Statistica Sinica Preprint No: SS-2020-0263	
Title	Consistency of survival tree and forest models: splitting bias and correction
Manuscript ID	SS-2020-0263
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0263
Complete List of Authors	Yifan Cui, Ruoqing Zhu, Mai Zhou and Michael Kosorok
Corresponding Author	Ruoqing Zhu
E-mail	rqzhu@illinois.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

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

Yifan Cui, Ruoqing Zhu, Mai Zhou, Michael Kosorok

University of Pennsylvania

University of Illinois at Urbana-Champaign

University of Kentucky

University of North Carolina at Chapel Hill

Abstract: Random survival forests and survival trees are popular models in statistics and machine learning. However, there is a general lack of understanding regarding the consistency, splitting rules, and influence of the censoring mechanism. In this study, we investigate the statistical properties of existing methods from several interesting perspectives. First, we show that traditional splitting rules with censored outcomes rely on a biased estimation of the within-node failure distribution. To exactly quantify this bias, we develop a concentration bound of the within-node estimation based on samples that are not independent and identically distributed, and apply it to the entire forest. Second, we analyze the entanglement between the failure and censoring distributions caused by univariate splits, and show that without correcting the bias at an internal node, survival tree and forest models can still enjoy consistency under suitable conditions. In particular, we demonstrate this property under two cases: a finite-dimensional

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

case, where the splitting variables and cutting points are chosen randomly, and a high-dimensional case, where the covariates are weakly correlated. Our results also apply to an independent covariate setting, which is commonly used in the random forest literature for high-dimensional sparse models. However, it may be unavoidable that the convergence rate depends on the total number of variables in the failure and censoring distributions. Third, we propose a new splitting rule that compares bias-corrected cumulative hazard functions at each internal node. We show that the rate of consistency of this new model depends only on the number of failure variables. We perform simulation studies to confirm that the proposed bias-correction can substantially benefit the prediction error.

Key words and phrases: Random Forests, Survival Analysis, Consistency, Adaptive Concentration, Bias Correction

1. Introduction

Random forests (Breiman, 2001) are among the most popular and powerful machine learning tools. The main advantage of tree-based models (Breiman et al., 1984) is their nonparametric nature. Although there has been a surge of research on understanding random forests, their theoretical properties have not been fully understood, even in regression settings. Lin and Jeon (2006) made one of the attempts to connect random forests to nearest neighbor predictors. Later on, a series of works, including Biau et al. (2008), Biau (2012), Genuer (2012), and Mentch and Hooker (2014),

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

established theoretical results on simplified tree-building processes or specific aspects of the model. More recently, Zhu et al. (2015) established consistency results based on an improved splitting rule criterion; Wager et al. (2014) and Athey et al. (2019) analyzed the confidence intervals induced from a random forest model; Linero (2016) established connections with Bayesian variable selection in a high-dimensional setting; Scornet et al. (2015) showed the consistency of the original random forest model on an additive model structure; and Wager and Walther (2015) studied the variance component of random forests and established corresponding concentration inequalities. For a comprehensive review of related topics, refer to Biau and Scornet (2016) and Athey et al. (2019).

In this study, we focus on the theoretical properties of a model in which the outcomes are right-censored (Fleming and Harrington, 2011). Censored survival data appear frequently in biomedical studies when the actual clinical outcome may not be directly observed, owing to early dropout or other reasons. Many random forest models have been developed for survival outcomes, including those of Hothorn et al. (2004), Hothorn et al. (2005), Ishwaran et al. (2008), Zhu and Kosorok (2012), Steingrimsson et al. (2016), Cui et al. (2020), and many others. However, there are few established theoretical results, despite the popularity of these methods in practice, es-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

pecially in genetic and clinical studies. For a general review of related topics, including single-tree-based survival models, refer to Bou-Hamad et al. (2011). To the best of our knowledge, the only consistency result to date is given by Ishwaran and Kogalur (2010), who considered the setting where all predictors are categorical. Some other results are established based on augmented outcomes that transform the problem to a fully observed regression model (Steingrimsson et al., 2017).

Our analysis provides insights into the consistency of survival forests and trees in general settings. In particular, we investigate whether existing methodologies enjoy consistency if the splitting rule is searched by comparing the survival distributions of the two potential child nodes. The answer is mixed, because a biased selection of the splitting rule may occur if there are marginal dependencies between the failure and censoring variables. We show that this drawback can be overcome in at least two general settings: a finite-dimensional case, if the splitting rule is data independent (Biau, 2012; Klusowski, 2018), and a high-dimensional case, if the marginal failure distribution signal is sufficiently large. However, the convergence rate may inevitably depend on the total dimension of the variables involved in both the failure and the censoring models. This phenomenon occurs even when the covariates are uniformly distributed, as long as the failure and censoring

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

times are not marginally independent. Such a result is surprising, given the results in the traditional parametric and semiparametric survival literature. However, it demonstrates the complexity of random forest models caused by the marginal splitting.

Motivated by the above results, we propose a new bias-correction procedure that actively selects the best splitting variable at each internal node, without the influence of the censoring distribution. This establishes a connection with existing methodology developments, such as Hothorn et al. (2005) and Steingrimsson et al. (2017), who convert censored observations to fully observed ones using the inverse probability of censoring weighting. However, our proposed splitting rule is much more general, in the sense that it compares the distributions of the failure times from the two potential child nodes, rather than focusing on the differences between the means. We further show that this new approach untangles the failure and censoring distributions, and improves the rate of convergence of tree and forest models so that the rate depends only on the number of important variables that define the failure distribution. Simulation studies are provided in the Supplementary Material that confirm that the proposed bias-correction can substantially benefit the prediction error.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

2. Survival tree and forest models

The essential element of tree-based survival models is recursive partitioning. A d-dimensional feature space $\mathcal{X}$ is partitioned into terminal nodes. For a single tree model, we denote the collection of these terminal nodes as $\mathscr{A} = \{\mathcal{A}_u\}_{u \in \mathcal{U}}$, where $\mathcal{U}$ is a set of indices, $\mathcal{X} = \bigcup_{u \in \mathcal{U}} \mathcal{A}_u$, and $\mathcal{A}_u \cap \mathcal{A}_l = \emptyset$ for any $u \neq l$. We also call $\mathscr{A}$ a partition of the feature space $\mathcal{X}$. In a traditional tree-building process (Breiman et al., 1984), binary splitting rules are used. Hence, all terminal nodes are hyper-rectangles.

Following the standard notation in the survival analysis literature, let $\mathcal{D}_n = \{X_i, Y_i, \delta_i\}_{i=1}^n$ be a set of n independent and identically distributed (i.i.d.) copies of the covariates, observed survival time, and censoring indicator, where the observed survival time $Y_i = \min(T_i, C_i)$, and $\delta_i = \mathbb{1}(T_i \leq C_i)$. We assume that each T_i follows a conditional distribution $F_i(t) = \text{pr}(T_i \leq t \mid X_i)$, where the survival function is denoted by $S_i(t) = 1 - F_i(t)$, the cumulative hazard function (CHF) $\Lambda_i(t) = -\log\{S_i(t)\}$, and the hazard function $\lambda_i(t) = \text{d}\Lambda_i(t)/\text{d}t$. The censoring time C_i follows the conditional distribution $G_i(t) = \text{pr}(C_i \leq t \mid X_i)$, where a noninformative censoring mechanism, $T_i \perp C_i \mid X_i$, is assumed.

In tree-based survival models, terminal node estimation is a crucial part. For any node $\mathcal{A}_u$, this can be obtained using the Kaplan–Meier

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

(KM) estimator (Kaplan and Meier, 1958) for the survival function or the Nelson–Aalen estimator (Nelson, 1969; Aalen, 1978) of the CHF based on the within-node samples. We focus on the following Nelson–Aalen estimator

$$\widehat{\Lambda}_{\mathcal{A}_u,n}(t) = \sum_{s\le t} \frac{\sum_{i=1}^{n} \mathbb{1}(\delta_i = 1)\mathbb{1}(Y_i = s)\mathbb{1}(X_i \in \mathcal{A}_u)}{\sum_{i=1}^{n} \mathbb{1}(Y_i \ge s)\mathbb{1}(X_i \in \mathcal{A}_u)}, \tag{2.1}$$

and the associated Nelson–Altshuler (NA) estimator (Altshuler, 1970) for the survival function when needed:

$$\widehat{S}_{\mathcal{A}_u,n}(t) = \exp\big\{ - \widehat{\Lambda}_{\mathcal{A}_u,n}(t)\big\}.$$

A survival tree model yields a collection of doublets $\{\mathcal{A}_u, \widehat{\Lambda}_{\mathcal{A}_u,n}\}_{u\in\mathcal{U}}$. In a survival forest model (Ishwaran et al., 2008; Zhu and Kosorok, 2012), a set of B trees are fitted. Hence, a collection of partitions $\{\{\mathcal{A}_u^b, \widehat{\Lambda}_{\mathcal{A}_u^b,n}\}_{u\in\mathcal{U}_b}\}_{b=1}^{B}$ is constructed. To facilitate better understanding, we provide a high-level outline (Algorithm **??** in the Supplementary Material **??**) of a survival forest model.

3. Biasedness of splitting rules

3.1 Within-node estimation

To begin our analysis, we start by investigating the KM and the NA estimators of the survival function. The main reason we revisit these classical methods is that they are popular for terminal node estimation in fitted

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

survival trees. The following lemma bounds their difference using an exact inequality, regardless of the underlying data distribution. The proof follows mostly from Cuzick (1985), and is given in the Supplementary Material.

Lemma 1. *Let $\widehat{S}_{KM}(t)$ and $\widehat{S}_{NA}(t)$ be the KM and the NA estimators, respectively, obtained using the same set of samples $\{Y_i, \delta_i\}_{i=1}^n$. Then, we have*

$$|\widehat{S}_{KM}(t) - \widehat{S}_{NA}(t)| < \widehat{S}_{KM}(t)\frac{4}{\sum_{i=1}^n \mathbb{1}(Y_i \geq t)},$$

for any observed failure time point t such that $\widehat{S}_{KM}(t) > 0$.

The above result suggests that calculating the difference between two KM curves is asymptotically the same as using the NA estimator, as long as we only calculate the curve up to a time point where the sample size is sufficiently large. For this purpose, Assumption 1 is always used throughout this paper. Note that similar assumptions are commonly used in the survival analysis literature, for example, $\text{pr}(T \geq \tau) > 0$ in Fleming and Harrington (2011), and $\text{pr}(C \geq \tau) > 0$ in Murphy et al. (1997), for some maximum study follow-up time τ. Then, with a large probability, $\widehat{S}_{NA}(t) = \widehat{S}_{KM}(t) + O(1/n)$ across all terminal nodes.

Assumption 1. *There exists a fixed maximum follow-up time $0 < \tau < \infty$*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

and a constant $M \in (0,1)$*, such that*

$$\text{pr}(Y \geq \tau \mid X = x) \geq M,$$

for all $x \in \mathcal{X}$*.*

3.2 A motivating example

Noticing that the splitting rule selection process is essentially comparing the survival curves computed from two potential child nodes, we take a closer look at this process. Most existing analyses of the KM estimator assume that the observations are i.i.d. (Breslow et al., 1974; Gill, 1980), or at least one set of the failure times or censoring times are i.i.d. (Zhou, 1991). However, this is almost always not true for tree-based methods at any internal node, because both T_i and C_i typically depend on the covariates. The question is whether this affects the selection of the splitting variable. We first use a simulation study to demonstrate this issue.

Consider the split at a particular node. We generate three random variables, $X^{(1)}$, $X^{(2)}$, and $X^{(3)}$, from a multivariate normal distribution with mean zero and covariance matrix Σ, where all diagonal elements of Σ are one, and the nonzero off-diagonal elements are $\Sigma_{12} = \Sigma_{21} = 0.8$. The failure distribution is exponential with mean $\exp(-1.25X^{(1)} - X^{(3)} + 2)$. We consider two censoring distributions for C: an exponential distribution with

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

mean two for all subjects, that is, independent of T; and an exponential distribution with mean equal to $\exp(-3X^{(2)})$. The splitting rule is searched for by maximizing the log-rank test statistic between the two potential child nodes $\{X^{(j)} \leq c, X \in \mathcal{A}\}$ and $\{X^{(j)} > c, X \in \mathcal{A}\}$, and the cutting point c is searched on the range of the variable. In an ideal situation, one would expect the best splitting rule to be constructed using $X^{(1)}$ with a large probability, because it carries the most signal. This is indeed the case shown in the first row of Table 1 for the i.i.d. censoring case, but not so much for the dependent censoring case. The simulation is done with $n = 1000$, and is repeated 1000 times. While this only demonstrates the splitting process on a single node, the consequence of this on the consistency of the entire tree is much more involved, because the entire tree structure can be altered by the censoring distribution. It is difficult to draw a definite conclusion at this point, but the impact of the censoring distribution is clearly demonstrated.

Table 1: Probability of selecting the splitting variable.

Censoring distribution	$X^{(1)}$	$X^{(2)}$	$X^{(3)}$
G_i identical	0.978	0.001	0.021
G_i depends on $X_i^{(2)}$	0.281	0.037	0.682

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

3.3 Survival estimation based on non-i.i.d. observations

It now seems impossible to analyze the consistency without exactly quantifying the within-node estimation performance. We look at two different quantities corresponding to the two scenarios used above. The first one is an averaged CHF within any node $\mathcal{A}$:

$$\Lambda_{\mathcal{A}}(t) = \frac{1}{\mu(\mathcal{A})} \int_{x \in \mathcal{A}} \Lambda(t \mid x) \mathrm{dP}(x), \tag{3.1}$$

where P is the distribution of X, and $\mu(\mathcal{A}) = \int_{x \in \mathcal{A}} \mathrm{dP}(x)$ is the measure of node $\mathcal{A}$. Clearly, in the first case, the censoring distribution is not covariate dependent. Thus, we are asymptotically comparing $\Lambda_{\mathcal{A}}(t)$ on the two child nodes, which results in the selection of the first variable. This should also be considered a rational choice because $X^{(1)}$ contains more signal at the current node.

In the second scenario, that is, the dependent censoring case, the within-node estimator $\widehat{\Lambda}_{\mathcal{A},n}(t)$ does not converge to $\Lambda_{\mathcal{A}}(t)$ in general, which can be inferred from the following theorem. As the main result of this section, Theorem 1 is interesting in its own right for understanding tree-based survival models, because it establishes a bound on the survival estimation under independent but non-identically distributed samples, which is a more general result than Zhou (1991). It quantifies exactly the estimation performance

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

for each potential child node; hence, it is also crucial for understanding splitting rules in general.

Theorem 1. *Let $\widehat{\Lambda}_n(t)$ be the Nelson–Aalen estimator of the CHF from a set of n independent samples $\{Y_i, \delta_i\}_{i=1}^n$ subject to right censoring, where the failure and censoring distributions (not necessarily identical) are given by F_i and G_i, respectively. Under Assumption 1, we have for $\epsilon_1 \leq 2$ and $n > 4/(\epsilon_1^2 M^4)$,*

$$\mathrm{pr}\left(\sup_{t<\tau}\left|\widehat{\Lambda}_n(t) - \Lambda_n^*(t)\right| > \epsilon_1\right) < 16(n+2)\exp\left\{\frac{-nM^4\epsilon_1^2}{1152}\right\}, \tag{3.2}$$

where

$$\Lambda_n^*(t) = \int_0^t \frac{\sum[1-G_i(s)]\mathrm{d}F_i(s)}{\sum[1-G_i(s)][1-F_i(s)]}. \tag{3.3}$$

The proof is deferred to the Supplementary Material **??**. Based on Theorem 1, if we restrict ourselves to any node $\mathcal{A}$, the difference between the within-node estimator $\widehat{\Lambda}_{\mathcal{A},n}(t)$ and

$$\Lambda_{\mathcal{A},n}^*(t) = \int_0^t \frac{\sum_{X_i\in\mathcal{A}}[1-G_i(s)]\mathrm{d}F_i(s)}{\sum_{X_i\in\mathcal{A}}[1-G_i(s)][1-F_i(s)]} \tag{3.4}$$

is bounded above, where $\Lambda_{\mathcal{A},n}^*(t)$ is some version of the underlying true cumulative hazard contaminated by the censoring distribution. Noting that $\Lambda_{\mathcal{A},n}^*(t)$ also depends on the sampling points X_i, we further develop Lemma

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

?? in the Supplementary Material to verify that $\Lambda^*_{\mathcal{A},n}(t)$ and its expected version $\Lambda^*_{\mathcal{A}}(t)$ are sufficiently close, where

$$\Lambda^*_{\mathcal{A}}(t) = \int_0^t \frac{E_{X\in\mathcal{A}}[1 - G(s \mid X)]\mathrm{d}F(s \mid X)}{E_{X\in\mathcal{A}}[1 - G(s \mid X)][1 - F(s \mid X)]}. \tag{3.5}$$

It is easy to see that the difference between $\Lambda^*_{\mathcal{A},n}(t)$ and $\Lambda_{\mathcal{A}}(t)$ vanishes if the F_i are identical within a node $\mathcal{A}$ (a sufficient condition). Note that this is what we are hoping for eventually at a terminal node.

$$\begin{aligned}\Lambda^*_{\mathcal{A},n}(t) &= \int_0^t \frac{\sum_{X_i\in\mathcal{A}}[1 - G_i(s)]}{\sum_{X_i\in\mathcal{A}}[1 - G_i(s)]} \frac{\mathrm{d}F(s)}{1 - F(s)} \\ &\quad \text{(if } F_i \equiv F \text{ for all } X_i \in \mathcal{A}) \\ &= \int_0^t \frac{\mathrm{d}F(s)}{1 - F(s)} = \frac{1}{\mu(\mathcal{A})} \int_{x\in\mathcal{A}} \int_0^t \frac{\mathrm{d}F(s)}{1 - F(s)} \mathrm{dP}(x) = \Lambda_{\mathcal{A}}(t). \end{aligned} \tag{3.6}$$

As we demonstrated in the simulation study above, comparing $\widehat{\Lambda}_{\mathcal{A},n}(t)$ between two child nodes may lead to a systematically different selection of splitting variables than using $\Lambda_{\mathcal{A}}(t)$, which is not known a priori. The main cause of the differences between these two quantities is that the NA estimator treats each node as a homogeneous group, which is typically not true. Another simple interpretation is that although the conditional independence assumption $T \perp C \mid X$ is satisfied, we have instead $T \not\perp C \mid \mathbb{1}(X^{(j)} < x)$ at an internal node. This tangling between the censoring and failure distributions makes it a very challenging problem because, at each internal node, we may select only one variable to split.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

4. Consistency of survival tree and forest

It becomes apparent now that this bias plays an important role in the asymptotic properties of survival tree and forest models. However, an important question we may ask is whether this affects the consistency of existing methodologies. To answer this question, we provide several analyses of consistency in different settings. Given that difficulty arises when the splitting rule is highly data dependent, we first investigate the consistency under random splitting rules and finite d (Section 4.2) to help our understanding. An analog of this result for regression and classification settings was proposed by Breiman (2004), and further analyzed by Lin and Jeon (2006), Biau et al. (2008), Biau (2012), Arlot and Genuer (2014), and many others. However, it is significantly more difficult when the splitting rule is data dependent, especially when the number of dimensions d is diverging with n. Note that because $\widehat{\Lambda}_n(t)$ is a biased estimator of the within-node averaged CHF, any marginal comparison splitting rule may falsely select a variable involved in the censoring mechanism. Furthermore, any dependencies in a high-dimensional setting may lead us to falsely select noise variables because they marginally carry signals. Hence, we investigate a high-dimensional setting where the noise variables have weak dependencies with the important variables (to be defined later). Under suitable condi-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

tions and some necessary modifications to the tree-build process, we show that a survival tree or forest model can still achieve consistency. However, the rate may be affected by the censoring distribution. To establish these results, we use the variance-bias breakdown, that is, for any x,

$$\left|\widehat{\Lambda}_n(t \mid x) - \Lambda(t \mid x)\right| \leq \left|\widehat{\Lambda}_n(t \mid x) - \Lambda_n^*(t \mid x)\right| + \left|\Lambda_n^*(t \mid x) - \Lambda(t \mid x)\right|,$$

and start by analyzing the variance component of a survival tree estimator.

4.1 Adaptive concentration bounds of survival trees and forests

We focus on quantifying survival forest models from a new angle, namely, the adaptive concentration (Wager and Walther, 2015) of each terminal node estimator to the true within-node expectation. In the sense of the variance-bias breakdown, the goal of this section is to quantify a version of the variance component of a tree-based model estimator. To be precise, with large probability, our main results bound $\left|\widehat{\Lambda}_{\mathcal{A},n}(t) - \Lambda_{\mathcal{A},n}^*(t)\right|$ across all possible terminal nodes in a fitted tree or forest. The adaptiveness comes from the fact that the target of the concentration is the censoring-contaminated version $\Lambda_{\mathcal{A},n}^*(t)$, which is defined adaptively for each node $\mathcal{A}$ with the observed samples, rather than as a fixed universal value. The results in this section have many implications. Because this bound is essentially the variance part of the estimator, we can focus on the bias to show consistencies.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

Although this may still pose challenges in specific situations, our later examples of consistency provide a framework that is largely applicable to most existing methods.

We start with some additional definitions and notations. Following our previous assumptions on the underlying data-generating model, we observe a set of n i.i.d. samples $\mathcal{D}_n$. We view each tree as a partition of the feature space, denoted by $\mathscr{A} = \{\mathcal{A}_u\}_{u\in\mathcal{U}}$, where $\mathcal{A}_u$ are non-overlapping hyper-rectangular terminal nodes. We first define a valid survival tree and forest estimators of the CHF. Roughly speaking, with certain constraints, these are all the possible survival tree or forest estimators resulting from a set of observed data. A tree partition $\mathscr{A}$ is $\{\alpha, k\}$-valid (Wager and Walther, 2015) if it satisfies two conditions: 1) for each splitting, the child node contains at least a fraction $\alpha \in (0, 0.5)$ of the training samples in its parent node; and 2) each terminal node contains at least k training examples. We denote the set of all $\{\alpha, k\}$-valid tree partitions by $\mathcal{V}_{\alpha,k}(\mathcal{D})$. In addition, we define the collection $\{\mathscr{A}^{(b)}\}_{b=1}^{B}$ as a valid forest partition if each of its tree partitions is valid. We denote the set of all such valid forest partitions as $\mathcal{H}_{\alpha,k}(\mathcal{D})$. A valid survival tree or forest estimator is induced from the corresponding valid set.

Definition 1 (Valid survival tree and forest)**.** Given the observed data $\mathcal{D}_n$,

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

a valid survival tree estimator of the CHF is induced by a valid partition $\mathscr{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$ with $\mathscr{A} = \{\mathcal{A}_u\}_{u\in\mathcal{U}}$:

$$\widehat{\Lambda}_{\mathscr{A},n}(t \mid x) = \sum_{u\in\mathcal{U}} \mathbb{1}(x \in \mathcal{A}_u)\widehat{\Lambda}_{\mathcal{A}_u,n}(t), \tag{4.1}$$

where each $\widehat{\Lambda}_{\mathcal{A}_u,n}(t \mid x)$ is defined by Equation (2.1). Furthermore, a valid survival forest $\widehat{\Lambda}_{\{\mathscr{A}_{(b)}\}_1^B,n}$ is defined as the average of B valid survival trees induced by a collection of valid partitions $\{\mathscr{A}_{(b)}\}_1^B \in \mathcal{H}_{\alpha,k}(\mathcal{D}_n)$,

$$\widehat{\Lambda}_{\{\mathscr{A}_{(b)}\}_1^B,n}(t \mid x) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\Lambda}_{\mathscr{A}_{(b)},n}(t \mid x). \tag{4.2}$$

We also define a censoring-contaminated survival tree and forest, which are asymptotic versions of the corresponding within-node average estimators of the CHF. Note that by Theorem 1, these averages are censoring-contaminated versions $\Lambda^*_{\mathcal{A},n}(t)$, but not the true averages $\Lambda_{\mathcal{A}}(t)$.

Definition 2 (Censoring-contaminated survival tree and forest)**.** Given the observed data $\mathcal{D}_n$ and $\mathscr{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$, the corresponding CHF of the censoring-contaminated survival tree is defined as

$$\Lambda^*_{\mathscr{A},n}(t \mid x) = \sum_{u\in\mathcal{U}} \mathbb{1}(x \in \mathcal{A}_u)\Lambda^*_{\mathcal{A}_u,n}(t), \tag{4.3}$$

where each $\Lambda^*_{\mathcal{A}_u,n}(t)$ is defined by Equation (3.4). Furthermore, let $\{\mathscr{A}_{(b)}\}_1^B \in \mathcal{H}_{\alpha,k}(\mathcal{D}_n)$. Then, the censoring-contaminated survival forest is given by

$$\Lambda^*_{\{\mathscr{A}_{(b)}\}_1^B,n}(t \mid x) = \frac{1}{B}\sum_{b=1}^{B} \Lambda^*_{\mathscr{A}_{(b)},n}(t \mid x). \tag{4.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

Our adaptive concentration bound result considers the quantity

$$\widehat{\Lambda}_{\mathscr{A},n}(t \mid x) - \Lambda^*_{\mathscr{A},n}(t \mid x),$$

for all valid partitions $\mathscr{A} \in \mathcal{V}_{\alpha,k}(\mathcal{D}_n)$. We first specify several regularity assumptions. The first assumption is a bound on the dependence of the individual features. Note that in the literature, uniform distributions are often assumed (Biau et al., 2008; Biau, 2012) on the covariates, which implies independence. To allow dependency between covariates, we assume the following assumption, which is also considered in Wager and Walther (2015). Without loss of generality, we assume the covariates are distributed on $[0,1]^d$.

Assumption 2. *The covariates $X \in [0,1]^d$ are distributed according to a density function $p(\cdot)$ satisfying $1/\zeta \leq p(x) \leq \zeta$, for all x and some $\zeta \geq 1$.*

We also set a restriction on the tuning parameter k, the minimum terminal node size, which may grow with n and the dimension d.

Assumption 3. *Assume that k is bounded below such that*

$$\lim_{n\to\infty} \frac{\log(n)\max\{\log(d), \log\log(n)\}}{k} = 0. \tag{4.5}$$

Then, we have the adaptive bound for our tree estimator in the following theorem. The proof is collected in the Supplementary Material **??**.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

Theorem 2. *Suppose the training data $\mathcal{D}_n$ satisfy Assumptions 1 and 2, and the rate of the sequence (n, d, k) satisfies Assumption 3. Then, all valid trees concentrate on a censoring-contaminated tree:*

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

with probability larger than $1 - 2/\sqrt{n}$, for some universal constant M_1.

The above theorem holds for all single-tree partitions in $\mathcal{V}_{\alpha,k}(\mathcal{D}_n)$. Consequently, we have a similar result for the forest estimator in Corollary **??** in the Supplementary Material **??**.

Remark 1. In a moderately high-dimensional setting, that is, $d \sim n$, the rate is $\log(n)/k^{1/2}$. In an ultrahigh-dimensional setting, for example, $\log(d) \sim n^{\vartheta}$, where $0 < \vartheta < 1$, the rate is close to $n^{\vartheta}/k^{1/2}$. The rate that k grows with n cannot be too slow, in order to achieve the bound in the ultrahigh-dimensional setting. This is quite intuitive because if k grows too slowly, then we are not able to bound all possible nodes.

The results established in this section essentially address the variance component in a fitted random forest. We chose not to use the true within-node population-averaged quantity $\Lambda^*_{\mathcal{A}}(t)$ (see Equation 3.5) or its single-tree and forest versions as targets of the concentration. This is because

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

such a result would require a bounded density function of the failure time T. However, when $f(t)$ is bounded, the results can be easily generalized to $\left|\widehat{\Lambda}_{\mathcal{A},n}(t) - \Lambda^*_{\mathcal{A}}(t)\right|$. Lemma 2 in Section 4.3 provides an analog of Theorem 1 in this situation.

With the above concentration inequalities established, we are now in a position to discuss the consistency results. We consider two specific scenarios: a finite-dimensional case, where the splitting rule is generated randomly, and a high-dimensional case, using the marginal difference of Nelson–Aalen estimators as the splitting rule.

4.2 Consistency under random splitting rules when d is finite

Assume that the dimension d of the covariate space is fixed and finite. At each internal node, we choose the splitting variable randomly and uniformly from all covariates (Biau, 2012; Klusowski, 2018). Once the splitting variable has been chosen, we choose the splitting point uniformly at random, such that both child nodes contain at least α proportion of the samples in the parent node. We bound the bias term

$$\sup_{t<\tau} E_X \left|\Lambda^*_{\{\mathscr{A}_{(b)}\}_1^B,n}(t \mid X) - \Lambda(t \mid X)\right|.$$

Note that in Section 4.1, we did not treat the tree and forest structures ($\mathscr{A}$ and $\{\mathscr{A}_{(b)}\}_1^B$) as random variables. Instead, they were treated as elements

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

of valid structure sets. However, in this section, once a particular splitting rule is specified, these structures become random variables associated with certain distributions induced from the splitting rule. When there is no risk of ambiguity, we inherit the notation $\widehat{\Lambda}_{\mathscr{A},n}$ to represent a tree estimator, where the randomness of $\mathscr{A}$ is understood as part of the randomness in the estimator itself. A similar strategy is applied to the forest version. Before presenting the consistency results, we make an additional smoothness assumption on the hazard function.

Assumption 4. *For any fixed time point t, the CHF $\Lambda(t \mid x)$ is L_1-Lipschitz continuous in terms of x, and the hazard function $\lambda(t \mid x)$ is L_2-Lipschitz continuous in terms of x; that is, $|\Lambda(t \mid x_1) - \Lambda(t \mid x_2)| \leq L_1||x_1 - x_2||$ and $|\lambda(t \mid x_1) - \lambda(t \mid x_2)| \leq L_2\|x_1 - x_2\|$, respectively, where $\|\cdot\|$ is the Euclidean norm.*

We are now ready to state our main consistency results for the proposed survival tree model. Theorem 3 provides the pointwise consistency result. The proof is collected in the Supplementary Material **??**.

Theorem 3. *Under Assumptions 1–4, the proposed survival tree model with a random splitting rule is consistent; that is, for each $x \in [0,1]^d$,*

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

with probability at least $1 - w_n$, *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 $c_2, c_4 \in (0,1)$, $c_3 = (1-2\alpha)/8$, *and* $c_1 = c_3(1-c_2)(1-c_4)/\log_{1-\alpha}(\alpha)$.

Remark 2. The first part $\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}}$ in the bound comes from the concentration results, and the second part $(k/n)^{c_1/d}$ comes from the bias. Note that the optimal rate is obtained by setting $k = n^{c_3/[c_3+d\log_{1-\alpha}(\alpha)/2]}$. Then the optimal rate is close to $n^{-c_3/[2c_3+d\log_{1-\alpha}(\alpha)]}$. If we further assume that we always split at the middle point at each internal node, then the optimal rate degenerates to $n^{-1/(d+2)}$, obtained by setting $k \sim n^{d/(d+2)}$.

The consistency result can be extended easily to survival forests with B trees. Theorem 4 presents an integrated version, which can be derived from Theorem 3. The proof is collected in the Supplementary Material **??**.

Theorem 4. *Under Assumptions 1–4, the proposed survival forest is consistent; that is,*

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

where w_n *is a sequence approaching zero as defined in Theorem 3,* $c_2, c_4 \in (0,1)$, $c_3 = (1-2\alpha)/8$, *and* $c_1 = c_3(1-c_2)(1-c_4)/\log_{1-\alpha}(\alpha)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

4.3 Consistency under adaptive splitting rules

We have so far established consistency results for both survival trees and forests for finite dimension d. In this section, we allow the dimension d to go to infinity with the sample size n, while the covariates are possibly correlated. Note that this is not a common setting in the literature, because it can be difficult to control the marginal distribution of a noise variable. We first provide several definitions. We assume that there are d_0 important variables involved in the failure time distribution, and denote $\mathscr{M}_F$ as the set of their indices; hence, $\mathscr{M}_F \subset \{1, \dots d\}$. Specifically, we have $T \perp X | X_{\mathscr{M}_F}$. The number of features involved in either the failure time or the censoring time distributions is $d_0 + d_1$, where d_0 and d_1 are fixed, and we denote $\mathscr{M}_{FC}$ as their indices. However, this does not imply that there are only d_1 variables for the censoring distribution, because the failure variables and censoring variables may share some common indices. For example, age, a commonly used demographic variable, can be informative for both the clinical outcome of interest (failure) and the loss to follow-up (censoring). However, splitting such variables is necessary, as long as they are involved in the failure distribution. The splits to avoid are on the variables in the set $\mathscr{M}_C$, defined as $\mathscr{M}_{FC} \setminus \mathscr{M}_F$, making $d_1 = |\mathscr{M}_C|$. However, our later analysis indicates that this may not be avoidable, because of the biasedness

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

caused by censoring. Lastly, we define the set of noise variables' indices as $\mathscr{M}_N = \{1, \dots d\} \setminus \mathscr{M}_{FC}$. We make the following assumption on the weak dependencies between the noise variables and the variables in $\mathscr{M}_{FC}$.

Assumption 5. *We assume that the conditional distribution of the failure and censoring covariates, $X_{\mathscr{M}_{FC}}$, has a weak dependency on any univariate noise variable, in the sense that for some constant $\gamma > 1$, we have*

$$\gamma^{-1} < \frac{p(X_{\mathscr{M}_{FC}} = x | X^{(j)} = x_1)}{p(X_{\mathscr{M}_{FC}} = x | X^{(j)} = x_2)} < \gamma,$$

for any $x_1, x_2 \in [0, 1]$, $(d_0 + d_1)$-dimensional vector x, and any $j \in \mathscr{M}_N$.

Assumption 5 relaxes the commonly used independent covariate assumption in the literature (Biau, 2012; Zhu et al., 2015; Scornet et al., 2015; Wager and Athey, 2018). However, this poses significant difficulties when evaluating the marginal signal carried by a noise variable, meaning that the difference between the two potential child nodes may not be zero. A large threshold is necessary to prevent the noise variables from entering the model, as shown in the later results. However, as the correlation reduces to zero, that is, $\gamma = 1$, the threshold will naturally degenerate to zero. We further need an additional assumption on the effect size of the failure variable.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

Assumption 6. *(Marginal signal of the failure distribution) Let $\mathcal{A}$ be any internal node, and let $j \in \mathscr{M}_F$ be an index of a variable that has never been split; that is, the range of $X^{(j)}$ in node $\mathcal{A}$ is $[0,1]$. Let $\mathcal{A}_j^+$ and $\mathcal{A}_j^-$ be defined as $\mathcal{A}_j^+(c) = \{X : X^{(j)} \geq c\}$ and $\mathcal{A}_j^-(c) = \{X : X^{(j)} < c\}$, respectively. We further define*

$$\ell^+(j,t,c) = \int_0^t \frac{E_{X\in\mathcal{A}_j^+(c)} f(s \mid X)}{E_{X\in\mathcal{A}_j^+(c)}[1-F(s \mid X)]}\,\mathrm{d}s,$$
$$\ell^-(j,t,c) = \int_0^t \frac{E_{X\in\mathcal{A}_j^-(c)} f(s \mid X)}{E_{X\in\mathcal{A}_j^-(c)}[1-F(s \mid X)]}\,\mathrm{d}s.$$

Then, there exists a time point $t_0 \in [0,\tau]$, a constant $c_0 \in (0,1)$, and a minimum effect size $\ell > 2(\gamma^2-\gamma^{-2})\tau L/M^2$, such that

$$M\ell^+(j,t_0,c_0) - M^{-1}\ell^-(j,t_0,c_0) > \ell, \quad \text{if} \quad \ell^+(j,t_0,c_0) > \ell^-(j,t_0,c_0),$$
$$\text{or} \quad M\ell^-(j,t_0,c_0) - M^{-1}\ell^+(j,t_0,c_0) > \ell, \quad \text{if} \quad \ell^+(j,t_0,c_0) < \ell^-(j,t_0,c_0),$$

where τ and M are defined in Assumption 1, γ is defined in Assumption 5, and L is an upper bound of $f(t|x)$ for all $x \in [0,1]^d$.

This assumption can be interpreted as follows. First, $\ell^+(j,t,c)$ and $\ell^-(j,t,c)$ are the averaged CHFs on the left- and right-hand sides, respectively, of a split at the jth variable. The constant M and its reciprocal can be understood as the minimum and maximum contamination, respectively, of the censoring distribution on these CHFs. The assumption requires that

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

the difference between these contaminated versions be sufficiently large at some time point t_0 and some cutting point c_0. Note that M is a lower bound of $\text{pr}(C \geq \tau \mid X = x)$. This essentially bounds below the signal size, regardless of any dependency structures between C and T. However, in some trivial cases, such as when the G_i are identical, the constant M can be removed from the assumption owing to the independence between T and C. A simplified version is provided in Assumption 7 in Section 5. Furthermore, ℓ can be an arbitrarily small constant if $\gamma = 1$, which is essentially the independent covariates case.

Another important observation of this assumption is that t_0 can be arbitrary. Hence, we essentially allow the CHFs of different subjects to cross each other. As a comparison, note that in many popular survival models, such as the Cox proportional hazard model, the CHF is a monotone function of X on the entire time domain. Hence, the survival curve of any subject can only be completely above or below that of another subject. However, when the survival curves cross each other, a log-rank test may not be effective (Fleming and Harrington, 2011; Eng and Kosorok, 2005). When we incorporate the splitting rule that detects the maximum differences on $[0, \tau)$, our model is capable of detecting nonmonotone signals of the CHF as a function of both X and t, making our approach more powerful than

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

the traditional log-rank test splitting rule.

Finally, to make the splitting rule concrete, we provide Algorithm ?? in the Supplementary Material ??, which marginally compares the estimated CHF over all time points, and uses the difference to select the best split. Based on this algorithm, Lemma ?? in the Supplementary Material ?? shows that our d-dimensional survival forest is equivalent to a $(d_0 + d_1)$-dimensional survival forest with probability larger than $1 - 3/\sqrt{n}$. This means that with a large probability, we do not split on the noise variable set $\mathscr{M}_N$. Note that $\Lambda^*_{\mathcal{A},n}(t)$ is an essential tool to prove Lemma ??. The intuition here is that when the failure distribution does not depend on the variable j, the quantity

$$\int_0^t \frac{E_{X\in\mathcal{A}_j^+(x)}[1-G(s\mid X)]\mathrm{d}F(s\mid X)}{E_{X\in\mathcal{A}_j^+(x)}[1-G(s\mid X)][1-F(s\mid X)]} - \int_0^t \frac{E_{X\in\mathcal{A}_j^-(x)}[1-G(s\mid X)]\mathrm{d}F(s\mid X)}{E_{X\in\mathcal{A}_j^-(x)}[1-G(s\mid X)][1-F(s\mid X)]}$$

is bounded by a small constant under weak dependency. However, this quantity will degenerate to zero as long as the dependency vanishes, that is, $\gamma = 1$, because $dF(s|X)/[1-F(s|X)]$ separates regardless of the censoring distribution. The proof is indeed beautiful and neat, and is deferred to the Supplementary Material ??.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

Note that $\Lambda^*_{\mathcal{A},n}(t)$ is a sample version of the asymptotic distribution of the terminal node $\mathcal{A}$. In Lemma 2, we show the bound of the difference of $\Lambda^*_{\mathcal{A},n}(t)$ and its integrated version $\Lambda^*_{\mathcal{A}}(t)$ across all valid nodes $\mathcal{A}$, where $\Lambda^*_{\mathcal{A}}(t)$ is as defined in Equation (3.5). The proof is given in the Supplementary Material **??**.

Lemma 2. *Assume Assumptions 1–3 hold and that the conditional density function $f(t \mid x)$ of the failure time T is bounded by L for all $x \in [0,1]^d$. The difference between $\Lambda^*_{\mathscr{A},n}(t)$ and $\Lambda^*_{\mathscr{A}}(t)$ is bounded by*

$$\sup_{t<\tau,\, x\in[0,1]^d,\, \mathscr{A}\in\mathcal{V}_{\alpha,k}(\mathcal{D}_n)} \left|\Lambda^*_{\mathscr{A},n}(t \mid x) - \Lambda^*_{\mathscr{A}}(t \mid x)\right| \leq M_2\sqrt{\frac{\log(n/k)[\log(dk)+\log\log(n)]}{k\log((1-\alpha)^{-1})}},$$

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

Based on Lemma **??**, provided in the Supplementary Material **??**, we essentially only split on (d_0+d_1) dimensions with probability larger than $1-3/\sqrt{n}$ on the entire tree. The consistency holds from Theorem 3. The following result shows the consistency of the proposed survival forest. The proof is almost identical to that of Theorem 4.

Theorem 5. *Under Assumptions 1–6, the proposed survival tree using the*

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

splitting rule specified in Algorithm **??** *is consistent; that is, for any* x*,*

$$\sup_{t<\tau} |\widehat{\Lambda}_{\mathscr{A},n}(t \mid x) - \Lambda(t \mid x)| = O\Bigg(\sqrt{\frac{\log(n/k)[\log\{(d_0+d_1)k\} + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + \Big(\frac{k}{n}\Big)^{\frac{c_1}{d_0+d_1}}\Bigg),$$

with probability at least $1 - w_n$*, where*

$$w_n = \frac{3}{\sqrt{n}} + (d_0+d_1)\Big[\exp\Big\{-\frac{c_2^2\log_{1/\alpha}(n/k)}{2(d_0+d_1)}\Big\} + \exp\Big\{-\frac{(1-c_2)c_3c_4^2\log_{1/\alpha}(n/k)}{2(d_0+d_1)}\Big\}\Big],$$

$c_2, c_4 \in (0,1)$*,* $c_3 = (1-2\alpha)/8$*, and* $c_1 = c_3(1-c_2)(1-c_4)/\log_{1-\alpha}(\alpha)$*.*

Consequently, the proposed survival forest is consistent; that is,

$$\lim_{B\to\infty} \sup_{t<\tau} E_X|\widehat{\Lambda}_{\{\mathscr{A}_{(b)}\}_1^B,n}(t \mid X) - \Lambda(t \mid X)| = O\Bigg(\sqrt{\frac{\log(n/k)[\log\{(d_0+d_1)k\} + \log\log(n)]}{k\log((1-\alpha)^{-1})}} + \Big(\frac{k}{n}\Big)^{\frac{c_1}{d_0+d_1}} + \log(k)w_n\Bigg).$$

Although we have developed a result where d can grow exponentially fast with n, the splitting rule implemented was not the same as in practice, because it essentially checks only the signal where the candidate variables have never been used. This is made possible by Assumption 6, which states that the difference between two potential child nodes resulting from $X^{(j)} < c_0$ and $X^{(j)} \geq c_0$ is sufficiently large. Once a variable is used, it is automatically included as a candidate in subsequent splits. The idea is similar to the protected variable set used in Zhu et al. (2015), where the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

protected set serves as the collection of variables that have been used in previous nodes.

Remark 3. We provide a consistency result for survival trees and forests under a weak dependency framework. We highlight that our results hold naturally if X is uniformly distributed with $\zeta = \gamma = 1$. Note that when the variables are uncorrelated, our results are still meaningful and constructive: the biasedness is not due to correlated variables, but to the entanglement between the failure time and the censoring time marginally. For example, if a censoring time shares one common variable with a failure time, then all other censoring variables also play a role in the limiting distribution, so there is no guarantee of splitting only on failure variables.

5. A bias-corrected survival forest

Recall that in Section 3, we investigated the estimation bias in a comparison of two potential child nodes. This is caused by ignoring the within-node heterogeneity of the censoring distribution while it is entangled with the failure distribution. A closer look at Equation (3.4) motivates us to use a weighted version of the Nelson–Aalen estimator to correct this bias and

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

estimate the true within-node averaged CHF. Hence, we consider

$$\widetilde{\Lambda}_{\mathcal{A},n}(t) = \sum_{s \le t} \frac{\sum_{i=1}^{n} \mathbb{1}(\delta_i = 1)\mathbb{1}(Y_i = s)\mathbb{1}(X_i \in \mathcal{A})/[1 - \widehat{G}(s|X_i)]}{\sum_{i=1}^{n} \mathbb{1}(Y_i \ge s)\mathbb{1}(X_i \in \mathcal{A})/[1 - \widehat{G}(s|X_i)]}, \quad (5.1)$$

where $\widehat{G}(s|X_i)$ is an estimated conditional censoring distribution function. Note that this estimator resembles a form of the inverse probably weighting strategy (Rotnitzky and Robins, 2005), which is studied extensively in the survival analysis and missing data literature (Robins and Rotnitzky, 1992). There have been many different forms of inverse probably weighted estimators under a variety of contexts. For example, Hothorn et al. (2005) uses $\delta_i/(1 - \widehat{G}(Y_i|X_i))$ as the subject-specific weight to fit regression random forests. One can also transform the censored observations into fully observed ones using, for example, Rubin and van der Laan (2007), and then fit a regression model using the complete data (Molinaro et al., 2004; Steingrimsson et al., 2016, 2017). Similar ideas have also been used for imputing censored outcomes (Zhu and Kosorok, 2012) when learning an optimal treatment strategy (Cui et al., 2017).

However, our proposal is fundamentally different from these existing methods. We estimate and compare an inverse probability weighted hazard function using the weight $\delta_i/(1 - \widehat{G}(s|X_i))$, and perform this repeatedly at each internal node. A key observation is that the comparison is over the entire domain of the survival time, instead of fitting regression forests based

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

on complete observations. This is a unique advantage, because the distribution function contains richer information than the within-node means. This makes our approach more sensitive in terms of detecting differences between two potential child nodes at all quantile levels of the survival time. It also advocates the goal of a typical survival analysis model, where the survival function is the target of interest, rather than the expected survival time. The intuition has a close connection with that of the sup-log-rank test statistic (Fleming and Harrington, 2011; Eng and Kosorok, 2005), which can be used to detect any distributional difference of T of the two potential child nodes. Furthermore, with the following modified algorithm, we can achieve an improved convergence rate that depends only on the size of $\mathscr{M}_F$.

Algorithm **??** can be modified accordingly to incorporate this new procedure. In particular, at each internal node $\mathcal{A}$, we use the weighted CHF estimator $\widetilde{\Lambda}_{\mathcal{A},n}(t)$ defined in Equation (5.1). We then pick the splitting point $\tilde{c}$ using the rule that both child nodes contain at least a proportion α of the samples at $\mathcal{A}$:

$$\tilde{c} = \arg\max_{c} \Delta_2(c),$$

where $\Delta_2(c) = \max_{t<\tau} \left|\widetilde{\Lambda}_{\mathcal{A}_j^+(c),n}(t) - \widetilde{\Lambda}_{\mathcal{A}_j^-(c),n}(t)\right|$, $\mathcal{A}_j^+(c) = \{X : X^{(j)} \geq c\}$, and $\mathcal{A}_j^-(c) = \{X : X^{(j)} < c\}$, $X^{(j)}$ is the jth dimension of X.

Note that the threshold of $\Delta_2(c)$ in this bias-corrected version is the

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

same as that used for $\Delta_1(c)$ in Algorithm **??**. The intuition is that after removing the censoring bias, the variables in $\mathscr{M}_C$ play the same role as the noise variables do in $\mathscr{M}_N$. In addition, the signal size in Assumption 6 can be relaxed as follows.

Assumption 7. *(Marginal signal of the failure distribution) Let* $\ell^+(j, t_0, c_0)$, $\ell^-(j, t_0, c_0)$, *and the effect size* l *be as defined in Assumption 6. Then, there exists a time point* t_0 *and a cutting point* c_0 *such that, for any* $j \in \mathscr{M}_F$,

$$\left|\ell^+(j, t_0, c_0) - \ell^-(j, t_0, c_0)\right| > \ell.$$

Note that this is essentially removing the censoring-contaminated part (M and its reciprocal) from Assumption 6. Of course, this is at the cost of plugging in a consistent estimator of the censoring distribution G to correct the bias. We need an additional assumption on the dependency structures.

Assumption 8. *We assume that the conditional distribution of the failure and censoring covariates* $X_{\mathscr{M}_{FC}}$ *has a weak dependency on any univariate censoring variable, in the sense that for constant* $\gamma > 1$, *we have*

$$\gamma^{-1} < \frac{p(X_{\mathscr{M}_{FC}\setminus\{j\}} = x | X^{(j)} = x_1)}{p(X_{\mathscr{M}_{FC}\setminus\{j\}} = x | X^{(j)} = x_2)} < \gamma,$$

for any $x_1, x_2 \in [0, 1]$, $(d_0 + d_1 - 1)$*-dimensional vector* x, *and any* $j \in \mathscr{M}_C$.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

This is an analog of Assumption 5 to further prevent the censoring variables from carrying strong marginal signals due to correlations. It degenerates to the commonly used independent covariate case when $\gamma = 1$. Finally, we show that consistency can be established based on our new model-fitting procedure, with the convergence rate depending only on the number of variables in $\mathscr{M}_F$.

Theorem 6. *Under Assumptions 1–5, 7, and 8, assuming that $\widehat{G}$ in Equation (5.1) is a consistent estimation of the censoring distribution, the proposed bias-corrected survival tree is consistent; that is, for any x,*

$$\begin{aligned}&\sup_{t<\tau} |\widehat{\Lambda}_{\mathscr{A},n}(t \mid x) - \Lambda(t \mid x)| \\ &= O\bigg(\sqrt{\frac{\log(n/k)[\log(d_0 k) + \log\log(n)]}{k \log((1-\alpha)^{-1})}} + \Big(\frac{k}{n}\Big)^{\frac{c_1}{d_0}}\bigg),\end{aligned}$$

with probability at least $1 - w_n$, where

$$w_n = \frac{3}{\sqrt{n}} + d_0 \exp\Big\{-\frac{c_2^2 \log_{1/\alpha}(n/k)}{2d_0}\Big\} + d_0 \exp\Big\{-\frac{(1-c_2)c_3 c_4^2 \log_{1/\alpha}(n/k)}{2d_0}\Big\},$$

$c_2, c_4 \in (0,1)$, $c_3 = (1-2\alpha)/8$, and $c_1 = c_3(1-c_2)(1-c_4)/\log_{1-\alpha}(\alpha)$. Consequently, the proposed survival forest is consistent; that is,

$$\begin{aligned}&\lim_{B\to\infty} \sup_{t<\tau} E_X |\widehat{\Lambda}_{\{\mathscr{A}_{(b)}\}_1^B,n}(t \mid X) - \Lambda(t \mid X)| \\ &= O\bigg(\sqrt{\frac{\log(n/k)[\log(d_0 k) + \log\log(n)]}{k \log((1-\alpha)^{-1})}} + \Big(\frac{k}{n}\Big)^{\frac{c_1}{d_0}} + \log(k) w_n\bigg).\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

In the Supplementary Material **??**, we perform simulation studies to show that by actively correcting the bias, the prediction error can be significantly reduced. Interestingly, as can be seen in the simulations, the biasedness is mainly caused by the splitting, rather than the terminal node estimation. This is intuitive and in line with our theory that the splitting bias-correction procedure can enjoy a potentially faster convergence rate than the that of non-bias-corrected version. One might not expect a good prediction if the trees are partitioned inefficiently, regardless of the terminal node estimation 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.

6. Discussion

In this paper, we have provided insights into survival forest and tree models, and developed several fundamental results analyzing the impact of splitting rules. We first investigated the within-node Nelson–Aalen estimator of the CHF and established a concentration inequality for independent but non-identically distributed samples. By introducing a new concept called censoring contamination, we exactly quantify the bias of existing splitting rules. Based on this, we also developed a concentration inequality that bounds

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

the variance component of survival trees and forests. We further analyzed how such bias affects consistency. In particular, we showed that for a commonly used marginal comparison splitting rule strategy, the convergence rate depends on the total number of variables involved in the failure and censoring distributions. However, by appropriately correcting the bias, the convergence rate depends only on the number of failure variables. Essentially, the new bias-correction procedure can be understood as untangling the failure and censoring distributions.

In addition to analyzing this entanglement, our result is based on a weak dependency structure that bounds the marginal signal of any noise variable. This is a generalization of the commonly used independent covariate setting in the literature. A univariate split has a disadvantage when dealing with noise variables, because if they are systematically selected in the splitting rule, the convergence rate will suffer. We believe that similar weak dependency assumptions are inevitable, because otherwise, any correlation structure may carry signals into the noise variables. It would be interesting to investigate whether more advanced splitting rules can overcome this drawback.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

Supplementary Materials

The online supplementary materials include all proofs, algorithms, and additional simulation scenarios.

Acknowledgements

The authors are thankful to the referees, associate editor, and editor for helpful comments which led to an improved manuscript.

References

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics 6*(4), 701–726.

Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Mathematical Biosciences 6*, 1–11.

Arlot, S. and R. Genuer (2014). Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*.

Athey, S., J. Tibshirani, S. Wager, et al. (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

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

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

Biau, G. and E. Scornet (2016). A random forest guided tour. *Test 25*(2), 197–227.

Bou-Hamad, I., D. Larocque, H. Ben-Ameur, et al. (2011). A review of survival trees. *Statistics Surveys 5*, 44–71.

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

Breiman, L. (2004). Consistency for a simple model of random forests.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and regression trees*. CRC press.

Breslow, N., J. Crowley, et al. (1974). A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics 2*(3), 437–453.

Cui, Y., M. R. Kosorok, E. Sverdrup, S. Wager, and R. Zhu (2020). Estimat-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

ing heterogeneous treatment effects with right-censored data via causal survival forests. *arXiv preprint arXiv:2001.09887*.

Cui, Y., R. Zhu, and M. Kosorok (2017). Tree based weighted learning for estimating individualized treatment rules with censored data. *Electron. J. Statist. 11*(2), 3927–3953.

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

Eng, K. H. and M. R. Kosorok (2005). A sample size formula for the supremum log-rank statistic. *Biometrics 61*(1), 86–91.

Fleming, T. R. and D. P. Harrington (2011). *Counting processes and survival analysis*, Volume 169. John Wiley & Sons.

Genuer, R. (2012). Variance reduction in purely random forests. *Journal of Nonparametric Statistics 24*(3), 543–562.

Gill, R. D. (1980). Censoring and stochastic integrals. *Statistica Neerlandica 34*(2), 124–124.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

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

Ishwaran, H. and U. B. Kogalur (2010). Consistency of random survival forests. *Statistics & Probability Letters 80*(13), 1056–1064.

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

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

Klusowski, J. M. (2018). Complete analysis of a random forest model. *arXiv preprint arXiv:1805.02587*.

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

Linero, A. R. (2016). Bayesian regression trees for high dimensional prediction and variable selection. *Journal of the American Statistical Association* (just-accepted).

Mentch, L. and G. Hooker (2014). Ensemble trees and CLTs: Statistical inference for supervised learning. *arXiv preprint arXiv:1404.6473*.

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

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.

Murphy, S., A. Rossini, and A. W. van der Vaart (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association 92*(439), 968–976.

Nelson, W. (1969). Hazard plotting for incomplete failure data(multiply censored data plotting on various type hazard papers for engineering information on time to failure distribution). *Journal of Quality Technology 1*, 27–52.

Robins, J. M. and A. Rotnitzky (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS epidemiology*, pp. 297–331. Springer.

Rotnitzky, A. and J. Robins (2005). Inverse probability weighted estimation in survival analysis. *Encyclopedia of Biostatistics 4*, 2619–2625.

Rubin, D. and M. J. van der Laan (2007). A doubly robust censoring unbiased transformation. *The international journal of biostatistics 3*(1).

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

Scornet, E., G. Biau, J.-P. Vert, et al. (2015). Consistency of random forests. *The Annals of Statistics 43*(4), 1716–1741.

Steingrimsson, J. A., L. Diao, A. M. Molinaro, and R. L. Strawderman (2016). Doubly robust survival trees. *Statistics in Medicine 35*(20), 3595–3612.

Steingrimsson, J. A., L. Diao, and R. L. Strawderman (2017). Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association* (just-accepted).

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

Wager, S., T. Hastie, and B. Efron (2014). Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research 15*(1), 1625–1651.

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

Zhou, M. (1991). Some properties of the Kaplan-Meier estimator for in-

Statistica Sinica: Preprint
doi:10.5705/ss.202020.0263

dependent nonidentically distributed random variables. *The Annals of Statistics*, 2266–2274.

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

Zhu, R., D. Zeng, and M. R. Kosorok (2015). Reinforcement learning trees. *Journal of the American Statistical Association 110*(512), 1770–1784.