Survival Trees for Current Status Data

Ce Yang C298YANG@UWATERLOO.CA
Liqun Diao L2DIAO@UWATERLOO.CA
Richard J. Cook RJCOOK@UWATERLOO.CA
Department of Statistics and Actuarial Science, University of Waterloo, 200 University Avenue West, Waterloo, ON, Canada N2L 3G1

Abstract

Current status data arise when the exact time of an event of interest is not known and the only available information about the time is whether the time is beyond a single assessment. When interest lies in prediction based on such data, we define observed data loss functions through censoring unbiased transformations and pseudo-observations to construct unbiased estimates of complete data loss functions, and we use these to fit regression trees and make predictions using current status data. The trees grown based on these methods are found have good properties empirically in terms of recovery of the true tree structure and event time prediction.

Keywords: Censoring unbiased transformations; current status data; prediction; pseudo-observations; regression trees; variable selection.

1. Introduction

The past several decades have seen considerable development and application of classification and regression tree (CART) (Breiman et al., 1984) algorithms. Recent developments in CART include the extensions to handle various types of outcomes, the use of a variety of loss functions, variations enabling the unbiased selection of covariates, and ensemble algorithms based on single trees; see Loh (2014) for a comprehensive review. Regression trees have been used extensively for prediction problems involving survival data, referred to as *survival trees*, where new splitting criteria and evaluation metrics have been required to deal with right-censored observations. In many settings, however, interest lies in the time of an event which is not directly observable but rather can only be detected to have occurred upon careful clinical examination, through the use of laboratory tests, or by imaging (e.g., radiographs) at particular points in time. Current status data arise when there is only a single assessment time and the event status is only known at this time. Some algorithms for predictive modeling have been developed to deal with interval-censored data, but to our knowledge, there is no predictive model approach developed for current status data. In this paper, we develop survival trees for current status data based on CART algorithm by constructing observed data loss functions which are consistent estimators of a complete data risk function. We demonstrate that survival trees built using the L_2 observed data loss functions are equivalent to the ones obtained by applying the complete data regression trees to the imputed event times. We discuss strategies to construct observed data loss functions and impute event times. The proposed methods are evaluated empirically and compared to the oracle tree built using uncensored event times, methods based on *ad hoc* approaches such as imputing the event times using the midpoint, or the right endpoint of the censoring interval, and the conditional inference tree approach proposed by Fu and Simonoff (2017) based on the log-rank score, which was originally designed for interval-censored data.

2. Survival Trees for Current Status Data

Here we provide a general framework for implementing the CART algorithm while accommodating a current status observation scheme of the response time of interest. We discuss two strategies for constructing observed data loss functions.

2.1. Notation and Preliminary Remarks

Let $T \in \mathbf{R}^+$ denote the event time of interest and $S(t|\boldsymbol{x}) = P(T > t|\boldsymbol{X} = \boldsymbol{x})$ denote the conditional survivor function of T given covariates $\boldsymbol{X} = \boldsymbol{x}$. Current status data, also known as type I interval-censored data, arises if there is a single random examination time U ($U > 0$), and it is only known whether or not the event time T exceeds U; we let $\Delta = I(T \leq U)$, so the observed data for a particular individual are $\boldsymbol{O} = (U, \Delta)$. We further assume that conditional on the covariates that are controlled for in the analysis, the event time is independent of the examination time as in Cook and Lawless (2019). With a sample of observations on n independent processes, we use a subscript i to label individuals and denote the *complete data* as $\boldsymbol{\mathcal{D}} = \{(T_i, \boldsymbol{X_i}')', i = 1, ..., n\}$ and the *observed data* as $\boldsymbol{\mathcal{O}} = \{(\boldsymbol{O_i}, \boldsymbol{X_i}')', i = 1, \ldots, n\}$.

2.2. NPMLE of Survivor Function with Current Status Data

We review of the nonparametric maximum likelihood estimator (NPMLE) of the marginal cumulative distribution function for a failure time based on current status data. Let $U_{(j)}$, $j = 1, \ldots, m$ denote the unique ordered elements of $\{0, U_1, \ldots, U_n\}$, let $n_j = \sum_{i=1}^n I(U_i = U_{(j)})$ denote the number of individuals who are assessed at $U_{(j)}$, and let $r_j = \sum_{i=1}^n \Delta_i I(U_i = U_{(j)})$ denote the number of individuals assessed and found to have failed at the examination $U_{(j)}$, $j = 1, \ldots, m$. The likelihood function can be written as

$$L(S(\cdot)) = \prod_{i=1}^{n} S(U_i)^{1-\Delta_i}[1 - S(U_i)]^{\Delta_i} = \prod_{j=1}^{m} F(U_{(j)})^{r_j}[1 - F(U_{(j)})]^{n_j - r_j}$$

where $F(t) = 1 - S(t)$. With the constraint of $F(U_{(1)}) \leq \ldots \leq F(U_{(m)})$, the optimization problem is equivalent to an isotonic regression (Robertson et al., 1988) problem involving data $\{r_1/n_1, \ldots, r_m/n_m\}$ with weights $\{n_1, \ldots, n_m\}$. According to the maximum-minimum formula for isotonic regression, the NPMLE is

$$\hat{F}(U_{(j)}) = \max_{v_1 \leq j} \min_{v_2 \geq j} \frac{\sum_{l=v_1}^{v_2} r_l}{\sum_{l=v_1}^{v_2} n_l}.$$

So the NPMLE of the survivor function $S(t) = P(T \geq t)$ has a closed form $\hat{S}(t) = 1 - \hat{F}(t)$ for current status data. In practice $\hat{S}(t)$ can be computed empirically via the pooled adjacent violators algorithm (PAVA) for isotonic regression (Barlow et al., 1972).

2.3. The Observed Data Loss Functions

We let $\Psi(\boldsymbol{X}) : \mathcal{X} \to \mathbf{R}$ be the real-valued function denoting a prediction rule, where $\mathcal{X}$ denotes the covariates space. The complete data loss function is $L(\boldsymbol{\mathcal{D}}, \Psi) = \frac{1}{n}\sum_{i=1}^n L(T_i, \Psi(\boldsymbol{X_i}))$,

and $\mathcal{R}(\Psi) = E[L(T, \Psi(\boldsymbol{X}))]$ is the complete data risk. Our goal is to build regression trees based on current status data, in which case the T_i are not observable and the complete data loss function $L(\boldsymbol{\mathcal{D}}, \Psi)$ cannot be calculated. To address this we define a class of *observed data loss functions* $L(\boldsymbol{\mathcal{O}}, \Psi)$ to be used in place of $L(\boldsymbol{\mathcal{D}}, \Psi)$ in the tree growing, pruning and cross-validation steps of the CART algorithm. We choose such functions so that the observed data loss function is unbiased or consistent for the complete data risk $\mathcal{R}(\Psi)$. We next propose two strategies of constructing observed data loss functions based on censoring unbiased transformations and methods based on pseudo-observations.

2.3.1. Censoring Unbiased Transformations (CUT).

Censoring unbiased transformations (CUT) have been utilized to deal with right-censored data (Fan and Gijbels, 1996; Rubin and van der Laan, 2007). More recently, Steingrimsson et al. (2019) considered the constructions of observed data loss functions using CUTs for building regression trees with right-censored data. Here we consider the use of CUT to facilitate the construction of regression trees with current status data. We begin by describing the general construction of CUTs.

Let $\mathcal{Y}$ be a scalar function of the complete data $(T, \boldsymbol{X}')'$ and $\mathcal{Y}^*$ be a scalar function of the observed data $(\boldsymbol{O}, \boldsymbol{X}')'$. We define $\mathcal{Y}^*$ as a CUT for $\mathcal{Y}$ if

$$E[\mathcal{Y}^*(\boldsymbol{O}, \boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}] = E[\mathcal{Y}(T, \boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}]$$

for every $\boldsymbol{x} \in \mathcal{X}$.

We set $\mathcal{Y}(T, \boldsymbol{X}) = L(T, \Psi(\boldsymbol{X}))$, and consider a general function $\mathcal{Y}^*(\boldsymbol{O}, \boldsymbol{X})$ as a CUT of $L(T, \Psi(\boldsymbol{X}))$. For a sample of n independent individuals the observed data loss function $L(\boldsymbol{\mathcal{O}}, \Psi)$ is constructed using the empirical average of the $\mathcal{Y}^*(\boldsymbol{O}, \boldsymbol{X})$ terms as

$$L_{CUT}(\boldsymbol{\mathcal{O}}, \Psi) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{Y}^*(\boldsymbol{O_i}, \boldsymbol{X_i}) \, . \tag{1}$$

The constructed observed data loss function (1) is thus an unbiased estimator for the complete data risk $\mathcal{R}(\Psi) = E[L(T, \Psi(\boldsymbol{X}))]$.

When building a CART with complete data, the default complete data loss for a continuous response is the L_2 loss and a piecewise constant prediction rule $\Psi(\boldsymbol{X}) = \sum_{k=1}^{K} \beta_k I(\boldsymbol{X} \in \mathcal{X}_k)$ is adopted, where $\{\mathcal{X}_1, \ldots, \mathcal{X}_k\}$ is a finite partition of the covariate space $\mathcal{X}$ and β_k is the predicted value if $\boldsymbol{X}$ falls into the kth partition $\mathcal{X}_k$, for $k = 1, \ldots, K$. When the complete data loss takes the form

$$L_2(T, \Psi(\boldsymbol{X})) = \sum_{k=1}^{K} I(\boldsymbol{X} \in \mathcal{X}_k)(T^2 - 2T\beta_k + \beta_k^2) \, , \tag{2}$$

the observed data loss function $L_{2,CUT}(\boldsymbol{\mathcal{O}}, \Psi)$ is built using the empirical average of its corresponding CUT given by

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I(\boldsymbol{X_i} \in \mathcal{X}_k) \left[\mathcal{Y}_2^*(\boldsymbol{O_i}, \boldsymbol{X_i}) - 2\mathcal{Y}_1^*(\boldsymbol{O_i}, \boldsymbol{X_i})\beta_k + \beta_k^2\right] \, , \tag{3}$$

where $\mathcal{Y}_1^*(\boldsymbol{O}, \boldsymbol{X})$ and $\mathcal{Y}_2^*(\boldsymbol{O}, \boldsymbol{X})$ are the CUTs for T and T^2, respectively. The expression in (3) has the same conditional expectation as the L_2 complete data loss (2) given covariates $\boldsymbol{X}$, so (3) is an unbiased estimator of the complete data risk $\mathcal{R}(\Psi) = E[L_2(T, \Psi(\boldsymbol{X}))]$.

The challenge then reduces to finding suitable functions $\mathcal{Y}_j^*(\boldsymbol{O}, \boldsymbol{X})$, $j = 1, 2$, the CUT for T^j, $j = 1, 2$, respectively. Since

$$E[T^j|\boldsymbol{X}] = \sum_{\delta=0}^{1} E[T^j|\Delta = \delta, \boldsymbol{X}]P(\Delta = \delta|\boldsymbol{X}),$$

the CUT of T^j can be constructed as

$$\mathcal{Y}_j^*(\boldsymbol{O}, \boldsymbol{X}) = \sum_{\delta=0}^{1} I(\Delta = \delta)E[T^j|\Delta = \delta, \boldsymbol{X}], \tag{4}$$

where

$$E[T^j|\Delta = 1, \boldsymbol{X}] = -\frac{\int_0^U t^j dS(t|\boldsymbol{X})}{1 - S(U|\boldsymbol{X})}, \tag{5}$$

$$E[T^j|\Delta = 0, \boldsymbol{X}] = -\frac{\int_U^\infty t^j dS(t|\boldsymbol{X})}{S(U|\boldsymbol{X})}. \tag{6}$$

Thus $\mathcal{Y}_j^*(\boldsymbol{O}, \boldsymbol{X})$ is a CUT of T^j since it has the same conditional expectation as T^j given covariates $\boldsymbol{X}$.

The conditional survivor function of T given covariates $\boldsymbol{X}$ can be estimated semiparametrically under a Cox proportional hazard model or nonparametrically using the conditional inference trees proposed by Fu and Simonoff (2017).

2.3.2. Use of Pseudo-Observations.

The jackknife pseudo-observation (PO) approach for incomplete data was introduced and originally used in standard regression settings (Quenouille, 1949; Tukey, 1958), but has been greatly promoted for applications to survival analysis in recent years. Andersen et al. (2003) applied the pseudo-observation approach for inferences based on multi-state models, Andersen et al. (2004) used the pseudo-observations in a regression of restricted mean survival time with right-censored data, and Han et al. (2014) used the pseudo-observations in a semiparametric regression for interval-censored responses. See Andersen and Perme (2010) for a recent review of the general theory and the range of applications of this approach; here we provide a brief introduction.

Suppose $\hat{\theta}$ is an estimator of a parameter of interest θ based on an i.i.d. sample of $T_1, \ldots, T_n$, and $\hat{\theta}^{(-i)}$ is a leave-one-out estimator of θ based on $T_1, \ldots, T_{i-1}, T_{i+1}, \ldots, T_n$. The ith pseudo-observation is constructed as

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}^{(-i)}, \tag{7}$$

for $i = 1, \ldots, n$. If $\hat{\theta}$ and $\hat{\theta}^{(-i)}$ are unbiased estimators of θ, the expectation of $\hat{\theta}_i$ is equal to θ and thus the empirical average of POs also gives an unbiased estimator of θ. In the

present context, we aim to construct an observed data loss function which is unbiased for the full data risk $\mathcal{R}(\Psi)$, so we set the quantity of interest to be $\theta = \mathcal{R}(\Psi)$.

Suppose that $\hat{\theta}$ is an estimator of θ based on the observed current status data $\boldsymbol{\mathcal{O}}$, and that $\hat{\theta}^{(-i)}$ is the corresponding leave-one-out estimator using $\boldsymbol{\mathcal{O}}^{(-i)} = \{(\boldsymbol{O_j}, \boldsymbol{X_j}')', j = 1, \ldots, i-1, i+1, \ldots, n\}$. The POs are obtained using (7), and they are used to further construct the observed data loss function

$$L_{PO}(\boldsymbol{\mathcal{O}}, \Psi) = \frac{1}{n} \sum_{i=1}^{n} \hat{\theta}_i.$$

We consider here the special case where the L_2 loss is specified along with piecewise-constant prediction rules. The observed data loss function $L_{2,PO}(\boldsymbol{\mathcal{O}}, \Psi)$ built using the empirical average of the POs is

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I(\boldsymbol{X_i} \in \mathcal{X}_k) \left(\hat{\theta}_{2i} - 2\hat{\theta}_{1i}\beta_k + \beta_k^2 \right), \tag{8}$$

where $\hat{\theta}_{2i}$ and $\hat{\theta}_{1i}$ are the POs for T_i^2 and T_i in the complete data loss (2). If $\hat{\theta}_j$ is an estimator of $E(T^j)$ and $\hat{\theta}_j^{(-i)}$ is the corresponding leave-one-out estimator, then $\hat{\theta}_{ji} = n\hat{\theta}_j - (n-1)\hat{\theta}_j^{(-i)}$ is the ith PO for $E(T^j)$, $j = 1, 2$.

We estimate $E(T^j) = -\int_0^\infty t^j dS(t)$ by replacing $S(t)$ with an estimate so that the POs for $E(T^j)$ can be written as

$$\hat{\theta}_{ji} = -n \int_0^\infty t^j d\hat{S}(t) + (n-1) \int_0^\infty t^j d\hat{S}^{(-i)}(t), \tag{9}$$

where $\hat{S}(\cdot)$ is NPMLE of the survivor function $S(\cdot)$ given in Section 2.2, and $\hat{S}^{(-i)}(\cdot)$ is the corresponding leave-one-out estimator excluding data from individual i.

2.3.3. Response Imputation.

The idea of response imputation was introduced in Steingrimsson et al. (2019) to facilitate a straightforward use of the observed data loss function for the CART algorithm when data are right-censored. Theorem 4.1 in Steingrimsson et al. (2019) implied that one can implement the L_2 observed data loss functions by applying the L_2 complete data CART algorithm with some imputed dataset $\mathcal{T} = \{(\hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i}), \boldsymbol{X_i}')', i = 1 \ldots, n\}$.

Note that using the complete data CART algorithm following imputation is equivalent to using the following imputed loss function

$$L_{2,I}(\mathcal{T}, \Psi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I(\boldsymbol{X_i} \in \mathcal{X}_k) \left[\hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i})^2 - \hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i})\beta_k + \beta_k^2 \right], \tag{10}$$

where $\hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i})$ is the imputed response for the ith subject in the dataset. Theorem 4.1 in Steingrimsson et al. (2019) showed that the CART algorithm makes decisions in the tree growing, pruning and cross-validation steps, which do not depend on the term $\hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i})^2$.

We extend the idea of response imputation to current status data. We can make the imputed loss function (10) to be equivalent to the L_2 CUT loss function (3) by letting $\hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i}) = \mathcal{Y}_1^*(\boldsymbol{O_i}, \boldsymbol{X_i})$; or implementing the L_2 PO loss function (8) by letting $\hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i}) = \hat{\theta}_{1i}$. The imputed values using CUT or POs are used as complete data in building CART and they lead to the same CART model as we implement CART algorithm with L_2 CUT or PO observed loss functions, respectively.

At this point, we comment that our methods serve as a natural extension of the regression tree to the current status data and they do not depend on the widely used proportional hazard assumptions.

3. Simulation Studies

In this section, we empirically evaluate our proposed methods via simulation and compare them to *ad hoc* imputation and the conditional inference tree approach of Fu and Simonoff (2017).

3.1. Data Generation

We considered a sample size $n = 500$ with 500 replications. We generate $(W_1, W_2, W_3, W_4, W_5)$ from a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$, which is of the form $\boldsymbol{\Sigma_1} = \boldsymbol{I}_{5\times5}$ or $\boldsymbol{\Sigma_2}$, the correlation matrix of an AR(1) model with parameter 0.9. They represent an independent and a highly correlated "autoregressive" dependence structure for the covariates, respectively. Based on these variables, we consider five covariates generated as follows:

- $X_1 = I(W_1 < 0)$ (binary);
- $X_2 = I(W_2 < Q_{0.25}) + 2I(Q_{0.25} \leq W_2 < Q_{0.5}) + 3I(Q_{0.5} \leq W_2 < Q_{0.75}) + 4I(Q_{0.75} \leq W_2)$, where Q_α is the α quantile of a standard normal distribution (ordinal);
- $X_3 = I(W_3 < Q_{0.25}) + 2I(Q_{0.25} \leq W_3 < Q_{0.75}) + 3I(Q_{0.75} \leq W_3)$ (nominal);
- $X_4 = e^{W_4}$ (continuous);
- $X_5 = W_5$ (continuous).

We suppose that the data structure has a tree form with three terminal nodes and the event time at each terminal node follows a Weibull distribution. Thus, we assume:

- *Node 1:* $T \sim$ Weibull(κ_1, λ_1) if $X_2 \leq 2$;
- *Node 2:* $T \sim$ Weibull(κ_2, λ_2) if $X_2 > 2$ and $X_4 > c$;
- *Node 3:* $T \sim$ Weibull(κ_3, λ_3) if $X_2 > 2$ and $X_4 \leq c$.

We let $c = 1$ if covariates are generated independently ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma_1}$) and $c = e^{0.611}$ if covariates are highly correlated ($\boldsymbol{\Sigma} = \boldsymbol{\Sigma_2}$). We found c to be $e^{0.611}$ for such highly correlated covariates to guarantee the proportion of subjects falling into three terminal nodes to be 50%, 25%, and 25%, respectively. Several constraints are imposed to determine the shape and scale

parameters of the Weibull distributions including (i) the median of the marginal distribution of T is 5; (ii) the 0.9 quantile of the distribution of T at the second terminal node is 10; (iii) $\kappa_1 = \kappa_2$ and $\kappa_3 = 3$; (iv) the means of the three terminal nodes are set to be 5μ, 4μ and 2μ, respectively. The node means are found as 7.48, 5.99, and 2.97, respectively.

We next generate the examination times. We let τ denote a maximum time of interest beyond which no assessments will be scheduled, and τ is set as the 95th quantile of the marginal distribution of T. We further let $U^* \sim \Gamma(\alpha, \beta)$. The examination time is set as $U = \min(U^*, \tau)$. We let $\rho = P(T < U)$ represent the proportion of individuals who fail at their examinations in the population. This set-up not only addresses the heterogeneity of timings of the examination times across subjects but also allows us to investigate the effect of informative assessments and the variability of the inspection times on performance by choosing different values of ρ and $\text{Var}(U^*)$, respectively. For each specified ρ and $\text{Var}(U^*)$, we can solve for α and β accordingly, followed by generating the assessments from the gamma distribution with an upper limit τ. Table 1 provides a summary of parameter configuration across various choices of ρ when $\text{Var}(U^*)$ is set to be 4.

Table 1: Parameter configuration for the distribution of the examination times when $\text{Var}(U^*)$ is 4.

ρ	α	β	$P(U^* > \tau)$
0.30	3.16	0.89	1.13×10^{-3}
0.50	7.13	1.34	2.2×10^{-3}
0.70	14.64	1.91	1.26×10^{-2}

3.2. Methods for Event Time Prediction

We propose regression trees for current status data based on the L_2 observed data loss functions using CUT in (3) and PO in (8). When predicting for event times, the L_2 CUT observed data loss function can be implemented with L_2 complete data regression trees based on CUT imputation and the imputed response is $\hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i}) = \mathcal{Y}_1^*(\boldsymbol{O_i}, \boldsymbol{X_i})$ in (4) with conditional means (5) and (6); the L_2 PO observed data loss function can be implemented with PO imputation and the imputed response $\hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i}) = \hat{\theta}_{1i}$ given in (9). Imputation based on the pseudo-observation (PO) $\hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i}) = \hat{\theta}_{1i}$ utilizes the linearly smoothed nonparametric estimator of the marginal survivor function of T obtained from the PAVA algorithm via the `gpava` function from the `R` package `isotone`. The form of the CUT imputation $\hat{T}(\boldsymbol{O_i}, \boldsymbol{X_i}) = \mathcal{Y}_1^*(\boldsymbol{O_i}, \boldsymbol{X_i})$ involves unknown conditional survivor function $S(\cdot|\boldsymbol{X})$, which is estimated semiparametrically under a Cox proportional hazard model (CUT_{Cox}) or nonparametrically using the conditional inference trees proposed by Fu and Simonoff (2017) (CUT_{Con}). The Cox model and conditional inference tree are built upon the current status data and implemented by expressing the current status data in the form of interval-censored data and using the functions `ic.sp` and `ICtree` from the `R` packages `icenReg` and `LTRCtrees`, respectively, originally developed for interval-censored

data. When using the package `LTRCtrees`, we can either use the conditional survivor function estimates obtained from the package `LTRCtrees` (CUT_{Con}) or directly estimate the conditional survivor functions by using the PAVA in each terminal node of the fitted conditional inference tree (CUT_{ConP}).

We aim to compare the performance of our proposed regression tree based on response imputation for predicting event times with the following benchmark methods:

- Oracle trees (O): the regression trees built on the uncensored event times T_i;
- Right imputation (R): When $T \leq U$, the imputed value takes U. If $T > U$, the imputed value is chosen as the time point at which the marginal survivor function estimate decreases to zero; in the case it does not decrease to zero, it is chosen as the time point at which the marginal survivor function estimate reduces to the minimal value.
- Midpoint imputation (M): When $T \leq U$, the imputed value takes $U/2$. When $T > U$, the imputed value takes the average of U and the right-imputed value;
- Conditional inference tree (CIT): conditional inference tree for interval-censored data proposed by Fu and Simonoff (2017).

For response imputation, oracle tree, midpoint and right endpoint imputation, the regression trees are built using the `rpart` function from the `R` package `rpart` with the argument `method = "anova"`.

3.3. Algorithm Evaluation Metrics

Various evaluation metrics are considered to assess the performance of the methods through the test dataset. The main attention is directed at prediction accuracy and the ability to recover the true tree structure. The prediction error (PE) reflects the prediction accuracy and is defined as

$$PE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (\mu_i - \hat{T}_i)^2 \ ,$$

where μ_i is the conditional expectation of T_i given X_i falling into a terminal node based on the true tree structure (i.e., $\lambda_1\Gamma(1+\frac{1}{\kappa_1})$, $\lambda_2\Gamma(1+\frac{1}{\kappa_2})$, $\lambda_3\Gamma(1+\frac{1}{3})$ in the three terminal nodes of the true tree, respectively) and $\hat{T}_i$ denotes the predicted event time of subject i calculated based on the fitted tree. The evaluation metrics for recover the true tree structure include:

- *Model Size:* The average size of the fitted model (i.e., the number of terminal nodes of the fitted tree). In our setting, the closer this is to 3 the better the algorithm performs.
- *Number of Predictors (# Predictors):* This is the average number of predictors (i.e., the mean number of unique covariates the tree splits on). In our setting, the closer to 2, the better the performance.

- *Percent Correct (% Correct):* This reflects the ability of the tree to split on the correct covariates, regardless of the splitting points and the order of splits. This is reported as the percentage of simulated samples for which the method split on both X_2 and X_4, so the higher the percentage, the better the performance.
- *Percent Without Noise (% w/o Noise):* The ability to avoid noise variables. This is reported as the percentage of simulated samples for which the method did not inappropriately split on X_1, X_3, and X_5, so higher percentages correspond to better performance.

3.4. Prediction and Structure Recovery

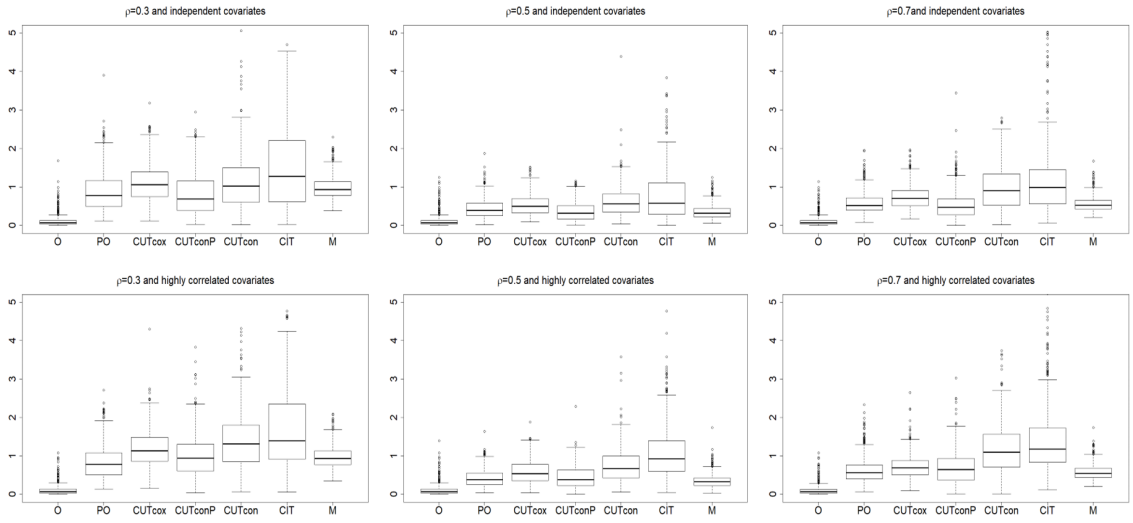

Figure 1: Prediction errors for predicting event times comparing proposed survival tree algorithms for current status data and the benchmarks under various settings.

Figure 1 summarizes the performance of the proposed CART algorithm based on PO, CUT_{Cox}, CUT_{ConP}, and CUT_{Con} compared to the benchmark approaches listed in Section 3.2. The set-ups with independent covariates are presented in the first row and the set-ups with highly correlated covariates are presented in the second row. The proportion of individuals who fail at their examination times are 30%, 50%, and 70% in the three columns, respectively, from left to right. Our proposed regression trees based on CUT_{ConP} had the best performance across the set-ups. The regression trees based on CUT_{Cox}, CUT_{Con} and PO also outperformed the conditional inference trees in all set-ups. All the methods perform worse than the oracle tree in predicting event times to a reasonable extent considering how much less information the current status data contains than the complete data. Furthermore, most tree algorithms deteriorate when the covariates are highly correlated and ρ is smaller. Simulations were repeated and led to similar results for assessments with lower

variance ($\text{Var}(U^*) = 1$, results not shown). However, when the assessments are less informative, the conditional inference trees are less stable and may produce extremely large PEs if some terminal nodes are full of right-censored individuals.

Table 2: Structure recovery measures comparing proposed survival trees algorithms for current status data and the benchmarks under various settings.

	Independent Covariates								Highly Correlated Covariates							
	O	PO	CUT_{cox}	CUT_{conP}	CUT_{con}	CIT	M	R	O	PO	CUT_{cox}	CUT_{conP}	CUT_{con}	CIT	M	R
							$\rho = 0.3$									
Model Size	3.11	3.12	6.02	3.23	3.18	2.90	3.13	3.15	3.09	3.17	6.43	3.82	3.64	3.46	3.11	3.13
# Predictors	2.05	2.03	2.56	2.08	2.05	1.85	2.05	2.03	2.03	2.05	2.85	2.46	2.35	2.35	2.03	2.03
% Correct	95.6	91.0	59.4	83.2	85.0	67.0	93.4	88.6	97.0	89.4	41.0	29.8	35.4	17.8	92.8	88.4
% w/o Noise	95.6	94.2	60.0	87.6	89.8	89.8	94.4	92.6	97.0	91.0	41.2	30.8	37.2	20.4	94.2	91.8
							$\rho = 0.5$									
Model Size	3.11	3.09	3.85	3.25	3.23	3.08	3.10	3.11	3.09	3.06	4.11	3.63	3.55	3.95	3.15	3.14
# Predictors	2.05	2.04	2.23	2.13	2.11	2.02	2.03	2.05	2.03	2.02	2.37	2.34	2.32	2.80	2.06	2.06
% Correct	95.6	95.6	81.4	88.2	89.8	82.8	96.8	95.6	97.0	97.4	70.2	65.2	67.4	15.8	94.6	93.4
% w/o Noise	95.6	96.0	81.6	88.6	90.2	90.4	96.8	95.6	97.0	97.6	70.4	65.2	67.4	16.4	94.6	93.4
							$\rho = 0.7$									
Model Size	3.11	3.12	3.41	3.19	3.20	2.91	3.14	3.10	3.09	3.15	3.53	3.78	3.71	3.73	3.09	3.10
# Predictors	2.05	2.02	2.12	2.07	2.08	1.84	2.06	2.03	2.03	2.01	2.20	2.43	2.38	2.59	2.03	2.01
% Correct	95.6	90.2	80.2	86.8	86.8	69.2	94.4	92.8	97.0	86.0	75.6	41.8	48.8	19.0	94.0	90.2
% w/o Noise	95.6	94.4	85.8	90.4	89.4	92.0	94.6	95.0	97.0	92.6	78.0	43.4	49.8	21.6	95.4	93.8

Table 2 summarizes the structure recovery performance of the proposed survival tree models and the benchmarks. The left half of the table corresponds to the set-ups with independent covariates and the right half contains those with highly correlated covariates. The proportion of individuals found to fail at the examination times are 30%, 50%, and 70% down the columns. The survival trees based on PO recover the underlying tree structure well and their results are comparable to those of the oracle trees in all set-ups, which is valuable as the current status data contains much less information than the complete data. When the covariates are independent, the conditional inference trees perform well; however, the conditional inference trees frequently fail to recover the underlying tree structure when the covariates are highly correlated as they tend to pick up some noise variables and build larger trees than the true tree structure. The survival trees based on CUT_{ConP} and CUT_{Con} perform comparably well to or slightly worse than the oracle trees when covariates are independent but their performance deteriorates in the set-ups with highly correlated covariates as their conditional survivor functions estimated by conditional inference trees and undermined by the compromised performance of conditional inference trees. It is noteworthy that the survival trees based on CUT_{ConP} and CUT_{Con} still perform consistently better than the conditional inference trees. The regression trees based on CUT_{Cox} do not recover the underlying tree structure as well as other survival trees and they suffer from higher computational costs. Similar results were found when the variability of assessments is set to be smaller.

The regression trees based on midpoint imputation have small PEs as illustrated in Figure 1 and both midpoint imputation and right imputation recover underlying tree structure very well across all set-ups as presented in Table 2. The PEs based on right imputation are

not shown in Figure 1 as the PEs are too large to fit in the figures of the current scale. It is noteworthy that the regression trees based on midpoint imputation lead to considerable larger PEs than the other methods when assessments are less informative ($\rho = 0.3$ and 0.5) and in the meanwhile assessment times are less variant ($\text{Var}(U^*) = 1$) (results not shown).

4. Discussion

Here we propose strategies to construct observed data loss functions in place of complete data loss in the CART algorithm when the available data arise from an independent current status observation scheme. We build L_2 complete data regression trees based on imputed responses using CUT and PO, which enables us to reveal influential covariates and make predictions. As shown in the simulation studies, our methods predict the event times more accurately than the conditional inference tree approach across a variety of assessment time models in terms of variance of the assessment times, the proportion of right-censored individuals, and dependence structures among the covariates. Our methods are shown to perform particularly well in recovering the underlying tree structures. Overall, when aiming to fit regression trees based on current status data, we recommend the use of the PO imputation approach and the CUT imputation based on the conditional survivor function estimated using the pooled adjacent violators algorithm for each of the terminal nodes of the conditional inference trees.

This work is based on the assumption that the inspection time is independent of the failure time given the covariates, but given the covariates are selected in a data-driven way this is essentially equivalent to a completely independent inspection time. If there is concern about covariate dependent inspection time model, one can consider the use of inverse density-weighted loss functions to ensure consistent estimation of the complete data loss function. In ongoing work, we are adapting these regression tree algorithms for current status data to accommodate ensemble methods such as random forests. Finally, we note that the equivalence between the constructed observed data loss and the imputed loss only holds under the L_2 loss function in the CART algorithm - an extension of these methods to deal with different loss functions is an area of future research.

References

P. K. Andersen and M. P. Perme. Pseudo-observations in survival analysis. *Statistical Methods in Medical Research*, 19(1):71–99, 2010.

P. K. Andersen, J. P. Klein, and S. Rosthoj. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. *Biometrika*, 90(1):15–27, 2003.

P. K. Andersen, M. G. Hansen, and J. P. Klein. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis*, 10(1):335–350, 2004.

R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical Inference Under Order Restrictions*. New York: Wiley, 1972.

L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*, volume 1. Taylor & Francis Group, 1984.

R.J. Cook and J.F. Lawless. Independence conditions and the analysis of life history studies with intermittent observation. *Biostatistics*, epub:1–27, 2019.

J. Fan and I. Gijbels. *Local Polynomial Modeling and Its Applications*. Chapman & Hall, 1996.

W. Fu and J. S. Simonoff. Survival trees for interval-censored survival data. *Statistics in Medicine*, 36(1):4831–4842, 2017.

S. Han, A. C. Andrei, and K. W. Tsui. A semiparametric regression method for interval-censored data. *Communications in Statistics - Simulation and Computation*, 43(1):18–30, 2014.

W.Y. Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3):329–348, 2014.

M. Quenouille. Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Series B*, 11:18–84, 1949.

T. Robertson, F. T. Wright, and R. Dykstra. *Order Restricted Statistical Inference*. John Wiley: New York., 1988.

D. Rubin and M. J. van der Laan. A doubly robust censoring unbiased transformation. *The International Journal of Biostatistics*, 3:Iss. 1, Article 4, 2007.

J. A. Steingrimsson, L. Diao, and R. L. Strawderman. Censoring unbiased regression trees and ensembles. *Journal of the American Statistical Association*, 114(1):370–383, 2019.

J. W. Tukey. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29:614, 1958.