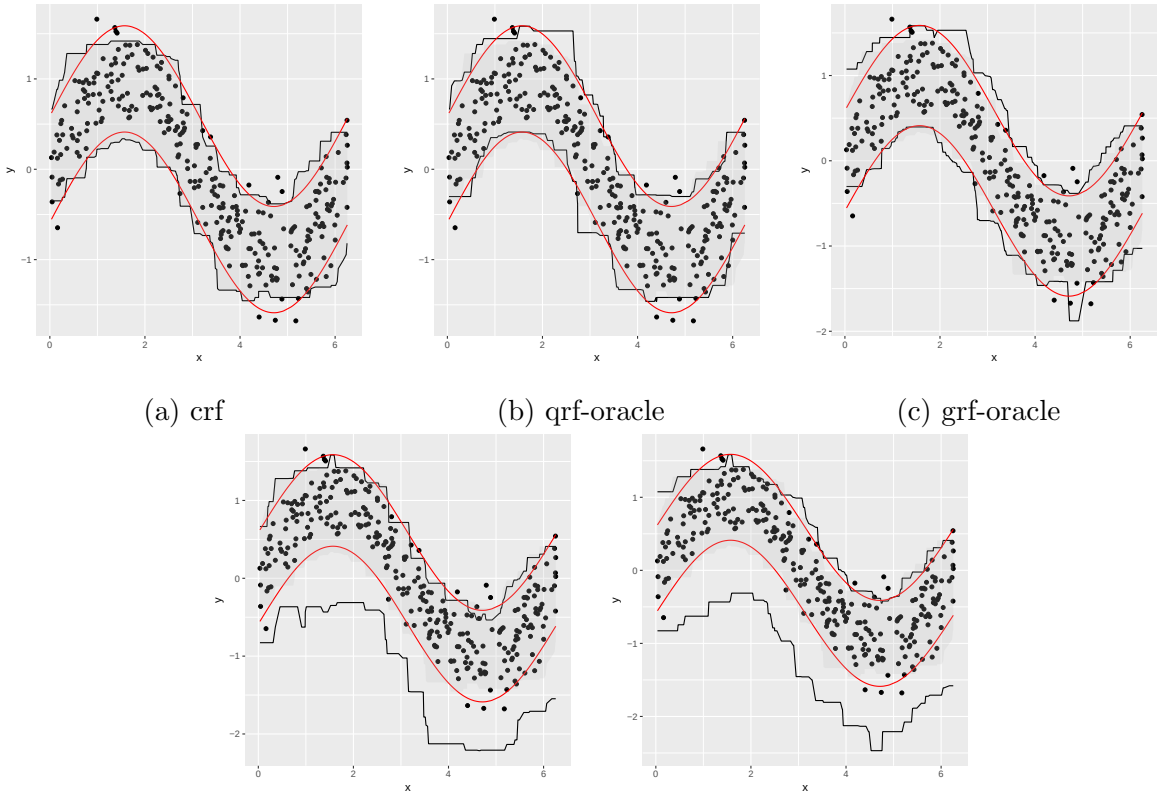
4.4 Prediction Intervals

All the forest methods can be used to get 95% prediction intervals by predicting the 0.025 and 0.975 quantiles of the true response variable. Then for any location $x \in \mathcal{X}$, a straightforward confidence interval will be $[Q(x; 0.025), Q(x; 0.975)]$. The result is illustrated in Figure 17 for the case of univariate censored sine model. For each data set, we bootstrap the data and calculate the 0.025 and 0.975 quantile for the out of bag points. Then for each node size, we repeat this process for 20 times and calculate the average coverage rate of the confidence intervals.

We observe that in all of the cases, our method *crf* and *qrf-oracle* give the coverage closest to 95%. As can be seen from Figure 15, both *qrf* and *grf* perform much worse on predicting lower quantiles. They tend to under-estimate the lower quantiles and hence make the confidence intervals much wider than the true ones. Nevertheless, as seen in Figure 17, out method has great coverage that is never below 95%, indicating certain efficiency of the proposed method.

5 Discussion

In this article, we introduced a censored quantile random 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 equations maintain the flexibility of general random forest approaches. The estimating equations are equipped with both censoring information as well as random forest information simultaneously.



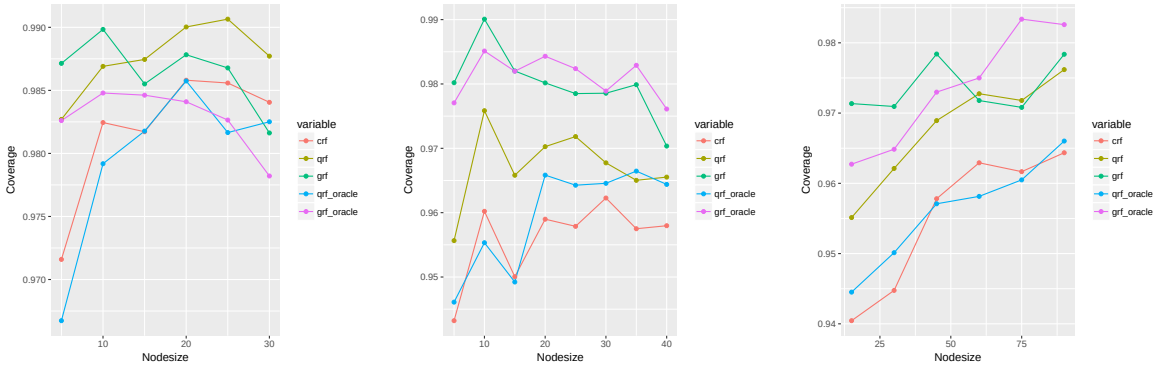
(a) crf (b) qrf-oracle (c) grf-oracle

(d) qrf (e) grf

Figure 16: Prediction intervals of the univariate censored since model.

The statistical mechanism, particularly asymptotic normality, of ensemble tree-based methods is still not fully understood. Some insightful discussions and first steps towards this goal can be found in Zhu and Kosorok (2012). As noted in Athey et al. (2018) even consistency properties of censored based methods present significant theoretical challenges. Generalization of the results of Athey et al. (2018) to the case of censored observations involves non-trivial extensions of the theoretical advancements introduced therein. A generalization would require not only adaptation to censoring but as well as to an infinite-dimensional nuisance parameter; current results focus on finite dimensional nuisance parameters. Once the latter generalization is achieved then, utilizing our estimating equations, will allow for the generalization of the former.

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



(a) BostonHousing (b) Ozone (c) Abalone

Figure 17: Average coverage of prediction intervals with different node sizes on the three real data sets: each one represented in one plot. The length of the confidence intervals is much larger for all of the other considered methods.

with right-censored observations is important and yet poorly understood research area. Equipping this literature with a fully non-parametric approach would lead to a significant broadening of the now more known parametric approaches. Our method appears to be a nice fit for this setting; observe that our estimating equations can be easily replaced with another kind that targets treatment effects directly.

References

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

Athey, S., J. Tibshirani, and S. Wager (2018). Generalized random forests. *Forthcoming in the Annals of Statistics*.

Backer, M. D., A. E. Ghouch, and I. V. Keilegom (2018). An adapted loss function for censored quantile regression. *Journal of the American Statistical Association 0*(ja), 1–40.

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

Bloniarz, A., A. S. Talwalkar, B. Yu, and C. Wu (2016). Supervised neighborhoods for distributed nonparametric regression. In *AISTATS*.

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

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

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

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

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

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

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

Heagerty, P. J. and Y. Zheng (2005). Survival model predictive accuracy and roc curves. *Biometrics 61*(1), 92–105.

Hoover, D. R., J. A. Rice, C. O. Wu, and L.-P. Yang (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika 85*(4), 809–822.

Hothorn, T., P. Bühlmann, S. Dudoit, A. Molinaro, and M. J. Van Der Laan (2005). Survival ensembles. *Biostatistics 7*(3), 355–373.

Hothorn, T., B. Lausen, A. Benner, and M. Radespiel-Tröger (2004). Bagging survival trees. *Statistics in medicine 23*(1), 77–91.

Huang, J., S. Ma, and H. Xie (2007). Least absolute deviations estimation for the accelerated failure time model. *Statistica Sinica*, 1533–1548.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer (2008). Random survival forests. *The annals of applied statistics*, 841–860.

Jin, Z., D. Lin, L. Wei, and Z. Ying (2003). Rank-based inference for the accelerated failure time model. *Biometrika 90*(2), 341–353.

Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association 53*(282), 457–481.

Koenker, R. et al. (2008). Censored quantile regression redux. *Journal of Statistical Software 27*(6), 1–25.

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

LeBlanc, M. and J. Crowley (1992). Relative risk trees for censored survival data. *Biometrics*, 411–425.

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

Li, A. H. and A. Martin (2017). Forest-type regression with general losses and robust forest. In *International Conference on Machine Learning*, pp. 2091–2100.

Li, G. and H. Doss (1995). An approach to nonparametric regression for life history data using local linear fitting. *The Annals of Statistics*, 787–823.

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

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

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

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

Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives 29*(3), 61–80.

Peng, L. and Y. Huang (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association 103*(482), 637–649.

Portnoy, S. (2003). Censored regression quantiles. *Journal of the American Statistical Association 98*(464), 1001–1012.

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

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

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

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

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

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

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

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

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

6 Appendix

Proof of Theorem 1. To get the candidate set $\mathcal{C}$, if we use the k-nearest neighbor estimator (15), then the first step is to sort n weights and choose the largest k elements. This is in general a $O(n\log(n))$ procedure. If we use the Beran estimator (16), then the time complexity is $O(n)$ because we need to find all the nonzero weights.

After we have the candidate set $\mathcal{C}$, evaluating $S_n(q;\tau)$ for all $q \in \mathcal{C}$ and finding the minimum is a $O(n|\mathcal{C}|)$ procedure. For (15), $|\mathcal{C}| = k$; and for (16), $|\mathcal{C}|$ is in the order of $m\log(n)^{p-1}$ by Lin and Jeon (2006). □

Proof of Theorem 2. Conditional on $X_1, \cdots, X_n$, the random variable $U_i = F(Y_i|X_i), i = 1, \cdots, n$ are i.i.d. uniform on $[0,1]$. By Condition 4, for a given X_i,

$$\mathbb{1}(Y_i \le q) = \mathbb{1}(U_i \le F(q|X_i)).$$

Then we can decompose

$$\begin{aligned}
& \sum_{i=1}^{n} w(X_i,x)\,\mathbb{1}(Y_i \le q) \\
= & \sum_{i=1}^{n} w(X_i,x)\,\mathbb{1}(U_i \le F(q|X_i)) \\
= & \sum_{i=1}^{n} w(X_i,x)\,\mathbb{1}(U_i \le F(q|x)) + \sum_{i=1}^{n} w(X_i,x)\Big\{\mathbb{1}(U_i \le F(q|X_i)) - \mathbb{1}(U_i \le F(q|x))\Big\}.
\end{aligned}$$

The difference between the empirical distribution function and the truth can then be bounded by

$$\begin{aligned}
& \left|\sum_{i=1}^{n} w(X_i,x)\,\mathbb{1}(Y_i \le q) - F(q|x)\right| \\
\le & \underbrace{\left|\sum_{i=1}^{n} w(X_i,x)\,\mathbb{1}(U_i \le F(q|x)) - F(q|x)\right|}_{\text{(I)}} \\
+ & \underbrace{\left|\sum_{i=1}^{n} w(X_i,x)\Big\{\mathbb{1}(U_i \le F(q|X_i)) - \mathbb{1}(U_i \le F(q|x))\Big\}\right|}_{\text{(II)}}.
\end{aligned}$$

For part (I), since U_i is uniform, we have

$$\sup_{q\in\mathbb{R}}\left|\sum_{i=1}^{n} w(X_i,x)\,\mathbb{1}(U_i\le F(q|x))-F(q|x)\right| = \sup_{z\in[0,1]}\left|\sum_{i=1}^{n} w(X_i,x)\,\mathbb{1}(U_i\le z)-z\right|$$

Now since $0\le w(X_i,x)\le 1/m$ and $\sum_{i=1}^{n} w(X_i,x)=1$, we have

$$\sum_{i=1}^{n} w(X_i,x)^2 \le \max_{i=1,\cdots,n} w(X_i,x) \le \frac{1}{m}\to 0$$

as $n\to\infty$, by Condition 2. Hence, by Chebyshev inequality, for every $z\in[0,1]$ and $x\in\mathcal{X}$,

$$\left|\sum_{i=1}^{n} w(X_i,x)\,\mathbb{1}(U_i\le z)-z\right| = o_p(1).$$

Then by Bonferroni's inequality,

$$\sup_{z\in[0,1]}\left|\sum_{i=1}^{n} w(X_i,x)\,\mathbb{1}(U_i\le z)-z\right| = o_p(1).$$

The proof of part (II)

$$\left|\sum_{i=1}^{n} w(X_i,x)\Big\{\mathbb{1}(U_i\le F(q|X_i))-\mathbb{1}(U_i\le F(q|x))\Big\}\right| = o_p(1)$$

follows the same argument of Theorem 1 and Lemma 2 in Meinshausen (2006) by invoking Condition 2. Finally, we notice that by Condition 5, $\sup_{q\in[-r,r]}|\hat{G}(q|x)-G(q|x)|=o(1)$ because $[-r,r]$ is compact. □

Proof of Theorem 3. By Van der Vaart (2000), we only need to show for any $\tau\in(0,1)$, $x\in\mathcal{X}$,

1. $\sup_{q\in[-r,r]}|S_n(q;\tau)-S(q;\tau)|=o_p(1)$.
2. For any $\epsilon>0$, $\inf\{|S(q;\tau)|:|q-q^*|\ge\epsilon, q\in[-r,r]\}>0$.
3. $S_n(q_n;\tau)=o_p(1)$.

Part 1 has been proved by Theorem 2. For part 2, note that

$$\begin{aligned} S(q;\tau) &= (1-\tau)G(q|x) - \mathbb{P}(Y > q|x) \\ &= (1-\tau)G(q|x) - \mathbb{P}(T > q|x)\mathbb{P}(C > q|x) \\ &= ((1-\tau) - \mathbb{P}(T > q|x))G(q|x) \\ &= (\mathbb{P}(T \leq q|x) - \tau)G(q|x). \end{aligned}$$

The second equality is because of the conditionally independency between T and C. Fix an $\epsilon > 0$, and denote

$$E = \{|S(q;\tau)| : |q - q^*| \geq \epsilon, q \in [-r, r]\}.$$

Since $0 < \tau < 1$, by Condition 4, there exists some $l > 0$ such that $G(q|x) \geq l$ and

$$|\mathbb{P}(T \leq q|x) - \tau| \geq l$$

for $q \in E$. Now for part 3, by the definition of q_n, we know

$$|S_n(q_n;\tau)| = \min_{q\in[-r,r]} |S_n(q;\tau)|.$$

Also by definition of q^*,

$$0 = |S(q^*;\tau)| = \min_{q\in[-r,r]} |S(q;\tau)|.$$

Then we get

$$\begin{aligned} |S_n(q_n;\tau)| &= |S_n(q_n;\tau)| - |S_n(q^*;\tau)| + |S_n(q^*;\tau)| - |S(q^*;\tau)| \\ &\leq |S_n(q^*;\tau) - S(q^*;\tau)| \\ &\leq \sup_{q\in[-r,r]} |S_n(q;\tau) - S(q;\tau)| \\ &= o_p(1) \end{aligned}$$

where the first inequality is because of the definition of q_n and the triangular inequality. □