Electronic Journal of Statistics
Vol. 10 (2016) 2685–2716
ISSN: 1935-7524
DOI: 10.1214/16-EJS1189

Tree-based censored regression with applications in insurance

Olivier Lopez[*]

Sorbonne Universités, UPMC Université Paris VI, CNRS FRE 3684, Laboratoire de Statistique Théorique et Appliquée, 4 place Jussieu 75005 PARIS, France
e-mail: olivier.lopez@upmc.fr

Xavier Milhaud[†]

Univ Lyon, UCBL, LSAF EA2429, ISFA, F-69007, LYON, France
CREST (LFA lab), 15, Boulevard Gabriel Péri, 92245 MALAKOFF CEDEX
e-mail: xavier.milhaud@univ-lyon1.fr

and

Pierre-E. Thérond[†]

Univ Lyon, UCBL, LSAF EA2429, ISFA, F-69007, LYON, France
Galea & Associés, 12 Avenue du Maine, 75015 PARIS, France
e-mail: pierre@therond.fr

Abstract: We propose a regression tree procedure to estimate the conditional distribution of a variable which is not directly observed due to censoring. The model that we consider is motivated by applications in insurance, including the analysis of guarantees that involve durations, and claim reserving. We derive consistency results for our procedure, and for the selection of an optimal subtree using a pruning strategy. These theoretical results are supported by a simulation study, and two applications involving insurance datasets. The first concerns income protection insurance, while the second deals with reserving in third-party liability insurance.

MSC 2010 subject classifications: Primary 62N01, 62N02, 62G08; secondary 91B30, 97M30.
Keywords and phrases: Survival analysis, censoring, regression tree, model selection, insurance.

Received October 2015.

Contents

[*]Olivier Lopez acknowledges partial financial support from the research program New Challenges for New Data of LCL and GENES.

[†]This research received partial support from ANR project *LoLitA: Dynamic population models for human longevity with lifestyle adjustments* and from the Research Chair *Data Analytics and Models in Insurance* (DAMI).

Introduction

In numerous applications of survival analysis, analyzing the heterogeneity of a population is a key issue. For example, in insurance, many risk evaluations are linked with the analysis of duration variables, such as lifetime, time between two claims, time between the opening of a claim and its closure. A strategic question is then to determine clusters of individuals which represent different levels of risk. Once such groups have been identified, it becomes possible to improve pricing, reserving or marketing targeting. In this paper, we show how to adapt CART methodology (Classification And Regression Trees) to a survival analysis context, with such applications in mind. The presence of censoring is a specific feature of data involving duration variables. Here, these variables appear naturally in the applications we consider, either because we are focusing on lifetimes, or because we are interested in quantities that are observed only when some event has occurred (typically, the final settlement of a claim). The procedure we develop is shown to be consistent, while its practical behavior is investigated through a simulation study and two real dataset analyses.

The CART procedure (Breiman et al. (1984)) is a natural candidate for dealing with such problems, since it simultaneously provides a regression analysis (which allows us to consider nonlinearity in the way the response depends on covariates) and a clustering of the population under study. Moreover, its tree-based algorithmic simplicity makes it easy to implement. It consists of successively splitting the population into less heterogeneous groups. A model selection step then allows us to select from this recursive partition a final subdivision into groups of observations of reasonable size, with simple classification rules to affect an individual to one of these classes. Tree-based methods have met with many successes in medical applications, due to the need for clinical researchers to define interpretable classification rules for understanding the prognostic structure of data (see e.g., Fan, Nunn and Su (2009), Gao, Manatunga and Chen (2004), Ciampi, Negassa and Lou (1995), Bacchetti and Segal (1995)). In survival analysis, a recent review of these methods can be found in Bou-Hamad, Larocque and Ben-Ameur (2011). Let us also mention Wey, Wang and Rudser (2014), who recently considered tree-based estimation of a censored quantile regression model, which extends the methodology of Wang and Wang (2009). For insurance applications, Olbricht (2012) highlighted their usefulness to approximate mortality curves in a reinsurance portfolio and compare them to German life tables in a nonparametric way, but based on fully observed data, which is not the case in the present paper.

As already mentioned, one of the most delicate problems when dealing with survival analysis is the presence of censoring in the data, and the necessity to correct the bias it introduces when using statistical methods. Our approach is based on the IPCW strategy ("Inverse Probability of Censoring Weighting"), see van der Laan and Robins (2003), Chapter 3.3. It consists in determining a weighting scheme that compensates the lack of complete observations in the sample. Therefore, our procedure is connected with the technique presented in Molinaro, Dudoit and van der Laan (2004). The main differences in our approach involve the specificity of the weighting scheme we consider (based on the Kaplan-Meier estimator of the censoring distribution) and the fact that we do not only focus on a duration (subject to censoring); our interest lies in the conditional distribution of a related variable, which is observed only if the duration is. This particular framework is motivated by applications in insurance where the final claim amount to be paid is known only after the claim has been settled, which can take several years in some cases. Another difference with Molinaro, Dudoit and van der Laan (2004) is that their approach requires modeling the conditional distribution of the censoring. In our case, no such model is required since we use weights based on a Kaplan-Meier estimator (Kaplan and Meier (1958)), and our strategy relies on Kaplan-Meier integrals (see e.g., Stute (1999), Gannoun et al. (2005) and Lopez, Patilea and Van Keilegom (2013) for the application of similar strategies to censored regression).

The paper is organized as follows. In Section 1, we describe specific details of the censored observations we consider. Section 2 is devoted to the description of the regression tree procedure, and its adaptation to the presence of censoring. Its consistency is shown in Section 3. A simulation study and two real data examples from the insurance field are respectively presented in Sections 4 and 5.

1. Observations and general framework

This section aims to describe the type of observations we have at our disposal (Section 1.1) and define the regression function we wish to estimate (Section 1.2). Section 1.3 is devoted to the nonparametric estimation of the distribution function of the variables involved in our model.

1.1. Censored observations

In the following, we are interested in a random vector $(M, T, \mathbf{X})$, where $M \in \mathbb{R}^p$, $T \in \mathbb{R}^+$ is a duration variable, and $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^d$ denotes a set of random covariates that may have an impact on T and/or M. The presence of censoring prevents the direct observation of (M, T), while $\mathbf{X}$ is always observed. Next, let us introduce a censoring variable $C \in \mathbb{R}^+$. For the sake of simplicity, we assume that T and C are continuous random variables. We also assume, for convenience but without loss of generality, that the components of M are all strictly positive. The variables that are observed instead of (M, T) are

$$\begin{aligned} Y &= \inf(T, C), \\ \delta &= \mathbf{1}_{T \leq C}, \\ N &= \delta M. \end{aligned}$$

The data is made up of i.i.d. replications $(N_i, Y_i, \delta_i, \mathbf{X}_i)_{1 \leq i \leq n}$. Compared to a classical censoring regression scheme, such as the one described for example in Stute (1993), the variables M_i correspond to quantities that are observed only when the individual i is fully observed. An illustration of such a case is described in Section 5.2, where T represents the time before a claim is fully settled, and M the total corresponding amount (only known at the end of the claim settlement process). The censored regression framework of Stute (1993) can be seen as a special case, taking $M = T$.

1.2. Regression function

Our aim is to understand the impact of $\mathbf{X}$, and possibly T, on M. More precisely, we wish to estimate a function

$$\pi_0 = \arg\min_{\pi \in \mathcal{P}} E\left[\phi(M, \pi(T, \mathbf{X}))\right], \qquad (1.1)$$

where $\mathcal{P}$ is a subset of an appropriate functional space and ϕ a loss function. In the following, we will restrict ourselves to real-valued functions π. Table 1 shows the different types of regression models corresponding to different possible choices of ϕ, and the corresponding set $\mathcal{P}$. These examples cover mean regression and quantile regression.

Table 1

Expressions for π_0 for some classical choices of ϕ and $\mathcal{P}$. The notation $L^p(\mathbb{R}^d)$ indicates a restriction to the set of functions $\pi(\boldsymbol{x}, t)$ which do not depend on t, and, for a random vector U, $q_{\tau,U}(u)$ denotes the τ-th conditional quantile of M with respect to U, that is, the value of m_u such that $\mathbb{P}(M \le m_u|U = u) = \tau$.

Function ϕ	$\mathcal{P}$	$\pi_0(t, \mathbf{x})$
$(m - \pi)^2$	$L^2(\mathbb{R}^d)$	$\pi_0(t, \mathbf{x}) = E[M\|\mathbf{X} = \mathbf{x}]$
	$L^2(\mathbb{R}^{d+1})$	$\pi_0(t, \mathbf{x}) = E[M\|T = t, \mathbf{X} = \mathbf{x}]$
$(m - \pi)(\tau - \mathbf{1}_{(m-u)\le 0})$	$L^1(\mathbb{R}^d)$	$\pi_0(t, \mathbf{x}) = q_{\tau,\mathbf{X}}(\mathbf{x})$
	$L^1(\mathbb{R}^{d+1})$	$\pi_0(t, \mathbf{x}) = q_{\tau,T,\mathbf{X}}(t, \mathbf{x})$

1.3. Estimation of the distribution function of $(M, T, \mathbf{X})$

In this framework, the empirical distribution function of $(M, T, \mathbf{X})$ cannot be computed, since M and T are not directly observed. Since most statistical methods rely on this nonparametric estimator, an effort should be made to finding an alternative estimator that takes censoring into account. Due to classical identifiability issues, an assumption on the way C depends on the variables $(M, T, \mathbf{X})$ must be specified. In the following, we assume that Assumption 1 below holds.

Assumption 1. *Assume that:*

1. *C is independent of (M, T),*
2. *$\mathbb{P}(T \le C \,|\, M, T, \mathbf{X}) = \mathbb{P}(T \le C \,|\, T)$.*

Under Assumption 1, observe that, for all functions $\psi \in L^1$,

$$E\left[\frac{\delta\psi(N, Y, \mathbf{X})}{1 - G(Y-)}\right] = E\left[\psi(M, T, \mathbf{X})\right], \tag{1.2}$$

where $G(t) = \mathbb{P}(C \le t)$. The function G is usually unknown. However, Assumption 1 ensures that it can be estimated consistently by the Kaplan-Meier estimator (see Kaplan and Meier (1958)), i.e.,

$$\hat{G}(t) = 1 - \prod_{Y_i \le t}\left(1 - \frac{\delta_i}{\sum_{j=1}^n \mathbf{1}_{Y_j \ge Y_i}}\right),$$

since T and C are independent, and $\mathbb{P}(T = C) = 0$ for continuous random variables (see Stute and Wang (1993) for consistency of Kaplan-Meier estimators). Therefore, a natural estimator of $F(m, t, \mathbf{x}) = \mathbb{P}(M \le m, T \le t, \mathbf{X} \le \mathbf{x})$ is

$$\hat{F}(m, t, \mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i \mathbf{1}_{N_i \le m, Y_i \le t, \mathbf{X}_i \le x}}{1 - \hat{G}(Y_i-)}, \tag{1.3}$$

while the integral

$$\int \psi(m, t, \mathbf{x}) d\hat{F}(m, t, \mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \frac{\delta_i \psi(N_i, Y_i, \mathbf{X}_i)}{1 - \hat{G}(Y_i-)}$$

is a consistent estimator of $E[\psi(M, T, \mathbf{X})]$ due to consistency of $\hat{G}$ and formula (1.2) under appropriate conditions. This type of approach can be linked with the IPCW method (van der Laan and Robins (2003), Chapter 3.3). In the case where $M = T$ (where we are only interested in time T), estimator (1.3) is the same as that defined by Stute (1993), due to a connection between $\hat{G}$ and the jumps of the Kaplan-Meier estimator of the distribution of T (see Satten and Datta (2001)).

Remark 1.1. *Assumption 1 is a natural extension of the identifiability condition considered by Stute (1993). Alternative assumptions have been proposed by several authors for censored regression. For example, Van Keilegom and Akritas (1999), Heuchenne and Van Keilegom (2010a), and Heuchenne and Van Keilegom (2010b) assume that T and C are independent conditionally on $\mathbf{X}$ (in the absence of an additional variable M). A special case of Assumption 1 is the situation where $(M, T, \mathbf{X})$ is independent of C. But, as shown in Stute (1993) (where Assumptions (i) and (ii) p. 91 are identical to ours in the case where $T = M$), Assumption 1 is more general. However, it still introduces constraints on the way C is allowed to depend on the covariates. An alternative would be to assume that (M, T) is independent of C conditionally on $\mathbf{X}$. A way to adapt our approach to this framework would be to replace the Kaplan-Meier estimator $\hat{G}$ by the conditional Kaplan-Meier estimator of Beran (1981) and Dabrowska (1989), as in Lopez (2011) (see also Lopez, Patilea and Van Keilegom (2013)). However, this complicates the procedure due to the introduction of kernel smoothing with respect to $\mathbf{X}$, with potentially erratic behavior when the dimension of the covariates d is high. We therefore restrain ourselves to the condition in Assumption 1, which is well-adapted to the practical applications we have in mind (see Section 5).*

Remark 1.2. *In practice, we use a learning sample to build the regression tree, and a validation sample to select the best-adapted subtree (further details in Section 2.3). Suppose that the learning sample is of size n, while there are v observations in the test sample. In this situation, the estimator $\hat{G}$ can be computed either from the learning sample (n observations) or from the whole sample ($n+v$ observations), this latter option leading to a slight modification in the definition of $\hat{G}$. As we will explain in Section 2.3, we use this second strategy in practice, which has no significant consequence in the theory, provided that v is at most of the same order as n.*

2. Adapting CART to survival data with Kaplan-Meier weights

This section is devoted to the description of our regression tree methodology, adapted to censoring. Section 2.1 explain the growing procedure, i.e., the successive partitions of the observations into elementary classes, while Section 2.2 shows the link between a subtree extracted from this procedure and an estimator of the regression function. Section 2.3 presents the pruning strategy for selecting our final estimator.

2.1. Growing the tree

The building procedure of a regression tree is based on the definition of a *splitting criterion* that furnishes partition rules at each step of the algorithm. More precisely, at each step s, a tree with L_s leaves is constituted, each of these leaves representing disjoint subpopulations of the initial n observed individuals. In our case, the rules used to create these populations are based on the values of Y and $\mathbf{X}$. More precisely, the leaves correspond to a partition of the space $\mathcal{T} = \mathbb{R}^+ \times \mathcal{X}$ into L_s disjoint sets $\mathcal{T}_1^{(s)}, ..., \mathcal{T}_{L_s}^{(s)}$. The individual i belongs to the subpopulation of the leaf l if $\tilde{\mathbf{X}}_i := (T_i, \mathbf{X}_i) \in \mathcal{T}_l^{(s)}$.

At step $s+1$, each leaf is likely to become a new node of the tree by making use of the splitting criterion. Let $\tilde{X}^{(j)}$ denote the j-th component of $\tilde{X}$. In the absence of censoring, to partition the subpopulation of the l-th leaf into two subpopulations, one determines, for each component $\tilde{X}^{(j)}$, the threshold $x_l^{(j)}$ that minimizes

$$\min_{(\pi,\pi')\in\Gamma^2} \left\{ \int \phi(m,\pi) \mathbf{1}_{\tilde{\mathbf{x}}\in\mathcal{T}_l^{(s)}} \mathbf{1}_{\tilde{x}^{(j)}\le x_l^{(j)}} d\hat{F}_n(m,t,\mathbf{x}) \right.$$
$$\left. + \int \phi(m,\pi') \mathbf{1}_{\tilde{\mathbf{x}}\in\mathcal{T}_l^{(s)}} \mathbf{1}_{\tilde{x}^{(j)}> x_l^{(j)}} d\hat{F}_n(m,t,\mathbf{x}) \right\} =: L_l(j, x_l^{(j)}), \tag{2.1}$$

where $\Gamma \subset \mathbb{R}$, $\tilde{\mathbf{x}} = (t, \mathbf{x})$, and $\hat{F}_n$ denotes the empirical distribution of $(M, T, \mathbf{X})$. The first term of (2.1) can be seen as an estimator of $E[\phi(M,\pi) \mid \tilde{\mathbf{X}} \in \mathcal{T}_l^{(s)}, \tilde{X}^{(j)} \le x_l^{(j)}]$, while the second term estimates $E[\phi(M,\pi) \mid \tilde{\mathbf{X}} \in \mathcal{T}_l^{(s)}, \tilde{X}^{(j)} > x_l^{(j)}]$. Then, one determines $j_0 = \arg\min_{j=1,..,d+1} L_l(j, x_l^{(j)})$. Next, the partition of the population of the l-th leaf is performed by separating the individuals having $\tilde{X}_i^{(j_0)} \le x_l^{(j_0)}$ from those for which $\tilde{X}_i^{(j_0)} > x_l^{(j_0)}$.

In our framework, the empirical distribution function $\hat{F}_n$ is unavailable. The idea is then to replace $\hat{F}_n$ in (2.1) by $\hat{F}$ defined in (1.3). In other words, in the previous regression tree procedure, the empirical means that we would use in the absence of censoring are replaced by weighted sums, with weight $W_{i,n} = \delta_i n^{-1}[1 - \hat{G}(Y_i-)]^{-1}$ being affected to the i-th observation in order to compensate the presence of censoring.

An important remark can be made in view of both the definition of the splitting criterion and the weights $W_{i,n}$. The splitting criterion consists of a rule which is based on the values of $\tilde{\mathbf{X}}$, whose first component T is unobserved for the censored individuals. Hence, under random censoring, this procedure cannot be understood as a rule to perform classification of all the observations in the sample; only uncensored individuals are classified. Nevertheless, the fact that the censored ones are not assigned to any leaf of the tree does not constitute an obstacle in view of performing the growing procedure. Indeed, if the i-th individual is censored, $W_{i,n} = 0$. Therefore, at each step, a censored observation could be assigned to any subpopulation without modifying the value of $L_l(j, x_l^{(j)})$. This does not mean that the information contained in censored observations is not used, since they play an important role in computing $\hat{G}$, and thus $W_{i,n}$.

To summarize, the aforementioned procedure thus produces clusters of individuals with rules to assign uncensored observations to one of them. The question about how to assign a censored observation should be considered separately; see the application in Section 5.2. The details of our modified CART algorithm (with censoring weights) are as follows:

Step 0: Compute the estimator $\hat{G}$ from the dataset with n individuals.

Step 1: initialization. Consider the tree with only one leaf ($L_1 = 1$), corresponding to the population composed of all n_U uncensored observations ($n_U \leq n$). Set $\mathcal{T}_1^{(1)} = \mathcal{T}$.

Step s: splitting. Consider the tree obtained at step $s-1$, with L_{s-1} leaves. Each leaf l corresponds to a set $T_l^{(s-1)}$ such that $\mathcal{T}_l^{(s-1)} \cap \mathcal{T}_{l'}^{(s-1)} = \emptyset$ and $\cup_l \mathcal{T}_l^{(s-1)} = \mathcal{T}$. The uncensored observations (denote e_l their number) such that $\tilde{\mathbf{X}} \in \mathcal{T}_l^{(s-1)}$ are assigned to leaf l. For each leaf l, with $1 \leq l \leq L_{s-1}$:

s.1 if $e_l = 1$ or if all observations have the same values of $\tilde{\mathbf{X}}$, do not split;

s.2 else, the leaf becomes a node in the next tree: determine the j_0 and $x_l^{(j_0)}$ that minimize $L_l(j, x_l^{(j)})$ and define $\mathcal{L}_l = \mathcal{T}_l^{(s-1)} \cap \{\tilde{X}^{(j_0)} \leq x_l^{(j_0)}\}$ and $\mathcal{U}_l = \mathcal{T}_l^{(s-1)} \cap \{\tilde{X}^{(j_0)} > x_l^{(j_0)}\}$.

Define a new collection of disjoint sets $\mathcal{T}_{l'}^{(s)}$ which consist of the sets $\mathcal{L}_l$, $\mathcal{U}_l$ for $1 \leq l \leq L_{s-1}$ (or $\mathcal{T}_l^{(s-1)}$ if the l-th leaf satisfied condition s.1). Set L_s the new number of leaves. Go to step $s+1$, unless $L_s = L_{s-1}$. The procedure stops when all leaves are in step s.1. This produces the *maximal tree* from which our final estimator is extracted.

2.2. From the tree to the regression function

Recall that our aim is to estimate the function π_0 in (1.1). Consider a subtree $\mathcal{S}$ of the maximal tree built from the algorithm in Section 2.1. We now describe how this subtree can be associated with an estimator of π_0. Let $K(\mathcal{S})$ denote the total number of leaves of $\mathcal{S}$. As previously explained, this subtree can be seen as a collection of rules (see Meinshausen (2009) for further formalization of this concept). By construction, a leaf l is associated with a set $\mathcal{T}_l$ (recall that the sets $\mathcal{T}_l$ are disjoint with their union equal to $\mathcal{T}$) and a rule $R_l(\tilde{\mathbf{x}}) = \mathbf{1}_{\tilde{\mathbf{x}} \in \mathcal{T}_l}$ that determines if an individual is affected or not to the corresponding cluster. This induces the following estimator of π_0:

$$\hat{\pi}^{\mathcal{S}}(t, \mathbf{x}) = \sum_{l=1}^{K(\mathcal{S})} \hat{\gamma}_l \, R_l(t, \mathbf{x}), \tag{2.2}$$

where

$$\hat{\gamma}_l = \arg\min_{\pi \in \Gamma} \int \phi(m, \pi) \, R_l(\tilde{\mathbf{x}}) \, d\hat{F}(m, t, \mathbf{x}).$$

The coefficient $\hat{\gamma}_l$ can be seen as an estimator of

$$\gamma_l = \arg\min_{\pi \in \Gamma} E[\phi(M, \pi) \,|\, \tilde{\mathbf{X}} \in \mathcal{T}_l].$$

Hence, defining

$$\pi^{\mathcal{S}}(t,\mathbf{x}) = \sum_{l=1}^{K(\mathcal{S})} \gamma_l \, R_l(t,\mathbf{x}),$$

$\pi^{\mathcal{S}}(t,\mathbf{x})$ can be seen as a piecewise-constant approximation of π_0, which tends to be closer to π_0 when the partition of $\mathcal{T}$ is finely spaced. On the other hand, $\hat{\pi}^{\mathcal{S}}$ should be close to $\pi^{\mathcal{S}}$ provided that the sets $\mathcal{T}_l$ are not too small. In view of estimating π_0, a crucial issue is thus to extract an appropriate subtree from the maximal tree, corresponding to a good compromise between a sharp partition of $\mathcal{T}$ and the necessity of having enough observations in each leaf to estimate well the coefficients γ_l. Achieving this is the aim of the pruning strategy developed in the following section.

Remark 2.1. *In the presence of right-censored observations, a classical difficulty is handling observations that are close to the right tail of the distribution. Indeed, little information is available on this part of the distribution due to the lack of large uncensored observations. Our procedure is impacted by this problem, which translates to a blowing up in the value of the weights when $1 - \hat{G}(Y_i-)$ tends to zero, that is, when Y_i is large.*

For this reason, a careful look at the leaves containing large observations is required. In the following, our theoretical results do not cover the case where weights may blow up. In other words, we exclude too-large uncensored observations from the procedure in order to avoid the instability they cause. This is a classical issue in censored regression, where, instead of π_0, one is often required to consider $\pi_0(\tau) = \arg\min_{\pi \in \mathcal{P}} E\left[\phi(M,\pi(T,\mathbf{X}))|T \leq \tau\right]$, where τ is strictly included in the support of Y, introducing a small bias.

2.3. Selection of a subtree: Pruning algorithm

Denote by $K_n \leq n$ the number of leaves of the maximal tree. The pruning strategy consists of selecting from the data a subtree $\hat{\mathcal{S}}$ with $\hat{K}$ leaves. Let S denote the set of subtrees of the maximal tree. The pruning strategy consists of determining $\hat{\mathcal{S}}(\alpha)$ such that

$$\hat{\mathcal{S}}(\alpha) = \underset{\mathcal{S}\in S}{\arg\min} \left\{ \int \phi(m,\hat{\pi}^{\mathcal{S}}(\mathbf{x},t)) d\hat{F}(m,t,\mathbf{x}) + \frac{\alpha K(\mathcal{S})}{n} \right\},$$

and to use $\hat{\pi}^{\hat{\mathcal{S}}(\alpha)}$ as a final estimator of π_0. We will denote $\hat{K}_\alpha$ the number of leaves in $\hat{\mathcal{S}}(\alpha)$. A penalty term proportional to $K(\mathcal{S})/n$ was first proposed by Breiman et al. (1984), see also Gey and Nedelec (2005). The procedure consists of starting with $\alpha = 0$, then progressively increasing its value, in order to determine a sequence $0 < \alpha_1 < ... < \alpha_{K_n}$ such that $\hat{K}_{\alpha_{j+1}} = \hat{K}_{\alpha_j}$. The existence of such a sequence has been proved by Breiman et al. (1984). Moreover, it follows from Breiman et al. (1984, p. 284–290) that $\mathcal{S}(\alpha_{j+1}) \subset \mathcal{S}(\alpha_j)$, and that $\mathcal{S}(\alpha) = \mathcal{S}(\alpha_j)$ for $\alpha_j \leq \alpha < \alpha_{j+1}$. Then, the question is to select the right α_j in this list. To this end, a test sample (see Remark 1.2) of size v is used. More precisely, let

$(N_i, Y_i, \delta_i, \mathbf{X}_i)_{n+1\leq i\leq n+v}$ denote the observations in the test sample. For all j, we compute

$$\mathcal{V}(\alpha_j) = \sum_{i=n+1}^{n+v} \frac{\delta_i\, \phi(N_i, \hat{\pi}^{K(\alpha_j)}(Y_i, \mathbf{X}_i))}{1 - \hat{G}(Y_i-)}, \tag{2.3}$$

and select α_{j_0} such that $\mathcal{V}(\alpha_j)$ is minimal. This procedure differs from the classical one by the introduction of the weights involving $\hat{G}$. Section 3.3 shows that this strategy remains valid in the presence of censoring.

Observe that different strategies may be used for computing the estimator $\hat{G}$ involved in (2.3). We choose to compute it once for all, i.e., using the whole sample $(N_i, Y_i, \delta_i, \mathbf{X}_i)_{i=1,\dots,n+v}$, and using this estimator both in the construction of the trees and in the validation step. Alternatively, one could use in the growing step an estimator $\hat{G}$ computed from the learning sample, and, in the validation step, another one computed from the test sample. We argue that such a strategy is likely to increase the instability of the procedure since the estimator $\hat{G}$ computed from the information contained in the test sample would usually have poorer performance (usually $v << n$). Therefore, taking an estimator $\hat{G}$ computed from the whole sample seems preferable, observing that correcting the presence of censoring and selecting the most appropriate tree are two separate problems.

Remark 2.2. *This selection criterion, in its uncensored version, has been shown to be consistent for selecting the best subtree in many cases, see Breiman et al. (1984) and Gey and Nedelec (2005). See also Molinaro, Dudoit and van der Laan (2004) for similar strategies. Optimality properties and practical evidence for some of these techniques can be found in van Der Laan, Dudoit and van der Vaart (2006), van Der Laan and Dudoit (2003), and Dudoit et al. (2003).*

3. Consistency of the CART weighted estimator

This section is devoted to the proof of the consistency of the tree procedure. The roadmap of the proof consists of the three following steps:

1. We consider the value of the criterion that we wish to optimize on each leaf of the tree. We provide a quasi-exponential bound for the deviations of the difference between this criterion and the limit that it is supposed to estimate. The result is presented in Theorem 1, Section 3.2.
2. Under some regularity assumptions on this criterion, the consistency of the parameters $\hat{\gamma}_l$ is obtained for each leaf of the tree, see Proposition 1 (Section 3.2). Next, the consistency of the global regression estimator $\hat{\pi}^{\mathcal{S}}$ is deduced in Corollary 1.
3. We show in Proposition 2 of Section 3.3 that the pruning strategy is legitimate, in the sense that it leads to, from a collection of subtrees, an estimator which achieves the best convergence rate, up to some smaller remainder terms. This is a consequence of the two previous steps, where the non-asymptotic results that are provided permit to easily track the effect of the size of the tree on estimation quality.

3.1. A bound on the deviations of the criterion

We consider in this section a tree with leaves $\mathcal{T}_l$ ($l = 1, \ldots, K$), where $\mathcal{T}_l$ is a random subdivision of $\mathcal{T}$ corresponding to the scheme defined in Section 2.1. Let

$$\begin{aligned} M_{n,l}(\gamma) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{1-\hat{G}(Y_i-)}\phi(N_i,\gamma)\mathbf{1}_{(Y_i,\mathbf{x}_i)\in\mathcal{T}_l}, \\ M_l(\gamma) &= \int \phi(m,\gamma)\mathbf{1}_{\tilde{\mathbf{x}}\in\mathcal{T}_l}dF(m,t,\mathbf{x}), \end{aligned}$$

and define the relative variation of $M_{n,l} - M_l$ around γ_l as

$$\Delta_l(\gamma,\gamma_l) = \frac{\{M_{n,l}(\gamma) - M_{n,l}(\gamma_l)\} - \{M_l(\gamma) - M_l(\gamma_l)\}}{|\gamma-\gamma_l|}.$$

The quantity $\Delta_l(\gamma,\gamma_l)$ is a way of measuring, in leaf l, a normalized variation of the error made by replacing the criterion M_l by its empirical counterpart. The cornerstone of our theoretical results is Theorem 1 below, which furnishes a bound for the deviations of Δ_l. Before stating the result, some assumptions on the regularity of the loss function are required.

Assumption 2. *There exists a constant $M < \infty$ such that, for all m,*

$$\sup_{(\pi,\pi')\in\Gamma^2}\frac{|\phi(m,\pi)-\phi(m,\pi')|}{|\pi-\pi'|} \leq M.$$

Assumption 2 holds provided that ϕ is continuously differentiable with respect to π, with uniformly bounded derivative. The second assumption that we need on ϕ requires us to introduce notation concerning covering numbers. For a class of functions $\mathcal{F}$ and a probability measure $\mathbb{Q}$, let $N(\varepsilon, L^2(\mathbb{Q}), \mathcal{F})$ denote the minimum number of $L^2(\mathbb{Q})$-balls of radius ε required to cover the set $\mathcal{F}$. In the following, for a class of functions $\mathcal{F}$ with envelope function $\mathcal{E}$ (by envelope, we mean that all functions in $\mathcal{F}$ are uniformly bounded by $\mathcal{E}$), we will use the following notation:

$$N_{\mathcal{E}}(\varepsilon,\mathcal{F}) = \sup_{\mathbb{Q}:\|\mathcal{E}\|_{L^2(\mathbb{Q})}<\infty} N(\varepsilon\|\mathcal{E}\|_{L^2(\mathbb{Q})}, L^2(\mathbb{Q}), \mathcal{F}).$$

Assumption 3. *Define the class of functions* $\Phi = \{m \to \frac{(\phi(m,\pi)-\phi(m,\pi'))}{(\pi-\pi')} : (\pi,\pi')\in\Gamma^2\}$.

Assume that, for some positive constants $\mathcal{C}_1$ and w,

$$N_M(\varepsilon,\Phi) \leq \mathcal{C}_1\left(\frac{1}{\varepsilon}\right)^w,$$

where we recall that the functions in Φ are bounded by M from Assumption 2.

Assumption 3 holds provided that the function ϕ is regular enough. For example, assume that ϕ is twice continuously differentiable with respect to π. Also assume that its second order derivative with respect to π is, for a fixed m, Hölder with Hölderian constant H_m. If this constant satisfies $E[H_m] < \infty$, it is easy to check that we are in the situation of Example 19.7 in van der Vaart (1998), and Assumption 3 holds.

We now state the main result of this section.

Theorem 1. *Let τ be such that $\mathbb{P}(Y \leq \tau) < 1$, and let $\mathfrak{T}_\tau \subset [0;\tau] \times \mathcal{X}$. Assume that $\mathbf{X}$ is a random vector with d continuous and k discrete components, where each discrete component takes at most m values. Then, under Assumptions 2 and 3, there exist positive constants $\mathcal{C}_j$ $(j = 1, \ldots, 5)$ such that*

$$\begin{aligned}\mathbb{P}\left(\sup_{l:\mathcal{T}_l \subset \mathfrak{T}_\tau} \sup_{\gamma \in \Gamma} |\Delta_l(\gamma, \gamma_l)| > x\right) \quad \leq \quad & 2\left\{\exp\left(-\mathcal{C}_1 n x^2\right) + \exp\left(-\mathcal{C}_2 n x\right)\right\} \\ & + 2.5\exp\left(-\mathcal{C}_3 n x^2 + \mathcal{C}_4 x\right) + u_n, \quad (3.1)\end{aligned}$$

with $u_n = O(\exp(-n))$, for $x \geq \mathcal{C}_5[kd\log m]^{1/2} n^{-1/2}$. Moreover, the constants $\mathcal{C}_j$ $(j = 1, \ldots, 5)$ do not depend on n nor (k, d, m).

The introduction of τ is required due to the erratic behavior of the Kaplan-Meier estimator at the right-hand side of the distribution. We therefore need to remove the observations that are too large, which is the purpose of considering only leaves such that $\mathcal{T}_l \subset \mathfrak{T}_\tau$. This type of truncation is classical in censored regression, see e.g., Sánchez Sellero, González Manteiga and Van Keilegom (2005), Heuchenne and Van Keilegom (2010b) and Lopez, Patilea and Van Keilegom (2013).

Sketch of the proof of Theorem 1. The probability (3.1) can be decomposed into

$$\begin{aligned}\mathbb{P}\left(\sup_{l:\mathcal{T}_l \subset \mathfrak{T}_\tau} \sup_{\gamma \in \Gamma} |\Delta_l(\gamma, \gamma_l)| > x\right) \quad \leq \quad & \mathbb{P}\left(\sup_{l:\mathcal{T}_l \subset \mathfrak{T}_\tau} \sup_{\gamma \in \Gamma} |\Delta_{l,C}(\gamma, \gamma_l)| > x/2\right) \\ & + \mathbb{P}\left(\sup_{l:\mathcal{T}_l \subset \mathfrak{T}_\tau} \sup_{\gamma \in \Gamma} |\Delta_l^*(\gamma, \gamma_l)| > x/2\right), \quad (3.2)\end{aligned}$$

where

$$\begin{aligned}\Delta_{l,C}(\gamma, \gamma_l) \quad &= \quad \frac{\{M_{n,l}(\gamma) - M_{n,l}(\gamma_l)\} - \{M_{n,l}^*(\gamma) - M_{n,l}^*(\gamma_l)\}}{|\gamma - \gamma_l|}, \\ \Delta_l^*(\gamma, \gamma_l) \quad &= \quad \frac{\{M_{n,l}^*(\gamma) - M_{n,l}^*(\gamma_l)\} - \{M_l(\gamma) - M_l(\gamma_l)\}}{|\gamma - \gamma_l|},\end{aligned}$$

with

$$M_{n,l}^*(\gamma) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i}{1 - G(Y_i-)} \phi(N_i, \gamma) \mathbf{1}_{(Y_i, \mathbf{X}_i) \in \mathcal{T}_l}.$$

This means that $\Delta_{l,C}$ corresponds to the replacement of $\hat{G}$ by G in the definition of $M_{n,l}$, while Δ_l^* corresponds to the deviation we would consider in a situation where the distribution of the censoring is known exactly.

The two probabilities in the decomposition (3.2) are studied separately in Lemmas 1 and 2 respectively. We proceed as follows:

1. Lemma 1 handles the replacement of $\hat{G}$ by G in the criterion (corresponding to $\Delta_{l,C}$) via the adaptation of the Dvóretsky-Kiefer-Wolfowitz inequality for the Kaplan-Meier estimator given by Bitouzé, Laurent and Massart (1999).
2. Lemma 2 is obtained through a concentration inequality due to Talagrand (Talagrand (1994)) to study the deviations of Δ_l^*, that is, of the criterion we would compute if we knew exactly the distribution of the censoring.

Using the notation in these two lemmas, the result follows by taking $\mathcal{C}_1 = \mathcal{B}_1/4$, $\mathcal{C}_2 = \mathcal{B}_2/2$, $\mathcal{C}_3 = A/4$, $\mathcal{C}_4 = B/2$, and $\mathcal{C}_5 = 2\mathcal{B}_3$. □

Remark 3.1. *The sequence u_n appears in Lemma 1 as $u_n = \mathbb{P}(E_n)$, where $E_n = \{\sup_{t<\tau} |\hat{G}(t) - G(t)| > c_G/2\}$ with $c_G = (1 - G(\tau))$. From the proof of Theorem 1,*

$$\mathbb{P}\left(\left\{\sup_{l:\mathcal{T}_l\subset\mathfrak{X}_\tau}\sup_{\gamma\in\Gamma}|\Delta_l(\gamma,\gamma_l)| > x\right\}\cap E_n^c\right) \leq 2\left\{\exp\left(-\mathcal{C}_1 nx^2\right)+\exp\left(-\mathcal{C}_2 nx\right)\right\} +2.5\exp\left(-\mathcal{C}_3 nx^2+\mathcal{C}_4 x\right). \quad (3.3)$$

Remark 3.2. *If $n+v$ observations are used to compute $\hat{G}$, n simply becomes $n+v$ in the third exponential term of (3.1), and u_n is replaced by u_{n+v}.*

3.2. Consistency of the regression tree

Consider a leaf $\mathcal{T}_l \subset \mathfrak{X}_\tau$. Once again, restraining ourselves to $\mathfrak{X}_\tau$ is required due to the poor performance of the Kaplan-Meier estimator near the tail of the distribution. Theorem 1 allows us to easily deduce consistency of $\hat{\gamma}_l$, up to adding some regularity assumptions on the function ϕ, which we now present.

Assumption 4. *$\phi(m,\gamma)$ is twice continuously differentiable with respect to γ for all m, and there exists a constant $\mathfrak{c} > 0$ such that*

$$\inf_{\gamma\in\Gamma,l}\left|\int \partial_\gamma^2\phi(m,\gamma)\mathbf{1}_{\tilde{x}\in\mathcal{T}_l}dF(m,t,\boldsymbol{x})\right| \geq \mathfrak{c}\mu_{\tilde{\boldsymbol{X}}}(\mathcal{T}_l),$$

where $\mu_{\tilde{\boldsymbol{X}}}(\chi) = \int \mathbf{1}_{\tilde{x}\in\chi}dF(m,t,\boldsymbol{x})$.

We also require some reasonable restrictions on the parameter space Γ.

Assumption 5. *Γ is compact, convex with non-empty interior, and for all $l = 1,\ldots,K$, γ_l belongs to the interior of Γ.*

By the definition of $\hat{\gamma}_l$, we have $M_{n,l}(\hat{\gamma}_l) - M_{n,l}(\gamma_l) \geq 0$, while $M_l(\hat{\gamma}_l) - M_l(\gamma_l) \leq 0$ by the definition of γ_l. Hence,

$$0 \leq \frac{-\{M_l(\hat{\gamma}_l) - M_l(\gamma_l)\}}{|\hat{\gamma}_l - \gamma_l|} \leq \Delta_l(\hat{\gamma}_l, \gamma_l) \leq \sup_{\gamma\in\Gamma} |\Delta_l(\gamma, \gamma_l)|.$$

Moreover, if follows from a second-order Taylor expansion and Assumptions 4 and 5 that

$$-\{M_l(\hat{\gamma}_l) - M_l(\gamma_l)\} \geq \frac{\mathfrak{c}\mu_{\tilde{\mathbf{X}}}(\mathcal{T}_l)\,|\hat{\gamma}_l - \gamma_l|^2}{2},$$

from which one deduces

$$|\hat{\gamma}_l - \gamma_l|\mu_{\tilde{\mathbf{X}}}(\mathcal{T}_l) \leq \frac{2\sup_{\gamma\in\Gamma}|\Delta_l(\gamma,\gamma_l)|}{\mathfrak{c}}. \tag{3.4}$$

The following Proposition 1 then easily follows from (3.4) and Theorem 1.

Proposition 1. *Under the conditions of Theorem 1 and Assumptions 4 and 5,*

$$\begin{aligned}\mathbb{P}\left(\sup_{l:\mathcal{T}_l\subset\mathfrak{T}_l} |\hat{\gamma}_l - \gamma_l|\mu_{\tilde{\mathbf{X}}}(\mathcal{T}_l) > x\right) \quad &\leq \quad 2\left\{\exp\left(-\mathcal{C}_1 n\mathfrak{c}^2x^2/4\right) + \exp\left(-\mathcal{C}_2 n\mathfrak{c}x/2\right)\right\} \\ &\quad\; +2.5\exp\left(-\mathcal{C}_3 n\mathfrak{c}^2x^2/4 + \mathcal{C}_4\mathfrak{c}x/2\right) + u_n,\end{aligned}$$

for $x \geq 2\mathcal{C}_5[kd\log m]^{1/2}\mathfrak{c}^{-1}n^{-1/2}$, where we have used the notation in Theorem 1, and where $\mu_{\tilde{\mathbf{X}}}$ is defined as in Assumption 4.

This Proposition means that in each leaf, the estimator $\hat{\gamma}_l$ is close to γ_l with high probability. Nevertheless, the term $\mu_{\tilde{\mathbf{X}}}(\mathcal{T}_l)$ shows that estimation performance in the leaf deteriorates when the leaf is "too small" (that is, when the selection rules define a region of the space $\mathcal{T}$ which has a small measure with respect to the distribution of $\tilde{\mathbf{X}}$). This is a classical issue when proving consistency of regression trees, see e.g., Condition 1 in Chaudhuri (2000) and Condition 1 in Chaudhuri and Loh (2002). Condition (3.5) in Corollary 1 below is clearly linked to this issue since, in a random design, $\mu_{\tilde{X}}(\mathcal{T}_l)$ represents in a sense the number of observations in $\mathcal{T}_l$.

Corollary 1. *Let $\mathfrak{T}'_\tau = \cup_{l:\mathcal{T}_l\subset\mathfrak{T}_\tau}\mathcal{T}_l$. Assume that, for all $\mathcal{T}_l \subset \mathfrak{T}_\tau$,*

$$\mu_{\tilde{\mathbf{X}}}(\mathcal{T}_l) \geq \mathfrak{m} > 0. \tag{3.5}$$

Define $\|\hat{\pi}^{\mathcal{S}} - \pi^{\mathcal{S}}\|_{2,\tau} = \left\{\int \left|\hat{\pi}^{\mathcal{S}}(t,\boldsymbol{x}) - \pi^{\mathcal{S}}(t,\boldsymbol{x})\right|^2 \mathbf{1}_{\tilde{\boldsymbol{x}}\in\mathfrak{T}'_\tau} dF(m,t,\boldsymbol{x})\right\}^{1/2}$ *and* $P(x) = \mathbb{P}(\|\hat{\pi}^{\mathcal{S}} - \pi^{\mathcal{S}}\|^2_{2,\tau} > x)$*. Then, for some positive constants $\mathcal{C}'_j$,*

$$\begin{aligned}P(x) \quad \leq \quad & K\left(2\left\{\exp\left(-\mathcal{C}'_1 nx\right) + \exp\left(-\mathcal{C}'_2 nx^{1/2}\right)\right\}\right. \\ &\left. +2.5\exp\left(-\mathcal{C}'_3 nx + \mathcal{C}'_4 x^{1/2}\right) + u_n\right),\end{aligned} \tag{3.6}$$

for $x \geq \mathcal{C}'_5 n^{-1}$. Moreover,

$$E\left[K(\mathcal{S})^{-1}\|\hat{\pi}^{\mathcal{S}} - \pi^{\mathcal{S}}\|^2_{2,\tau}\right] = O(1/n). \tag{3.7}$$

Proof. We have

$$\int \left|\hat{\pi}^{\mathcal{S}}(t,\mathbf{x}) - \pi^{\mathcal{S}}(t,\mathbf{x})\right|^2 \mathbf{1}_{\tilde{\mathbf{x}}\in\mathfrak{T}'_\tau} dF(m,t,\mathbf{x}) \leq \sum_{l=1}^{K} [|\hat{\gamma}_l - \gamma_l|\mu_{\tilde{\mathbf{X}}}(\mathcal{T}_l)]^2 \frac{\mathbf{1}_{\mathcal{T}_l\subset\mathfrak{T}_\tau}}{\mathfrak{m}},$$

since the intersection of $\mathcal{T}_l$ and $\mathcal{T}_{l'}$ is empty for $l \neq l'$, and using (3.5). Equation (3.6) then follows from Proposition 1.

To show (3.7), observe, following Remark 3.1, that $P^{(n)}(x) := \mathbb{P}(\{\|\hat{\pi}^{\mathcal{S}} - \pi^{\mathcal{S}}\|^2_{2,\tau} > x\} \cap E_n) = P(x) - 2.5u_n$, where $E_n^c = \{\sup_{t<\tau} |\hat{G}(t) - G(t)| > c_G/2\}$. Then, since $\|\hat{\pi}^{\mathcal{S}} - \pi^{\mathcal{S}}\|^2_{2,\tau}$ is bounded (say by a finite constant $\mathcal{A}$),

$$E\left[K(\mathcal{S})^{-1}\|\hat{\pi}^{\mathcal{S}} - \pi^{\mathcal{S}}\|^2_{2,\tau}\right] \leq \int_0^\infty P^{(n)}(x)dx + \mathcal{A}\mathbb{P}(E_n),$$

and the result follows since $\mathbb{P}(E_n) = 2.5u_n$. □

3.3. Consistency of the pruning strategy

The next result shows that penalizing the subtree $\mathcal{S}$ by a factor $\alpha K(\mathcal{S})/n$ is a relevant strategy. This idea already seems reasonable in view of (3.7). Indeed, $\int \phi(m,\hat{\pi}^{\mathcal{S}})d\hat{F}(m,t,\mathbf{x})$ is, due to the regularity assumptions on ϕ (Assumption 4), of the same order as $\|\hat{\pi}^{\mathcal{S}} - \pi^{\mathcal{S}}\|^2_{2,\tau}$, which is of order $K(\mathcal{S})/n$. Penalizing by $\alpha K(\mathcal{S})/n$ can then be interpreted as compensating the structural decrease in $\|\hat{\pi}^{\mathcal{S}} - \pi^{\mathcal{S}}\|_{2,\tau}$ when $K(\mathcal{S})$ increases. Proposition 2 below confirms this.

Proposition 2. *Let $S = (\mathcal{S}_1, \ldots, \mathcal{S}_{K_n})$ denote a sequence of subtrees all satisfying the assumptions of Corollary 1, and with $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \ldots \subset \mathcal{S}_{K_n}$. Let*

$$K_0 = \underset{K=1,\ldots,K(n)}{\arg\min} \int \phi(m, \pi^{\mathcal{S}_K}(t,\boldsymbol{x}))dF(m,t,\mathbf{x}).$$

Define $\hat{\pi}^{\hat{\mathcal{S}}(\alpha)}$ as the estimator selected using the pruning strategy with parameter α. Let

$$\Delta(K) = -\int [\phi(m,\pi^{\mathcal{S}_{K_0}}(t,\boldsymbol{x})) - \phi(m,\pi^{\mathcal{S}_K}(t,\boldsymbol{x}))]\, dF(m,t,\mathbf{x}).$$

Assume that

$$\inf_{K<K_0} \Delta(K) - \alpha[K-K_0]n^{-1} \geq \mathcal{C}_6^{-1} n^{-1}\log n, \tag{3.8}$$

for some absolute constant $\mathcal{C}_6 > 0$, and $\sup_{\gamma,m} |\partial^2_\gamma \phi(m,\gamma)| \leq \mathcal{B}$ for some finite constant $\mathcal{B}$. Then, if $\mathcal{C}_6$ is small enough, under the assumptions of Corollary 1,

$$\left(\frac{E\left[\|\hat{\pi}^{\hat{\mathcal{S}}(\alpha)} - \pi_0\|^2_{2,\tau}\right]}{K_0}\right)^{1/2} = \frac{\|\pi^{K_0} - \pi_0\|_{2,\tau}}{K_0^{1/2}} + O(n^{-1/2}),$$

where the $O(n^{-1})$-term does not depend on K_0.

The proof of Proposition 2 is postponed to the appendix. It introduces an optimal choice of complexity K_0 for the selected tree. It is optimal in the sense that K_0 minimizes $\int \phi(t, \pi^K) dF(t, m, \mathbf{x})$ over K, that is, the unknowable criterion that would be optimized if we knew the distribution F. Proposition 2 means that the penalization strategy gives approximately the same performance as we would have if we knew the optimal complexity K_0. Indeed, the L^2-norm of the error is of order $K_0 n^{-1}$, plus some approximation term (the distance between π^{K_0} and π_0).

4. Simulations

We investigate here the practical behaviour of tree-based estimators for censored data via simulations. For the sake of simplicity, we consider the case where we are interested in the distribution of the lifetime T, thus focusing on estimating $\pi_0(\mathbf{x}) = E[T \mid \mathbf{X} = \mathbf{x}]$. Consider the following simulation scheme (see the parameter values in Table 2):

1. draw $n + v$ i.i.d. replications $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ of the covariate, with $\mathbf{X}_i \sim \mathcal{U}(0, 1)$;
2. draw $n + v$ i.i.d. lifetimes $(T_1, \ldots, T_n)$ following an exponential distribution such that $T_i \sim \mathcal{E}(\beta = \alpha_1 \mathbb{1}_{\mathbf{X}_i \in [a,b[} + \alpha_2 \mathbb{1}_{\mathbf{X}_i \in [b,c[} + \alpha_3 \mathbb{1}_{\mathbf{X}_i \in [c,d[} + \alpha_4 \mathbb{1}_{\mathbf{X}_i \in [d,e]})$.
   (Notice that there thus exist four subgroups in the whole population.)
3. draw $n + v$ i.i.d. censoring times, Pareto-distributed: $C_i \sim \mathcal{P}areto(\lambda, \mu)$;
4. from the simulated lifetimes and censoring times, get for all i the actual observed lifetime $Y_i = \inf(T_i, C_i)$ and the indicator $\delta_i = \mathbf{1}_{T_i \leq C_i}$;
5. compute the estimator $\hat{G}$ from the entire generated sample $(Y_i, \delta_i)_{1 \leq i \leq n+v}$.

Descriptive statistics corresponding to various simulated datasets (of different sizes) are available in Table 3. To each simulated sample, we fit a regression tree with the algorithm in Section 2.1, and prune it using the strategy in Section 2.3.

TABLE 2
Parameters involved in the simulation scheme.

Group-specific means				Component probabilities				Censorship rate		
α_1	α_2	α_3	α_4	$[a, b[$	$[b, c[$	$[c, d[$	$[d, e]$	10%	30%	50%
0.08	0.05	0.16	0.5	$[0, 0.3[$	$[0.3, 0.6[$	$[0.6, 0.8[$	$[0.8, 1]$	(λ, μ)	(λ, μ)	(λ, μ)
12.5	20	6.25	2	30%	30%	20%	20%	(80,1.03)	(20,1.2)	(14,2)

TABLE 3
Descriptive statistics of simulated datasets.

Sample	Group-specific exposure				Sample
size n	Group 1	Group 2	Group 3	Group 4	mean
100	35%	28%	17%	20%	11.08
500	26.8%	31.6%	20%	21.6%	11.37
1 000	30.1%	28.7%	20.6%	20.6%	11.33
5 000	31.42%	29.96%	19.5%	19.12%	11.53
10 000	30.25%	30.19%	19.79%	19.77%	11.52

Table 4

Mean weighted squared errors w.r.t. the censoring rate and sample size.

% of	Sample	Group-specific MWSE				Global
censored	size	Group 1	Group 2	Group 3	Group 4	MWSE
observations	n	MWSE	MWSE	MWSE	MWSE	
10%	100	0.19516	0.42008	0.17937	0.30992	***1.10454***
	500	0.03058	0.07523	0.03183	0.06029	***0.19796***
	1 000	0.01509	0.03650	0.01517	0.02619	***0.09306***
	5 000	0.00295	0.00714	0.00289	0.00530	***0.01804***
	10 000	0.00105	0.00378	0.00117	0.00292	***0.00910***
30%	100	0.20060	0.43664	0.17448	0.29022	***1.10765***
	500	0.03736	0.07604	0.04301	0.06584	***0.22217***
	1 000	0.01748	0.04095	0.01535	0.02674	***0.10043***
	5 000	0.00319	0.00758	0.00291	0.00547	***0.01904***
	10 000	0.00117	0.00372	0.00125	0.00292	***0.00930***
50%	100	0.19784	0.45945	0.17387	0.28363	***1.11476***
	500	0.04906	0.08993	0.05301	0.06466	***0.25668***
	1 000	0.02481	0.05115	0.01788	0.03004	***0.12387***
	5 000	0.00520	0.00867	0.00389	0.00516	***0.02299***
	10 000	0.00153	0.00407	0.00162	0.00308	***0.01057***

Then, we compute the weighted squared errors given by $WSE_i = \delta_i n^{-1}[1 - \hat{G}(Y_i-)]^{-1}(\hat{\gamma}_{l(i)} - \pi_0(\mathbf{X}_i))^2$, where the i^{th} observation belongs to leaf $l(i)$, where we know that $\pi_0(\mathbf{X}_i) = 1/\beta$.

In order to gain some robustness in our results, we repeated 5000 times the simulation scheme above to compute empirical means of WSE_i, leading to the $MWSE$. We also considered different values for (λ, μ) in the censoring process so as to measure the impact of censoring on the procedure's performance (see Table 2 for these values for the Pareto distribution). The performance of the procedure is shown in Figure 1 and Table 4. Clearly, the strength of the censoring has an impact on performance. One can also observe that the performance in the group with the highest mean (Group 2) is worse than in the others, which is linked with the fact that largest observations are more likely to be censored. However, the hierarchy of the groups in term of performance cannot be entirely summarized with respect to the typical size of the lifetimes (see Group 4 which has a lower mean, but performs worse than Group 1).

5. Applications to real-life insurance datasets

In this section, we consider two applications in insurance. The first, described in Section 5.1, focuses on the prediction of a duration variable only (duration in a state of disability). The second, in Section 5.2, is dedicated to claim reserving, and illustrates the need to introduce a supplementary variable M. In this situation, the key issue is to predict the claim amount, this being known only after some time T, subject to censoring.

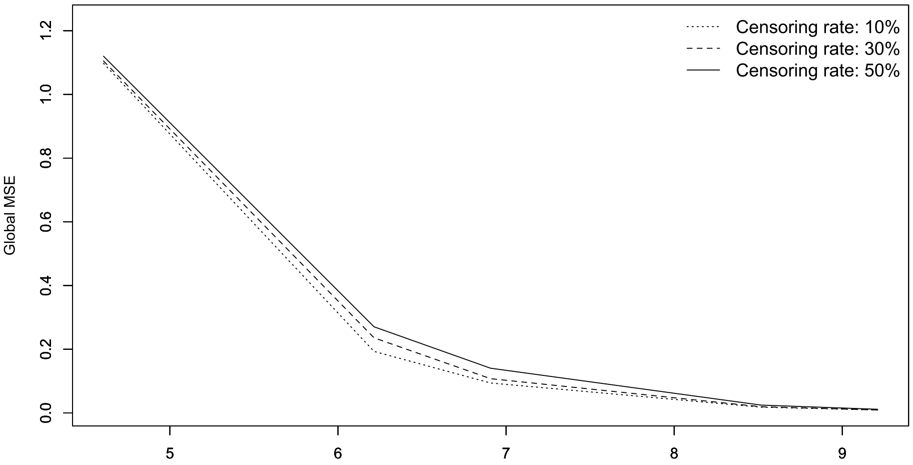

FIG 1. *MWSE as a function of the sample size (n=100, 500, 1 000, 5 000, 10 000).*

5.1. Income protection insurance

The real-life database we consider reports the claims of income protection guarantees over six years. It consists of 83 547 claims, with the following information for each claim: a policyholder ID, cause (sickness or accident), gender (male or female), socio-professional category (SPC: manager, employee or miscellaneous), age at the claim date, duration in the disability state (perhaps right-censored), commercial network (3 kinds of brokers). All insurance contracts considered have a common deductible of 30 days.

Here, the censoring rate equals 7.2%, the mean observed duration in the disability state is about 100 days (beyond the deductible of 30 days), with a median of 42 days. There is strong dispersion among the observed durations, the standard deviation being 162 days. Our goal is to find a segmentation into several classes of homogeneous individuals, and to predict the duration in the disability state in each class.

To begin, we compute the Cox proportional-hazards model with the (discretized) age at the claim date as covariate, since the recovery rates used in the calculation of technical provisions for this kind of guarantee depends on the age range at the claim date. This adjustment leads us to consider the high predictive power of this variable. However, the proportional hazards assumption is thoroughly rejected by all classical statistical tests (likelihood ratio, Wald and log-rank tests). Nevertheless, the obtained results are retained, to enable a comparison with those from the tree approach. We thus try to explain the disability duration by *sex*, *SPC*, *commercial network*, *age at the claim date* (5 pre-determined classes, due to local prudential regulation) and *cause* of disability. The final tree (after pruning) is given in Figure 2. We see in Table 5 the significant differences between tree and Cox estimates. These differences can be

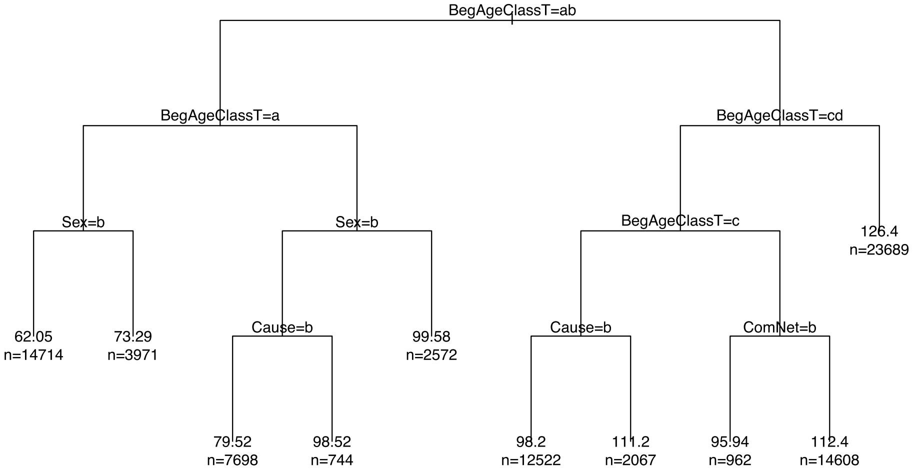

Fig 2. *Disability duration in terms of sex, SPC, commercial network, age and cause.*

Table 5

Estimates of expected disability time (days) depending on age.

Classes	Mean Age	Tree	Cox
a	26.83	64.44	80.01
b	34.19	85.48	96.35
c	39.57	100.04	110.19
d	45.05	111.38	126.03
e	51.29	126.40	146.28

explained by two phenomena resulting from using the Cox proportional-hazards model:

- our approach directly targets the duration expectation while Cox partial-likelihood is focused on estimating the hazard rate;
- the estimation of the baseline hazard is very sensitive to the longer durations (mainly concentrated in class e), which affect the estimates of all other classes (whereas our estimation is expected to be less sensitive to this phenomenon for classes a to d).

These differences reinforce the interest of such an approach to incorporate heterogeneity in the reserving process of an insurance portfolio.

More generally, the predictive performance of duration models for censored survival endpoints can be assessed using various techniques including time-dependent ROC curves using independent test data (hereafter denoted by $ROC(t)$, see e.g., Heagerty, Lumley and Pepe (2000) and Heagerty and Zheng (2005)). Figure 3 illustrates such ROC curves (at $t = 15, 100, 110$, which correspond respectively to the first quartile, mean and third quartile of observed lifetimes), obtained from previously-built models. The only difference lies in age, which is considered here as a continuous covariate to benefit from one of the strengths of tree-based procedures, i.e., eliciting good cut-points for continuous

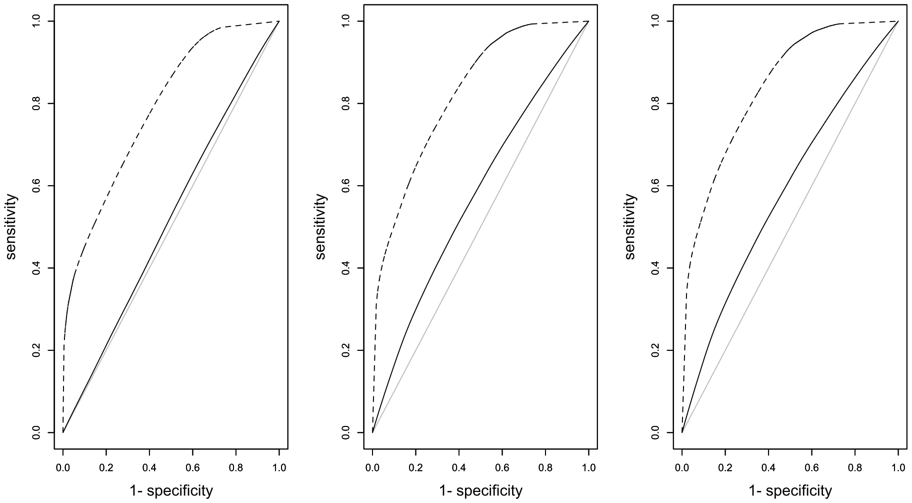

Fig 3. *Dynamic ROC curves at $t = 15, 100, 110$ (from left to right). The dotted line corresponds to the CART model and the black line to the Cox model.*

Table 6
Dynamic Area Under Curve $AUC(t)$.

	t	15	40	100	110
$AUC(t)$	CART	0.787	0.802	0.824	0.839
	Cox	0.518	0.531	0.576	0.585

covariates and capturing potential nonlinearity. Table 6 gives the value of the AUC (Area Under Curve) at various time points, corresponding to previous durations to which were added the median of observed lifetimes. The tree approach seems significantly better than the Cox one at predicting lifetime, with an excellent mean AUC of 80%. Once again and in this more general framework, these results prove the interest of using trees as opposed to the Cox model for prediction, whatever the duration threshold under study.

5.2. Reserving in third-party liability insurance

This real-life database was extracted in the 2000s by an international insurance company, and reports about 650 claims related to medical malpractice insurance during seven successive years. The initial dataset contains information about various dates concerning the claims (date for reporting, opening or closing the case, etc.), contract features, and some data on associated payments. These payments encompass indemnity payments and ALAE (Allocated Loss Adjustment Expenses), where ALAE are assignable to specific claims and represent fees paid to outside attorneys used to defend the claims. After some pre-processing, one can compute useful quantities for our purposes, especially (potentially censored) development times and total payments. Here T_i is the “lifetime” of a claim, that

Table 7

Statistics on the information selected for our application.

		Statistical indicators					
	Type	Median	Mean	Std.	Min.	Max.	# categories
Insurance type	categorical						2
Specialty	categorical						41
Class	categorical						19
Report date	date				N	N+7	
Area	categorical						30
Closed without payments	boolean						2
Closed without indemnity	boolean						2
Time before opening (days)	continuous	1164	1223	614	2	4728	
Time before declaration	continuous	734	724	560	0	4657	
Reopen status	boolean						2
Cancel status	boolean						2
Reserves	continuous	0	44170	138867	0	1062000	
Development time	continuous	419	606	506	0	2249	
Observed payments	continuous	2617	41810	152319	0	1557000	

is, the time between its issue date and the claim settlement date. The consorship C_i is the delay between the claim issue date and the extraction date of the database, and M_i is the total amount of the i^{th} claim. The latter is observed only if the claim has been fully settled (32% of the observations are censored). In this setting, it is reasonable to assume that C_i does not depend on $(M_i, T_i, \mathbf{X}_i)$, but this would clearly be wrong in the case of covariates depending on the claim issue date. Table 7 summarizes some descriptive statistics about the covariates that are used when running the weighted CART algorithm to explain the response M_i. As could be expected in this type of business, the data are highly skewed; for instance, many declared claims are assigned no payments because the company is still waiting for a court decision before paying. A parametric model would then be quite difficult to fit, which emphasizes the interest of using such techniques.

As we have already mentioned, a key issue is to predict the future coming expenses related to claims that are still under payment. Typically, computing

$$M^*(N_i, Y_i, \delta_i, \mathbf{X}_i) := E[M_i \mid N_i,\, Y_i,\, \delta_i,\, \mathbf{X}_i],$$

would give the best L^2-approximation of the amount M_i based on the information available on claim i. Our aim is then to produce an estimator $\hat{M}$ of this ideal (but unattainable) predictor. Of course, M^* is known if $\delta_i = 1$, that is, $M^*(m, y, 1, \mathbf{x}) = m$, but the key issue is to predict it for unsettled claims ($\delta_i = 0$). For such claims, rewrite

$$\begin{aligned} M^*(m, y, 0, \mathbf{x}) &= E[M \mid M > m, T > y, \mathbf{X} = \mathbf{x}] \\ &= \frac{E[M\, \mathbb{1}(M > m, T > y) \mid \mathbf{X} = \mathbf{x}]}{\mathbb{P}(M > m, T > y \mid \mathbf{X} = \mathbf{x})}, \end{aligned} \tag{5.1}$$

and introduce $Z_1(m, y) = \mathbb{1}(M > m, T > y)$, and $Z_2(m, y) = M\, Z_1$.

In view of (5.1), we have to estimate the quantities $\pi_{0,1}^{m,y}(\mathbf{x}) = E[Z_1 | \mathbf{X} = \mathbf{x}]$ and $\pi_{0,2}^{m,y}(\mathbf{x}) = E[Z_2 | \mathbf{X} = \mathbf{x}]$. Each of these are estimated using the CART procedure described in Section 2. Hence, for each censored claim, we use two regression trees to compute a prediction $\hat{M}_i$, obtained as the ratio $\hat{M}_i = \hat{\pi}_{0,2}^{N_i,Y_i}(\mathbf{X}_i)/\hat{\pi}_{0,1}^{N_i,Y_i}(\mathbf{X}_i)$. Note that, for each censored claim, the trees we compute are different since the values of Y_i and N_i are. We now determine a reserve to be constituted by summing the $\hat{M}_i$. To check that the proposed amount is reasonable, we can compare the values of $\hat{M}_i$ with the prediction of experts that are present in the database. The aggregated results are presented in Tables 8 and 9.

The predictions are highly overdispersed for both “expert” and “tree” reserves (see Table 8) but, as mentioned earlier, this is not surprising from a business point of view. We observe that our regression tree approach produces reserve amounts which are significantly higher than the reserves made by the experts, except for lower amounts. We argue that this has to be linked with the fact that the expert reports are made close to the opening of the claim. In our approach, we use posterior information: if a claim is open for a long time, our procedure tends to predict an higher final value (claims with long duration before settlement are more likely to be associated with larger amounts). This difference justifies the use in practice of our technique as a second diagnosis, complementary to expert judgment. Finally, notice in Table 9 that the gap between the two reserves is not necessary increasing when increasing the level of information. For instance, the tree global reserve is 1.22 times bigger than the expert one when considering two thirds of the censored observations (from the

Table 8

Descriptive statistics of the reserves (in US$) computed from both approaches (tree estimators and expert's judgment) and for different quantile levels.

	Expert reserves		Tree reserves	
	Mean	Std	Mean	Std
Quantiles				
0-25%	52 193	45 324	55 566	45 446
0-33%	58 703	68 427	95 198	118 237
0-50%	84 251	134 878	122 293	109 460
0-66%	112 551	188 676	145 005	108 844
0-75%	115 216	196 829	**209 696**	**478 048**
0-90%	150 790	224 863	**308 190**	**494 322**
0-99%	144 239	218 913	**343 892**	**500 388**

Table 9

Reserve gaps (reserves by tree estimators minus reserves following experts' judgments) for different level of information, going from the lowest censored observation up to the x-th percentile of censored observations.

Reserve gap	total US$	in %	mean	std	min.	max.
Censored data:						
0-25%	158 496	**+6%**	2 911	116 233	-170 288	205 082
0-33%	2 262 728	+38%	321 655	669 894	-170 288	2 262 728
0-50%	3 576 000	**+31%**	1 383 074	1 626 570	-170 288	4 522 203
0-66%	4 024 335	**+22%**	2 175 544	2 024 009	-170 288	5 870 474
0-75%	13 321 685	+45%	2 660 409	2 484 695	-170 288	13 321 685
0-90%	26 600 691	+51%	**5 779 725**	7 587 193	-170 288	**27 079 860**
0-99%	37 135 400	**+58%**	**8 216 874**	10 594 793	-170 288	**37 135 400**

minimum to the 66-th percentile of the censored observations), whereas it is 1.31 times bigger with half.

Conclusion

In this paper, we defined a regression tree procedure well-adapted to the presence of incomplete observations due to censoring, and proved its consistency. The framework that we considered is motivated by the field of survival analysis, but also allows us to consider related applications, such as claim reserving in insurance. In such types of problem, a duration is present (and subject to censoring), but also an additional variable (the claim amount) that is observed only if the observation is uncensored. We presented two practical applications of this technique that demonstrate its feasibility and interest. The next step is to extend the procedure we have developed to random forests, in the same spirit as Hothorn et al. (2006). Indeed, regression tree procedures, although they produce an easily understandable model at the end, are known for their sensitivity to changes in the dataset. This investigation is left for future work, but we would like to emphasize the role of ensemble methods in improving the predictive abilities of the technique we described in the present paper.

Appendix A: Main lemmas

Lemmas 1 and 2 below are the key results required to prove Theorem 1.

Lemma 1. *Under Assumption 2, we have*

$$\mathbb{P}\left(\sup_{l:\mathcal{T}_l\in\mathfrak{T}_\tau}\sup_{\gamma\in\Gamma}|\Delta_{l,C}(\gamma,\gamma_l)|>x\right)\leq 2.5\left\{\exp\left(-nAx^2+Bn^{1/2}x\right)+u_n\right\},$$

with $u_n=O(\exp(-n))$, *and* A *and* B *two positive constants.*

Proof. Since $\mathcal{T}_l\in\mathfrak{T}_\tau$, we have that $\mathbf{1}_{\tilde{\mathbf{x}}\in\mathcal{T}_l}=0$ if $t>\tau$. Let $c_G=(1-G(\tau))$ and $c_F=(1-F(\tau))$. We have $c_F>0$ and $c_G>0$.Therefore, we have

$$\sup_{l:\mathcal{T}_l\subset\mathfrak{T}_\tau}\sup_{\gamma\in\Gamma}|\Delta_{l,C}(\gamma,\gamma_l)|\leq\sup_{t<\tau}\frac{\left|\hat{G}(t)-G(t)\right|}{1-\hat{G}(t)}\times\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_iM}{1-G(Y_i-)},$$

where we have used Assumption 2. Since $(1-G)$ is bounded away from zero, the empirical mean on the right-hand side is bounded by Mc_G^{-1}. On the other hand,

$$\mathbb{P}\left(\sup_{t<\tau}\frac{\left|\hat{G}(t)-G(t)\right|}{1-\hat{G}(t)}>y\right)\leq\mathbb{P}\left(\sup_{t<\tau}\left|\hat{G}(t)-G(t)\right|>c_G/2\right)$$
$$+\mathbb{P}\left(\sup_{t<\tau}\left|\hat{G}(t)-G(t)\right|\leq c_G/2,\sup_{t<\tau}\frac{\left|\hat{G}(t)-G(t)\right|}{1-\hat{G}(t)}>y\right).$$

On the event $\{\sup_{t<\tau}\left|\hat{G}(t)-G(t)\right| \leq c_G/2\}$, we have

$$\begin{aligned}\sup_{t<\tau}\frac{\left|\hat{G}(t)-G(t)\right|}{1-\hat{G}(t-)} &= \sup_{t<Y_{(n)}}\frac{\left|\hat{G}(t)-G(t)\right|}{1-G(t)+\{G(t)-\hat{G}(t-)\}}\\ &\leq \frac{\sup_{t<\tau}\left|\hat{G}(t)-G(t)\right|}{c_G/2}.\end{aligned}$$

Moreover,

$$\mathbb{P}\left(\sup_{t<\tau}c_F\left|\hat{G}(t)-G(t)\right|>z\right)\leq\mathbb{P}\left(\sup_{t<\tau}(1-F(t))|\hat{G}(t)-G(t)|>z\right),$$

and the probability on the right-hand side can be bounded by $2.5\exp(-2nz^2+\mathcal{C}n^{1/2}z)$, for some absolute constant $\mathcal{C}>0$, using the Dvoretsy-Kiefer-Wolfowitz inequality for the Kaplan-Meier estimator proved in Bitouzé, Laurent and Massart (1999). Hence the result follows, with $A=c_F^2c_G^4[2M]^{-1}$, $B=\mathcal{C}c_Fc_G^2[2M]^{-1}$, and $u_n=\exp(-n^{1/2}c_Fc_G[\mathcal{C}+n^{1/2}c_Fc_G]/2)$. □

Lemma 2. *Assume that* $\mathbf{X}$ *is a random vector with* d *continuous components and* k *discrete components, where each discrete component takes at most* m *values. Then, under Assumptions 2 and 3, there exist strictly positive constants* $\mathcal{B}_1$, $\mathcal{B}_2$ *and* $\mathcal{B}_3$, *such that*

$$\mathbb{P}\left(\sup_{l:\chi_l\in\mathfrak{T}_\tau}\sup_{\gamma\in\Gamma}|\Delta^*(\gamma,\gamma_l)|>x\right)\leq 2\left\{\exp\left(-\mathcal{B}_1nx^2\right)+\exp\left(-\mathcal{B}_2nx\right)\right\},$$

for $x\geq\mathcal{B}_3[kd\log m]^{1/2}n^{-1/2}$, *where* $\mathcal{B}_j$ *for* $j=1,2,3$ *depend on* M, w, *and* $c_G=(1-G(\tau))$.

Proof. Let

$$\mathcal{F}=\left\{(n,y,\mathfrak{d},\mathbf{x})\to\frac{\mathfrak{d}\{\phi(m,\gamma)-\phi(m,\gamma')\}\mathbf{1}_{(y,\mathbf{x})\in\chi}}{\{1-G(y-)\}(\gamma-\gamma')}:\gamma\in\Gamma,\chi\in E_\tau\right\},\tag{A.1}$$

with E_τ denoting the set of subsets of $\mathfrak{T}_\tau$ of the type $\prod_{j=1}^{d+1}[x_{j-};x_{j+}]$. From Lemma 3,

$$N_{Mc_G^{-1}}(\varepsilon,\mathcal{F})\leq 2^{w+4(d+1)(d+2)}C_1m^k\left(\frac{\tilde{K}}{\varepsilon}\right)^{w+4d(d+1)},$$

where $c_G=(1-G(\tau))$ as in the proof of Lemma 1. As in Proposition C1 in Appendix C, introduce a sequence of i.i.d. Rademacher variables $(\varepsilon_i)_{1\leq i\leq n}$, independent from $(N_i,Y_i,\delta_i,\mathbf{X}_i)_{1\leq i\leq n}$, and define

$$Z=E\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(N_i,Y_i,\delta_i,\mathbf{X}_i)\varepsilon_i\right|\right].$$

Since

$$n \sup_{l:\mathcal{T}_l\subset\mathfrak{T}_\tau} \sup_{\gamma\in\Gamma} |\Delta^*(\gamma,\gamma_l)| \leq \sup_{f\in\mathcal{F}} \Bigg| \sum_{i=1}^{n} \{ f(N_i, Y_i, \delta_i, \mathbf{X}_i) \\ - \int f(n, y, \mathfrak{d}, \tilde{\mathbf{x}}) d\mathbb{P}(n, y, \mathfrak{d}, \tilde{\mathbf{x}}) \} \Bigg|,$$

we get, from Proposition C1,

$$\mathbb{P}\left(n \sup_{l:\mathcal{T}_l\subset\mathfrak{T}_\tau} \sup_{\gamma\in\Gamma} |\Delta^*(\gamma,\gamma_l)| > \mathcal{A}_1(Z+y) \right) \\ \leq 2\left\{ \exp\left(-\frac{\mathcal{A}_2 y^2}{n\sigma_{\mathcal{F}}^2} \right) + \exp\left(-\frac{c_G \mathcal{A}_2 y}{M} \right) \right\},$$

with $\sigma_{\mathcal{F}}^2 \leq M^2 c_G^{-2}$. It follows from Proposition C2 that

$$Z \leq \tilde{\mathcal{A}}[kd\log m]^{1/2} n^{1/2},$$

for some constant $\tilde{\mathcal{A}}$. Hence, for $y > \tilde{\mathcal{A}}[kd\log m]^{1/2} n^{1/2}$, we get

$$\mathbb{P}\left(n \sup_{l} \sup_{\gamma\in\Gamma} |\Delta^*(\gamma,\gamma_l)| > 2\mathcal{A}_1 y \right) \leq 2\left\{ \exp\left(-\frac{\mathcal{A}_2 c_G^2 y^2}{nM^2} \right) + \exp\left(-\frac{c_G\mathcal{A}_2 y}{M} \right) \right\}.$$

The result follows by applying this inequality to $y = nx/(2\mathcal{A}_1)$, with $\mathcal{B}_1 = \mathcal{A}_2 c_G^2 [4\mathcal{A}_1^2 M^2]^{-1}$, $\mathcal{B}_2 = \mathcal{A}_2 c_G [2\mathcal{A}_1 M]^{-1}$, and $\mathcal{B}_3 = 2\mathcal{A}_1\tilde{\mathcal{A}}$. □

Appendix B: Technical lemmas

B.1. Covering numbers

This section is devoted to the computation of covering numbers of classes of functions that appear naturally in the proof of Theorem 1.

Lemma 3. *Let $\mathcal{F}$ denote the class of functions defined in (A.1). Then, assuming that X is a random vector with d continuous components and k discrete components, where each discrete component takes at most m values,*

$$N_{Mc_G^{-1}}(\varepsilon, \mathcal{F}) \leq 2^{w+4(d+1)(d+2)} C_1 m^k \left(\frac{\tilde{K}}{\varepsilon} \right)^{w+4d(d+1)},$$

where $\tilde{K}$ is a constant depending only on $c_G = (1 - G(\tau))$, and w is defined in Assumption 3.

Proof. We combine Lemma 4 and Assumption 3 using Lemma A.1 in Einmahl and Mason (2000). This shows that the class

$$\mathcal{G} = \left\{ (m, \tilde{x}) \to \frac{(\phi(m,\pi) - \phi(m,\pi'))}{(\pi - \pi')} \mathbf{1}_{\tilde{x}\in\chi_l} : (\pi,\pi') \in \Gamma\times\Gamma, \chi_l \in E \right\},$$

satisfies

$$N_M(\varepsilon, \mathcal{G}) \leq 2^{w+4(d+1)(d+2)} C_1 m^k \left(\frac{K}{\varepsilon}\right)^{w+4(d+1)(d+2)}.$$

Multiplying the class $\mathcal{G}$ by some fixed bounded function (that is, $(\mathfrak{d}, y) \to \mathfrak{d}[1 - G(y-)]^{-1}$) barely changes the covering number, leading to

$$N_{Mc_G^{-1}}(\varepsilon, \mathcal{F}) \leq 2^{w+4(d+1)(d+2)} C_1 m^k \left(\frac{\tilde{K}}{\varepsilon}\right)^{w+4d(d+1)},$$

with $\tilde{K} = 2Kc_G^{-1}$, since $1 - G(y-) \geq c_G$ for $y \leq \tau$. □

Lemma 4. *Assume that* $\mathbf{X}$ *is a random vector with* d *continuous components and* k *discrete components, where each discrete component takes at most* m *values. Then, letting* $F_\tau = \{\tilde{\boldsymbol{x}} \to \mathbf{1}_{\tilde{\boldsymbol{x}} \in \chi} : \chi \in E_\tau\}$, *we have*

$$N_1(\varepsilon, F_\tau) \leq m^k \left(\frac{K}{\varepsilon}\right)^{4(d+1)(d+2)},$$

for some universal constant K.

Proof. Without loss of generality, we can assume that the first d variables in $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)}, X^{(d+1)}, \ldots, X^{(d+k)})$ are continuous, while the k other variables are discontinuous with at most m values each. Let $\{x_1^{(j)}, \ldots, x_m^{(j)}\}$ denote these values for variable $X^{(j)}$ for $j > d$. A set χ_l is of the form

$$(t, \mathbf{x}) \in \chi_l \Longleftrightarrow \left\{ \begin{array}{lcl} \alpha_0 & < t \leq \beta_0 & \\ \alpha_1 & < x^{(1)} \leq & \beta_1 \\ & \vdots & \\ \alpha_d & < x^{(d)} \leq & \beta_d \\ & x^{(d+1)} = & x_{g_{d+1}}^{(d+1)} \\ & \vdots & \\ & x^{(d+k)} = & x_{g_{d+k}}^{(d+k)} \end{array} \right. ,$$

with $g := (g_{d+1}, \ldots, g_{d+k}) \in \{1, \ldots, m\}^k$. For any $g \in \{1, \ldots, m\}^k$, let $E_{g,\tau} = E_\tau \cap \{(t, \mathbf{x}) \in \chi_l : (x^{(d+1)}, \ldots, x^{(d+k)}) = (x_{g_{d+1}}^{(d+1)}, \ldots, x_{g_{d+k}}^{(d+k)})\}$. Let $\mathcal{H}_d$ be the family of subsets of $\mathbb{R}^{d+1}$ which are projections on $\mathbb{R}^{d+1}$ of sets of E_τ (that is, we keep only the first d coordinates). Clearly, for any probability measure $\mathcal{Q}$,

$$N_1(\varepsilon, F_\tau, L^2(\mathcal{Q})) \leq \sum_{g \in \{1, \ldots, m\}^k} N_1(\varepsilon, F_{g,\tau}, L^2(\mathcal{Q})), \tag{B.1}$$

where $F_{g,\tau} = \{(t, x) \to \mathbf{1}_{(t,x) \in \chi} : \chi \in E_{g,\tau}\}$, and $N_1(\varepsilon, F_{g,\tau}, L^2(\mathcal{Q})) = N_1(\varepsilon, \mathcal{H}_d, L^2(\mathcal{Q}))$. Moreover, a set $H \in \mathcal{H}_d$ can be expressed as

$$H = \cap_{j=0,\ldots,d} \left(\{y \in \mathbb{R}^d :< y, e_j > \leq \beta_j\} \cap \{y \in \mathbb{R}^d :< y, e_j > \leq \alpha_j\}^c\right),$$

where A^c denotes the complement of a set A, e_j denotes the vector of $\mathbb{R}^{d+1}$ with all components equal to zero except the $(j+1)$-th one, and $< \cdot, \cdot >$ the scalar product in $\mathbb{R}^{d+1}$. It follows from Example 8.4 in van der Vaart and Wellner (1996), combined with points (i) and (ii) in Proposition 8.2 in the same (stability properties of VC-classes), that $\mathcal{H}_d$ is a VC-class of sets (see van der Vaart and Wellner (1996) for definition), with VC-index $2(d+1)(d+2)$. As a consequence,

$$N_1(\varepsilon, \mathcal{H}_d, L^2(\mathcal{Q})) \leq \left(\frac{K}{\varepsilon}\right)^{4(d+1)(d+2)},$$

for some universal constant K (see Dudley (1999)), and the result follows from (B.1). □

B.2. Proof of Proposition 2

Observe that, for $K > K_0$, $\pi^{\mathcal{S}_K} = \pi^{\mathcal{S}_{K_0}}$. Hence,

$$\begin{aligned}\|\hat{\pi}^{\hat{S}(\alpha)} - \pi^{\mathcal{S}_{K_0}}\|_{2,\tau}^2 = \|\hat{\pi}^{\mathcal{S}(K_0)} - \pi^{\mathcal{S}_{K_0}}\|_{2,\tau}^2 \mathbf{1}_{K_\alpha = K_0} + \sum_{K=1}^{K_0-1} \|\hat{\pi}^{\mathcal{S}_K} - \pi^{\mathcal{S}_{K_0}}\|_{2,\tau}^2 \mathbf{1}_{K_\alpha = K} \\ + \sum_{K=K_0+1}^{k_{max}} \|\hat{\pi}^{\mathcal{S}_K} - \pi^{\mathcal{S}_K}\|_{2,\tau}^2 \mathbf{1}_{K_\alpha = K}. \qquad \text{(B.2)}\end{aligned}$$

Following the proof of Corollary 1, one has $K^{-2}E[\|\hat{\pi}^{\mathcal{S}_K} - \pi^{\mathcal{S}_K}\|_{2,\tau}^4] = O(1/n^2)$. Hence, from the Cauchy-Schwarz inequality,

$$\begin{aligned}E\left[\frac{1}{K_0}\sum_{K=K_0+1}^{k_{max}} \|\hat{\pi}^{\mathcal{S}_K} - \pi^{\mathcal{S}_K}\|_{2,\tau}^2 \mathbf{1}_{K_\alpha = K}\right] \\ \leq \left(\sum_{k=K_0+1}^{k_{max}} \frac{K}{K_0} \mathbb{P}(K_\alpha = K)^{1/2}\right) \times O(n^{-1}).\end{aligned}$$

Rewrite

$$\begin{aligned}\sum_{k=K_0+1}^{k_{max}} K\mathbb{P}(K_\alpha = K)^{1/2} = K_0 \sum_{k=K_0+1}^{k_{max}} \mathbb{P}(K_\alpha = K)^{1/2} \\ + \sum_{k=K_0+1}^{k_{max}} [K - K_0]\mathbb{P}(K_\alpha = K)^{1/2}.\end{aligned}$$

Due to Lemma 5 below, we have

$$K_0^{-1} \sum_{k=K_0+1}^{k_{max}} K\mathbb{P}(K_\alpha = K)^{1/2} = O(1). \qquad \text{(B.3)}$$

Next, since there exists a finite constant $\mathcal{A}$ such that $\|\hat{\pi}^{\mathcal{S}_K}-\pi^{\mathcal{S}_{K_0}}\|_{2,\tau}^2 \leq \mathcal{A}$, we have

$$E\left[\sum_{K=1}^{K_0-1}\|\hat{\pi}^{\mathcal{S}_K}-\pi^{\mathcal{S}_{K_0}}\|_{2,\tau}^2\mathbf{1}_{K_\alpha=K}\right]\leq \mathcal{A}\sum_{K=1}^{K_0-1}\mathbb{P}(K_\alpha=K).$$

We now use Lemma 5 to deduce that

$$K_0^{-1}\sum_{K=1}^{K_0-1}\mathbb{P}(K_\alpha=K)=O(n^{-1}). \tag{B.4}$$

From (B.2), Corollary 1, and the combination of (B.3) and (B.4), we get

$$E\left[\frac{1}{K_0}\|\hat{\pi}^{\hat{\mathcal{S}}(\alpha)}-\pi^{\mathcal{S}_{K_0}}\|_{2,\tau}^2\right]=O(n^{-1}),$$

and the result follows from the fact that $\|\hat{\pi}^{\hat{\mathcal{S}}(\alpha)}-\pi_0\|_{2,\tau}\leq\|\hat{\pi}^{\hat{\mathcal{S}}(\alpha)}-\pi^{\mathcal{S}_{K_0}}\|_{2,\tau}+\|\hat{\pi}^{\mathcal{S}_{K_0}}-\pi_0\|_{2,\tau}$. We now state our auxiliary Lemma 5.

Lemma 5. *Under the Assumptions of Proposition 2, we have*

$$\frac{\mathbb{P}\left(K_\alpha=K\right)}{K}=\begin{cases} O(n^{-1}) & \text{if } K<K_0,\\ O(\exp(-\mathcal{C}_6'[K-K_0])) & \text{if } K>K_0,\end{cases}$$

for some positive constant $\mathcal{C}_6'<\infty$.

Proof. On the event $\{K_\alpha=K\}$, we have

$$\int\phi(m,\hat{\pi}^{\mathcal{S}_{K_0}}(\mathbf{x},t))d\hat{F}(m,t,\mathbf{x})-\int\phi(m,\hat{\pi}^{\mathcal{S}_K}(\mathbf{x},t))d\hat{F}(m,t,\mathbf{x})+\frac{\alpha[K_0-K]}{n}\geq 0. \tag{B.5}$$

We decompose the left-hand side of (B.5) into $A_1(K_0)-A_1(K)-\Delta(K)+A_2(K)-A_2(K_0)$, where

$$\begin{aligned} A_1(K) &= \int[\phi(m,\hat{\pi}^{\mathcal{S}_K}(\mathbf{x},t))-\phi(m,\pi^{\mathcal{S}_K}(\mathbf{x},t))]d[\hat{F}(m,t,\mathbf{x})-F(m,t,\mathbf{x})],\\ A_2(K) &= \int[\phi(m,\hat{\pi}^{\mathcal{S}_K}(\mathbf{x},t))-\phi(m,\pi^{\mathcal{S}_K}(\mathbf{x},t))]dF(m,t,\mathbf{x}).\end{aligned}$$

We have, due to the regularity of ϕ,

$$\begin{aligned} |A_1(K)| &\leq \mathcal{B}\|\hat{\pi}^{\mathcal{S}_K}-\pi^{\mathcal{S}_K}\|_{2,\tau}^2,\\ |A_2(K)| &\leq \mathcal{B}\|\hat{\pi}^{\mathcal{S}_K}-\pi^{\mathcal{S}_K}\|_{2,\tau}^2.\end{aligned}$$

We can distinguish two cases, depending whether $K<K_0$ or $K>K_0$.

A bound for $K<K_0$.

In this case, $\Delta(K)>0$, and

$$\mathbb{P}(K_\alpha=K) \quad\leq\quad \mathbb{P}\left(2\mathcal{B}\|\hat{\pi}^{\mathcal{S}_K}-\pi^{\mathcal{S}_K}\|>\frac{\Delta(K)-\alpha[K_0-K]/n}{2}\right)$$

$$+\mathbb{P}\left(2\mathcal{B}\|\hat{\pi}^{\mathcal{S}_{K_0}}-\pi^{\mathcal{S}_{K_0}}\|>\frac{\Delta(K)-\alpha[K_0-K]/n}{2}\right).$$

Hence,

$$\mathbb{P}(K_\alpha=K)/K=O\left(\exp\left(-\min(\mathcal{C}_6'\{\Delta(K)-\alpha[K_0-K]/n\},1)n\right)\right),$$

for some positive constant $\mathcal{C}_6'$ from Corollary 1. Using (3.8), we have

$$\mathbb{P}(K_\alpha=K)/K=O\left(\exp\left(-\min(\mathcal{C}_6'\mathcal{C}_6^{-1}[\log n]/n,1)n\right)\right),$$

and the result follows if $\mathcal{C}_6\leq\mathcal{C}_6'$.

A bound for $K>K_0$.
In this case, $\Delta(K)=0$, and $\alpha[K_0-K]/n<0$, and

$$\begin{aligned}\mathbb{P}(K_\alpha=K) \quad\leq\quad & \mathbb{P}\left(2\mathcal{B}\|\hat{\pi}^{\mathcal{S}_K}-\pi^{\mathcal{S}_K}\|>\alpha[K-K_0]/[2n]\right)\\ &+\mathbb{P}\left(2\mathcal{B}\|\hat{\pi}^{\mathcal{S}_{K_0}}-\pi^{\mathcal{S}_{K_0}}\|>\alpha[K-K_0]/[2n]\right).\end{aligned}$$

From Corollary 1, $\dfrac{\mathbb{P}(K_\alpha=K)}{K}=O\left(\exp(-\mathcal{C}_6'\alpha[K-K_0])\right)$ for some constant $\mathcal{C}_6'>0$. □

Appendix C: Concentration inequality

The following inequality was proved initially by Talagrand (1994). See also Einmahl and Mason (2005).

Proposition C1. *Let $(U_i)_{1\leq i\leq n}$ denote i.i.d. replications of a random vector U, and let $(\varepsilon_i)_{1\leq i\leq n}$ denote a vector of i.i.d. Rademacher variables (that is, $\mathbb{P}(\varepsilon_i=-1)=\mathbb{P}(\varepsilon_i=1)=1/2$) independent from $(U_i)_{1\leq i\leq n}$. Let $\mathcal{F}$ be a pointwise measurable class of functions bounded by a finite constant M_0. Then, for all u,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left\|\sum_{i=1}^n\{f(U_i)-E[f(U)]\}\right\|>\mathcal{A}_1\left\{E\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(U_i)\varepsilon_i\right|\right]+u\right\}\right)$$
$$\leq 2\left\{\exp\left(-\frac{\mathcal{A}_2u^2}{n\sigma^2_{\mathcal{F}}}\right)+\exp\left(-\frac{\mathcal{A}_2u}{M_0}\right)\right\},$$

with $\sigma^2_{\mathcal{F}}=\sup_{f\in\mathcal{F}}Var(f(U))$, and where $\mathcal{A}_1$ and $\mathcal{A}_2$ are universal constants.

The difficulty in using Proposition C1 comes from the need to control the symmetrized quantity $E\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n f(U_i)\varepsilon_i\right|\right]$. Proposition C2 is due to Einmahl and Mason (2005) and permits this control via some assumptions on the class of functions $\mathcal{F}$ considered.

Proposition C2. *Let $\mathcal{F}$ be a pointwise measurable class of functions bounded by M_0 such that, for some constants $\mathcal{C},\nu\geq 1$, and $0\leq\sigma\leq M_0$, we have*

(i) $\mathcal{N}_{M_0}(\varepsilon,\mathcal{F})\leq\mathcal{C}\varepsilon^{-\nu}$, *for* $0<\varepsilon<1$,

(ii) $\sup_{f\in\mathcal{F}} E\left[f(U)^2\right] \leq \sigma^2$,
(iii) $M_0 \leq \frac{1}{4\nu}\sqrt{n\sigma^2/\log(C_1 M_0/\sigma)}$, *with* $C_1 = \max(e, \mathcal{C}^{1/\nu})$.

Then, for some absolute constant $\mathcal{A}$,

$$E\left[\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n} f(U_i)\varepsilon_i\right|\right] \leq \mathcal{A}\sqrt{\nu n\sigma^2\log(C_1 M_0/\sigma)}.$$

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive comments that contributed to improving the paper.

References

Bacchetti, P. and Segal, M. R. (1995). Survival trees with time-dependent covariates: application to estimating changes in the incubation period of AIDS. *Lifetime Data Analysis* **1** 35–47.

Beran, R. (1981). Nonparametric regression with randomly censored survival data, Technical Report, University of California, Berkeley.

Bitouzé, D., Laurent, B. and Massart, P. (1999). A Dvoretzky-Kiefer-Wolfowitz type inequality for the Kaplan-Meier estimator. *Ann. Inst. H. Poincaré Probab. Statist.* **35** 735–763. MR1725709 (2000j:62143)

Bou-Hamad, I., Larocque, D. and Ben-Ameur, H. (2011). A review of survival trees. *Statistics Surveys* **5** 44–71. MR3018509

Breiman, L., Friedman, J., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall. MR0726392

Chaudhuri, P. (2000). Asymptotic consistency of median regression trees. *JSPI* **91** 229–238. MR1814782

Chaudhuri, P. and Loh, W.-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli* **8** 561–576. MR1935647

Ciampi, A., Negassa, A. and Lou, Z. (1995). Tree-structured prediction for censored survival data and the Cox model. *Journal of Clinical Epidemiology* **48** 675–689.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *Ann. Statist.* **17** 1157–1167. MR1015143 (90h:62089)

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge Studies in Advanced Mathematics. MR1720712

Dudoit, S., van Der Laan, M. J., Keles, S., Molinaro, A., Sinisi, S. E. and Teng, S. L. (2003). Loss-based estimation with cross-validation: Applications to microarray data analysis and motif finding.

Einmahl, U. and Mason, D. M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.* **13** 1–37. MR1744994 (2001b:62030)

Einmahl, U. and Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33** 1380–1403. MR2195639 (2006j:62041)

Fan, J., Nunn, M. E. and Su, X. (2009). Multivariate exponential survival trees and their application to tooth prognosis. *CSDA* **53** 1110–1121. MR2657075

Gannoun, A., Saracco, J., Yuan, A. and Bonney, G. E. (2005). Nonparametric quantile regression with censored data. *Scand. J. Statist.* **32** 527–550. MR2232341 (2007b:62034)

Gao, F., Manatunga, A. K. and Chen, S. (2004). Identification of prognostic factors with multivariate survival data. *CSDA* **45** 813–824. MR2054888

Gey, S. and Nedelec, E. (2005). Model selection for CART regression trees. *IEEE Transactions on Information Theory* **51** 658–670. MR2236074

Heagerty, P. J., Lumley, T. and Pepe, M. S. (2000). Time-Dependent ROC Curves for Censored SurvivalData and a Diagnostic Marker. *Biometrics* **56** 337-344.

Heagerty, P. J. and Zheng, Y. (2005). Survival Model Predictive Accuracy and ROC Curves. *Biometrics* **61** 92-105. MR2135849

Heuchenne, C. and Van Keilegom, I. (2010a). Estimation in nonparametric location-scale regression models with censored data. *Ann. Inst. Statist. Math.* **62** 439–463. MR2608458 (2011e:62282)

Heuchenne, C. and Van Keilegom, I. (2010b). Goodness-of-fit tests for the error distribution in nonparametric regression. *Comput. Statist. Data Anal.* **54** 1942–1951. MR2640298

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. and Van Der Laan, M. J. (2006). Survival ensembles. *Biostatistics* **7** 355-373.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457–481. MR0093867 (20 ##387)

Lopez, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Communications in Statistics: Theory and Methods* **40** 2639–2660. MR2860770

Lopez, O., Patilea, V. and Van Keilegom, I. (2013). Single index regression models in the presence of censoring depending on the covariates. *Bernoulli* **19** 721–747. MR3079294

Meinshausen, N. (2009). Forest garrote. *Electronic Journal of Statistics* **3** 1288–1304. MR2566188

Molinaro, A. M., Dudoit, S. and van der Laan, M. J. (2004). Tree-based multivariate regression and density estimation with right-censored data. *JMVA* **90** 154–177. MR2064940

Olbricht, W. (2012). Tree-based methods: a useful tool for life insurance. *European Actuarial Journal* **2** 129–147. MR2954472

Sánchez Sellero, C., González Manteiga, W. and Van Keilegom, I. (2005). Uniform representation of product-limit integrals with applications. *Scand. J. Statist.* **32** 563–581. MR2232343 (2007i:62078)

Satten, G. A. and Datta, S. (2001). The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. *Amer. Statist.* **55** 207–210. MR1947266

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.* **45** 89–103. MR1222607 (94d:62117)

Stute, W. (1999). Nonlinear censored regression. *Statist. Sinica* **9** 1089–1102. MR1744826

Stute, W. and Wang, J. L. (1993). The strong law under random censorship. *Ann. Statist.* **21** 1591–1607. MR1241280 (94j:62092)

Talagrand, M. (1994). Sharper bounds for Gaussian and empirical processes. *Ann. Probab.* **22** 28–76. MR1258865 (95a:60064)

van Der Laan, M. J. and Dudoit, S. (2003). Unified Cross-Validation Methodology for Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples.

van Der Laan, M. J., Dudoit, S. and van der Vaart, A. W. (2006). The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions* **24** 373–395. MR2305113

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. *Springer Series in Statistics*. Springer-Verlag, New York. MR1958123 (2003m:62003)

van der Vaart, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge. MR1652247

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. *Springer Series in Statistics*. Springer-Verlag, New York. MR1385671 (97g:60035) MR1385671

Van Keilegom, I. and Akritas, M. G. (1999). Transfer of tail information in censored regression models. *Ann. Statist.* **27** 1745–1784. MR1742508 (2001b:62082)

Wang, H. J. and Wang, L. (2009). Locally weighted censored quantile regression. *JASA* **104** 1117–1128. MR2562007

Wey, A., Wang, L. and Rudser, K. (2014). Censored quantile regression with recursive partitioning based weights. *Biostatistics* **15** 170–181.