A Theorems and Proofs

A.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 equation 9 and equation 10) 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 $X_i \in R_x$ equation 9 or with $w(X_i,x) > 0$ equation 10 will be jump points, and hence, we can refine $\mathcal{C} = \{Y_i : X_i \in R_x\}$ for equation 9 or $\mathcal{C} = \{Y_i : w(X_i,x) > 0\}$ for equation 10. We then have the following theorem.

Theorem 3. *For a fixed test point x, depending on whether $G(q|X)$ is estimated by equation 9 or equation 10, the time complexity for Algorithm 1 is $O(n\max\{k,\log(n)\})$ or $O(nm\log(n)^{p-1})$, respectively.*

Proof of Theorem 3. To get the candidate set $\mathcal{C}$, if we use the k-nearest neighbor estimator equation 9, 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 equation 10, 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 equation 9, $|\mathcal{C}| = k$; and for equation 10, $|\mathcal{C}|$ is in the order of $m\log(n)^{p-1}$ by Lin and Jeon (2006). □

A.2 Proof of Theorem 1

Proof. When the conditions 1 to 4 are satisfied, by Theorem 3 in Athey et al. (2019) or Theorem 1 in Meinshausen (2006), we have

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

Note that $\sum_{i=1}^n w(X_i,x) = 1$ and $0 \le w(X_i,x) \le 1/m$. For convenience, we suppress the dependency on x and denote $F_n(q) = \sum_{i=1}^n w(X_i,x)\mathbb{1}\{Y_i \le q\}$ and $F(q) = \mathbb{P}(Y \le q|x)$. Because F is continuous, choose $q_0 < q_1 < \ldots < q_n$ from $\mathcal{B}$ such that $F(q_j) - F(q_{j-1}) = 1/n$. Then for any $q \in \mathcal{B}$, there exists $j \in \{1,\ldots,n\}$ such that $q \in [q_{j-1}, q_j]$, and hence $F_n(q) - F(q) \le F_n(q_j) - F(q_{j-1}) = F_n(q_j) - F(q_j) + 1/n$. Similarly, $F_n(q) - F(q) \ge F_n(q_{j-1}) - F(q_{j-1}) - 1/n$. Therefore, we have

$$\sup_{q\in\mathcal{B}} |F_n(q) - F(q)|x)| \le \max_{j=1,\ldots,n} |F_n(q_j) - F(q_j)| + 1/n.$$

Then by Bonferroni's inequality, we have

$$\sup_{q\in\mathcal{B}} |F_n(q) - F(q)|x)| = o_p(1).$$

Combined with Condition 4, we have the expected result. □

A.3 Proof of Theorem 2

Proof of Theorem 2. 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 $\epsilon > 0$, $\inf\{|S(q;\tau)| : |q - q^*| \ge \epsilon, q \in [-r,r]\} > 0$. Here, q^* stands for the true τth quantile of T.
3. $S_n(q_n;\tau) = o_p(1)$.

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

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

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

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

Since $0 < \tau < 1$, by Condition 2, there exists some $l > 0$ such that $G(q|x) \ge l$ and

$$|\mathbb{P}(T \le q|x) - \tau| \ge l$$

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

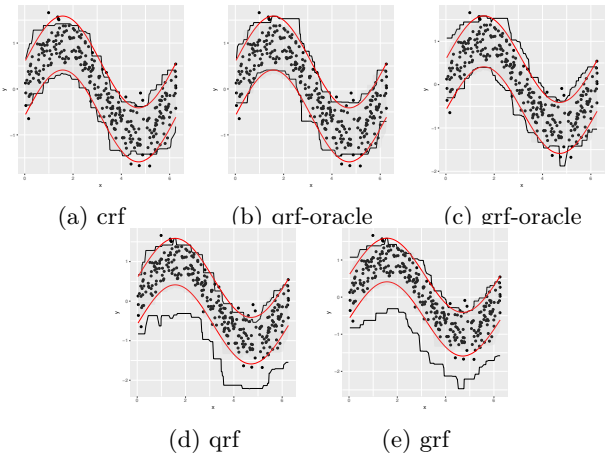
$$\begin{aligned} &|S_n(q_n;\tau)| \\ &= |S_n(q_n;\tau)| - |S_n(q^*;\tau)| + |S_n(q^*;\tau)| - |S(q^*;\tau)| \\ &\le |S_n(q^*;\tau) - S(q^*;\tau)| \\ &\le \sup_{q\in[-r,r]} |S_n(q;\tau) - S(q;\tau)| \\ &= o_p(1) \end{aligned}$$

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

B More Experiments

B.1 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 7 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.



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

(d) qrf (e) grf

Figure 7: Prediction intervals of the univariate censored since model. We observe that in all of the cases, our method *crf* and *qrf-oracle* give the coverage closest to 95%. 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.

B.2 One-dimensional Sine-curve Model

Since the proposed method *crf* is nonparametric and does not rely on any parametric assumption, it can be used to estimate quantiles for any general model $T = f(X) + \epsilon$. Hence we set $f(x) = \sin(x)$ and

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

where $X \sim \text{Unif}(0, 2\pi)$ and $\epsilon \sim \mathcal{N}(0, 0.3^2)$. The censoring variable $C \sim 1 + \sin(X) + \text{Exp}(\lambda = 0.2)$ depends on the covariates, and the censoring level is about 25%. The results are in Figure 8.

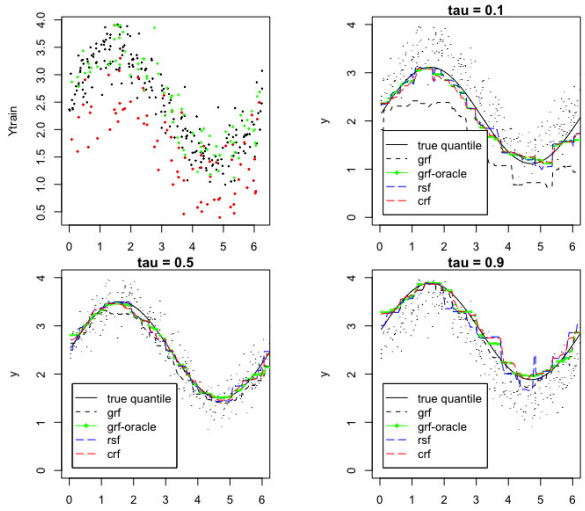


Figure 8: One-dimensional Sine model results.

Again, the proposed model *crf* produces almost identical quantile predictions compared with *grf-oracle*. Especially when $\tau = 0.1$, the grf result (blue dotted curve) severely deviates from the true quantile, while *crf* still predicts the correct quantile and performs as good as the oracle *grf-oracle*.

B.3 Censoring Level

In this section, we investigate the impact of censoring levels on the predictions of different forest algorithms. The results are summarized in Figure 9.

We use the same multi-dimensional AFT data as in section 5.1.1 but vary the parameter λ of the Exponential noise from 0.10 to 0.24, where larger λ means higher censoring level. As can be seen from Figure 9, the performance of both vanilla qrf and grf deteriorate fast when more data are censored. Meanwhile, all the censoring forest algorithms are quite robust to censoring, with almost flat quantile loss curves. Among them, the proposed methods, crf-quantile and crf-generalized still outperform rsf on almost all the censoring levels.

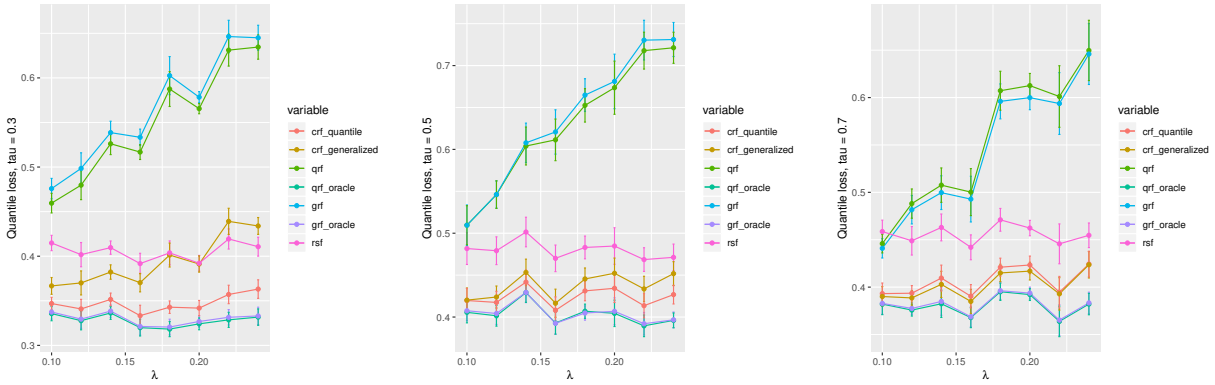


Figure 9: Comparison of the quantile losses of different forest algorithms under various censoring levels.