Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Journal of Multivariate Analysis 90 (2004) 154–177

Journal of Multivariate Analysis

http://www.elsevier.com/locate/jmva

Tree-based multivariate regression and density estimation with right-censored data

Annette M. Molinaro,[1] Sandrine Dudoit,[*,2] and Mark J. van der Laan[3]

Division of Biostatistics, University of California, Berkeley, 140 Earl Warren Hall, #7360, Berkeley, CA 94720-7360, USA

Received 1 July 2003

Abstract

We propose a unified strategy for estimator construction, selection, and performance assessment in the presence of censoring. This approach is entirely driven by the choice of a loss function for the full (uncensored) data structure and can be stated in terms of the following three main steps. (1) First, define the parameter of interest as the minimizer of the expected loss, or risk, for a full data loss function chosen to represent the desired measure of performance. Map the full data loss function into an observed (censored) data loss function having the same expected value and leading to an efficient estimator of this risk. (2) Next, construct candidate estimators based on the loss function for the observed data. (3) Then, apply cross-validation to estimate risk based on the observed data loss function and to select an optimal estimator among the candidates. A number of common estimation procedures follow this approach in the full data situation, but depart from it when faced with the obstacle of evaluating the loss function for censored observations. Here, we argue that one can, and should, also adhere to this estimation road map in censored data situations.

Tree-based methods, where the candidate estimators in Step 2 are generated by recursive binary partitioning of a suitably defined covariate space, provide a striking example of the

[*] Corresponding author. Fax: + 510-643-5163.

E-mail addresses: molinaro@stat.berkeley.edu (A.M. Molinaro), sandrine@stat.berkeley.edu (S. Dudoit), laan@stat.berkeley.edu (M.J. van der Laan).

URLs: http://www.stat.berkeley.edu/˜molinaro, http://www.stat.berkeley.edu/˜sandrine, http://www.stat.berkeley.edu/˜laan.

[1] A. M. was supported by a grant from the Institute for Scientific Computing Research at Lawrence Livermore National Laboratory.

[2] S. D.'s research was supported in part by Grant R01 LM007609-01 from the National Institutes of Health.

[3] M vdL.'s research was supported in part by Grant R01 GM67233 from the National Institutes of Health.

doi:10.1016/j.jmva.2004.02.003

chasm between estimation procedures for full data and censored data (e.g., regression trees as in CART for uncensored data and adaptations to censored data). Common approaches for regression trees bypass the risk estimation problem for censored outcomes by altering the node splitting and tree pruning criteria in manners that are specific to right-censored data. This article describes an application of our unified methodology to tree-based estimation with censored data. The approach encompasses univariate outcome prediction, multivariate outcome prediction, and density estimation, simply by defining a suitable loss function for each of these problems. The proposed method for tree-based estimation with censoring is evaluated using a simulation study and the analysis of CGH copy number and survival data from breast cancer patients.

AMS 2000 subject classifications: 62H99; 62N02; 62G99; 62G05; 62G07; 62G08; 62C99; 62P10

Keywords: CART; Censored data; Comparative genomic hybridization; Cross-validation; Density estimation; Loss function; Microarray; Model selection; Multivariate outcome; Prediction; Regression tree; Risk estimation; Survival analysis

1. Introduction

1.1. *Estimation road map for censored data*

Our general strategy for estimator construction, selection, and performance assessment is entirely driven by the choice of a *loss function* for the full, uncensored data structure. Censored data can be handled simply by replacing the full data loss function by an observed data loss function with the same expectation. Our proposed estimation road map for censored data can be stated in terms of the following three main steps:

(1) *Definition of the parameter of interest in terms of a loss function for the observed data.* For the full data structure, define the parameter of interest as the minimizer of the expected loss, or *risk*, for a loss function chosen to represent the desired measure of performance (e.g., squared error loss in regression trees). Apply the general *estimating function* methodology of van der Laan and Robins [28] to map the *full, uncensored* data loss function into an *observed, censored* data loss function having the same expected value and leading to an efficient estimator of this risk.
(2) *Construction of candidate estimators based on a loss function for the observed data.* Define a finite collection of candidate estimators for the parameter of interest based on a sieve of increasing dimension approximating the complete parameter space (e.g., recursive binary partitioning of the covariate space as in regression trees). For each element of the sieve, the candidate estimator is defined as a minimizer of empirical risk for the observed data loss function (e.g., within-node sample mean for the squared error loss).
(3) *Cross-validation for estimator selection and performance assessment based on a loss function for the observed data.* Use *cross-validation* to estimate risk based on

the observed data loss function and to select an optimal estimator among the candidates in Step 2. This step relies on the unified cross-validation methodology of van der Laan and Dudoit [25] and their finite sample and asymptotic optimality results concerning cross-validation estimator selection for general data generating distributions, loss functions (possibly depending on a nuisance parameter), estimators (e.g., from linear regression, regression trees), and cross-validation procedures (e.g., V-fold, Monte-Carlo cross-validation).

As described below, a number of common estimation procedures follow this approach in the full data situation, but depart from it when faced with the obstacle of evaluating the loss function in the presence of censoring. Tree-based methods, where the candidate estimators in Step 2 are generated by recursive binary partitioning of a suitably defined covariate space, provide a striking example of the chasm between estimation procedures for full data and censored data: regression trees for uncensored data [4] vs. adaptations to censored data [1,2,5,6,10,16,22]. Here, we argue that one can, and should, also adhere to the above estimation road map in censored data situations. All that is required is to replace the full (uncensored) data loss function by an observed (censored) data loss function with the same expected value, i.e., the same risk. This key step can be achieved using the general estimating function methodology of van der Laan and Robins [28]. Note that we use the term estimation in a broad sense, to provide a unified treatment of multivariate outcome prediction and density estimation based on censored data. Each of these problems can be dealt with according to the road map by the choice of a suitable loss function.

The present article introduces a general loss-based methodology for estimator construction, selection, and performance assessment with cross-validation, in the context of tree-structured estimation with censored data. We focus on the choice of a loss function (i.e., Step 1 of the road map) and refer to van der Laan and Dudoit [25] for details on the general methodology for generating candidates and for cross-validation selection (i.e., Steps 2 and 3, respectively). The remainder of this section reviews the literature on survival trees. Our proposed methodology for tree-based multivariate regression and density estimation with censored data is described in Section 2. The approach is evaluated in Section 3 via a simulation study and the analysis of CGH copy number and survival data from breast cancer patients. Finally, Section 4 summarizes our findings and discusses ongoing work.

Our unified loss-based estimation methodology with cross-validation is discussed in detail in van der Laan and Dudoit [25]. A less technical and shorter overview is given in Dudoit et al. [8]. Special cases and applications are described in a collection of related articles: estimator selection and performance assessment based on uncensored data [7]; estimator selection with censored data [14]; likelihood-based cross-validation [26]; deletion/substitution/addition (or D/S/A) algorithms for generating candidate estimators [18,23]; supervised detection of regulatory motifs in DNA sequences [15].

1.2. Review of tree-based estimation

Tremendous amounts of clinical and genomic data are currently being collected in the hopes of finding significant diagnostic and prognostic factors for diseases such as cancer. A common scenario in medical studies is that in which hundreds, possibly thousands, of covariates are recorded for each patient along with a time to event. Examples of the event of interest can be recurrence of chronic illness, death from disease, or drop in a bodily measurement (e.g., white blood cell count). In addition to clinical, epidemiological, and histological variables, the covariates may include microarray measurements of transcript (i.e., mRNA) levels for thousands of genes or of DNA copy number for thousands of chromosomal regions. By the completion of a study, some patients may have dropped out, been lost to follow-up, or not had the particular event. In this situation, the last date of follow-up is recorded and referred to as the censored time to event. One objective in these studies is to build predictors for the time to event based on the measured covariates and identify which of the covariates are integral in affecting this outcome.

Over the past three decades, numerous non-parametric and semi-parametric approaches have been suggested to deal with censored data. In *tree-based estimation* procedures, candidate estimators are generated by recursive binary partitioning of a suitably defined covariate space into *nodes* and an estimator is returned for each set in the final partition, i.e., each *terminal node* or *leaf*. Regression trees were first introduced by Morgan and Sonquist [19] in their *automatic interaction detection* (AID) program. The methodology was then generalized and formalized in the monograph on *classification and regression trees* (CART) by Breiman et al. [4]. There are three main aspects to tree-structured estimation: (i) the *node splitting* rule for generating partitions of the covariate space, i.e., generating the candidate estimators (cf. Step 2 of the road map); (ii) the selection of a 'right-sized' tree, by *tree pruning* with *cross-validation* (cf. Step 3 of the road map); (iii) estimation of the parameter of interest within each node (cf. Step 1 of the road map). Solutions to each of these problems typically involve optimization of a loss-based criterion.

As suggested above, the CART methodology of Breiman et al. [4] can be formulated in terms of the three main steps of our general road map. In the special case of regression trees for continuous outcomes, the loss function is the squared error loss, or quadratic loss, and the parameter of interest is the conditional expected value of an outcome given covariates. The loss function enters at two key stages of the tree building process: node splitting and tree pruning with cross-validation, corresponding to Steps 2 and 3 of the road map, respectively. Specifically, the CART candidate estimators are generated by recursive binary partitioning of the covariate space using a splitting rule based on the decrease of within-node mean squared error (MSE). The result is a collection of candidate estimators, starting with a single-node tree and running up to a tree with numerous terminal nodes (i.e., maximal exploratory tree). After growing a large tree, the loss function is used again for pruning and for selecting a right-sized tree among the generated sequence of trees using cross-validation risk estimation. The survival trees discussed in Breiman [1,2] can also be viewed within this framework. In this context, the outcome is a right-censored survival time and parameters of interest may include the conditional expected value and median of the (log) survival time given covariates and the conditional survival function given covariates. Corresponding full data loss functions are the squared error, absolute error, and negative log-likelihood loss functions, respectively. However, an immediate difficulty arises with censored data when evaluating the loss function at the splitting and pruning stages. Common approaches for tree-based regression and density estimation bypass the risk estimation problem for censored outcomes by altering the splitting and pruning criteria in manners that are specific to right-censored survival times. As described next, some of these proposals deviate from the estimation road map in essential ways.

Previously proposed modifications to regression trees, often referred to as *survival trees*, fall into two categories based on their use of within-node homogeneity or between-node heterogeneity measures. Included in the first category are: Breiman [1,2], Davis and Anderson [6], Gordon and Olshen [10], and LeBlanc and Crowley [16]. These approaches inherit the fundamental basis of CART, in the sense that they rely on splitting rules which optimize a loss-based *within-node homogeneity* criterion and use cost-complexity pruning and cross-validation to select a right-sized tree from the sequence of candidate trees. However, they each propose a different loss function to accommodate censored survival data. Davis and Anderson [6] base their split function on the negative log-likelihood of an exponential model; Gordon and Olshen [10] use L^p, L^p Wasserstein, and Hellinger distances for within-node Kaplan–Meir estimates of the survival distribution; and LeBlanc and Crowley [16] use the first step of a full likelihood estimation procedure for a Cox proportional hazards model with the same baseline hazard for each node implied by the partition of the covariate space. In the recent work of Breiman [1,2] on survival trees and survival forests, the time-covariate space is partitioned by seeking splits that maximize the increase in the observed data log-likelihood for a constant hazards model within each node. In the case of random forests, maximal trees are grown until only one uncensored observation is left in each node and aggregated over bootstrap samples. The effects of covariates over time are traced by monitoring correlations of the conditional cumulative hazard function with individual covariates, based on only uncensored observations. Most of the methods described above thus rely on a negative log-likelihood loss function, with the explicit or implicit goal of estimating the conditional survival function given covariates, and differ mainly in their choice of model for the observed data likelihood within nodes. By partitioning the time-covariate space, rather than only the covariate space, the survival trees of Breiman [1,2] seem to provide the least parametric estimation procedure. The choice of loss function is discussed further in Section 2.2, below.

In the second class of survival trees, Ciampi et al. [5] and Segal [22] employ two-sample log-rank test statistics as *between-node heterogeneity* measures. This approach leads to alternative methods for splitting and pruning and thus deviates markedly from standard tree methodology and our proposed road map.

Hothorn et al. [12] consider bagging the survival trees produced by the aforementioned procedures, with the aim of generating improved estimators of the conditional survival function. Given bootstrap partitions of the covariate space, a Kaplan–Meier estimator of the survival function is produced for each learning set observation based on bootstrap-aggregated nodes, i.e., based on the union, over bootstrap survival trees, of nodes containing the given observation. Performance is assessed using the Brier score, which relies on the assumption of independent survival and censoring times [11].

In essence, existing survival tree methods all have in common that they bypass direct evaluation of the loss function in splitting and pruning, by replacing the full data loss-based criteria inherent in regression trees with alternatives specific to censored outcomes. In general, the splitting and pruning criteria seem to be chosen based on convenience for handling censored data and do not reduce to the preferred choice for uncensored data. That is, rather than specifying a loss function based on a parameter of interest as in the uncensored data case (e.g., squared error loss for conditional expected value of survival time), the choice of a loss function seems to be dictated by the ability to evaluate it on censored observations. In principle, one could be interested in other parameters than the conditional survival density (corresponding to the negative log-likelihood loss function used in the above approaches), such as the conditional mean or median survival times, or the conditional survival function evaluated at a single point. In such cases, one should employ a different loss function, which is specific to the parameter of interest. Finally, existing methods do not provide adequate means for evaluating the overall performance of the resulting estimators: due to the inability to evaluate arbitrary loss functions for censored observations, risk estimates are often based on only uncensored data. Discarding censored observations could potentially lead to serious biases in performance assessment (the implications of omitting censored data in risk estimation are discussed in Section 2.2). This general difficulty in evaluating risk for censored observations results in a discontinuity between the full and observed data worlds.

It is our intention to follow the loss-based estimation road map of Section 1.1 and derive estimators that link the full and censored data worlds with the following two requirements. First, when applied to uncensored observations, the censored data methodology should reduce to the full data methodology for estimator construction, selection, and performance assessment. Second, in order to allow for informative censoring and a gain in efficiency, we wish to have the ability to build estimators using other covariates than (possibly in addition to) those used to define the parameter of interest. Neither of these two requirements nor this methodology have been adopted by the aforementioned approaches. In contrast to these modifications, which depart from the standard tree building framework and the estimation road map, we propose to use the general estimating function methodology of van der Laan and Robins [28] to map the full data loss function into an observed, censored data loss function having the same expected value and leading naturally to an efficient estimator of this risk. This observed data loss function is then used for tree building and performance assessment.

2. Tree-based estimation with right-censored data

This section elaborates on the main steps of our general approach to loss-based estimator construction, selection, and performance assessment with cross-validation. We emphasize the choice of a loss function and illustrate the methodology in the context of tree-structured estimators for censored data. In tree-based estimation procedures such as CART [4], the candidate estimators in Step 2 of the road map are generated by recursive binary partitioning of a suitably defined covariate space. Univariate outcome prediction, multivariate outcome prediction, and density estimation can be handled within the same framework simply by specifying a suitable full data loss function for each of these problems. The estimating function methodology of van der Laan and Robins [28] is applied to yield observed data loss functions for node splitting, tree pruning, and cross-validation performance assessment in the presence of censoring. The rest of the tree building procedure is retained and the reader is referred to Breiman et al. [4] for details.

2.1. Model

2.1.1. Full data structure

In the full data world, let $\{X(t) : t \in \mathbb{R}^+\}$ be a multivariate stochastic process, indexed by time t. Let T denote either a fixed endpoint of this stochastic process or a random *survival time*, and let $Z \equiv \log T$. The *full data structure* is defined as $X \equiv \bar{X}(T) = \{X(t) = (R(t), L(t)) : 0 \leqslant t \leqslant T\}$, where $R(t) \equiv \mathrm{I}(T \leqslant t)$, $L(t)$ is the covariate process, and T is now a function of X. Denote the distribution of the full data structure X by $F_{X,0}$. The covariate process $L(t)$ may contain time-dependent and time-independent covariates. Denote the time-independent, or baseline, covariates by $L(0)$. If T is fixed, then let $Z(t)$, $t \in \{t_0 = 0, \ldots, t_{m-1} = T\}$, be an m-dimensional outcome process of interest included in $X(t)$.

2.1.2. Observed data structure

In the observed data world, one rarely sees all of the relevant variables in the process $X = \bar{X}(T) = \{X(t) : 0 \leqslant t \leqslant T\}$. Rather, one observes the full data process $X(t)$ only up to the minimum, $\tilde{T} \equiv \min(T, C)$, of the survival time T and a univariate *censoring variable* C. This missing, or *censored*, survival data situation can be due to drop out or the end of follow-up. The *observed data structure* can be written as $O \equiv (\tilde{T}, \Delta, \bar{X}(\tilde{T}))$, where $\Delta \equiv \mathrm{I}(T \leqslant C)$ is the *censoring indicator*, equal to one for uncensored observations and to zero for censored observations. The censoring process is denoted by $A(t) \equiv \mathrm{I}(C < t)$. By convention, if T occurs prior to C, set $C = \infty$; thus, C is always observed and one can rewrite the observed data structure as $O = (C, \bar{X}(C))$. The random variable O has a distribution $P_0 = P_{F_{X,0},G_0}$, indexed by the full data distribution, $F_{X,0}$, and the conditional distribution, $G_0(\cdot\,|X)$, of the censoring variable C given X. Because what one observes about X is determined by C, $G_0(\cdot\,|X)$ is referred to as the *censoring* or *coarsening mechanism*.

The survivor function for the censoring mechanism is denoted by $\bar{G}_0(c \mid X) \equiv Pr_0(C > c \mid X)$ and referred to as *censoring survivor function*.

We assume that for $c < T$, the Lebesgue *hazard function*, λ_0, corresponding to the censoring mechanism given the full data X, satisfies

$$\lambda_0(c|X) = Pr_0(C = c \mid C \geqslant c, X) = m(c, \bar{X}(c)) = m(o) \tag{1}$$

for some measurable function, m. This assumption on the censoring mechanism, referred to as *coarsening at random* (CAR), holds if the censoring distribution only depends on the observed process $\bar{X}(c)$. If X does not include time-dependent covariates (i.e., $L = L(0)$), then, under CAR, the censoring time C is conditionally independent of the survival time T, given baseline covariates $L(0)$. An important consequence of CAR is that it implies the following factorization for the density of the observed data $O = (C, \bar{X}(C))$ (with respect to a dominating measure satisfying CAR itself), into an F_X-part and a G-part, $p_{F_X,G}(o) = p_{F_X}(o)h(o)$, where $h(o)$ is the density $g_{C \mid X}(c \mid x)$ and $p_{F_X}(o) = f_{F_X}(\bar{X}(t))\,|_{t=c}$ only depends on the measure F_X. Denote by $\mathscr{G}(CAR)$ the set of all conditional distributions G of C given X satisfying CAR. Gill et al. [9], van der Laan and Robins [28] (Section 1.2.3, in particular), and Robins and Rotnitzky [20] provide further, thorough explanations of CAR.

2.2. *Definition of parameter of interest in terms of loss function*

2.2.1. *Full data loss function*

In the full data world, assume that we have a sample, or *learning set*, of n independent and identically distributed (i.i.d.) observations, $X_1, \ldots, X_n$, from the distribution $F_{X,0}$. The parameter of interest, ψ_0, is a mapping $\psi : \mathscr{S} \to \mathbb{R}$, from a covariate space $\mathscr{S}$ into the real line $\mathbb{R}$. The space $\mathscr{S}$ is typically a subset of $\mathbb{R}^d$, corresponding to a d-dimensional vector $W \subseteq L(0)$ of covariates measured at baseline; $\mathscr{S}$ could also refer to other variables, such as the survival time T in survival function estimation or a time index t in multivariate outcome prediction. Denote the parameter space by $\mathbf{\Psi}$. The parameter ψ_0 is defined in terms of a *loss function*, $L(X, \psi)$, as (one of) the minimizer(s) of the expected loss, or *risk*. That is, ψ_0 is such that

$$\begin{aligned} E_{F_{X,0}}[L(X, \psi_0)] &= \int L(x, \psi_0)\, dF_{X,0}(x) \\ &\equiv \min_{\psi \in \mathbf{\Psi}} \int L(x, \psi)\, dF_{X,0}(x) = \min_{\psi \in \mathbf{\Psi}} E_{F_{X,0}}[L(X, \psi)]. \end{aligned} \tag{2}$$

Note that we do not require uniqueness of the risk minimizer, rather, we simply assume that there is a loss function whose risk is minimized by the parameter of interest ψ_0. To simplify notation, we use the subscript 0 to refer to parameters of the underlying data generating distributions $F_{X,0}$ and G_0, that is, write $E_{F_{X,0}}[L(X, \psi)] = E_0[L(X, \psi)]$. The purpose of the loss function L is to quantify performance. Thus, depending on the parameter of interest, there could be numerous loss functions from which to choose. When the parameter of interest is the conditional mean, $\psi_0(W) =$

$E_0[Z \mid W]$, a common choice of loss function is the squared error loss, $L(X,\psi) = (Z - \psi(W))^2$. As described in Sections 2.2.3–2.2.5 below, we focus on three types of full data loss functions, for the purposes of univariate outcome prediction, multivariate outcome prediction, and density estimation, respectively.

2.2.2. Observed data loss function

In the observed data world, one has a learning set of n i.i.d. observations, $O_1, \ldots, O_n$, from the right-censored data structure, $O_i \sim P_0 = P_{F_{X,0},G_0}$. Let the empirical distribution of $O_1, \ldots, O_n$ be denoted by P_n. The goal remains to find an estimator for a parameter ψ_0 defined in terms of the risk for a full data loss function $L(X,\psi)$, e.g., a predictor of the log survival time Z based on covariates W. An immediate problem is that a loss function such as the quadratic loss, $L(X,\psi) = (Z - \psi(W))^2$, cannot be evaluated for an observation O with censored survival time (i.e., $Z = \log T$ is unobserved for $\Delta = 0$). Risk estimators based on only uncensored observations, such as $\frac{1}{n}\sum_i L(X_i,\psi)\Delta_i$, are biased for $E_0[L(X,\psi)]$ and, in particular, estimate the quantity $E_0[L(X,\psi)\bar{G}_0(T \mid X)]$ which is not minimized by the parameter of interest ψ_0.

Our proposed general solution is to replace the full (uncensored) data loss function by an observed (censored) data loss function with the same expected value, i.e., the same risk. The general *estimating function* methodology of van der Laan and Robins [28] can be used to link the observed data world to the full data world. Specifically, the methodology allows full data estimating functions, $D(X)$, to be mapped into observed data estimating functions, $IC(O \mid G, Q, D)$, indexed by *nuisance parameter* G and, possibly, $Q = Q(F_X)$. The abbreviation IC stands for *influence curve*, as in [28]. The estimating functions satisfy

$$E_{P_0}[IC(O \mid G, Q, D)] = E_{F_{X,0}}[D(X)] \quad \text{if } G = G_0 \text{ or } Q = Q_0 = Q(F_{X,0}).$$

In our specific application, the full data estimating function is the loss function, $D(X) = L(X,\psi)$, and the risk for a given estimator ψ is viewed as the full data parameter of interest, $\theta_0 = E_0[D(X)] = E_0[L(X,\psi)]$. Observed data loss functions are obtained from the estimating functions IC, that is, $L(O,\psi \mid \eta_0) = IC(O \mid G_0, Q_0, L(\cdot,\psi))$ is an observed data loss function with the same risk as the full data loss function $L(X,\psi)$, where η_0 denotes the nuisance parameters (G_0, Q_0),

$$\int L(o,\psi \mid \eta_0)\, dP_0(o) = \int L(x,\psi)\, dF_{X,0}(x). \tag{3}$$

Inverse probability of censoring weighted loss function: The *inverse probability of censoring weighted* (IPCW) estimating function was introduced by Robins and Rotnitzky [20]. Its name derives from the fact that the full data function $D(X)$ is weighted by the inverse of a censoring probability. This estimating function is defined as

$$IC(O \mid G, D) \equiv D(X)\frac{\Delta}{\bar{G}(T|X)}, \tag{4}$$

where $\bar{G}$ is a conditional survivor function for the censoring time C given full data X and $\Delta = \mathrm{I}(T \leqslant C)$ is the censoring indicator. Given that

$$E_0[\Delta|X] = Pr_0(C \geqslant T|X) = \bar{G}_0(T|X) > 0, \quad F_{X,0}\text{-a.e.},$$

one has

$$E_0\left[\frac{D(X)\Delta}{\bar{G}_0(T|X)}\right] = E_0\left[E_0\left[\frac{D(X)\Delta}{\bar{G}_0(T|X)}\middle| X\right]\right] = E_0[D(X)].$$

This suggests the IPCW observed data loss function, $L(O, \psi \mid \eta_0) = IC(O \mid G_0, L(\cdot, \psi))$, with nuisance parameter $\eta_0 = G_0$. The corresponding risk estimator is the empirical mean

$$\hat{\theta}_n \equiv \frac{1}{n}\sum_{i=1}^{n} L(O_i, \psi \mid \eta_n) = \frac{1}{n}\sum_{i=1}^{n} L(X_i, \psi)\frac{\Delta_i}{\bar{G}_n(T_i|X_i)}, \tag{5}$$

where η_n represents $\bar{G}_n$, an estimator of the nuisance parameter $\bar{G}_0$ derived under the CAR assumption for the censoring mechanism, i.e., by considering censoring mechanisms $G \in \mathcal{G}(CAR)$. For such models, the estimator $\bar{G}_n(T|X)$ is a function of $O = (C, \bar{X}(C))$ and thus the resulting risk estimator $\hat{\theta}_n$ depends only on the observed data structure, $O_1, \ldots, O_n$. If a Cox proportional hazards model is assumed for the censoring mechanism G, then $\lambda(t \mid X) = \lambda_0(t)\exp(\beta^{\mathrm{T}} J(t))$, where $J(t) = f(L(t))$ is a set of covariates extracted from the process $\bar{L}(t) = \{L(s) : 0 \leqslant s \leqslant t\}$ for some given $\mathbb{R}^k$-valued function f. Standard software can then be employed to obtain maximum (partial) likelihood estimators of the baseline hazard function and the regression coefficients β (e.g., `coxph` function in R).

The IPCW estimating function provides a consistent risk estimator under the following conditions: (i) $\bar{G}_0(T|X) > \delta > 0$, $F_{X,0}$-a.e., for some $\delta > 0$, and (ii) $\bar{G}_n$ is a consistent estimator for $\bar{G}_0$. When there are no time-dependent covariates (i.e., $L = L(0)$), let α^L denote the right endpoint of the support of the distribution $F_{T|L}(\cdot \mid L)$, of the survival time T given time-independent covariates L. Then, under CAR, condition (i) holds if $\bar{G}_0(\alpha^L \mid L) > 0$, a.e. in L.

An alternative to the IPCW estimating function is the *doubly robust inverse probability of censoring weighted* (DR-IPCW) estimating function. Under an identifiability condition, the DR-IPCW loss function ensures consistent risk estimation if at least one of the two nuisance parameters, G or Q, are consistently estimated and asymptotic efficiency if both are consistently estimated [28]. This double robustness property thus allows misspecification of either the censoring mechanism or of part of the full data generating distribution. The DR-IPCW observed data loss function is further discussed in [17]. A third, more parametric approach for estimating the full data risk of a given estimator could be based on the F_X-part of the observed data likelihood. For example, one could assume a parametric model for F_X and estimate the full data risk by maximum likelihood. The three estimation approaches (IPCW, DR-IPCW, and maximum likelihood) are contrasted in Section 1.2 of [28].

We stress that in the absence of censoring, i.e., when $\Delta \equiv 1$ and $C \equiv \infty$, both the IPCW and DR-IPCW observed data loss functions reduce to the full data loss function, $L(O, \psi \mid \eta_0) = L(X, \psi)$. This ensures that the censored and full data estimators coincide when there is no censoring. In addition, to allow for informative censoring and a gain in efficiency, one may estimate the nuisance parameter $\bar{G}_0$ in the IPCW and DR-IPCW loss functions using other covariates than those defining the parameter ψ. The methodology is illustrated below for three types of loss functions using the simple IPCW estimating function; one can proceed similarly for the DR-IPCW estimating function.

2.2.3. *Univariate outcome prediction*

One is concerned with predicting a univariate outcome Z, such as the log survival time $Z = \log T$, based on a vector of covariates $W \subseteq L(0)$. The parameter of interest in regression trees (continuous outcome Z) is typically the conditional expectation, $\psi_0(W) = E_0[Z \mid W]$, corresponding to the *squared error* (i.e., *quadratic* or L^2) loss function, $L(X, \psi) = (Z - \psi(W))^2$. Another parameter of interest could be the conditional median, $\psi_0(W) = \text{Median}_0[Z \mid W]$, corresponding to the *absolute error* (i.e., L^1) loss function, $L(X, \psi) = |Z - \psi(W)|$.

The IPCW observed data loss function for the quadratic loss is

$$L(O, \psi \mid \eta_n) = (Z - \psi(W))^2 \frac{\Delta}{\bar{G}_n(T|X)}, \tag{6}$$

where $\Delta = \mathrm{I}(T \leqslant C)$ is the censoring indicator and $\eta_n = G_n$ is an estimator of the nuisance parameter $\eta_0 = G_0$, corresponding to the conditional survivor function $\bar{G}_0$ for the censoring time C given full data X. Under the coarsening at random (CAR) assumption and when there are no time-dependent covariates (i.e., $L = L(0)$), one can estimate $\bar{G}_0(\cdot|X)$ by $\bar{G}_n(\cdot|L(0))$, a function only of the observed data structure O. As described in Section 3, survival trees can be grown using the IPCW observed data loss function by specifying suitable weights for individual observations.

In classification trees (polychotomous outcome Z), the parameter of interest involves the class conditional probabilities, $Pr_0(z|W)$. For the *indicator* loss function, $L(X, \psi) = \mathrm{I}(Z \neq \psi(W))$, the optimal parameter is $\psi_0(W) = \text{argmax}_z\, Pr_0(z \mid W)$, the class with maximum probability given covariates W. One could also use a loss function which incorporates differential misclassification costs. Note that in the standard CART methodology, Breiman et al. [4] favor replacing the indicator loss function in the splitting rule by measures of node impurity, such as the entropy, Gini, or twoing indices [4, Chapter 4]. The indicator loss function is still used for pruning and performance assessment. It turns out that the *entropy* criterion corresponds to the negative log-likelihood loss function, $L(X, \psi) = -\log \psi(X)$, and parameter of interest $\psi_0(X) = Pr_0(Z|W)$. Likewise, the *Gini* criterion corresponds to the loss function $L(X, \psi) = 1 - \psi(X)$, with parameter of interest $\psi_0(X) = \mathrm{I}(Z = \text{argmax}_z\, Pr_0(z \mid W))$. These modifications thus fall within our framework and amount to using different loss functions for the same parameter at different stages of the tree building process.

2.2.4. Multivariate outcome prediction

Consider an m-variate outcome, such as the time-dependent outcome process $Z(t)$ included in $X(t)$, with $t \in \{t_0 = 0, \ldots, t_{m-1} = T\}$ and T fixed. Here, the parameter of interest is the $m \times 1$ conditional mean vector, $\psi_0(\cdot, W) = E_0[Z(\cdot) \mid W]$. A corresponding loss function can be defined as $L(X, \psi) = (Z(\cdot) - \psi(\cdot, W))^\top \Omega(W)(Z(\cdot) - \psi(\cdot, W))$, for a symmetric matrix function $\Omega(W)_{m \times m}$. A natural choice for $\Omega(W)$ is the inverse of the conditional covariance matrix, $\Sigma(W)$, of the outcome process $Z(t)$ given covariates W, $\Sigma(W) = E_0[(Z(\cdot) - E_0[Z(\cdot) \mid W])(Z(\cdot) - E_0[Z(\cdot) \mid W])^\top \mid W]$. Such a loss function takes into account the dependence structure among responses. As in the univariate outcome case, the corresponding IPCW observed data loss function is

$$L(O, \psi \mid \eta_n) = (Z(\cdot) - \psi(\cdot, W))^\top \Omega(W)(Z(\cdot) - \psi(\cdot, W)) \frac{\Delta}{\bar{G}_n(T|X)}. \tag{7}$$

Note that using this type of loss function for regression trees amounts to creating partitions of the time-covariate space using transformed outcomes $\Omega(W)^{1/2} Z(\cdot)$, where different choices of $\Omega(W)$ correspond to different notions of distance. Although risk is minimized by the conditional mean vector $\psi_0(\cdot, W) = E_0[Z(\cdot) \mid W]$ for arbitrary $\Omega(W)$, different choices of $\Omega(W)$ lead to estimators with different properties. In practice, one may work with a matrix $\Omega(W)$ that is diagonal, constant in W, or has a particular parametric representation. Previous approaches for multivariate outcome prediction in the context of linear regression have relied on canonical analysis to perform regression on transformed versions of the outcome [3].

2.2.5. Density estimation

The parameter of interest is the joint density, $\psi_0(W, T) = f_0(W, T)$, and the loss function is the negative log-likelihood, $L(X, \psi) = -\log \psi(W, T)$ (cf. Kullback–Leibler divergence). Again, the corresponding IPCW observed data loss function is simply

$$L(O, \psi \mid \eta_n) = -\log \psi(W, T) \frac{\Delta}{\bar{G}_n(T|X)}. \tag{8}$$

The resulting joint density estimator can then be used to obtain the conditional survivor or hazard functions given covariates W.

As in previously proposed survival tree methods, one could also use as loss function the negative log-likelihood for the observed data

$$L(O, f) = -\log p_f(o), \tag{9}$$

where $p_f(o) = p_{F_X}(o) = f_{F_X}(\bar{X}(t)) \mid_{t=c}$ is the F_X-part of the observed data likelihood under CAR and f denotes the joint density corresponding to F_X (Section 2.1.2). Indeed, the risk $E_0[L(O, f)]$ is minimized at the true underlying density $f = f_0$. Different procedures consider different models for f within each node [1,2,6,16].

One should keep in mind the following issues when choosing between the IPCW loss function $L(O, f \mid \eta_0)$ and the observed data negative log-likelihood loss function $L(O, f)$. Firstly, the choice $L(O, f)$ corresponds to minimizing the risk $E_0[L(O, f)] =$

$-\int \log p_f(o)\, dP_0(o)$, which involves the underlying data generating distribution $P_0 = P_{F_{X,0},G_0}$, while we might only be concerned with the risk $E_0[L(X,f)] = -\int \log f(x)\, dF_{X,0}(x)$, which does not depend on G_0. Secondly, unlike the IPCW or DR-IPCW loss functions, the $L(O,f)$ choice has the advantage that it does not require estimating the nuisance parameter η_0. Thirdly, in order to handle censored observations in likelihood calculations, methods based on the observed data loss function $L(O,f)$ assume coarsening at random, i.e., independence of the survival and censoring times given covariates. For example, for a within-node Cox proportional hazards model, consistent parameter estimation relies on independence of the survival and censoring times given covariates in the model. The implications of the coarsening at random assumption depend on the complexity of the model under consideration and should be most problematic in the early stages of procedures based on forward selection, such as tree estimators. However, such modeling assumptions are made for the purpose of generating candidate estimators and the final selected estimator may still be a good estimator of the density f_0.

2.3. *Constructing piecewise constant candidate estimators based on censored data*

In general, it is not feasible to consider all possible candidate estimators ψ in the parameter space $\boldsymbol{\Psi}$ and, in Step 2 of the road map, one generates a sequence of candidates according to some search procedure. Tree-based estimators correspond to one such procedure, analogous to forward selection (node splitting) followed by backward deletion (tree pruning). Define a *sieve*, $\{\boldsymbol{\Psi}_k\}$, of subspaces $\boldsymbol{\Psi}_k \subseteq \boldsymbol{\Psi}$ of increasing dimension approximating the complete parameter space $\boldsymbol{\Psi}$

$$\boldsymbol{\Psi}_k \equiv \left\{ \psi_{I,\beta}(\cdot) = \sum_{j \in I} \beta_j \phi_j(\cdot) : \beta, I, \, |I| \leqslant k \right\}, \tag{10}$$

where $I \subset \mathbb{N}$ denote finite *index sets* and $\beta = (\beta_1, \ldots, \beta_{|I|}) \in \mathbb{R}^{|I|}$ are corresponding *regression coefficients*. The *basis functions* ϕ_j are set indicators, $\phi_j(s) \equiv \mathrm{I}(s \in S_j)$, and the subsets $S_j \subseteq \mathcal{S}, j \in I$, of the covariate space $\mathcal{S}$ are disjoint ($S_j \cap S_{j'} = \emptyset,\ j \neq j'$) and exhaustive ($\mathcal{S} = \cup_{j \in I} S_j$). The goal is to identify for each k the parameter $\psi_{0k} \in \boldsymbol{\Psi}_k$ with minimum risk, $\psi_{0k} \equiv \mathrm{argmin}_{\psi \in \boldsymbol{\Psi}_k} E_0[L(X,\psi)] = \mathrm{argmin}_{\psi \in \boldsymbol{\Psi}_k} E_0[L(O,\psi \mid \eta_0)]$. In practice, one seeks the empirical analogue, $\hat{\psi}_k$, which minimizes the *empirical risk*, i.e., the *resubstitution error*,

$$\hat{\psi}_k \equiv \mathrm{argmin}_{\psi \in \boldsymbol{\Psi}_k} \int L(o, \psi \mid \eta_n)\, dP_n(o), \tag{11}$$

where η_n represents an estimator of the nuisance parameter η_0 derived under the CAR assumption for the censoring mechanism.

Tree-structured estimators such as CART [4] do not search over all index sets I with $|I| \leqslant k$, but rather approximate the minimum by recursive binary partitioning of the covariate space $\mathcal{S}$ according to a loss-based node splitting rule. In this setting, the S_j correspond to terminal nodes and k indexes the 'size' of the tree, measured by the number of terminal nodes $|I| = k$ (or by the complexity parameter α, as described in Section 2.4, below); in particular, for $k = 1$, S_1 is the root node, $\mathscr{S}$. Thus, as detailed next, trees tackle the optimization problem in Eq. (11) in two steps: generation of the index sets I by a forward partitioning algorithm and minimization over coefficients β for a given index set I.

2.3.1. Within-node estimation: minimizing risk over coefficients β for a given index set I

Given an index set I, the node coefficients β are defined as the minimizers $\hat{\beta}_I$ of the empirical risk

$$\begin{aligned}\hat{\beta}_I &\equiv \mathrm{argmin}_\beta \int L(o, \psi_{I,\beta} | \eta_n)\, dP_n(o) \\ &= \mathrm{argmin}_\beta \sum_{i=1}^{n} L(X_i, \psi_{I,\beta}) \frac{\Delta_i}{\bar{G}_n(T_i|X_i)}.\end{aligned} \tag{12}$$

This generally involves solving the following estimating equation:

$$0 = \sum_{i=1}^{n} \frac{d}{d\beta} L(X_i, \psi_{I,\beta}) \frac{\Delta_i}{\bar{G}_n(T_i|X_i)}.$$

The resulting estimator is denoted by $\hat{\psi}_I$. Below are solutions for each of the three loss functions defined in Sections 2.2.3–2.2.5.

Univariate outcome prediction: For the quadratic loss function, $L(X_i, \psi_{I,\beta}) = (Z_i - \psi_{I,\beta}(W_i))^2 = (Z_i - \beta_j)^2$, if $W_i \in S_j$. Hence

$$\hat{\beta}_I = \mathrm{argmin}_\beta \sum_{j \in I} \sum_{i=1}^{n} \mathrm{I}(W_i \in S_j)(Z_i - \beta_j)^2 \frac{\Delta_i}{\bar{G}_n(T_i|X_i)}$$

and

$$\hat{\beta}_{I,j} = \frac{1}{\sum_{i=1}^{n} \mathrm{I}(W_i \in S_j) \frac{\Delta_i}{\bar{G}_n(T_i|X_i)}} \sum_{i=1}^{n} \mathrm{I}(W_i \in S_j) \frac{\Delta_i}{\bar{G}_n(T_i|X_i)} Z_i, \quad j \in I.$$

Thus, the coefficients $\hat{\beta}_I$ are weighted means of the outcome in nodes $S_j, j \in I$. In the absence of censoring $(\Delta_i \equiv 1, C_i \equiv \infty)$, $\hat{\beta}_{I,j}$ reduces to the standard regression tree prediction, that is, to the average outcome in node S_j.

Multivariate outcome prediction: In this setting, $\psi_{I,\beta}(t, W_i) = \beta_j$ if $(t, W_i) \in S_j$, thus, the same observation O_i can contribute to different nodes depending on time t. For the quadratic loss function, $L(X_i, \psi_{I,\beta}) = (Z_i(\cdot) - \psi_{I,\beta}(\cdot, W_i))^\top \Omega(W_i)(Z_i(\cdot) - \psi_{I,\beta}(\cdot, W_i))$ and $\psi_{I,\beta}(\cdot, W_i)$ can be rewritten as

$$\psi_{I,\beta}(\cdot, W_i) = \sum_{j \in I} \mathrm{I}((\cdot, W_i) \in S_j)\beta_j = \tilde{W}_i(I)\beta$$

for an $m \times k$ matrix of indicators, $\tilde{W}_i(I)$, with (t,j)th entry equal to $\mathrm{I}((t, W_i) \in S_j)$ and row sums of one. Thus, the $k \times 1$ vector $\hat{\beta}_I$ has the form of a *generalized* *least squares estimator*

$$\begin{aligned}\hat{\beta}_I &= \operatorname{argmin}_\beta \sum_{i=1}^{n} (Z_i(\cdot) - \tilde{W}_i(I)\beta)^\top \Omega(W_i)(Z_i(\cdot) - \tilde{W}_i(I)\beta)\frac{\Delta_i}{\bar{G}_n(T_i \mid X_i)} \\ &= \left(\sum_{i=1}^{n} \frac{\Delta_i}{\bar{G}_n(T_i \mid X_i)} \tilde{W}_i(I)^\top \Omega(W_i)\tilde{W}_i(I)\right)^{-1} \\ &\quad \times \left(\sum_{i=1}^{n} \frac{\Delta_i}{\bar{G}_n(T_i \mid X_i)} \tilde{W}_i(I)^\top \Omega(W_i) Z_i(\cdot)\right).\end{aligned}$$

To see this, one can define a stacked $nm \times 1$ outcome vector, $Z = [Z_1(\cdot), \ldots, Z_n(\cdot)]^\top$, a stacked $nm \times k$ design matrix of indicators, $\tilde{W}(I) = [\tilde{W}_1(I), \ldots, \tilde{W}_n(I)]^\top$, and an $nm \times nm$ block diagonal matrix $\Omega(W)$ based on the $\Omega(W_i)$ and the IPCW weights. The risk criterion then becomes the standard generalized least squares criterion

$$\hat{\beta}_I = \operatorname{argmin}_\beta (Z - \tilde{W}(I)\beta)^\top \Omega(W)(Z - \tilde{W}(I)\beta).$$

Density estimation: For the negative log-likelihood loss function, $L(X_i, \psi_{I,\beta}) = -\log \psi_{I,\beta}(W_i, T_i) = -\log \beta_j$, if $(W_i, T_i) \in S_j$, hence

$$\hat{\beta}_I = \operatorname{argmin}_{\left\{\beta:\beta_j \geqslant 0, \sum_j \beta_j = 1\right\}} - \sum_{j \in I} \sum_{i=1}^{n} \mathrm{I}((W_i, T_i) \in S_j) \frac{\Delta_i}{\bar{G}_n(T_i|X_i)} \log \beta_j$$

and

$$\hat{\beta}_{I,j} = \frac{\sum_{i=1}^{n} \mathrm{I}((W_i, T_i) \in S_j)\Delta_i/\bar{G}_n(T_i|X_i)}{\sum_{j \in I}\sum_{i=1}^{n} \mathrm{I}((W_i, T_i) \in S_j)\Delta_i/\bar{G}_n(T_i|X_i)}, \quad j \in I.$$

The coefficients $\hat{\beta}_I$ are simply weighted proportions of observations falling in each node S_j, $j \in I$. Details on within-node likelihood calculations for other types of observed data log-likelihood loss functions are given in [1,2,6,16].

2.3.2. Node splitting: minimizing risk over index sets I

In tree-based estimation, the index sets I are obtained by recursive binary partitioning of the covariate space $\mathscr{S}$. Specifically, a new index set I' is obtained from the current I by considering all possible binary splits of each mother node S_j into a left and a right daughter node, $S_{L(j)}$ and $S_{R(j)}$, respectively. The split which results in the maximal decrease in empirical risk yields the new index set I', that is, one seeks I' that maximizes the empirical risk difference between candidates $\hat{\psi}_I$ and $\hat{\psi}_{I'}$

$$\int L(o, \hat{\psi}_I \mid \eta_n)\, dP_n(o) - \int L(o, \hat{\psi}_{I'} \mid \eta_n)\, dP_n(o).$$

In the univariate outcome prediction problem with the squared error loss, the risk difference for the split of node S_j into nodes $S_{L(j)}$ and $S_{R(j)}$ simplifies to

$$\begin{aligned}\int L(o,\hat{\psi}_I \mid \eta_n)\, dP_n(o) - \int L(o,\hat{\psi}_{I'} \mid \eta_n)\, dP_n(o) \\ = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(W_i \in S_j)\frac{\Delta_i}{\bar{G}_n(T_i|X_i)}(Z_i - \hat{\beta}_{I,j})^2 \\ - \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(W_i \in S_{L(j)})\frac{\Delta_i}{\bar{G}_n(T_i|X_i)}(Z_i - \hat{\beta}_{I',L(j)})^2 \\ - \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(W_i \in S_{R(j)})\frac{\Delta_i}{\bar{G}_n(T_i|X_i)}(Z_i - \hat{\beta}_{I',R(j)})^2,\end{aligned}$$

where $\hat{\beta}_I$ and $\hat{\beta}_{I'}$ are weighted node averages as derived in Section 2.3.1. Similarly, for density estimation, the risk difference only depends on observations in the mother node S_j. For multivariate outcome prediction, however, the same observation can contribute to different nodes. In basic CART, candidate splits of a node S_j into daughter nodes $S_{L(j)}$ and $S_{R(j)}$ are generated based on the values of individual covariates. For example, in the case of an ordered variable W, one considers binary partitions of S_j according to whether or not $W \leqslant a$, where the cut-offs a are chosen to be halfway between consecutive, distinct values of W in the learning set. For polychotomous variables, possible subsets of the categories are considered. Details and extensions (e.g., splits based on linear combinations and Boolean combinations of explanatory variables) are discussed in Chapters 4 and 5 of Breiman et al. [4].

2.4. Cross-validation for estimator selection and performance assessment with censored data

The approaches described in Section 2.3 can be used to construct a sequence of candidate tree estimators, $\hat{\psi}_k$, $k \in \{1, \ldots, K(n)\}$, up to a maximal tree, $\psi_{max} = \psi_{K(n)}$. Here, k indexes the size of the tree, measured by the number of terminal nodes. The size $K(n)$ of the maximal tree is typically determined by criteria such as minimal terminal node size for continuous outcomes or terminal node homogeneity (purity) for polychotomous outcomes. Cross-validation can then be applied to estimate risk for the candidates $\hat{\psi}_k$, based on the observed data loss function, and to select an optimal estimator among these.

In the standard CART methodology, once a maximal tree is grown, a *minimal cost-complexity pruning* algorithm is applied to generate a new sequence of candidate estimators indexed by a complexity parameter α. Specifically, a cost-complexity measure $R_\alpha(\psi)$ is defined for each candidate tree ψ as

$$R_\alpha(\psi) \equiv \int L(o,\psi \mid \eta_n)\, dP_n(o) + \alpha \mid \psi \mid, \tag{13}$$

where $|\psi|$ denotes the number of terminal nodes in the tree. Minimal cost-complexity pruning is applied to yield a nested decreasing sequence of subtrees $\{\hat{\psi}_\alpha\}$ as candidate estimators and cross-validation is used to select the complexity parameter α which minimizes risk [4, Chapter 3]. Note that CART's approach for generating candidate estimators can be viewed as forward selection (splitting) all the way to a maximal tree, followed by backward elimination (pruning), where the stopping rule in backward elimination is determined by cross-validation.

The unified cross-validation methodology of van der Laan and Dudoit [25] can be readily applied to extend the CART framework for pruning and performance assessment to multivariate outcome prediction and density estimation with censored data. All that is required is to replace the full data loss function used in CART by one of the observed (censored) data loss functions described in Section 2.2. van der Laan and Dudoit [25] derive finite sample and asymptotic optimality results concerning the cross-validation selector for general data generating distributions, loss functions (possibly depending on a nuisance parameter, η_0), estimators, and cross-validation procedures (e.g., V-fold, Monte-Carlo cross-validation). The asymptotic optimality result states that the cross-validation selector $\hat{k}$ performs asymptotically as well as an optimal benchmark selector, $\tilde{k} = \operatorname{argmin}_k \int L(o, \hat{\psi}_k \mid \eta_0)\, dP_0(o)$, based on the unknown data generating distribution P_0. That is,

$$\frac{\int L(o,\hat{\psi}_{\hat{k}} \mid \eta_0)\, dP_0(o) - \int L(o,\psi_0 \mid \eta_0)\, dP_0(o)}{\int L(o,\hat{\psi}_{\tilde{k}} \mid \eta_0)\, dP_0(o) - \int L(o,\psi_0 \mid \eta_0)\, dP_0(o)}$$
$$= \frac{\int L(x,\hat{\psi}_{\hat{k}})\, dF_{X,0}(x) - \int L(x,\psi_0)\, dF_{X,0}(x)}{\int L(x,\hat{\psi}_{\tilde{k}})\, dF_{X,0}(x) - \int L(x,\psi_0)\, dF_{X,0}(x)} \to 1 \quad \text{in probability,}$$

provided that, as $n \to \infty, p_n \to 0$, $\log(K(n))/np_n$ and $\int (\bar{G}_n - \bar{G}_0)^2 (T \mid X)\, dF_{X,0}$ both converge to zero faster than the rate at which the estimator $\hat{\psi}_{\tilde{k}}$ converges to the parameter ψ_0 in risk distance, i.e., faster than $\int L(x,\hat{\psi}_{\tilde{k}})\, dF_{X,0}(x) - \int L(x,\psi_0)\, dF_{X,0}(x) \to 0$, where p_n denotes the proportion of observations in the validation sets (see van der Laan and Dudoit [25] for full statements and proofs of the results).

3. Simulation study and data analysis

To evaluate our proposed data-adaptive loss-based estimation methodology and demonstrate its application to tree-structured estimation with censored data, we present the following results from a simulation study (Section 3.1) and analysis of breast cancer survival and CGH copy number data (Section 3.2).

3.1. Simulation study

The proposed survival tree approach based on the IPCW loss function was compared to that of LeBlanc and Crowley [16], which is implemented as a default for censored data in the R `rpart` function [13,24]. The loss function for the survival trees of LeBlanc and Crowley [16] is based on the observed data negative log-likelihood for a Cox proportional hazards model with the same baseline hazard for each node. Risk estimates used in splitting and pruning are based on the first step of a full likelihood estimation procedure. Trees based on the IPCW loss function can be grown using the `rpart` function, by setting the `method` argument to "`anova`" and by providing the IPCW weights for individual observations through the `weights` argument. Tree selection by cross-validation requires minor modifications to `rpart` (see details in [17]). The censoring survivor function, $\bar{G}_0$, used in the IPCW loss function, is estimated separately for each training set. In what follows, Method 1 and 2 refer, respectively, to the survival trees of LeBlanc and Crowley [16] and to trees grown using the proposed IPCW loss function. The two approaches differ in the choice of loss function for node splitting and tree pruning and thus lead to two different partitions of the covariate space, i.e., to different assignments of observations to terminal nodes. Given such a final partition, we consider two survival estimation methods for the terminal nodes: the IPCW mean survival and the Kaplan–Meier (KM) median survival approach. These two types of estimators correspond to full data parameters defined in terms of the squared and absolute error loss functions, respectively. The two different loss functions and the two different within-node estimation methods thus produce *four* different predictors of survival (namely, Method 1 with IPCW mean, Method 1 with KM median, Method 2 with IPCW mean, Method 2 with KM median), which were compared by simulation as described below.

Simulation model for full and observed data structures: The following model was considered for the full data structure: $Z \equiv \log T = W^2 + \varepsilon$, where W and ε are independent random variables with $W \sim U(0,1)$, $\varepsilon \sim N(0,\sigma^2)$, and $\sigma = 0.25$. Thus, $E_0[Z|W] = Median_0[Z|W] = W^2$ and the conditional survival function is given by $S_0(z \mid W) = Pr_0(Z > z \mid W) = 1 - \Phi((z - W^2)/\sigma)$, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. Censoring times C were simulated using a mixture of three uniform distributions: $Cens_1 \sim U(\min(Z), cut.dat)$, $Cens_2 \sim U(cut.dat, \max(Z))$, and $Cens_3 \sim U(\max(Z), \max(Z) + 2)$, where $\min(Z)$ and $\max(Z)$ refer, respectively, to the minimum and maximum of a random sample of Z's. The mixing proportions for $Cens_1$ and $Cens_2$ were fine-tuned to achieve the desired level of censoring, $Pr_0(\Delta = 0) = Pr_0(C < Z)$, while $Cens_3$ ensured that $Pr_0(\bar{G}_0(Z|W) > 0.1) = 1$, a condition for the IPCW method (Section 2.2.2). The censoring survivor function, $\bar{G}_0$, used in the IPCW loss function, was estimated separately for each training set, by fitting a Cox proportional hazards model to the survival time T and covariate W.

Simulation study design: The simulation study consisted of the following five steps, repeated $B = 100$ times for each of four sample sizes, $n = 250$, 600, 1250, and 6000.

First step: A learning set was generated from the above model with 20% censoring. *Second step*: Both Method 1 and 2 were applied to the learning set, resulting in two partitions of the covariate space. For each method, five-fold cross-validation was employed to select the 'best' tree (for Method 1, the default $1 - SE$ rule in `rpart` was used). *Third step*: For each terminal node in Method 1 and 2 trees, two survival estimators, the IPCW mean survival time and the KM median survival time, were computed. *Fourth step*: A large, independent test set, of size $N = 5000$, was generated from the full data distribution and partitioned according to both 'best' trees. *Fifth step*: For each test set observation, in each of the two trees, predicted survival times were obtained using the two different within-node estimation methods (resulting in a total of four predicted survival times for each test case).

Test set risk estimates were computed for each of the *four* predictors, using the L^2 loss function for the IPCW mean within-node estimation method and the L^1 loss function for the KM median estimation method. Within each sample size, the four test set risk estimates were averaged over the $B = 100$ repetitions. Method 1 and 2 were compared by forming the ratio of Method 2's average risk to that of Method 1, separately for each of the two within-node estimation methods. Ratios of average test set risk are displayed in Table 1 for both the KM median and IPCW mean estimation methods; ratios less than one correspond to improved accuracy for Method 2, i.e., for trees based on the new IPCW loss function. The results illustrate the impact on accuracy of the choice of loss function used for node splitting and tree pruning. As expected, when the parameter of interest is the conditional mean survival, the risk is smaller for partitions generated by Method 2 ("IPCW Mean" column). The IPCW loss function also corresponds to lower risk when interest is in estimating the median survival. The difference in risk decreases with increasing sample size.

Table 1
Simulation study

Sample size, n	Ratios of average risk	
	KM median	IPCW mean
250	0.9422	0.8838
600	0.9524	0.9062
1250	0.9629	0.9244
6000	0.9767	0.9533

Comparison of survival trees grown with Method 1 (`rpart`'s default) and Method 2 (proposed IPCW loss function). Ratios of average risk for Method 2 to 1 are displayed for the KM median and IPCW mean within-node estimation methods for four sample sizes, n. Individual entries of the table are ratios of average test set risk, $(1/B)\sum_{b=1}^{B} \int L(x, \hat{\psi}_n^b)\, dP_N^b(x)$, where P_n^b and P_N^b denote, respectively, the learning set and test set empirical distributions in the bth simulation, $\hat{\psi}_n^b$ refers to one of the four survival predictors based on the bth simulated learning set P_n^b, $N = 5000$, and $B = 100$. For the KM median within-node estimation method (column 2), L is the absolute error loss, and for the IPCW mean within-node estimation method (column 3), L is the squared error loss.

3.2. Breast cancer survival and CGH copy number data analysis

Our censored regression tree method was also applied to a dataset from a *comparative genomic hybridization* (CGH) study of breast cancer patients. Data were collected on 152 patients, all with initial occurrences of breast cancer; 52 subsequently recurred. Time to event (in years) was defined as time to recurrence. Patients with no recurrence at the time of death or of final follow-up are censored. According to these definitions, the censoring percentage is 66%. Explanatory variables include epidemiological variables (e.g., age at diagnosis, race), histopathological variables (e.g., tumor stage, grade), and DNA copy number measures from a CGH microarray with 2254 bacterial artificial chromosomes (BACs). Details on CGH and on the particular dataset are described in a forthcoming manuscript by members of the UCSF Comprehensive Cancer Center (Waldman et al., in preparation).

The 152 observations were split at random into a learning set and a test set of 128 and 24 (i.e., five sixths and one sixth) observations, respectively, while retaining the appropriate level of censoring. Trees were grown using the learning set and their overall performance assessed on the test set. Five-fold cross-validation of the learning set was used to select the 'best' tree (again, retaining the appropriate level of censoring). The censoring survivor function, $\bar{G}_0$, used in the IPCW loss function, was estimated separately for each of the five training sets in the cross-validation, by fitting a Cox proportional hazards model to the epidemiological and histopathological variables. For each training set, a maximal exploratory tree was grown using the R `rpart` function with the `weights` argument set to the IPCW estimates corresponding to the particular training observations and with `cp=0` [24]. The training set estimate of $\bar{G}_0$ was maintained in the IPCW loss function and used on the validation set to evaluate the candidate subtrees. The risk for each subtree was then averaged over the five validation sets. The minimum cross-validated risk was achieved for a two-node tree, i.e., with only one split.

A tree was then grown with the entire learning set and the resulting predictor was assessed using the independent test set. The possible numbers of splits for this tree were 0, 2, 3, or 4, with corresponding test set IPCW mean squared error 2.530699, 2.349634, 2.535852, and 2.701512. Since cost complexity pruning does not always return a sequence of trees corresponding to all possible numbers of splits, one needs to choose between zero and two splits. This could be done in principle by testing whether the risk difference between the two trees is equal to 0 using a standard t-statistic. The full learning set tree is shown in Fig. 1, with filled circles for the two-split subtree. Each terminal node is described by the IPCW mean log survival time (in years) and the number of observations. The legend in the bottom left corner indicates the chromosomal location of each BAC. The first two splits are based on BACs that fall in chromosomal regions known to contain genes related to breast cancer (personal communication with Joe Gray and Fred Waldman). The default `rpart` method selected only the root node; the `rpart` maximal tree was based on different variables (BACs) than those used in the IPCW loss function trees.

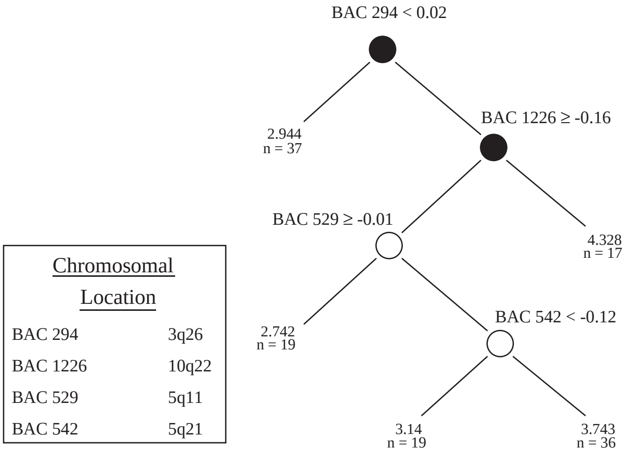

Fig. 1. *Breast cancer survival and CGH copy number data analysis.* Survival tree built from the learning set of 128 patients, using the IPCW squared error loss function. Each terminal node is described by the IPCW mean log survival time (in years) and the number of observations.

4. Discussion

We have described an application of our general loss-based estimation methodology with cross-validation [25] to tree-structured estimation with censored data. The approach encompasses univariate outcome prediction, multivariate outcome prediction, and density estimation, simply by defining a suitable loss function for each of these problems. Censored data are handled by mapping the full, uncensored data loss function into an observed, censored data loss function having the same expected value [28]. This approach reconciles censored and full data estimation methods, in the sense that standard full data estimators are recovered as special cases of censored data estimators. In addition, the IPCW and DR-IPCW loss functions allow for informative censoring and can be used for any type of prediction problem, including standard linear regression, logic regression, and bagging and boosting procedures [14,21]. Previously proposed survival tree methods, such as those of Breiman [1,2], Davis and Anderson [6], and LeBlanc and Crowley [16], correspond to different choices for the observed data negative log-likelihood loss function.

The simulation study of Section 3 illustrated that the choice of loss function used for node splitting and tree pruning can have a significant impact on accuracy. It also showed that gains in accuracy can be obtained by using a loss function that is specific to the parameter of interest. Analysis of a breast cancer survival and CGH dataset using trees built with the IPCW squared error loss function identified two BACs known to be implicated in breast cancer. However, this preliminary analysis also highlighted limitations of single trees based on microarray measures: they typically involve a very small number of splits and therefore only provide limited biological insight. Improved prediction accuracy and more information on chromosomal regions related to breast cancer survival may be obtained from aggregation methods such as bagging and boosting. We are also exploring more aggressive procedures for generating candidate estimators (see below), that include "OR" statements, in addition to the "AND" statements of tree estimators, and that are more specific to CGH data [18].

Tree-structured estimators correspond to one particular approach for generating the candidates in Step 2 of the road map, analogous to forward selection (node splitting) followed by backward elimination (tree pruning). Current problems in genomics (e.g., DNA microarray experiments, genetic mapping using SNPs) involve the analysis of high-dimensional datasets with complex interactions among variables. In this setting, it is particularly important to perform an efficient search of the parameter space to generate a good sequence of candidate estimators. Dudoit et al. [8], van der Laan and Dudoit [25], van der Laan et al. [27], Molinaro and van der Laan [18], and Sinisi and van der Laan [23] discuss more general sieves and more aggressive search strategies based on deletion/substitution/addition (or D/S/A) algorithms capable of revealing high-order interactions among variables. Other ongoing efforts include deriving loss-based measures of variable importance and the development of software implementing the new methodology.

Section 2.2 alluded to the fact that a given parameter of interest, ψ_0, can arise as the risk minimizer for a number of different loss functions, say $L_1, \dots, L_m$ (e.g., different loss functions for classification trees in Section 2.2.3; different choices of quadratic loss function for multivariate outcome prediction in Section 2.2.4; different models for the negative log-likelihood loss function in density estimation in Section 2.2.5). Natural questions then include: choosing suitable full data loss functions for generating candidate estimators and for overall performance assessment and, given a particular choice of loss function, obtaining an efficient estimator of the corresponding risk. While the later question was discussed in Sections 2.2.2 and 2.4, the former deserves further study. Although risk is minimized by the same parameter ψ_0 for each L_j, i.e., $\int L_j(x, \psi_0)\, dF_{X,0}(x) = \min_{\psi \in \mathbf{\Psi}} \int L_j(x, \psi)\, dF_{X,0}(x)$, $\forall j = 1, \dots, m$, different choices for the loss function lead to estimators of ψ_0 with different properties. In particular, minimizing the empirical risk for a loss function L_1 could yield an estimator with lower risk for a second loss function L_2, than the empirical risk minimizer for L_2, i.e., one can have $\int L_2(x, \hat{\psi}_1)\, dF_{X,0}(x) \leqslant \int L_2(x, \hat{\psi}_2)\, dF_{X,0}(x)$, where $\hat{\psi}_j = \operatorname{argmin}_{\psi \in \mathbf{\Psi}} \int L_j(x, \psi)\, dP_n(x)$, $j = 1, 2$ (cf. generalized least squares estimation). In other words, it may be advantageous to use a different loss function for generating candidate estimators and for overall performance assessment. One could envisage employing a collection of loss functions, $L_1, \dots, L_m$, to generate candidate estimators and then applying cross-validation to select among these candidates using another loss function L^* for overall performance assessment. We are further investigating the loss function selection issue.

Software

The new tree-based estimation methods for censored data will be implemented in an R software package to be released on CRAN (`cran.r-project.org/`). In the meantime, sample R code for growing trees using the IPCW loss function can be found in the appendix of Molinaro et al. [17] (`www.bepress.com/ucbbiostat/paper135/`) and downloaded from `www.stat.berkeley.edu/~molinaro`.

Acknowledgments

We would like to thank Joe Gray, Dan Moore, and Fred Waldman, from the Comprehensive Cancer Center at the University of California, San Francisco, for graciously providing the CGH dataset, biological insight, and fruitful discussions. We are also grateful to Terry Therneau and Elizabeth Atkinson for their thorough explanation of the R `rpart` package.

References

[1] L. Breiman, Software for the masses, in: Wald Lectures, Meeting of the Institute of Mathematical Statistics, Banff, Canada, 2002, URL: `www.stat.berkeley.edu/~breiman`.

[2] L. Breiman, How to Use Survival Forests, Department of Statistics, University of California, Berkeley, 2003, URL: `www.stat.berkeley.edu/~breiman`.

[3] L. Breiman, J.H. Friedman, Predicting multivariate responses in multiple linear regression, J. Roy. Statist. Soc. Ser. B 59 (1) (1997) 3–54.

[4] L. Breiman, J.H. Friedman, R. Olshen, C.J. Stone, Classification and Regression Trees, The Wadsworth Statistics/Probability Series, Wadsworth International Group, Belmont, CA, 1984.

[5] A. Ciampi, J. Thiffault, J.P. Nakache, B. Asselain, Stratification by stepwise regression, correspondence analysis and recursive partition, Comput. Statist. Data Anal. 4 (1986) 185–204.

[6] R. Davis, J. Anderson, Exponential survival trees, Statist. Med. 8 (1989) 947–961.

[7] S. Dudoit, M.J. van der Laan, Asymptotics of cross-validated risk estimation in model selection and performance assessment, Technical report 126, Division of Biostatistics, University of California, Berkeley, 2003, URL: `www.bepress.com/ucbbiostat/paper126/`.

[8] S. Dudoit, M.J. van der Laan, S. Keleş, A.M. Molinaro, S.E. Sinisi, S.L. Teng, Loss-based estimation with cross-validation: applications to microarray data analysis, SIGKDD Explor. Microarray Data Min Special Issue, 2004, to appear. URL: www.bepress.com/ucbbiostat/paper137.

[9] R.D. Gill, M.J. van der Laan, J.R. Robins, Coarsening at random: characterizations, conjectures and counter-examples, in: D.Y. Lin, T.R. Fleming (Eds.), Proceedings of the First Seattle Symposium in Biostatistics, 1995, Springer Lecture Notes in Statistics, Springer, Berlin, 1997, pp. 255–294.

[10] L. Gordon, R. Olshen, Tree-structured survival analysis, Cancer Treatment Rep. 69 (1985) 1062–1069.

[11] E. Graf, C. Schmoor, W. Sauerbrei, M. Schumacher, Assessment and comparison of prognostic classification schemes for survival data, Statist. Med. 18 (1999) 2529–2545.

[12] T. Hothorn, B. Lausen, A. Benner, M. Radespiel-Tröger, Bagging survival trees, Statist. Med. 23 (1) (2004) 77–91.

[13] R. Ihaka, R.C. Gentleman, R: a language for data analysis and graphics, J. Comput. Graphical Statist. 5 (1996) 299–314.

[14] S. Keleş, M.J. van der Laan, S. Dudoit, Asymptotically optimal model selection method for regression on censored outcomes, Bernoulli, 2004, to appear. URL: www.bepress.com/ucbbiostat/paper124/.

[15] S. Keleş, M.J. van der Laan, S. Dudoit, M.B. Eisen, B. Xing, Supervised detection of regulatory motifs in DNA sequences, Statist. Appl. Genetics Mol. Bio. 2(1) (2003b) Article 5. URL: www.bepress.com/sagmb.

[16] M. LeBlanc, J. Crowley, Relative risk trees for censored survival data, Biometrics 48 (1992) 411–425.

[17] A.M. Molinaro, S. Dudoit, M.J. van der Laan, Tree-based multivariate regression and density estimation with right-censored data, Technical report 135, Division of Biostatistics, University of California, Berkeley, 2003. URL: www.bepress.com/ucbbiostat/paper135/.

[18] A.M. Molinaro, M.J. van der Laan, A Deletion/Substitution/Addition algorithm for partitioning the covariate space in prediction, Technical report, Division of Biostatistics, University of California, Berkeley, 2004, in preparation.

[19] M. Morgan, J.A. Sonquist, Problems in the analysis of survey data and a proposal, J. Amer. Statist. Assoc. 58 (1963) 415–434.

[20] J. Robins, A. Rotnitzky, Recovery of information and adjustment for dependent censoring using surrogate markers, chapter AIDS Epidemiology, Methodological issues, Bikhauser, Basel, 1992.

[21] I. Ruczinski, C. Kooperberg, M. LeBlanc, Logic regression, J. Comput. Graphical Statist. 12 (3) (2003) 474–511 URL: biostat.jhsph.edu/ iruczins/publications/publications.html.

[22] M. Segal, Regression trees for censored data, Biometrics 44 (1988) 35–48.

[23] S.E. Sinisi, M.J. van der Laan, Loss-based cross-validated Deletion/Substitution/Addition algorithms in estimation, Technical report 143, Division of Biostatistics, University of California, Berkeley, 2004, URL: www.bepress.com/ucbbiostat/paper143.

[24] T. Therneau, E. Atkinson, An introduction to recursive partitioning using the rpart routine, Technical report 61, Section of Biostatistics, Mayo Clinic, Rochester, 1997.

[25] M.J. van der Laan, S. Dudoit, Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive ε-net estimator: finite sample oracle inequalities and examples, Technical report 130, Division of Biostatistics, University of California, Berkeley, 2003, URL: www.bepress.com/ucbbiostat/paper130/.

[26] M.J. van der Laan, S. Dudoit, S. Keleş, Asymptotic optimality of likelihood-based cross-validation, Statis. Appl. Genetics Mol. Bio. 3(1) (2004) Article 4. URL: www.bepress.com/sagmb.

[27] M.J. van der Laan, S. Dudoit, A.W. van de Vaart, The cross-validated adaptive ε-net estimator, Technical report 142, Division of Biostatistics, University of California, Berkeley, 2004. URL: www.bepress.com/ucbbiostat/paper142.

[28] M.J. van der Laan, J.M. Robins, Unified Methods for Censored Longitudinal Data and Causality, Springer, New York, 2003.