Consistency of survival tree and forest models: splitting bias and correction

Yifan Cui, Ruoqing Zhu, Mai Zhou, Michael Kosorok

National University of Singapore

University of Illinois at Urbana-Champaign

University of Kentucky

University of North Carolina at Chapel Hill

Supplementary Material

The following table provides a summary of Appendices.

The following table provides a summary of notation used in the proofs.

Basic Notation	
T	Failure time
C	Censoring time
Y	$= \min(T, C)$: observed time
δ	$= \mathbb{1}(T \leq C)$: censoring indicator
F_i, f_i	Survival distribution of i-th observation, $f_i = dF_i$
G_i	Censoring distribution of i-th observation
$\mathcal{A}$	A node, internal or terminal
$\mathscr{A}$	$= \{\mathcal{A}_u\}_{u \in \mathcal{U}}$, the collection of all terminal nodes in a single tree
$\Lambda(t\|x)$	Cumulative hazard function (CHF)
$\widehat{\Lambda}_n$, $\widehat{\Lambda}_{\mathcal{A},n}$, $\widehat{\Lambda}_{\mathscr{A},n}$	NA estimator on a set of samples, a node $\mathcal{A}$, or an entire tree $\mathscr{A}$
Λ^*_n, $\Lambda^*_{\mathcal{A},n}$, $\Lambda^*_{\mathscr{A},n}$	Censoring contaminated averaged CHF on a set of samples, a node $\mathcal{A}$, or the entire tree $\mathscr{A}$
Λ^*, $\Lambda^*_{\mathcal{A}}$, $\Lambda^*_{\mathscr{A}}$	Population versions of Λ^*_n, $\Lambda^*_{\mathcal{A},n}$ and $\Lambda^*_{\mathscr{A},n}$, respectively
$\widetilde{\Lambda}_n$, $\widetilde{\Lambda}_{\mathcal{A},n}$, $\widetilde{\Lambda}_{\mathscr{A},n}$	Biased correct NA estimator on a set of samples, a node $\mathcal{A}$, or an entire tree $\mathscr{A}$
k	Minimum leaf size
α	Minimum proportion of observations contained in child node
B	Number of trees in a forest
$\mathcal{V}_{\alpha,k}(\mathcal{D})$	Set of all $\{\alpha, k\}$ valid partitions on the feature space $\mathcal{X}$
$\mathcal{H}_{\alpha,k}(\mathcal{D})$	Set of all $\{\alpha, k\}$ valid forests on the feature space $\mathcal{X}$
$\mathcal{R}$	Approximation node
$\mathscr{R}$	The set of approximation nodes
$N(t)$	Counting process
$K(t)$	At-risk process
$\mu(\mathcal{R})$, $\mu(\mathcal{A})$	The expected fraction of training samples inside $\mathcal{R}, \mathcal{A}$
$\#\mathcal{R}$, $\#\mathcal{A}$	The number of training samples inside $\mathcal{R}, \mathcal{A}$
$\mathscr{M}_F, \mathscr{M}_C, \mathscr{M}_N$	Set of indices of failure variables, censoring variables, noise variables, respectively

Constants	
$d(d_0, d_1)$	Dimension of (failure, censoring) covariates
τ	The positive constant as the upper bound of Y
M	Lower bound of $\mathrm{pr}(Y \geq \tau \mid X)$
ζ	A constant used in Assumption 2
L	Bound of the density function $f(t)$
L_1, L_2	Lipschitz constant of Λ and λ
ℓ, ℓ'	Minimum effect size of marginal signal of the failure distribution
γ	Bound for weak dependency

S1

Algorithm 1: Pseudo algorithm for survival forest models

```
   Input: Training set $\mathcal{D}_n$, terminal node size $k$, number of trees $B$;
1  for $b = 1$ to $B$ do
2      Initiate $\mathcal{A} = \mathcal{X}$, a bootstrap sample $\mathcal{D}_n^b$ of $\mathcal{D}_n$, $\mathcal{K}_b = \emptyset$, $u = 1$;
3      At a node $\mathcal{A}$, if $\sum_{X_i \in \mathcal{D}_n^b} \mathbb{1}(X_i \in \mathcal{A}) < k$, proceed to Line 5. Otherwise,
        construct a splitting rule such that $\mathcal{A} = \mathcal{A}_{\text{left}} \cup \mathcal{A}_{\text{right}}$, where
        $\mathcal{A}_{\text{left}} \cap \mathcal{A}_{\text{right}} = \emptyset$. ;
4      Send the two child nodes $\mathcal{A}_{\text{left}}$ and $\mathcal{A}_{\text{right}}$ to Line 3 separately;
5      Conclude the current node $\mathcal{A}$ as a terminal node $\mathcal{A}_u^b$, calculate $\widehat{\Lambda}_{\mathcal{A}_u^b,n}$ using
        the within-node data, and update $\mathcal{K}_b = \mathcal{K}_b \cup \{u\}$ and $u = u + 1$;
6  end
7  return $\{\{\mathcal{A}_u^b, \widehat{\Lambda}_{\mathcal{A}_u^b,n}\}_{u \in \mathcal{K}_b}\}_{b=1}^B$
```

Algorithm 2: A marginal splitting rule for survival forest

```
1 At any internal node $\mathcal{A}$ containing at least $2k$ training samples, we pick a
    splitting variable $j \in \{1, \ldots, d\}$ uniformly at random;
2 We then pick the splitting point $\tilde{c}$ using the following rule such that both child
    nodes contain at least proportion $\alpha$ of samples at $\mathcal{A}$:
        $$\tilde{c} = \arg\max_{c} \Delta_1(c),$$
    where $\Delta_1(c) = \max_{t<\tau} \left|\widehat{\Lambda}_{\mathcal{A}_j^+(c),n}(t) - \widehat{\Lambda}_{\mathcal{A}_j^-(c),n}(t)\right|$, $\mathcal{A}_j^+(c) = \{X : X^{(j)} \geq c\}$,
    and $\mathcal{A}_j^-(c) = \{X : X^{(j)} < c\}$, $X^{(j)}$ is the $j$-th dimension of $X$;
3 If the variable $j$ has already been used along the sequence of splitting rules
    leading up to $\mathcal{A}$, or the following inequality holds for some constant $M_3$:
        $$\Delta_1(\tilde{c}) \geq (\gamma^2 - \gamma^{-2}) \frac{\tau L}{M^2} + M_3 \sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}},$$
    we split at $\tilde{c}$ along the $j$-th variable, where $L$ and $M_3$ can be regarded as
    tuning parameters obtained by cross-validation. If not, we randomly sample
    another variable out of remaining feasible variables and proceed to Step 2).
    When there is no remaining variable, we randomly select an index out of $d$
    and split at $\tilde{c}$.
```

S2

Proof of Lemma 1. For simplicity, we prove the results for the case when there are no ties in the failure time. The proof follows mostly Cuzick (1985). Let $n_1 > n_2 > \ldots > n_k \geq 1$ be the sequence of counts of the at-risk sample size, i.e., $n_j = \sum_{i=1}^{n} \mathbb{1}(Y_i \geq t_j)$, where t_i is the ith ordered failure time. Then the Kaplan–Meier estimator at any observed failure time point t_j can be expressed as $\widehat{S}_{KM}(t_j) = \prod_{i=1}^{j}(n_i - 1)/n_i$, while the Nelson–Altshuler estimator at the same time point is $\widehat{S}_{NA}(t_j) = \exp\{-\sum_{i=1}^{j} 1/n_i\}$. We first apply the Taylor expansion of e^{-n_i} for $n_i \geq 1$:

$$1 - 1/n_i < e^{-n_i} < 1 - 1/n_i + 1/(2n_i^2) \leq 1 - 1/(n_i + 1).$$

Thus we can bound the Nelson–Altshuler estimator with

$$\widehat{S}_{KM}(t_j) < \widehat{S}_{NA}(t_j) < \prod_{i=1}^{j} n_i/(n_i + 1).$$

To bound the difference between the two estimators, note that for $n_j \geq 2$,

$$\begin{aligned} \left|\widehat{S}_{KM}(t_j) - \widehat{S}_{NA}(t_j)\right| &< \left|\widehat{S}_{KM}(t_j) - \textstyle\prod_{i=1}^{j} n_i/(n_i + 1)\right| \\ &= \widehat{S}_{KM}(t_j)\left|1 - \textstyle\prod_{i=1}^{j} \frac{n_i/(n_i+1)}{(n_i-1)/n_i}\right| \\ &\leq \widehat{S}_{KM}(t_j) \textstyle\sum_{i=1}^{j}(n_i^2 - 1)^{-1} \\ &\leq 2\widehat{S}_{KM}(t_j) \textstyle\sum_{i=1}^{j} n_i^{-2} \\ &\leq 4\widehat{S}_{KM}(t_j)/n_j. \qquad \text{(S2.1)} \end{aligned}$$

Now note that both the Kaplan–Meier and the Nelson–Altshuler estimators stay constant within (t_i, t_{i+1}), and this bound applies to the entire interval $(0, t_k)$ for $n_k \geq 2$. $\square$

Proof of Theorem 1. Recall the counting process

$$N(s) = \sum_{i=1}^{n} N_i(s) = \sum_{i=1}^{n} \mathbb{1}(Y_i \leq s, \delta_i = 1),$$

and the at risk process

$$K(s) = \sum_{i=1}^{n} K_i(s) = \sum_{i=1}^{n} \mathbb{1}(Y_i \geq s).$$

We prove the theorem based on the following key results.

Lemma S1. *Provided Assumption 1 holds, for arbitrary $\epsilon > 0$ and n such that $\frac{1}{n} \leq \frac{\epsilon^2}{2}$, we have*

$$\mathrm{pr}(\sup_{t\leq\tau} |\frac{1}{n} \sum_{i=1}^{n} \{K_i(s) - E[K_i(s)]\}| > \epsilon) \leq 8(n+1) \exp\{\frac{-n\epsilon^2}{32}\},$$
$$\mathrm{pr}(\sup_{t\leq\tau} |\frac{1}{n} \sum_{i=1}^{n} \{N_i(t) - E[N_i(t)]\}| > \epsilon) \leq 8(n+1) \exp\{\frac{-n\epsilon^2}{32}\}.$$

Lemma S2. *Provided Assumption 1 holds, for any $\epsilon > 0$, we have*

$$\mathrm{pr}(\sup_{t\leq\tau} |\int_0^t (\frac{1}{K(s)} - \frac{1}{E[K(s)]}) dN(s)| > \epsilon) \leq 8(n+2) \exp\Big\{ - \frac{n \min(\epsilon^2 M^4, M^2)}{128} \Big\},$$

where n satisfies $\frac{1}{n} < \min(\frac{\epsilon^2}{2}, \frac{\epsilon^2 M^4}{4})$ and M is defined in Assumption 1.

Lemma S3. *Provided Assumption 1 holds, for any $\epsilon > 0$, we have*

$$\mathrm{pr}(\sup_{t\leq\tau} |\int_0^t \frac{d\{N(s) - E[N(s)]\}}{E[K(s)]}| > \epsilon) \leq 8(n+1) \exp\Big\{ - \frac{n\epsilon^2 M^2}{228} \Big\},$$

where n satisfies $\frac{1}{n} \leq \frac{\epsilon^2}{2}$ and M is defined in Assumption 1.

The proof of Lemma S1 follows pages 14–16 in Pollard (2012). The proofs of Lemma S2 and S3 are presented below. Now we are ready to prove Theorem 1. Note that

$$\begin{aligned}
&\mathrm{pr}\big(\sup_{t<\tau}|\widehat{\Lambda}_n(t) - \Lambda_n^*(t)| > \epsilon_1\big)\\
= &\,\mathrm{pr}\Big(\sup_{t<\tau}|\widehat{\Lambda}_n(t) - \int_0^t \frac{dE[N(s)]}{E[K(s)]}| > \epsilon_1\Big)\\
\leq &\,\mathrm{pr}\Big(\sup_{t\leq\tau}\Big|\int_0^t \Big[\frac{1}{K(s)} - \frac{1}{E[K(s)]}\Big]dN(s)\Big| > \frac{\epsilon_1}{2}\Big)\\
&+\mathrm{pr}\Big(\sup_{t\leq\tau}\Big|\int_0^t \frac{d\{N(s) - E[N(s)]\}}{E[K(s)]}\Big| > \frac{\epsilon_1}{2}\Big).
\end{aligned}$$

By Lemma S2, the first term is bounded by $8(n+2)\exp\{-\frac{n\min(\epsilon_1^2M^4, 4M^2)}{512}\}$. By Lemma S3, the second term is bounded by $8(n+1)\exp\{-\frac{n\epsilon_1^2M^2}{1152}\}$. The sum of these two terms is further bounded by $16(n+2)\exp\{-\frac{n\epsilon_1^2M^4}{1152}\}$ for any $\epsilon_1 \leq 2$ and $n > \frac{4}{\epsilon_1^2M^4}$. This completes the proof. □

Proof of Lemma S2. For any $t \leq \tau$,

$$\begin{aligned}
&\Big|\int_0^t (\frac{1}{K(s)} - \frac{1}{E[K(s)]})dN(s)\Big|\\
\leq &\int_0^t \frac{|E[K(s)] - K(s)|}{K(s)E[K(s)]}dN(s)\\
\leq &\int_0^t \frac{\sup_{0<r\leq\tau}|E[K(r)] - K(r)|}{K(s)E[K(s)]}dN(s).
\end{aligned} \tag{S2.2}$$

Thanks to Hoeffding's inequality, we have

$$\mathrm{pr}\Big(\big|K(\tau) - E[K(\tau)]\big| > \frac{nM}{2}\Big) < 2\exp\big\{-\frac{nM^2}{2}\big\}.$$

Then (S2.2) is further bounded by

$$\frac{n}{(nM)^2/2}\sup_{0<t\leq\tau}\big|E[K(t)]-K(t)\big|.$$

Combining with Lemma S1, we have

$$\begin{aligned}
&\text{pr}\Big(\sup_{t\leq\tau}\Big|\int_0^t(\frac{1}{K(s)}-\frac{1}{E[K(s)]})dN(s)\Big|>\epsilon\Big)\\
&\leq \text{pr}\Big(\frac{2}{nM^2}\sup_{t\leq\tau}|E[K(t)]-K(t)|>\epsilon\Big)\\
&\leq 8(n+2)\exp\Big\{-\frac{n\min(\epsilon^2M^4,M^2)}{128}\Big\},
\end{aligned}$$

for any n satisfying $\frac{1}{n}<\min(\frac{\epsilon^2}{2},\frac{\epsilon^2M^4}{4})$. This completes the proof. □

Proof of Lemma S3. For any $t\leq\tau$, we utilize integration by parts to obtain

$$\begin{aligned}
&\Big|\int_0^t\frac{1}{EK(s)}d\{N(s)-E[N(s)]\}\Big|\\
&=\Big|\frac{N(s)-E[N(s)]}{E[K(s)]}\Big|_0^t-\int_0^t\{N(s)-E[N(s)]\}d\Big\{\frac{1}{E[K(s)]}\Big\}\Big|\\
&\leq 2\sup_{t\leq\tau}\big|N(t)-E[N(t)]\big|\frac{1}{E[K(\tau)]}+\sup_{t\leq\tau}\big|N(t)-E[N(t)]\big|\int_0^\tau d\Big\{\frac{1}{E[K(s)]}\Big\}\\
&\leq\frac{3}{M}\sup_{t\leq\tau}\frac{1}{n}\big|N(t)-E[N(t)]\big|.
\end{aligned}$$

Thanks to Lemma S1, we now have

$$\begin{aligned}
&\text{pr}\Big(\sup_{t\leq\tau}\Big|\int_0^t\frac{d\{N(s)-E[N(s)]\}}{E[K(s)]}\Big|>\epsilon\Big)\\
&\leq \text{pr}\Big(\frac{3}{nM}\sup_{t\leq\tau}\big|N(t)-E[N(t)]\big|\Big)\\
&\leq 8(n+1)\exp\Big\{-\frac{n\epsilon^2M^2}{288}\Big\},
\end{aligned}$$

where n satisfies $\frac{1}{n} \leq \frac{\epsilon^2}{2}$. This completes the proof. □

S3

Preliminary. The proof of Theorem 2 uses two main mechanisms: the concentration bound results we established in Theorem 1 to bound the variations in each terminal node, and a construction of a parsimonious set of rectangles, namely $\mathscr{R}$, defined in Wager and Walther (2015). We first introduce some notation. Denote the rectangles $\mathcal{R} \in [0,1]^d$ by

$$\mathcal{R} = \bigotimes_{j=1}^{d} [r_j^-, r_j^+], \quad \text{where} \quad 0 \leq r_j^- < r_j^+ \leq 1 \quad \text{for all} \quad j = 1, \cdots, d.$$

The Lebesgue measure of rectangle $\mathcal{R}$ is $\lambda(\mathcal{R}) = \prod_{j=1}^{d}(r_j^+ - r_j^-)$. Here we define the expected fraction of training samples and the number of training samples inside R, respectively, as follows:

$$\mu(\mathcal{R}) = \int_{\mathcal{R}} f(x)dx, \#\mathcal{R} = |\{i : X_i \in \mathcal{R}\}|.$$

We define the support of rectangle $\mathcal{R}$ as $S(\mathcal{R}) = \{j \in 1, \ldots, d : r_j^- \neq 0 \text{ or } r_j^+ \neq 1\}$.

Lemma S4 below shows that with high probability there are enough observations larger than or equal to τ on the rectangle $\mathcal{R}$.

Lemma S4. *Provided Assumption 1 holds, the number of observations larger than or equal to τ on all $\mathcal{R} \in \mathscr{R}$ is larger than* $\left(1 - \sqrt{\frac{4\log(|\mathscr{R}|\sqrt{n})}{kM}}\right)kM$

with probability larger than $1 - 1/\sqrt{n}$.

Proof. For one $\mathcal{R} \in \mathscr{R}$, by the Chernoff bound, with probability larger than $1 - \exp\big\{ - \frac{c^2 \#\mathcal{R} M}{2})\big\} \geq 1 - \exp\big\{ - \frac{c^2 k M}{4})\big\}$, the number of observations larger than or equal to τ on $\mathcal{R}$ is larger than $(1-c)kM$, where $0 < c < 1$ is a constant. Thus with probability larger than $1 - 1/\sqrt{n}$, the number of observations larger than or equal to τ on every $\mathcal{R} \in \mathscr{R}$ is larger than $\Big(1 - \sqrt{\frac{4\log(|\mathscr{R}|\sqrt{n})}{kM}}\Big)kM$. □

Proof of Theorem 2. We first establish a triangle inequality by picking some element $\mathcal{R}$ in the set $\mathscr{R}$ such that it is a close approximation of $\mathcal{A}$ and $\mathcal{R} \subseteq \mathcal{A}$.

$$\begin{aligned} &\sup_{t<\tau,\, \mathcal{A}\in\mathscr{A},\, \mathscr{A}\in\mathcal{V}} \big|\widehat{\Lambda}_{\mathcal{A},n}(t) - \Lambda^*_{\mathcal{A},n}(t)\big| \\ &\qquad \leq \sup_{t<\tau,\, \mathcal{A}\in\mathscr{A},\, \mathscr{A}\in\mathcal{V}} \inf_{\mathcal{R}\in\mathscr{R}} \big|\widehat{\Lambda}_{\mathcal{A},n}(t) - \widehat{\Lambda}_{\mathcal{R}}(t)\big| \\ &\qquad + \sup_{t<\tau,\, \mathcal{R}\in\mathscr{R},\, \#\mathcal{R}\geq k/2} \big|\widehat{\Lambda}_{\mathcal{R},n}(t) - \Lambda^*_{\mathcal{R},n}(t)\big| \\ &\qquad + \sup_{t<\tau,\, \mathcal{A}\in\mathscr{A},\, \mathscr{A}\in\mathcal{V}} \inf_{\mathcal{R}\in\mathscr{R}} \big|\Lambda^*_{\mathcal{R},n}(t) - \Lambda^*_{\mathcal{A},n}(t)\big|. \end{aligned} \tag{S3.1}$$

Here, we have $\#\mathcal{R} \geq k/2$ in the sub-index of the second term because $\#\mathcal{A} \geq k$ and from Theorem 10 in Wager and Walther (2015), $\#\mathcal{A} - \#\mathcal{R} \leq 3\zeta^2 \#\mathcal{A}/\sqrt{k} + 2\sqrt{3\log(|\mathscr{R}|)\#\mathcal{A}} + O(\log(|\mathscr{R}|)) = o(k)$ for any possible $\mathcal{A}$ with probability larger than $1 - 1/\sqrt{n}$.

We now bound each part of the right hand side of the above inequality. Note that we always select a close approximation of $\mathcal{A}$ from the set $\mathscr{R}$.

With a slight abuse of notation, we let the subject index i first run through the observations within $\mathcal{R}$ and then through the observations in $\mathcal{A}$ but not in $\mathcal{R}$. This can always be done since $\mathcal{R} \subseteq \mathcal{A}$. Thus we have

$$\begin{aligned}
&\sup_{t<\tau,\,\mathcal{A}\in\mathscr{A},\,\mathscr{A}\in\mathcal{V}} \big|\widehat{\Lambda}_{\mathcal{R},n}(t) - \widehat{\Lambda}_{\mathcal{A},n}(t)\big| \\
&\le \sup_{t<\tau,\mathcal{A}\in\mathscr{A},\mathscr{A}\in\mathcal{V}} \Big| \sum_{s\le t} \frac{[\Delta N(s)]_{\mathcal{R}}}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s)} - \sum_{s\le t} \frac{[\Delta N(s)]_{\mathcal{R}} + [\Delta N(s)]_{\mathcal{A}\setminus\mathcal{R}}}{\sum_{i=1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s)} \Big| \\
&= \sup_{t<\tau,\mathcal{A}\in\mathscr{A},\mathscr{A}\in\mathcal{V}} \Big| \sum_{s\le t} \frac{[\Delta N(s)]_{\mathcal{R}}}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s)} \\
&\qquad\qquad\qquad - \sum_{s\le t} \frac{[\Delta N(s)]_{\mathcal{R}} + [\Delta N(s)]_{\mathcal{A}\setminus\mathcal{R}}}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s)} \Big| \\
&\le \sup_{t<\tau,\mathcal{A}\in\mathscr{A},\mathscr{A}\in\mathcal{V}} \Big\{ \sum_{j=\#\mathcal{R}+1}^{\#\mathcal{A}} \frac{\Delta N(s_j)}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s_j)} \\
&\qquad + \sum_{j=1}^{\#\mathcal{R}} \Big[\frac{\Delta N(s_j)}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j)} - \frac{\Delta N(s_j)}{\sum_{i=1}^{\#\mathcal{R}} \mathbb{1}(Y_i \ge s_j) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}} \mathbb{1}(Y_i \ge s_j)} \Big] \Big\},
\end{aligned}$$

where $N(s) = \sum_{i=1}^{n} N_i(s) = \sum_{i=1}^{n} \mathbb{1}(Y_i \le s, \delta_i = 1)$. By Lemma S4 the first term is bounded by

$$\begin{aligned}
&\frac{\#\mathcal{A} - \#\mathcal{R}}{\Big(1 - \sqrt{\frac{4\log(|\mathscr{R}|\sqrt{n})}{kM}}\Big) kM} \\
&\le \frac{1}{\Big(1 - \sqrt{\frac{4\log(|\mathscr{R}|\sqrt{n})}{kM}}\Big) M} \Big[\frac{6\zeta^2}{\sqrt{k}} + 2\sqrt{\frac{6\log(|\mathscr{R}|)}{k}} + O\Big(\frac{\log(|\mathscr{R}|)}{k}\Big) \Big],
\end{aligned}$$

and the second term is bounded by

$$\begin{aligned}
&\sum_{j=1}^{\#\mathcal{R}}\Big[\frac{\Delta N(s_j)}{\sum_{i=1}^{\#\mathcal{R}}\mathbb{1}(Y_i \ge s_j)} - \frac{\Delta N(s_j)}{\sum_{i=1}^{\#\mathcal{R}}\mathbb{1}(Y_i \ge s_j) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}}\mathbb{1}(Y_i \ge s_j)}\Big] \\
&\le \sum_{j=1}^{\#\mathcal{R}} \frac{\Delta N(s_j)\sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}}\mathbb{1}(Z_i \ge s_j)}{\big[\sum_{i=1}^{\#\mathcal{R}}\mathbb{1}(Y_i \ge s_j)\big]\big[\sum_{i=1}^{\#\mathcal{R}}\mathbb{1}(Y_i \ge s_j) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}}\mathbb{1}(Y_i \ge s_j)\big]} \\
&\le \sum_{j=1}^{\#\mathcal{R}} \frac{\Delta N(s_j)(\#\mathcal{A} - \#\mathcal{R})}{\big[\sum_{i=1}^{\#\mathcal{R}}\mathbb{1}(Y_i \ge s_j)\big]\big[\sum_{i=1}^{\#\mathcal{R}}\mathbb{1}(Y_i \ge s_j) + \sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}}\mathbb{1}(Y_i \ge s_j)\big]} \\
&\le \frac{\#\mathcal{R}(\#\mathcal{A} - \#\mathcal{R})}{\Big(1 - \sqrt{\frac{4\log(|\mathscr{R}|\sqrt{n})}{kM}}\Big)^2 k^2 M^2} \\
&\le \frac{2}{\Big(1 - \sqrt{\frac{4\log(|\mathscr{R}|\sqrt{n})}{kM}}\Big)^2 M^2}\Big[\frac{6\zeta^2}{\sqrt{k}} + 2\sqrt{\frac{6\log(|\mathscr{R}|)}{k}} + O\Big(\frac{\log(|\mathscr{R}|)}{k}\Big)\Big].
\end{aligned}$$

Combining these two terms, the first part of Equation (S3.1) is bounded by

$$\frac{3}{\Big(1 - \sqrt{\frac{4\log(|\mathscr{R}|\sqrt{n})}{kM}}\Big)^2 M^2}\Big[\frac{6\zeta^2}{\sqrt{k}} + 2\sqrt{\frac{6\log(|\mathscr{R}|)}{k}} + O\Big(\frac{\log(|\mathscr{R}|)}{k}\Big)\Big], \tag{S3.2}$$

with probability larger than $1 - 1/\sqrt{n}$. For the second part, by Theorem 1,

$$\sup_{t<\tau, \mathcal{R}\in\mathscr{R}, \#\mathcal{R}\ge k/2} \big|\widehat{\Lambda}_{\mathcal{R},n}(t) - \Lambda^*_{\mathcal{R},n}(t)\big| \le \frac{\{1728\log(n)\}^{1/2}}{k^{1/2}M^2}, \tag{S3.3}$$

with probability larger than $1 - 1/\sqrt{n}$. The third part of Equation (S3.1) is bounded by

$$
\begin{aligned}
&\sup_{t<\tau,\mathcal{A}\in\mathscr{A},\mathscr{A}\in\mathcal{V}} \left|\Lambda^*_{\mathcal{A},n}(t)-\Lambda^*_{\mathcal{R},n}(t)\right| \\
\leq &\sup_{t<\tau,\mathcal{A}\in\mathscr{A},\mathscr{A}\in\mathcal{V}} \Big|\int_0^t \frac{\sum_{i=1}^{\#\mathcal{A}}\{1-G_i(s)\}dF_i(s)}{\sum_{i=1}^{\#\mathcal{A}}\{1-G_i(s)\}\{1-F_i(s)\}} - \int_0^t \frac{\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}dF_i(s)}{\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}\{1-F_i(s)\}}\Big| \\
\leq &\sup_{t<\tau,\mathcal{A}\in\mathscr{A},\mathscr{A}\in\mathcal{V}} \int_0^t \Big| \frac{\big[\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}dF_i(s)\big]\big[\sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}}\{1-G_i(s)\}\{1-F_i(s)\}\big]}{\big[\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}\{1-F_i(s)\}\big]\big[\sum_{i=1}^{\#\mathcal{A}}\{1-G_i(s)\}\{1-F_i(s)\}\big]} \\
&\qquad - \frac{\big[\sum_{i=\#\mathcal{R}+1}^{\#\mathcal{A}}\{1-G_i(s)\}dF_i(s)\big]\big[\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}\{1-F_i(s)\}\big]}{\big[\sum_{i=1}^{\#\mathcal{R}}\{1-G_i(s)\}\{1-F_i(s)\}\big]\big[\sum_{i=1}^{\#\mathcal{A}}\{1-G_i(s)\}\{1-F_i(s)\}\big]}\Big| \\
\leq &\sup_{t<\tau,\mathcal{A}\in\mathscr{A},\mathscr{A}\in\mathcal{V}} \tau\frac{\#\mathcal{A}(\#\mathcal{A}-\#\mathcal{R})}{\#\mathcal{R}\#\mathcal{A}M^4} \\
\leq &\frac{2\tau}{M^4}\{\frac{3\zeta^2}{\sqrt{k}}+2\sqrt{\frac{3\log(|\mathscr{R}|)}{k}}+O(\frac{\log(|\mathscr{R}|)}{k})\}. \qquad \text{(S3.4)}
\end{aligned}
$$

Combining inequalities (S3.2), (S3.3) and (S3.4) and Corollary 8 in Wager and Walther (2015), we obtain the desired adaptive concentration bound. With probability larger than $1-2/\sqrt{n}$, we have

$$
\begin{aligned}
&\sup_{t<\tau,\mathcal{A}\in\mathscr{A},\mathscr{A}\in\mathcal{V}} |\widehat{\Lambda}_{\mathcal{A},n}(t)-\Lambda^*_{\mathcal{A},n}(t)| \\
\leq &\frac{3}{\Big(1-\sqrt{\frac{4\log(|\mathscr{R}|\sqrt{n})}{kM}}\Big)^2M^2}\Big[\frac{6\zeta^2}{\sqrt{k}}+2\sqrt{\frac{6\log(|\mathscr{R}|)}{k}}+O\Big(\frac{\log(|\mathscr{R}|)}{k}\Big)\Big] \\
&\qquad + \frac{(1728\log n)^{1/2}}{k^{1/2}M^2}+\frac{2\tau}{M^4}\Big\{\frac{3\zeta^2}{\sqrt{k}}+2\sqrt{\frac{3\log(|\mathscr{R}|)}{k}}+O\Big(\frac{\log(|\mathscr{R}|)}{k}\Big)\Big\} \\
\leq &M_1\Big[\sqrt{\frac{\log(|\mathscr{R}|)}{k}}+\sqrt{\frac{\log(n)}{k}}\Big] \leq M_1\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}},
\end{aligned}
$$

where M_1 is an universal constant. This completes the proof of Theorem 2. □

Corollary S1. *Suppose Assumptions 1-3 hold. Then all valid forests concentrate on the censoring contaminated forest with probability larger than* $1 - 2/\sqrt{n}$*,*

$$\sup_{t<\tau,\, x\in[0,1]^d,\, \{\mathscr{A}_{(b)}\}_1^B \in \mathcal{H}_{\alpha,k}(\mathcal{D}_n)} \left| \widehat{\Lambda}_{\{\mathscr{A}_{(b)}\}_1^B, n}(t \mid x) - \Lambda^*_{\{\mathscr{A}_{(b)}\}_1^B, n}(t \mid x) \right| \\ \leq M_1 \sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}},$$

for some universal constant M_1*.*

Proof of Corollary S1. Since for any $\mathscr{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$ we have

$$\sup_{t<\tau,\, x\in[0,1]^d} \left| \widehat{\Lambda}_{\mathscr{A},n}(t \mid x) - \Lambda^*_{\mathscr{A},n}(t \mid x) \right| \leq \; M_1 \sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}},$$

and furthermore if $\liminf_{n\to\infty}(d/n) \to \infty$, for any $\mathscr{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$,

$$\sup_{t<\tau,\, x\in[0,1]^d} \left| \widehat{\Lambda}_{\mathscr{A},n}(t \mid x) - \Lambda^*_{\mathscr{A},n}(t \mid x) \right| \leq \; M_1 \sqrt{\frac{\log(n)\log(d)}{k \log((1-\alpha)^{-1})}}.$$

By the definition of $\mathcal{H}_{\alpha,k}(\mathcal{D}_n)$, any $\{\mathscr{A}_{(b)}\}_1^B$ belonging to $\mathcal{H}_{\alpha,k}(\mathcal{D}_n)$ is an element of $\mathcal{V}_{\alpha,k}(\mathcal{D}_n)$. Hence we have,

$$\sup_{t<\tau,\, x\in[0,1]^d,\, \{\mathscr{A}_{(b)}\}_1^B \in \mathcal{H}_{\alpha,k}(\mathcal{D}_n)} \left| \widehat{\Lambda}_{\{\mathscr{A}_{(b)}\}_1^B, n}(t \mid x) - \Lambda^*_{\{\mathscr{A}_{(b)}\}_1^B, n}(t \mid x) \right| \\ \leq M_1 \sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}},$$

for some universal constant M_1. □

S4

Proof of Theorem 3. In order to show consistency, we first show that each terminal node is small enough in all d dimensions. Let m be the lower bound of the number of splits on the terminal node $\mathcal{A}$ containing x, and m_i be the number of splits on the i-th dimension. Then we have

$$n\alpha^m = k, \quad m = \log_{1/\alpha}(n/k) = \frac{\log n - \log k}{\log(1/\alpha)} \ \text{ and } \ \sum_{i=1}^{d} m_i = m.$$

The lower bound of the number of splits on the i-th dimension m_i has distribution $Binomial(m, \frac{1}{d})$. By the Chernoff bound on each dimension,

$$\mathrm{pr}\Big(m_i > \frac{(1-c_2)m}{d}\Big) > 1 - \exp\Big\{-\frac{c_2^2 m}{2d}\Big\}$$

with any $0 < c_2 < 1$. Then, by Bonferroni,

$$\mathrm{pr}\Big(\min m_i > \frac{(1-c_2)m}{d}\Big) > 1 - d\exp\Big\{-\frac{c_2^2 m}{2d}\Big\}.$$

Suppose we are splitting at the i-th dimension on a specific internal node with ν observations. Recall the splitting rule is choosing the splitting point randomly between the $\max((k+1), \lceil \alpha\nu \rceil)$-th, and $\min((n-k-1), \lfloor (1-\alpha)\nu \rfloor)$-th observations. Without loss of generality, we consider splitting between the $\lceil \alpha\nu \rceil$-th and $\lfloor (1-\alpha)\nu \rfloor$-th observations. The event that the splitting point is between q_α and $q_{1-\alpha}$ happens with probability larger than c_3, where q_α is the α-th quantile of the i-th component of X conditional on the current internal node and previous splits. Here $c_3 = (1-2\alpha)/8$ and is just a lower bound. Since with probability larger than 1/4, the $\lfloor \frac{\alpha+0.5}{2}\nu \rfloor$-th order statistic is larger than α and the $\lceil \frac{1.5-\alpha}{2}\nu \rceil$-th order statistic is less than $1-\alpha$ for large enough ν, where ν is known to be larger than $2k$. So with probability larger than c_3, the splitting point is between q_α and $q_{1-\alpha}$. Thanks to Assumption 2, conditioning on the current internal node and previous splits, $\max_{x^{(j)}} p(x^{(j)})/\min_{x^{(j)}} p(x^{(j)}) < \zeta^2$, where $p(x^{(j)})$ is the marginal distribution of $x^{(j)}$. So with probability larger than c_3, the splitting point falls into the interval $[\alpha/\zeta^2, 1-\alpha/\zeta^2]$.

The number of splits which partition the parent node to two child nodes with proportion of length between both α and $1-\alpha$ on the i-th dimension of the terminal node $\mathcal{A}$ is denoted by m^* and is $Binomial(m_i, c_3)$. By the Chernoff bound, for any $0 < c_4 < 1$,

$$\mathrm{pr}\big(m^* \geq (1-c_4)c_3 m_i\big) \geq 1 - \exp\Big\{-\frac{c_4^2 c_3 m_i}{2}\Big\}.$$

If we denote the length of the i-th dimension on the terminal node $\mathcal{A}$ as l_i,

$$\mathrm{pr}\big(l_i \leq (1-\alpha/\zeta^2)^{(1-c_4)c_3 m_i}\big) \geq 1 - \exp\Big\{-\frac{c_4^2 c_3 m_i}{2}\Big\}.$$

Furthermore, by combining the d dimensions together, we obtain

$$\mathrm{pr}\big(\max_i l_i \leq (1-\alpha/\zeta^2)^{(1-c_4)c_3 \min_i m_i}\big) \geq 1 - d\exp\Big\{-\frac{c_4^2 c_3 \min_i m_i}{2}\Big\},$$

and then

$$\max_{x_1,x_2\in\mathcal{A}} ||x_1 - x_2|| \leq \sqrt{d}(1-\alpha/\zeta^2)^{\frac{c_3(1-c_4)(1-c_2)m}{d}},$$

with probability larger than $1-d\exp\big\{-\frac{c_2^2 m}{2d}\big\}-d\exp\big\{-\frac{(1-c_2)c_3c_4^2 m}{2d}\big\}$. Hence, for any observation x_j inside the node $\mathcal{A}$ containing x, by Assumption 4, we have

$$\sup_{t<\tau}|F(t \mid x)-F(t \mid x_j)| \leq L_1\sqrt{d}(1-\alpha/\zeta^2)^{\frac{c_3(1-c_4)(1-c_2)m}{d}},$$

$$\sup_{t<\tau}|f(t \mid x)-f(t \mid x_j)| \leq (L_1^2+L_2)\sqrt{d}(1-\alpha/\zeta^2)^{\frac{c_3(1-c_4)(1-c_2)m}{d}},$$

where $f(\cdot \mid x)$ and $F(\cdot \mid x)$ denote the true density function and distribution function at $x \in \mathcal{A}$, respectively. Then $\Lambda^*_{\mathcal{A},n}(t)$ has the upper and lower bounds

$$\int_0^t \frac{f(s \mid x)+b_1}{1-F(s \mid x)-b_2}\,ds \quad \text{and} \quad \int_0^t \frac{f(s \mid x)-b_1}{1-F(s \mid x)+b_2}\,ds,$$

respectively, where

$$b_1=(L_1^2+L_2)\sqrt{d}(1-\alpha/\zeta^2)^{\frac{c_3(1-c_4)(1-c_2)m}{d}}, \text{ and } b_2=L_1\sqrt{d}(1-\alpha/\zeta^2)^{\frac{c_3(1-c_4)(1-c_2)m}{d}}.$$

Hence, $|\Lambda^*_{\mathcal{A},n}(t)-\Lambda(t \mid x)|$ has the bound

$$\int_0^t \frac{b_1(1-F(s \mid x))+b_2 f(s \mid x)}{(1-F(s \mid x)-b_2)(1-F(s \mid x))}ds \leq M_2\tau\sqrt{d}(1-\alpha/\zeta^2)^{\frac{c_3(1-c_4)(1-c_2)m}{d}},$$

for any $t<\tau$, where M_2 is some constant depending on L_1 and L_2. Hence, for the terminal node $\mathcal{A}$ containing x, we bound the bias by

$$\sup_{t<\tau}|\Lambda^*_{\mathcal{A},n}(t)-\Lambda(t \mid x)| \leq M_2\tau\sqrt{d}(1-\alpha/\zeta^2)^{\frac{c_3(1-c_4)(1-c_2)m}{d}},$$

with probability larger than $1-d\exp\big\{-\frac{c_2^2 m}{2d}\big\}-d\exp\big\{-\frac{(1-c_2)c_3c_4^2 m}{2d}\big\}$. Combining this with the adaptive concentration bound result from Theorem

2, for each x, we further have

$$\sup_{t<\tau} |\widehat{\Lambda}_{\mathscr{A},n}(t \mid x) - \Lambda(t \mid x)| = O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + \big(\frac{k}{n}\big)^{\frac{c_1}{d}}\Big),$$

with probability larger than $1 - w_n$, where

$$w_n = \frac{2}{\sqrt{n}} + d\exp\Big\{-\frac{c_2^2\log_{1/\alpha}(n/k)}{2d}\Big\} + d\exp\Big\{-\frac{(1-c_2)c_3c_4^2\log_{1/\alpha}(n/k)}{2d}\Big\},$$

and $c_1 = \frac{c_3(1-c_2)(1-c_4)}{\log_{1-\alpha}(\alpha)}$. This completes the proof of point-wise consistency.

□

Proof of Theorem 4. From Theorem 3, we need to establish the bound of $|\widehat{\Lambda}_{\mathscr{A},n}(t \mid x) - \Lambda(t \mid x)|$ under an event with small probability w_n. Noticing that $\widehat{\Lambda}_{\mathscr{A},n}(t \mid x)$ is simply the Nelson-Aalen estimator of the CHF with at most k terms, for any $t < \tau$, we have

$$\widehat{\Lambda}_{\mathscr{A},n}(t \mid x) \leq \frac{1}{k} + \ldots + \frac{1}{1} = O(\log(k)),$$

which implies that

$$|\widehat{\Lambda}_{\mathscr{A},n}(t \mid x) - \Lambda(t \mid x)| \leq O(\log(k)).$$

Then we have

$$\begin{aligned}&\sup_{t<\tau} E_X\big|\widehat{\Lambda}_n(t \mid X) - \Lambda(t \mid X)\big| \\ &= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + \big(\frac{k}{n}\big)^{\frac{c_1}{d}} + \log(k)w_n\Big),\end{aligned}$$

which leads to the following bounds:

$$\begin{aligned}
&\sup_{t<\tau} E_X|\widehat{\Lambda}_{\{\mathscr{A}_{(b)}\}_1^B,n}(t \mid X) - \Lambda(t \mid X)| \\
&= \lim_{B\to\infty} \sup_{t<\tau} E_X|\frac{1}{B}\sum_{b=1}^{B}\widehat{\Lambda}_{\mathscr{A}_{(b)},n}(t \mid X) - \frac{1}{B}\sum_{b=1}^{B}\Lambda(t \mid X)| \\
&\le \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B}\sup_{t<\tau} E_X|\widehat{\Lambda}_{\mathscr{A}_{(b)},n}(t \mid X) - \Lambda(t \mid X)| \\
&= O\Big(\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + \big(\frac{k}{n}\big)^{\frac{c_1}{d}} + \log(k)w_n\Big). \ \square
\end{aligned}$$

S5

Lemma S5. *Under Assumption 1 and assume that the density function of the failure time $f_i(t) = dF_i(t)$ is bounded above by L for each i. The difference between $\Lambda^*_{\mathcal{A},n}(t)$ and $\Lambda^*_{\mathcal{A}}(t)$ is bounded by*

$$\sup_{t<\tau}\big|\Lambda^*_{\mathcal{A},n}(t) - \Lambda^*_{\mathcal{A}}(t)\big| \le \sqrt{\frac{4\tau^2L^2\log(4\sqrt{n})}{M^2n}},$$

with probability larger than $1 - 1/\sqrt{n}$.

Proof. By Hoeffding's inequality, we have for each $s \le t$,

$$\mathrm{pr}\Big(\Big|\frac{1}{n}\sum_{X_i\in\mathcal{A}}[1-G_i(s)]f_i(s) - E_X\{[1-G(s \mid X)]f(s \mid X)\}\Big| \ge \sqrt{\frac{L^2\log(4\sqrt{n})}{2n}}\ \Big) \le \frac{1}{2\sqrt{n}},$$

and

$$\text{pr}\Big(\Big|\frac{1}{n}\sum_{X_i\in\mathcal{A}}[1-G_i(s)][1-F_i(s\mid X)] - E_X\{[1-G(s\mid X)][1-F(s\mid X)]\}\Big| \\ \geq \sqrt{\frac{\log(4\sqrt{n})}{2n}}\ \Big) \leq \frac{1}{2\sqrt{n}}.$$

After combining the above two inequalities, we have

$$\sup_{t<\tau}\left|\Lambda^*_{\mathcal{A},n}(t) - \Lambda^*_{\mathcal{A}}(t)\right| \leq \sqrt{\frac{4\tau^2L^2\log(4\sqrt{n})}{M^2n}},$$

with probability larger than $1-1/\sqrt{n}$. □

Proof of Lemma 2. In a similar way as done for Lemma S5, for each $s\leq t$,

$$\text{pr}\Big(\Big|\frac{1}{n}\sum_{X_i\in\mathcal{A}}[1-G_i(s)]f_i(s) - E_X\{[1-G(s\mid X)]f(s\mid X)\}\Big| \\ \geq \sqrt{\frac{L^2\log(4\sqrt{n}|\mathcal{R}|)}{2n}}\Big) \leq \frac{1}{2\sqrt{n}},$$

and

$$\text{pr}\Big(\Big|\frac{1}{n}\sum_{X_i\in\mathcal{A}}[1-G_i(s)][1-F_i(s\mid X)] - E_X\{[1-G(s\mid X)][1-F(s\mid X)]\}\Big| \\ \geq \sqrt{\frac{\log(4\sqrt{n}|\mathcal{R}|)}{2n}}\Big) \leq \frac{1}{2\sqrt{n}}.$$

Thus, with probability larger than $1/\sqrt{n}$,

$$\begin{aligned}|\Lambda^*_{\mathcal{A},n}(t) - \Lambda^*_{\mathcal{A}}(t)| &\leq \sqrt{\frac{4\tau^2L^2\log(4\sqrt{n}|\mathcal{R}|)}{M^2n}} \\ &\leq M_2\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}},\end{aligned}$$

for all $t<\tau$ and all $\mathcal{A}\in\mathscr{A}, \mathscr{A}\in\mathcal{V}_{\alpha,k}(\mathcal{D}_n)$, where M_2 is some universal constant depending on L and M. □

Lemma S6. *Under the marginal screening splitting rule given in Algorithm 2, uniformly across all internal nodes, we essentially only split on* $(d_0 + d_1)$ *dimensions with probability larger than* $1 - 3/\sqrt{n}$ *on the entire tree.*

Proof of Lemma S6. We prove the lemma in two parts. In the first part, we show that the event that a survival tree ever splits on a noise variable has with probability smaller than $3/\sqrt{n}$. In the second part, we prove that if a failure variable is randomly selected and has never been used in the upper level of the tree, then the probability that the proposed survival tree splits on this variable is at least $1 - 3/\sqrt{n}$.

We start with defining $\Delta^*(c) = \max_{t<\tau} \left|\Lambda^*_{\mathcal{A}_j^+(c)}(t) - \Lambda^*_{\mathcal{A}_j^-(c)}(t)\right|$. Then for any noise variable j,

$$\begin{aligned}
\Delta^*(c) &= \max_{t<\tau} \left|\Lambda^*_{\mathcal{A}_j^+(c)}(t) - \Lambda^*_{\mathcal{A}_j^-(c)}(t)\right| \\
&= \max_{t<\tau} \left| \int_0^t \frac{E_{X\in\mathcal{A}_j^+(c)}[1-G(s \mid X)]\mathrm{d}F(s \mid X)}{E_{X\in\mathcal{A}_j^+(c)}[1-G(s \mid X)][1-F(s \mid X)]} \right. \\
&\qquad\qquad \left. - \int_0^t \frac{E_{X\in\mathcal{A}_j^-(c)}[1-G(s \mid X)]\mathrm{d}F(s \mid X)}{E_{X\in\mathcal{A}_j^-(c)}[1-G(s \mid X)][1-F(s \mid X)]} \right| \\
&= \max_{t<\tau} \left| \int_0^t \frac{\int_{x^{(j)}\geq c}\int_{\mathcal{A}^{d-1}}[1-G(s \mid x^{(-j)})]\mathrm{d}F(s \mid x^{(-j)})p(x^{(-j)}|x^{(j)})p(x^{(j)})dx^{(-j)}dx^{(j)}}{\int_{x^{(j)}\geq c}\int_{\mathcal{A}^{d-1}}[1-G(s \mid x^{(-j)})][1-F(s \mid x^{(-j)})]p(x^{(-j)}|x^{(j)})p(x^{(j)})dx^{(-j)}dx^{(j)}} \right. \\
&\quad \left. - \int_0^t \frac{\int_{x^{(j)}< c}\int_{\mathcal{A}^{d-1}}[1-G(s \mid x^{(-j)})]\mathrm{d}F(s \mid x^{(-j)})p(x^{(-j)}|x^{(j)})p(x^{(j)})dx^{(-j)}dx^{(j)}}{\int_{x^{(j)}< c}\int_{\mathcal{A}^{d-1}}[1-G(s \mid x^{(-j)})][1-F(s \mid x^{(-j)})]p(x^{(-j)}|x^{(j)})p(x^{(j)})dx^{(-j)}dx^{(j)}} \right| \\
&= \max_{t<\tau} |(I) - (II)|,
\end{aligned}$$

where $\mathcal{A}^{d-1}$ refers to integrating over d dimensions except variable j on the internal node $\mathcal{A}$ and $x^{(-j)}$ refers to $d-1$ dimensions of x except the coordinate j. Without loss of generality, we assume that $(I) > (II)$ when the maximum is achieved. By Assumption 5,

$$(I) < \int_0^t \frac{\int_{x^{(j)}\geq c}\int_{\mathcal{A}^{d-1}}[1-G(s \mid x^{(-j)})]\mathrm{d}F(s \mid x^{(-j)})p_{\mathcal{A}}(x^{(-j)})\gamma p(x^{(j)})dx^{(-j)}dx^{(j)}}{\int_{x^{(j)}\geq c}\int_{\mathcal{A}^{d-1}}[1-G(s \mid x^{(-j)})][1-F(s \mid x^{(-j)})]p_{\mathcal{A}}(x^{(-j)})\gamma^{-1} p(x^{(j)})dx^{(-j)}dx^{(j)}},$$

$$(II) > \int_0^t \frac{\int_{x^{(j)}<c} \int_{\mathcal{A}^{d-1}} [1 - G(s \mid x^{(-j)})] \mathrm{d}F(s \mid x^{(-j)}) p_{\mathcal{A}}(x^{(-j)}) \gamma^{-1} p(x^{(j)}) dx^{(-j)} dx^{(j)}}{\int_{x^{(j)}<c} \int_{\mathcal{A}^{d-1}} [1 - G(s \mid x^{(-j)})][1 - F(s \mid x^{(-j)})] p_{\mathcal{A}}(x^{(-j)}) \gamma p(x^{(j)}) dx^{(-j)} dx^{(j)}},$$

where $p_{\mathcal{A}}(x^{(-j)})$ refers to the marginal distribution of $x^{(-j)}$ on the internal node $\mathcal{A}$. So $\Delta^*(c)$ is further bounded by

$$\begin{aligned}
&\int_0^t \frac{\gamma \int_{x^{(j)} \geq c} p(x^{(j)}) dx^{(j)} \int_{\mathcal{A}^{d-1}} [1 - G(s \mid x^{(-j)})] \mathrm{d}F(s \mid x^{(-j)}) p_{\mathcal{A}}(x^{(-j)}) dx^{(-j)}}{\gamma^{-1} \int_{x^{(j)} \geq c} p(x^{(j)}) dx^{(j)} \int_{\mathcal{A}^{d-1}} [1 - G(s \mid x^{(-j)})][1 - F(s \mid x^{(-j)})] p_{\mathcal{A}}(x^{(-j)}) dx^{(-j)}} \\
&- \int_0^t \frac{\gamma^{-1} \int_{x^{(j)} < c} p(x^{(j)}) dx^{(j)} \int_{\mathcal{A}^{d-1}} [1 - G(s \mid x^{(-j)})] \mathrm{d}F(s \mid x^{(-j)}) p_{\mathcal{A}}(x^{(-j)}) dx^{(-j)}}{\gamma \int_{x^{(j)} < c} p(x^{(j)}) dx^{(j)} \int_{\mathcal{A}^{d-1}} [1 - G(s \mid x^{(-j)})][1 - F(s \mid x^{(-j)})] p_{\mathcal{A}}(x^{(-j)}) dx^{(-j)}} \\
\leq &(\gamma^2 - \gamma^{-2}) \Lambda^*_{\mathcal{A}}(\tau) \leq (\gamma^2 - \gamma^{-2}) \frac{\tau L}{M^2}.
\end{aligned}$$

From the adaptive concentration bound result and Lemma 2, we have, for an arbitrary $x \in [0, 1]^d$ and a valid partition $\mathscr{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$,

$$\max_{t<\tau} \left| \widehat{\Lambda}_{\mathscr{A},n}(t \mid x) - \Lambda^*_{\mathscr{A}}(t \mid x) \right| \leq M_3 \sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}},$$

with probability larger than $1 - 3/\sqrt{n}$, where $M_3 = \max(M_1, M_2)$. Hence

$$\Delta_1(c) \leq (\gamma^2 - \gamma^{-2}) \frac{\tau L}{M^2} + M_3 \sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}},$$

with probability larger than $1 - 3/\sqrt{n}$ uniformly over all possible nodes with at least $2k$ observations and all noise variables. Thus only with probability less than $3/\sqrt{n}$ will the proposed survival tree split on a noise variable.

To prove the second argument, suppose $\mathcal{A}$ is the current node and $X^{(j)}$ is an important variable. Since we choose the splitting point $\tilde{c}$ which maximizes $\Delta_1(c)$, the empirical signal is larger than that of c_0. Without loss of generality, we consider a cutoff point of c_0 and $\ell^+(j, t_0, c_0) \geq \ell^-(j, t_0, c_0)$.

Hence we are interested in

$$\begin{aligned}\Delta^*(c_0) =& \max_{t<\tau} \left|\Lambda^*_{\mathcal{A}_j^+(c_0)}(t) - \Lambda^*_{\mathcal{A}_j^-(c_0)}(t)\right| \\ =& \max_{t<\tau} \left| \int_0^t \frac{E_{\mathcal{A}_j^+(c_0)}[1 - G(s \mid X)]\mathrm{d}F(s \mid X)}{E_{\mathcal{A}_j^+(c_0)}[1 - G(s \mid X)][1 - F(s \mid X)]} \right. \\ & \left. - \int_0^t \frac{E_{\mathcal{A}_j^-(c_0)}[1 - G(s \mid X)]\mathrm{d}F(s \mid X)}{E_{\mathcal{A}_j^-(c_0)}[1 - G(s \mid X)][1 - F(s \mid X)]} \right|.\end{aligned}$$

Since $1-G(\tau)$ is bounded away from 0 by our assumption with $1-G(\tau) \geq M$, the above expression can be further bounded below by

$$\begin{aligned}& \Delta^*(c_0) \\ \geq & M \int_0^{t_0} \frac{E_{\mathcal{A}_j^+(c_0)}\mathrm{d}F(s \mid X)}{E_{\mathcal{A}_j^+(c_0)}[1 - F(s \mid X)]} - M^{-1} \int_0^{t_0} \frac{E_{\mathcal{A}_j^-(c_0)}\mathrm{d}F(s \mid X)}{E_{\mathcal{A}_j^-(c_0)}[1 - F(s \mid X)]} \\ = & M\ell^+(j, t_0, c_0) - M^{-1}\ell^-(j, t_0, c_0) > \ell.\end{aligned}$$

Then, by the adaptive concentration bound results above, $\Delta^*(c_0)$ has to be close enough to $\Delta_1(c_0)$. Thus we have

$$\Delta_1(\tilde{c}) \geq \Delta_1(c_0) > \ell - M_3\sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k\log((1-\alpha)^{-1})}},$$

with probability larger than $1-3/\sqrt{n}$ uniformly over all possible nodes and all signal variables. □

Proof of Theorem 5. The results follow by Lemma S6 and Theorem 3.

Proof of Theorem 6.

The proof essentially requires that uniformly across all internal nodes, we essentially only split on d_0 dimensions with probability larger than $1 - 3/\sqrt{n}$ on the entire tree. We first define $\widetilde{\Lambda}^*_{\mathcal{A}}(t)$ on a node $\mathcal{A}$,

$$\widetilde{\Lambda}^*_{\mathcal{A}}(t) = \int_0^t \frac{E_{X\in\mathcal{A}}[1 - G(s \mid X)]\mathrm{d}F(s \mid X)/[1 - \widehat{G}(s \mid X)]}{E_{X\in\mathcal{A}}[1 - G(s \mid X)][1 - F(s \mid X)]/[1 - \widehat{G}(s \mid X)]}.$$

For any noise variables, $\Delta^*(c) \leq (\gamma^2 - \gamma^{-2})\tau L/M^2$ follows the proof of Lemma S6. For any censoring but not failure variable j, we have that

$$\begin{aligned}
\Delta^*(c) = & \max_{t<\tau} |\widetilde{\Lambda}^*_{\mathcal{A}_j^+(c)}(t) - \widetilde{\Lambda}^*_{\mathcal{A}_j^-(c)}(t)| \\
= & \max_{t<\tau} \Bigg| \int_0^t \frac{E_{\mathcal{A}_j^+(c)}[1 - G(s \mid X)]/[1 - \widehat{G}(s \mid X)]\mathrm{d}F(s \mid X)}{E_{\mathcal{A}_j^+(c)}[1 - G(s \mid X)]/[1 - \widehat{G}(s \mid X)][1 - F(s \mid X)]} \\
& \qquad - \int_0^t \frac{E_{\mathcal{A}_j^-(c)}[1 - G(s \mid X)]/[1 - \widehat{G}(s \mid X)]\mathrm{d}F(s \mid X)}{E_{\mathcal{A}_j^-(c)}[1 - G(s \mid X)]/[1 - \widehat{G}(s \mid X)][1 - F(s \mid X)]} \Bigg| \\
= & \max_{t<\tau} |(I) - (II)|.
\end{aligned}$$

Without loss of generality, we assume that $(I) > (II)$ when the maximum is achieved. Since $\lim_{n\to\infty} \mathrm{pr}(\sup_{t<\tau} |\widehat{G}(t|X = x) - G(t|X = x)| > \epsilon) = 0$ for any $0 < \epsilon < M$ and x, the following inequalities

$$\frac{1 - G(t|X = x)}{1 - \widehat{G}(t|X = x)} > \frac{1 - G(t|X = x)}{1 - \widehat{G}(t|X = x) + \epsilon} > \frac{M}{M + \epsilon},$$

$$\frac{1 - G(t|X = x)}{1 - \widehat{G}(t|X = x)} < \frac{1 - G(t|X = x)}{1 - \widehat{G}(t|X = x) - \epsilon} < \frac{M}{M - \epsilon},$$

hold. Then $\Delta^*(c)$ is further bounded by

$$\begin{aligned}&\frac{M+\epsilon}{M-\epsilon}\int_0^t \frac{E_{\mathcal{A}_j^+(c)}\mathrm{d}F(s\mid x)}{E_{\mathcal{A}_j^+(c)}[1-F(s\mid x)]}-\frac{M-\epsilon}{M+\epsilon}\int_0^t \frac{E_{\mathcal{A}_j^-(c)}\mathrm{d}F(s\mid x)}{E_{\mathcal{A}_j^-(c)}[1-F(s\mid x)]}\\ \leq&[\frac{M+\epsilon}{M-\epsilon}\gamma^2-\frac{M-\epsilon}{M+\epsilon}\gamma^{-2}]\Lambda_{\mathcal{A}}(\tau)\leq[\frac{M+\epsilon}{M-\epsilon}\gamma^2-\frac{M-\epsilon}{M+\epsilon}\gamma^{-2}]\frac{\tau L}{M},\end{aligned}$$

where the first inequality holds from Assumption 8.

For any failure variable j,

$$\begin{aligned}\Delta^*(c_0)=&\max_{t<\tau}\left|\widetilde{\Lambda}^*_{\mathcal{A}_j^+(c_0)}(t)-\widetilde{\Lambda}^*_{\mathcal{A}_j^-(c_0)}(t)\right|\\ \geq&\Bigg|\int_0^{t_0}\frac{E_{\mathcal{A}_j^+(c_0)}[1-G(s\mid X)]/[1-\widehat{G}(s\mid X)]\mathrm{d}F(s\mid X)}{E_{\mathcal{A}_j^+(c_0)}[1-G(s\mid X)]/[1-\widehat{G}(s\mid X)][1-F(s\mid X)]}\\ &-\int_0^{t_0}\frac{E_{\mathcal{A}_j^-(c_0)}[1-G(s\mid X)]/[1-\widehat{G}(s\mid X)]\mathrm{d}F(s\mid X)}{E_{\mathcal{A}_j^-(c_0)}[1-G(s\mid X)]/[1-\widehat{G}(s\mid X)][1-F(s\mid X)]}\Bigg|.\end{aligned}$$

Without loss of generality, we assume that $\ell^+(j,t_0,c_0)>\ell^-(j,t_0,c_0)$. We have that

$$\begin{aligned}&\Delta^*(c_0)\\ \geq&\int_0^{t_0}\frac{E_{\mathcal{A}_j^+(c_0)}[1-G(s\mid X)]/[1-G(s\mid X)+\epsilon]\mathrm{d}F(s\mid X)}{E_{\mathcal{A}_j^+(c_0)}[1-G(s\mid X)]/[1-G(s\mid X)-\epsilon][1-F(s\mid X)]}\\ &-\int_0^{t_0}\frac{E_{\mathcal{A}_j^-(c_0)}[1-G(s\mid X)]/[1-G(s\mid X)-\epsilon]\mathrm{d}F(s\mid X)}{E_{\mathcal{A}_j^-(c_0)}[1-G(s\mid X)]/[1-G(s\mid X)+\epsilon][1-F(s\mid X)]}\\ \geq&\frac{M-\epsilon}{M+\epsilon}\int_0^{t_0}\frac{E_{\mathcal{A}_j^+(c_0)}\mathrm{d}F(s\mid X)}{E_{\mathcal{A}_j^+(c_0)}[1-F(s\mid X)]}-\frac{M+\epsilon}{M-\epsilon}\int_0^{t_0}\frac{E_{\mathcal{A}_j^-(c_0)}\mathrm{d}F(s\mid X)}{E_{\mathcal{A}_j^-(c_0)}[1-F(s\mid X)]}\\ =&\frac{M-\epsilon}{M+\epsilon}\ell^+(j,t_0,c_0)-\frac{M+\epsilon}{M-\epsilon}\ell^-(j,t_0,c_0),\end{aligned}$$

with probability going to 1. Combined with the adaptive concentration bound results,

$$\max_{t<\tau} \left| \widetilde{\Lambda}_{\mathscr{A},n}(t \mid x) - \widetilde{\Lambda}^*_{\mathscr{A}}(t \mid x) \right| \leq M_3 \sqrt{\frac{\log(n/k)[\log(dk) + \log\log(n)]}{k \log((1-\alpha)^{-1})}},$$

we have

$$\Delta_2(\tilde{c}) \geq \Delta_2(c_0) > \ell - o_p(1),$$

uniformly over all possible nodes and all signal variables. □

S7

To fully understand the impact of bias-correction, we consider a set of simulation studies. There are many existing implementations of random survival forests, including R packages `randomForestSRC` (Ishwaran and Kogalur, 2019), `party` (Hothorn et al., 2006), `ranger` (Wright and Ziegler, 2017), etc. However, it is difficult to compare across different packages as they may utilize certain tuning parameters slightly differently. It would not be possible to investigate the sole impact of bias-correction if these subtle differences are involved. Hence, we turn to make our own implementation of survival forest modeling with and without the bias-correction, while ensuring all other mechanisms remain the same.

Furthermore, we note that there are two possible ways to make a bias-correction based on our previous analysis. First, and most apparently, we can incorporate $\widetilde{\Lambda}_{\mathcal{A},n}$ in $\Delta_2(c)$ to search for a better splitting rule. Alternatively, we may apply a regular splitting rule and use $\widetilde{\Lambda}_{\mathcal{A},n}$ only in the terminal node estimation to correct the bias. Based on our analysis, the second approach would not improve the convergence rate because the tree structure is already built on $\mathscr{M}_{FC}$ variables, while the first approach has the potential for improvement. Hence, contrasting these two approaches would allow us to investigate the importance of changing the entire tree structure through bias-correction. We consider four different algorithms out of the combination of these two choices: 1) (C-C) bias-corrected splitting rule and terminal node estimation; 2) (C-N) bias-corrected splitting rule without correcting the terminal node estimation; 3) (N-C) correcting only the terminal node estimation; and 4) (N-N) do not perform any bias-correction. Note again that we implement all four methods under the same algorithm framework that assures all other tuning parameters remain the same.

We consider two data generating scenarios, each with dependent censoring and independent censoring mechanisms. For the first scenario, we consider the setting in Section 3.2. Let $d = 3$ and $X^{(1)}$, $X^{(2)}$ and $X^{(3)}$ from a multivariate normal distribution with mean 0 and variance Σ, where the diagonal elements of Σ are all 1, and the only nonzero off diagonal element is $\Sigma_{12} = \Sigma_{21} = \rho = 0.8$. T is exponential distribution with mean $\exp(-1.25X^{(1)} - X^{(3)} + 2)$. For dependent censoring, the censoring time follows an exponential distribution with mean $\exp(-3X^{(2)})$; For independent censoring, the censoring time follows an exponential distribution with mean 2. For the second scenario, we consider a setting where the covariates are independent. we let $d = 10$ and draw X from a multivariate normal distribution with mean 0. Survival times are drawn independently from an accelerated failure time model, $\log(T) = X^{(1)} + X^{(2)} + X^{(3)} + \epsilon_1$, where ϵ_1 is generated from a standard normal distribution. For dependent censoring, the censoring time shares the variable $X^{(3)}$ with the failure time T, and follows $\log(C) = -1 + 2X^{(3)} + X^{(4)} + X^{(5)} + \epsilon_2$; For marginal independent censoring, the censoring time follows $\log(C) = -1 + X^{(4)} + X^{(5)} + 2X^{(6)} + \epsilon_3$, where ϵ_2 and ϵ_3 are generated from a standard normal distribution which is independent of ϵ_1.

For each setting, we used a training dataset with sample size $n = 400$. A testing dataset with size 800 was used to evaluate the mean squared error of the estimated conditional survival functions (Zhu and Kosorok, 2012). The censoring distributions were estimated from standard survival forests. Each simulation was repeated 500 times. Note that since all methods were implemented under the same code, we fixed the tuning parameters and only investigated the influence of bias-correction. Tuning parameters in the survival trees were chosen as follows. According to Ishwaran et al. (2008), the number of covariates considered at each splitting was set to $\lceil \sqrt{d} \, \rceil$. The minimal number of observed failures in each terminal node was set to 10 and 25, respectively. The total number of trees was set to be 100.

The simulation results are summarized in Table 1. In both scenarios, we clearly see that the bias-corrected splitting rule (C-C and C-N columns) has significantly improved the performance compared with N-C and N-N. This shows that by selecting a better variable to split, the tree structure can be corrected to reduce the prediction accuracy. For scenario 2, which has a more complicated censoring structure, the improvement of the proposed bias-corrected splitting rule is more significant than Scenario 1, with the average mean square error decreasing approximately from 47.7 to 42. Hence, the performance of bias-correction may also depend on the complexity of the censoring distribution and the accuracy of its estimation. The standard error is comparable among all four methods.

Interestingly, we want to highlight that the biasedness is mainly caused by the splitting bias rather than terminal node estimation, i.e., C-C is similar to C-N, and N-C is similar to N-N. This is intuitive and in line with our theory that the splitting bias-correction procedure can enjoy a potentially faster convergence rate than the non-bias-corrected version. One might not expect a good prediction if trees are partitioned inefficiently regardless of what kind of terminal node estimation is used. After the tree is constructed, there is not much room to correct the bias if previous splits were chosen on noise or censoring variables.

Table 1: Simulation results: Mean and (standard deviation) of mean squared error

	Censoring	C-C	C-N	N-C	N-N
Scenario 1	Dependent	21.62 (7.46)	21.68 (7.50)	23.32 (7.18)	23.29 (7.17)
	Independent	8.42 (2.01)	8.41 (1.98)	8.42 (2.02)	8.43 (2.04)
Scenario 2	Dependent	42.02 (5.37)	41.97 (5.34)	47.76 (5.76)	47.75 (5.77)
	Independent	35.18 (3.80)	35.22 (3.77)	36.11 (3.55)	36.14 (3.59)

C-C/C-N/N-C/N-N refer to the configurations of correcting or not correcting the bias, while the first letter refers to the splitting rule correction, and the second letter refers to the terminal node correction.

Bibliography

Cuzick, J. (1985). Asymptotic properties of censored linear rank tests. *The Annals of Statistics*, 133–141.

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

Ishwaran, H. and U. Kogalur (2019). *Random Forests for Survival, Regression, and Classification (RF-SRC).* R package version 2.8.0.

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

Pollard, D. (2012). *Convergence of stochastic processes.* Springer Science & Business Media.

Wager, S. and G. Walther (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

Wright, M. N. and A. Ziegler (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software 77*(1), 1–17.

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