Lifetime Data Anal (2017) 23:671–691
DOI 10.1007/s10985-016-9372-1

L_1 splitting rules in survival forests

Hoora Moradian[1] · **Denis Larocque**[1] · **François Bellavance**[1]

Received: 31 May 2015 / Accepted: 17 June 2016 / Published online: 5 July 2016

Abstract The log-rank test is used as the split function in many commonly used survival trees and forests algorithms. However, the log-rank test may have a significant loss of power in some circumstances, especially when the hazard functions or when the survival functions cross each other in the two compared groups. We investigate the use of the integrated absolute difference between the two children nodes survival functions as the splitting rule. Simulations studies and applications to real data sets show that forests built with this rule produce very good results in general, and that they are often better compared to forests built with the log-rank splitting rule.

Keywords Survival data · Right-censored data · Ensemble methods · Random forests · Survival forests

1 Introduction

There have been numerous studies on time-to-event data in a wide range of research areas. One important feature of survival data is that some observations are censored, that is these observations are incomplete since the event has not yet occurred at the time of data collection. In these situations, we must adequately incorporate all the available information to optimize the prediction models. Parametric models

Electronic supplementary material The online version of this article (doi:10.1007/s10985-016-9372-1) contains supplementary material, which is available to authorized users.

✉ Denis Larocque
denis.larocque@hec.ca

[1] Department of Decision Sciences, HEC Montréal, 3000 chemin de la Côte–Sainte–Catherine, Montreal, QC H3T 2A7, Canada

(Gamma, Weibull, etc.…), and semi-parametric ones such as the Cox proportional hazard model, can be useful and have been discussed in details in the literature (Hosmer Jr et al. 2011). However, (semi)parametric models have the important limitation that the functional link between the survival time and the covariates must be specified in advance. This is why more flexible nonparametric methods, like survival forests, that let the data automatically find the structure of the model, are useful alternatives (e.g., Hothorn et al. 2006a; Ishwaran et al. 2008; Zhu and Kosorok 2012).

Tree-based methods were originally developed to model the relation between covariates and either a categorical or a continuous outcome. The Classification and Regression Tree paradigm (CART) is widely popular (Breiman et al. 1984). Survival trees, introduced by Gordon and Olshen (1985), are an adaptation of the tree paradigm to right censored data. A variety of split rules have been suggested for survival trees so far. First, Gordon and Olshen (1985) used the idea of imposing homogeneity in each node through the use of a Wasserstein distance between the Kaplan–Meier estimators of the two survival functions. Even though this test does not require any underling assumption, it has not been used much in later works. Wilcoxon–Gehan statistics and Kolmogorov–Smirnov test are other metrics to maximize the heterogeneity between two children nodes that were proposed (Ciampi et al. 1988). However, the log-rank statistic proposed by Ciampi et al. (1986) gained the most popularity. The reader can refer to Bou-Hamad et al. (2011) for a review on various splitting statistics proposed in the literature.

As is well known, single trees, despite being a very powerful descriptive tool, can be unstable predictive tools. Ensemble methods constructed from trees as base learners such as random forests (Breiman 2001) can improve the predictive performance through additional randomization. The reader can refer to Siroky (2009) and Verikas et al. (2011) that provided recent surveys on random forests or to Rokach (2009) for a discussion on ensemble methods, in general. A similar argument can be applied to the survival data settings; a combination of survival trees generally leads to higher predictive accuracy. For more in-depth discussions on survival trees and forests, the reader can refer to Bou-Hamad et al. (2011) for a comprehensive overview, and to Boulesteix et al. (2012) and Chen and Ishwaran (2012) for an overview of the subject in genomics and bioinformatics.

Perhaps, the most popular random forest technique for survival data is the one proposed by Ishwaran et al. (2008), called random survival forest (RSF). It is implemented in their R package `randomForestSRC` (Ishwaran and Kogalur 2014). In this method, an ensemble of cumulative hazard function is built by averaging the Nelson–Aalen cumulative hazard function of each survival tree. It uses the log-rank statistic (Segal 1988; Leblanc and Crowley 1993) as the default splitting rule. Other available splitting rules in `randomForestSRC` are log-rank score (Hothorn and Lausen 2003) and random splitting rule (Cutler and Zhao 2001; Lin and Jeon 2006). According to Ishwaran et al. (2008), RSF is the only survival forest technique that adheres to all random forest principles introduced by Breiman (2001). They also provided a built-in new missing data handling algorithm which deals with two problems not addressed by the previous missing data methods for forests: (i) the biasedness of out-of-bag estimates of prediction error and (ii) the inability to predict on test data sets including missing values. Further, Ishwaran and Kogalur (2010) proved uniform consistency of RSF under the assumption that all variables are categorical. Ishwaran et al. (2010) came up with a regularization strategy for RSF, applicable to survival data where the sample size is small and the number of covariates is large. In their method, the importance of a covariate is measured by the tree depth at which the first split on that covariate happens, a concept called “the minimal depth of a maximal subtree”. This technique is useful for both variable selection and variable importance ranking. Ishwaran et al. (2011) provided a follow-up for the use of this method through the `randomSurvivalForest` package, the older version of `randomForestSRC`. They discussed ways to select the tuning parameters of random forest as well as a weighted variable selection technique in order to better regularize the forest. Chen and Ishwaran (2013) studied the use of the minimal depth concept through the `randomSurvivalForest` package in high–dimensional genomic data for effective pathway selection and suggested a “pathway hunting” algorithm for extremely high–dimensional data. Recently, Zhu and Kosorok (2012) proposed a nonparametric regression technique called recursively imputed survival tree (RIST) suitable for right-censored data. In this method, through calculation of the conditional survival distribution, censoring information of observations is retained and then, through recursive imputation and refitting steps, conditional failure information is constantly updated leading to higher predictive accuracy of the final model. The authors suggest three to five steps of imputation to get the best performance. In their implementation, the best split is again obtained through the log-rank test statistic.

From the above discussion, it appears that the log-rank test is routinely used in the various implementations of survival forests. This means that the best split is chosen as the one that makes the two children nodes the most significantly different according to this test. However, it is well known that the log-rank test may have a significant loss of power in some circumstances, especially when the hazard functions or when the survival functions cross each other in the two compared groups (Lin and Wang 2004; Lin and Xu 2010). This means that if the goal is to accurately estimate the conditional survival function, then using the log-rank test as splitting criterion may not be adequate. This is why we propose and investigate a splitting rule which works directly with the survival function, defined by:

$$L_1 = (n_L n_R) \int_t |\hat{S}_L(t) - \hat{S}_R(t)| dt, \tag{1}$$

where $\hat{S}_L(t)$, $\hat{S}_R(t)$, n_L and n_R are the Kaplan–Meier survival function estimates and the number of observations in the left and right node, respectively. We call it L_1 splitting rule. The L_1 splitting rule is related to the test statistic proposed by Lin and Xu (2010).

The rest of the paper is organized as follows. Section 2 describes the data setting and the proposed method. The results from a simulation study are presented in Sect. 3. It aims at comparing the proposed method to traditional and popular methods. Section 4 pursues the comparison with real data sets. Section 5 concludes and provides directions for further work.

2 Survival forest approach and splitting criterion

We have data on N independent subjects. For each subject i, observations are in the form of (τ_i, δ_i, x_i) where τ_i is the observed survival time, δ_i is the censoring index which takes a value of 0 if i is right censored and a value of 1 if i has experienced the event of interest, and x_i is a vector of covariates. Note that only time-invariant covariates are considered in this paper. The true time-to-event and the true censoring times for subject i are denoted by U_i and V_i, respectively. We have $\tau_i = \min(U_i, V_i)$ and assume that U_i and V_i are independent given x_i. The survival function for subject i is denoted by $S_i(t) = P(U_i > t)$. We use this simplified notation but it should be obvious that τ_i, δ_i, U_i and V_i depend on x_i.

We assume that the reader is familiar with the CART paradigm (Breiman et al. 1984) and the basic random forest method (Breiman 2001). Basically, a forest is a collection of large unpruned trees built on bootstrap samples from the original data. Moreover, at each node of any tree, a random subset of the predictors are selected at random and the best split is obtained from them. The final forest prediction is the average of the predictions from the individual trees.

The main focus of this paper is to investigate the use of a new splitting criterion to build the trees in the forest. Assume that we are at a given node of a tree and we want to split it in two children nodes. If x is continuous (or at least ordinal), the possible splits take the form $C = I(x \leq c)$. If x is categorical, the possible splits take the form $C = x \in \{c_1, \ldots, c_q\}$ where $\{c_1, \ldots, c_q\}$ is a subset of the possible values of x. Once the best split is found, observations with $C = 0$ go to the left node and the ones with $C = 1$ go to the right node. We saw in the introduction that, typically, the log-rank test with the two children nodes acting as the two samples is used as the splitting criterion. However, the log-rank test has a low power for detecting differences between the two groups in some situations. For the testing problem, Lin and Xu (2010) proposed a new method that has greater power than the log-rank test under a variety of situations. To avoid introducing unnecessary notation, suppose we want to test the equality of the survival functions in two children nodes, L (left) and R (right). Their test is based on

$$\begin{aligned}\Delta &= \int_0^{\tau} |\hat{S}_L(t) - \hat{S}_R(t)| dt \\ &= \sum_{j|t_j<\tau} |\hat{S}_L(t_j) - \hat{S}_R(t_j)|(t_{j+1} - t_j)\end{aligned}$$

where S_L and S_R are the Kaplan–Meier estimators of the survival function in nodes L and R, $t_1 < t_2 < \cdots < t_k$ are the pooled distinct event times, and τ is the last time point by which the areas under the survival curves can be calculated for both groups. To perform a formal test, Lin and Xu (2010) use the standardized statistic$\Delta^* = (\Delta - \hat{E}(\Delta))/(\widehat{Var}(\Delta))^{1/2}$ where $\hat{E}(\Delta)$ and $\widehat{Var}(\Delta)$ are suitable estimates of the mean and variance of Δ. However, since the splitting criterion has to be evaluated a large number of times when building a forest, we proceed to a simplification, detailed in Appendix, to speed up computations. With this simplification, we consider

$$L_1^* = \sqrt{n_L n_R}\Delta = \sqrt{n_L n_R} \int_t |\hat{S}_L(t) - \hat{S}_R(t)| dt \tag{2}$$

as the splitting criterion, where n_L and n_R are the left and right node sizes. We call it the L_1^* splitting criterion, as opposed to the L_1 splitting criterion given by $L_1 = (n_L n_R)\Delta$ and introduced in (1). As we will see in the next sections, both versions provide good results but the L_1 criterion was slightly better in the cases considered in this paper. For completeness, we will report the results for both splitting criteria. A more complete discussion about these two criteria appears in the concluding remarks section. To use any of these criteria for tree building, we simply compute it with the two groups formed by the left and right nodes for each candidate binary split. The one with the maximal value is the best split.

The forest algorithm can be described as follows:

1. Draw B bootstrap samples from the original data.
2. For each bootstrap sample, grow a tree with the L_1 (or L_1^*) splitting criterion. At each node, randomly select k out of p covariates where $k \leq p$ and is a user-specified parameter. Splitting ends when a stopping criterion is reached; for instance, when a node has less than a predetermined number of observations. No pruning is performed.
3. To compute the estimated survival function of an observation with covariate vector x, use a "similarity" weighting scheme as in Hothorn et al. (2006a). More precisely, send x in all B trees and collect all the observations that end in the same terminal nodes. Note that some observations may appear more than one time. $\hat{S}(t|x)$ is then the Kaplan–Meier estimate of this pooled set of observations.

Evaluating the performance of any model on a given data set with survival data is not a straightforward task because of the censoring. One approach is to use the Brier score (Graf et al. 1999) and other criteria derived from it. The `R` package `pec` (Mogensen et al. 2012) can be useful for that matter. We will use one of them, the integrated Brier score, in Sect. 4 when we analyse real data sets. We will also use the C-index (Harrell et al. 1982) as a complement measure. Evaluating how close an estimated survival curve is to the actual one is a lot easier in a simulation setting where the latter is known. In this case, well known criteria like the integrated absolute error and the integrated square error, as defined in the next section, can readily be used.

3 Simulation study

In this section, we investigate the performance of our proposed method through a simulation study. Nine methods are compared, seven forest methods and two benchmarks. They are:

A forest where trees are build with the proposed L_1 splitting rule. It is denoted by L_1-forest. A forest where trees are build with the proposed L_1^* splitting rule. It is denoted by L_1^*-forest. A forest where trees are build with the log-rank splitting rule. It is denoted by RFsrc1. A forest where trees are build with the log-rank score splitting rule. It is denoted by RFsrc2. A forest where trees are build with the random splitting rule. It is denoted by RFsrc3. A forest built with three-step imputation in the RIST method (Zhu and Kosorok 2012). It is denoted by RIST3. A forest built with five-step imputation in the RIST method (Zhu and Kosorok 2012). It is denoted by RIST5. A Cox model where the covariates are entered linearly with main effects only. This is the first benchmark. It is denoted by Cox. A Kaplan–Meier estimator, which does not use the covariates. This is the second benchmark. It is denoted by KM.

The L_1-forest and L_1^*-forest are implemented in Fortran and callable from `R` (R Core Team 2014). The `R` package `randomForestSRC` (Ishwaran and Kogalur 2014) was used for RFsrc1, RFsrc2 and RFsrc3. The R code generously made available by the authors Zhu and Kosorok (2012) was used for RIST3 and RIST5. Note that in case of no censoring in the data, no imputation is required with the RIST method. Therefore, a simple forest was used in these scenarios. We denote it by RIST0. The `R` package `survival` (Therneau 2014) was used for both the Cox model and the Kaplan–Meier estimator.

In all seven forest methods, 100 trees are grown and the number of covariates tried at each split is set to the integer part of $\sqrt{p}$, as suggested by Ishwaran et al. (2008). As a stopping criterion, the minimum number of observations in a terminal node is 3, the default value in `randomForestSRC`.

To evaluate the performance of these methods, two commonly used criteria were employed to measure how well the survival function is estimated. Assume that S is the true survival function and that $\hat{S}$ is the estimated survival function. The two criterion are the Integrated Absolute Error (IAE) and the Integrated Square Error (ISE) defined by:

$$\text{IAE} = \int_t |S(t) - \hat{S}(t)| dt \tag{3}$$

and

$$\text{ISE} = \int_t (S(t) - \hat{S}(t))^2 dt. \tag{4}$$

Since the results for the ISE were very similar, only the ones for the IAE are reported.

3.1 Simulation design

In the main simulation study, five Data Generating Processes (DGPs) are used to generate artificial data. For each DGP, six different censoring proportion ranging from 0 to 50 % are considered, namely, 0, 10, 20, 30, 40 and 50 %. Thus, overall 30 scenarios are investigated.

Each model is fitted with a training sample of size 500. Then the performance of the fitted models is evaluated with an independent test set of size 1000. Each simulation is repeated 500 times. Here are a detailed description of the DGPs. In all cases, the parameter α controls the proportion of censoring. The values of α which produce the desired censoring proportions were found empirically.

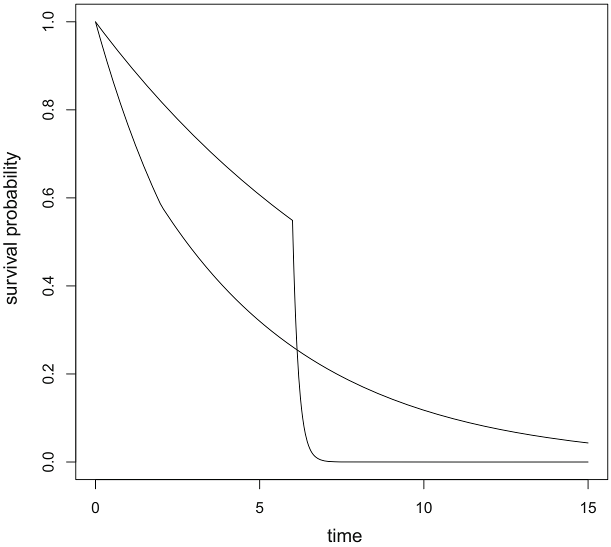

Fig. 1 The two survival curves in the terminal nodes of the tree for DGP 1

3.1.1 DGP 1

The model is a tree with two equally likely terminal nodes with respective survival functions illustrated in Fig. 1. Ten iid uniform covariates on the interval (0,1) are available, $X_1, \ldots, X_{10}$, but the response is only related to X_1. The censoring times are uniformly distributed on the interval (0,α). The hazard function is presented in Appendix.

3.1.2 DGP 2

This DGP is slightly more complex than DGP 1. It is a tree with four equally likely terminal nodes with respective survival functions illustrated in Fig. 2. Again, ten iid uniform covariates on the interval (0,1) are available, $X_1, \ldots, X_{10}$. The response is related to X_1 and X_2. The censoring times are uniformly distributed on the interval (0,α). The hazard function is presented in Appendix.

3.1.3 DGP 3

It is an altered version of scenario 2 from Sect. 4.1 of Zhu and Kosorok (2012). Ten iid uniform covariates on the interval (0,1) are available, $X_1, \ldots, X_{10}$. Survival times are drawn from an exponential distribution with mean μ where $\mu = 10|\sin(X_1\pi - 1)| + 3|X_2 - 0.5| + X_3$. The censoring times are uniformly distributed on the interval (0,α).

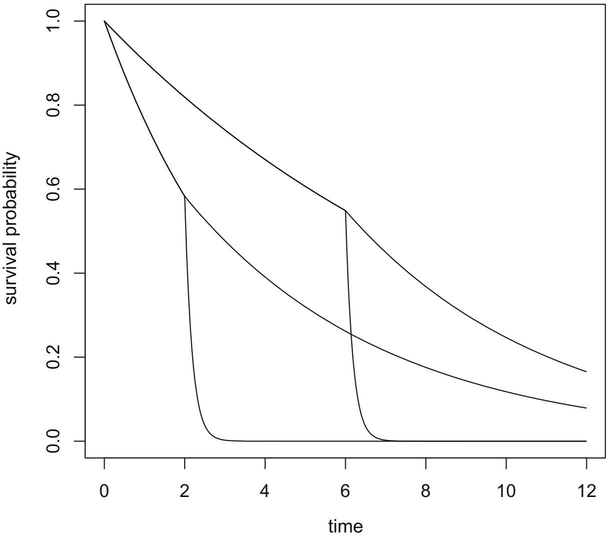

Fig. 2 The four survival curves in the terminal nodes of the tree for DGP 2

3.1.4 DGP 4

It is adapted from scenario 3 in Sect. 4.1 of Zhu and Kosorok (2012). Twenty-five covariates $X_1, \ldots, X_{25}$ are generated from a multivariate normal distribution with covariance matrix $\sigma_{ij} = 0.75^{|i-j|}$. The survival time follows a gamma distribution with shape parameter $\mu = 0.5 + 0.3|\sum_{i=11}^{15} X_i|$ and scale parameter of 2. The censoring times are uniformly distributed on the interval (0,α).

3.1.5 DGP 5

This is a dependent censoring DGP. It is adapted from scenario 1 in Sect. 4.1 of Zhu and Kosorok (2012). Twenty-five covariates $X_1, \ldots, X_{25}$ are generated from a multivariate normal distribution with covariance matrix $\sigma_{ij} = 0.9^{|i-j|}$. The survival time follows an exponential distribution with mean of $\mu = 0.1|\sum_{i=11}^{20} X_i|$. The censoring times are drawn from an exponential distribution with mean μ/α.

3.2 Simulation results

We first present a global summary of the results in Table 1. For each scenario with some censoring (that is, excluding the 0 % censoring case), we are comparing nine methods. For each individual data set, we ranked these methods from 1 to 9 with respect to the IAE criterion (3) evaluated on the test set. The rank of one was given to the method with the lowest value of the IAE, hence the best one for this data set. Table 1 reports the average ranks over all 12,500 simulation runs with censoring. Namely, over the 500 repetitions × 5 proportions of censoring × 5 DGPs. We see that the L_1 and L_1^*

Table 1 Average ranks (smaller is better), according to the IAE criterion, of the nine methods over all individual data sets (12,500) with censoring

Name	Cox	KM	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
Average rank	5.60	7.80	5.15	7.84	6.66	3.64	3.37	2.12	2.78

forests obtained the best results overall. The L_1-forest came in 2.12th place among the nine methods, on average, while the L_1^*-forest came in 2.78th place. The two RIST methods have the next best average ranks, followed by RFsrc1. Not surprisingly, the KM method which does not use the covariates, comes in last.

The detailed results for all DGPs are summarized in Figs. 1–5 of the Supplementary Material. As an overall summary, only results for DGP 1, DGP 3 and DGP 4 at 10 and 40 % censoring proportions are presented in Fig. 3. As expected, the L_1-type forests are the best performing method in terms of IAE in the first two DGPs with crossing survival functions. RIST comes in second place for DGP 1 while RIST and RFsrc1 also perform well for DGP 2. The first two DGPs were designed explicitly to exhibit the advantage of the proposed methods when the survival functions are crossing. But what is even more interesting is that the L_1-type forests also do very well for the other DGPs, that do not involve crossing survival functions. For DGP 3, the RFsrc1 is the best for censoring proportions up to 20 %, then the L_1^*-forest does better when the censoring proportion reaches 40 %. For DGP 4 and 5, RIST performs best followed closely by the L_1-type forests. Hence the L_1-type forests seem to be generally competitive in a wide variety of situations.

3.3 Additional simulations

We present briefly the results of some additional simulations following reviewers suggestions. Complete results with seven additional figures are available in the Supplementary Material.

Firstly, the performance of the methods is investigated with a smaller training sample and with additional noise covariates. The same 30 scenarios are investigated. But the training sample size is divided by two (hence it is 250) and the number (which depends on the DGP) of noise covariates is multiplied by 5 in each DGP. The added noise covariates are iid uniform covariates on the interval (0,1). Table 2 presents the average ranks of the nine methods for these modified DGPs. The results are very similar to those obtained before. Again, L_1 type forests obtained the best results overall, followed by the two RIST methods and then by RFsrc1.

Secondly, since all covariates are continuous in the main simulation, a few scenarios with binary covariates are investigated here. Only DGP 2 is considered. Recall that only two covariates, X_1 and X_2, are related to the response. In the main simulation, X_1 and X_2 are uniformly distributed on (0,1). In the first variation, X_2 is still uniformly distributed on (0,1) but X_1 is now a binary covariate taking values 0 and 1 with probability 1/2. In the second variation, both X_1 and X_2 are binary covariates taking

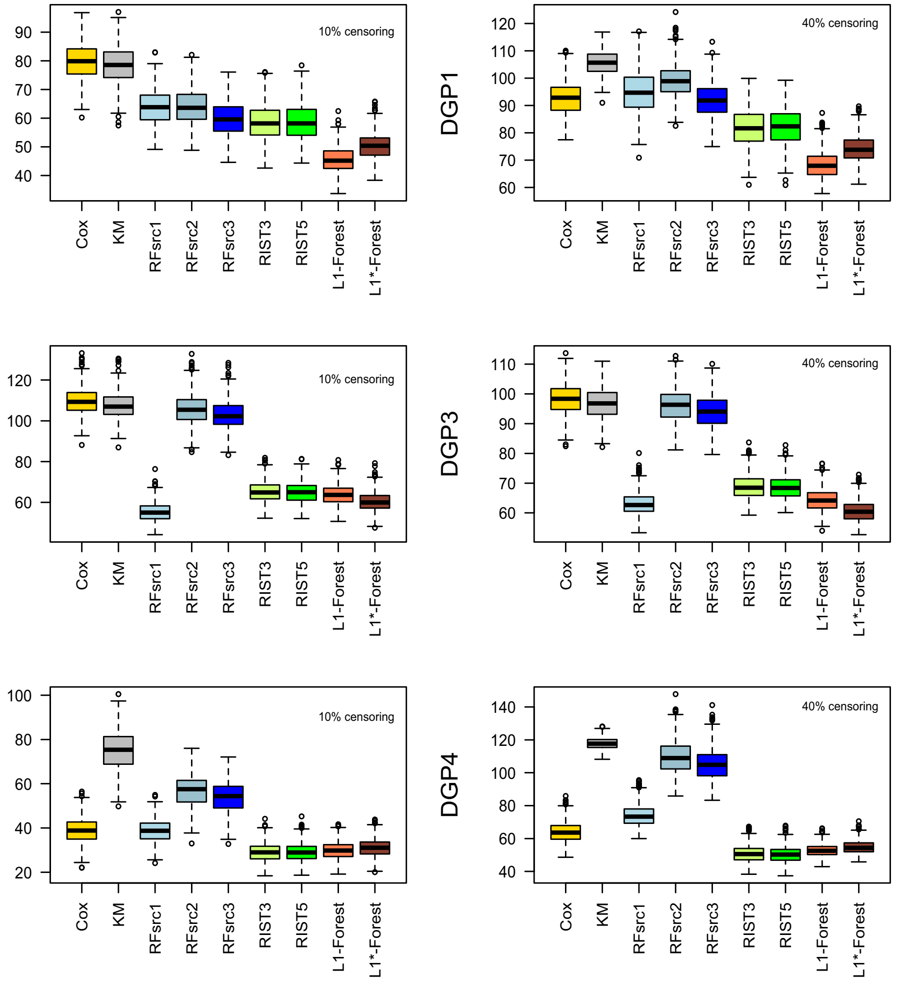

Fig. 3 IAE of methods for DGP 1, DGP 3 and DGP 4 at 10 and 40 % censoring proportions

values 0 and 1 with probability 1/2. Table 3 presents the average ranks, as before, for DGP 2 only, for these three situations. For the original setup where both X_1 and X_2 are continuous (top part of the table), the L_1 type forests are the best but the Cox model comes in third place, followed by the two RIST methods. Then, when X_1 is binary and X_2 continuous (middle part), the Cox model comes in between the L_1 type forests, followed again by the two RIST methods. Finally, when both X_1 and X_2 are binary (bottom part), then the Cox model has the best performance followed by the L_1-forest. But this time both RIST perform better than the L_1^*-forest. Again, more complete results are presented in the Supplementary Material which suggest that the IAE of the forest methods did not really degrade when moving from the continuous to the binary covariates. Rather, it is the performance of the Cox model that improved.

Table 2 Average ranks (smaller is better), according to the IAE criterion, of the nine methods over all individual data sets (12,500) with censoring, for the modified DGPs (smaller n and larger p)

Method	Cox	KM	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
Average rank	7.90	6.26	4.57	7.39	6.85	3.76	3.71	2.26	2.30

Table 3 Average ranks (smaller is better), according to the IAE criterion, of the nine methods over all individual data sets (2,500) with censoring for DGP 2

X_1 and X_2 continuous (original setup)									
Name	Cox	KM	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
Ranking	3.75	7.18	5.42	7.58	6.77	5.17	5.19	1.65	2.24
X_1 binary and X_2 continuous									
Name	Cox	KM	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
Ranking	2.63	7.25	6.69	7.70	6.39	4.39	4.50	2.39	3.04
X_1 and X_2 binary									
Name	Cox	KM	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
Ranking	2.42	6.76	7.15	7.57	5.65	4.07	4.08	2.99	4.31

The original setup and two variations are presented

4 Real data sets

In this section, we compare the performance of the same methods used in the simulation study with six real data sets: The Primary Biliary Cirrhosis (PBC) data, the CSL liver chirrosis data, the German Breast Cancer (GBC) Study Group data, the Wisconsin Breast Cancer Prognostic (WPBC) data, the Veteran data, and the National Wilm's Tumor Study (NWTCO) data. A brief description of these data sets is presented in Table 4.

The PBC data is described in the monograph by Fleming and Harrington (1991). We use all twelve covariates used by Bou-Hamad et al. (2011) plus copper, sgot and stage. The same 312 patients who participated in the randomized trial are used here. Missing values are replaced by the median as in Bou-Hamad et al. (2011) and Fleming and Harrington (1991). The CSL data was obtained by Schlichting et al. (1983) and is provided in the `timereg` package (Scheike et al. 2009). In this example, we only use the six time–invariant covariates. Records are grouped by id variable so the number of observations used is 446. The GBC data (Schumacher et al. 1994) is obtained from the package `mfp` (Ambler and Benner 2014). The data contains 686 observations and eight covariates. There is no missing data. The WPBC data is available in the UCI machine learning repository (Bache and Lichman 2013). There are 198 observations in the data. However, four missing values are replaced by the median as in the PBC data. Thirty-two covariates are used in this example. The Veteran data (Kalbfleisch and Prentice 1980) is obtained from the `randomForestSRC` package (Ishwaran and Kogalur 2014). There are 137 observations with no missing values. It contains six covariates. Finally, the NWTCO data (Breslow and Chatterjee 1999) is available in the package `survival` (Therneau 2014). The four relevant covariates, instit, histol, age and stage, are used here. The data consists of 4088 observations and no missing values.

We use the same parameters as in the simulation section to build the forests. Namely, 100 trees are grown, the number of covariates tried at each split is set to the integer part of $\sqrt{p}$, and the minimum number of observations in a terminal node is 3.

Since the true survival function is not known, we can not compute the IAE (or ISE) as we did with the artificial data sets. Instead, our primary criterion is the integrated Brier score (Graf et al. 1999). Let $\hat{S}(t|x)$ denote the estimated survival function, estimated by any model, at time t for a subject with covariate vector x. Let $\hat{G}$ denote the Kaplan–Meier estimate of the censoring distribution. The Brier score at any time t is computed as

$$\begin{aligned}\mathrm{BS}(t) = \frac{1}{N}\sum_{i=1}^{n}\Bigg(& (\hat{S}(t|x_i)^2 I(\tau_i \le t \textit{ and } \delta_i = 1)\hat{G}^{-1}(\tau_i) \\ & + (1 - \hat{S}(t|x_i))^2 I(\tau_i > t)\hat{G}^{-1}(t)\Bigg).\end{aligned}$$

The integrated Brier score is given by

$$\mathrm{IBS} = \frac{1}{\max(\tau_i)}\int_0^{\max(\tau_i)} BS(t)dt.$$

Lower values of IBS indicate better performances. Basically, the IBS is an integrated weighted squared distance between the estimated survival function and the empirical survival curve. The inverse weighting scheme is used to adjust for censoring. It is thus similar in spirit to the IAE used in the previous section.

Table 4 Description of the data sets

Name	# Covariates	Sample size	% Censoring	Source
PBC	15	312	60	Fleming and Harrington (1991)
CSL	6	446	39	Schlichting et al. (1983) in `timereg` package (Scheike et al. 2009)
GBC	8	686	56	Schumacher et al. (1994) in `mfp` package (Ambler and Benner 2014)
WPBC	32	198	76	(Bache and Lichman 2013)
Veteran	6	137	7	Kalbfleisch and Prentice (1980) in `randomForestSRC` package (Ishwaran and Kogalur 2014)
NWTCO	4	4088	85	Breslow and Chatterjee (1999) in `survival` package (Therneau 2014)

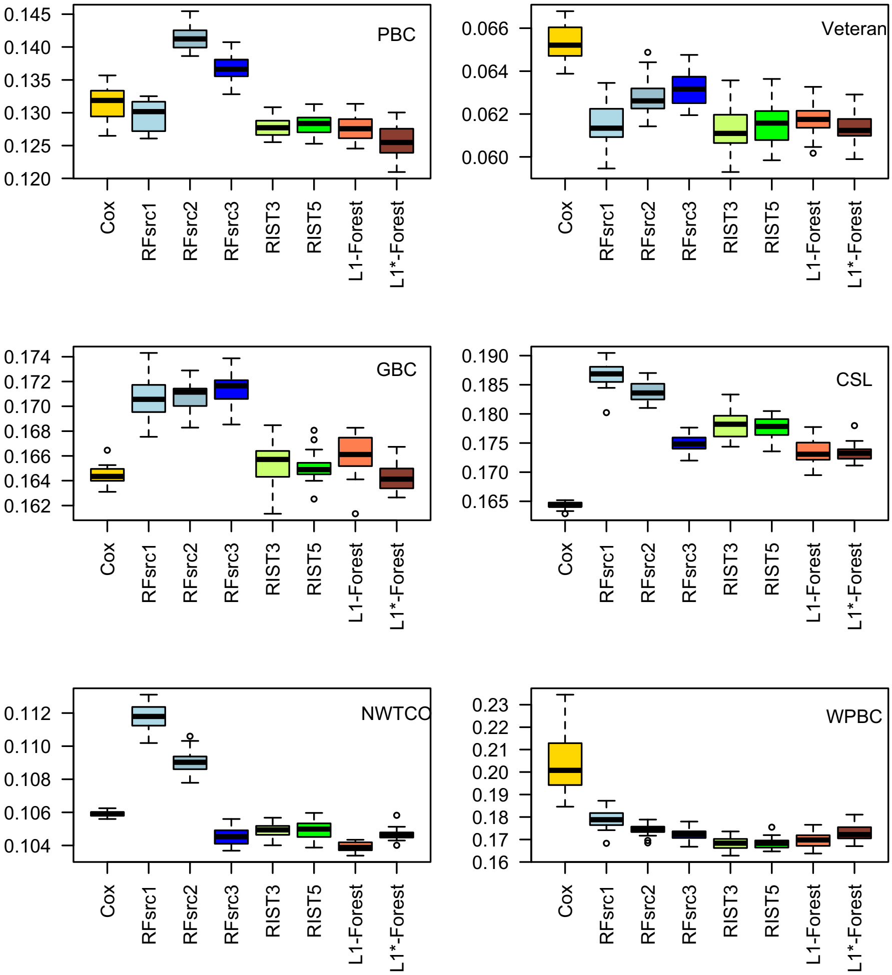

Fig. 4 Integrated Brier score, across 20 runs of 10-fold cross-validation, for the real data sets

Table 5 Ranking of methods based on % increase of median IBS with respect to best method

Name	Cox	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
PBC	5.08	3.73	12.55	8.89	1.78	2.29	1.66	0.00
Veteran	6.72	0.39	2.47	3.38	0.00	0.77	1.05	0.22
GBC	0.12	3.92	4.26	4.58	0.96	0.46	1.20	0.00
CSL	0.00	13.72	11.73	6.41	8.46	8.22	5.32	5.43
NWTCO	1.95	7.62	4.96	0.62	1.02	1.07	0.00	0.71
WPBC	19.30	6.27	3.78	2.45	0.04	0.00	0.90	2.33
Average	5.52	5.90	6.62	4.36	2.03	2.14	1.69	1.44

Figure 4 illustrate the results for the IBS across 20 runs of 10-fold cross-validation for each data set. Note that the results for the Kaplan–Meier method are not included because it is so much worse than the other methods that the plots would have been distorted. Table 5 provide another look at these results by computing the % increase of the median IBS with respect to best method for each data set. The median here is the one over the 20 runs of 10-fold cross-validation. We see in Table 5 that the L_1 type forests are globally the best according to the average % increase in IBS. Indeed, the IBS of L_1^*-forest is only 1.44 % greater than the IBS of the best method, on average. It is even the best one for two of the six data sets. Moreover, the IBS of L_1-forest is 1.69 % greater than the IBS of the best method, on average. It is also the best method for the NWTCO data. The two RIST methods have the next best performance with average percent increases just above 2 %. Hence, we get the same top four methods as the ones we obtained with the artificial data sets. The fact that the L_1 type forests are always competitive for all data sets is clearly seen in Fig. 4.

Another popular criterion to evaluate a model with survival data is the C-index (Harrell et al. 1982). We use it as a complement measure here because we think the IBS is more appropriate when the goal is to estimate the survival function. The C-index is a concordance measure that evaluates if the predictions from a model are ranked in the same way as the observed times. Following Ishwaran et al. (2008), let $\hat{H}(t|x)$ denote the estimated cumulative hazard function, estimated by any model, at time t for a subject with covariate vector x. Let $t_1 < t_2 < \cdots < t_m$ be the distinct event times and let $\hat{H}_i = \sum_{l=1}^{m} \hat{H}(t_l|x_i)$. The C-index, using the usable pairs, is computed as following

$$\mathrm{CI} = \frac{\sum_{i<j}(I(t_i < t_j)I(\hat{H}_i > \hat{H}_j)\delta_i + I(t_i > t_j)I(\hat{H}_i < \hat{H}_j)\delta_j)}{\sum_{i<j}(I(t_i < t_j)\delta_i + I(t_i > t_j)\delta_j)}.$$

Higher values of CI indicate better performances. Figure 5 illustrate the results for the CI across 20 runs of 10-fold cross-validation for each data set. Table 6 reports the % decrease of the median CI with respect to best method for each data set. We see in Table 6 that the Cox model has the best performance according to the CI, followed by the two RIST methods and the L_1 type forests. The average percent decrease in CI of these methods are all below 2 %. As a matter of fact, all methods do fairly well except maybe for RFsrc1 which is a bit further apart. Hence, it seems that the CI is a less discriminating criterion compared to the IBS, for these data sets. Zhu and Kosorok (2012) also report that the C-index is not as sensitive as other measurements and even that its interpretability is sometimes unclear. This may be partly explained by the fact that the C-index is uniquely a discrimination measure. That is, it measures if the predicted survival times are in the right order. The IBS is a discrimination and calibration measure. The calibration aspect measures the similarity between the actual and predicted survival curves; see De Bin et al. (2014). Hence for these data sets, it seems that all methods do fairly well in terms of discrimination, but the L_1 type forests perform better in terms of calibration.

To conclude this section, we will take a closer look at the GBC data analysed in details in Sauerbrei and Royston (1999). We will refer to this paper by SR for

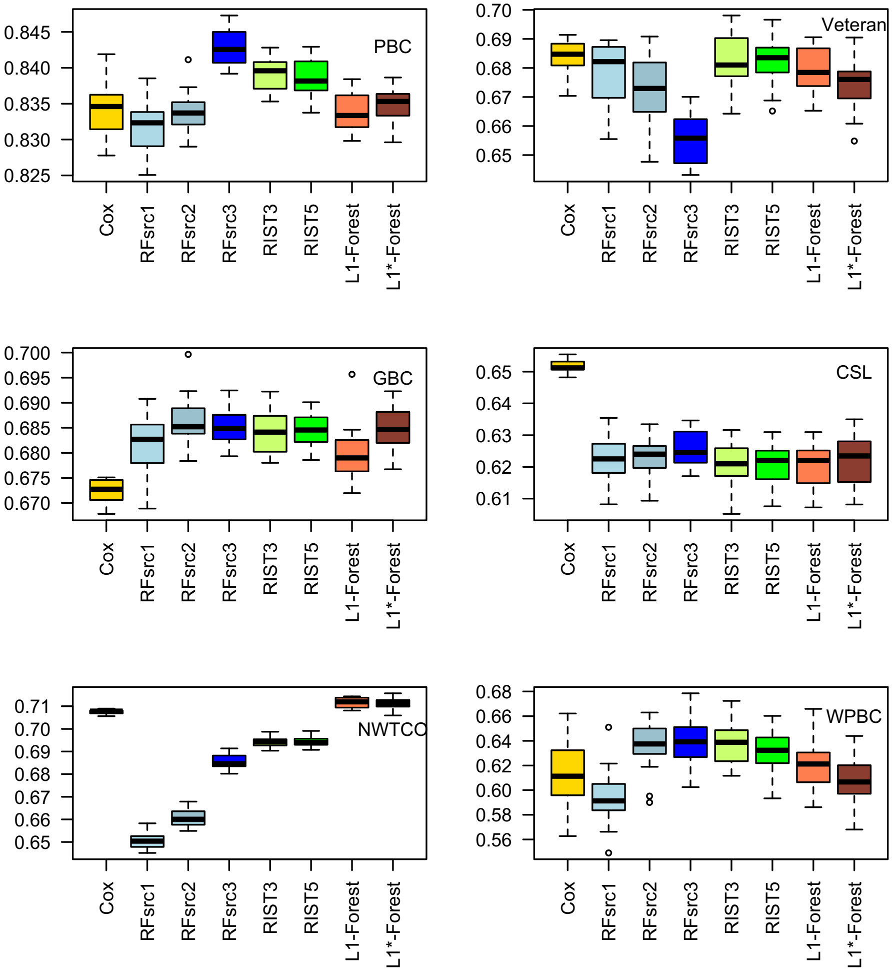

Fig. 5 C-index, across 20 runs of 10-fold cross-validation, for the real data sets

Table 6 Ranking of methods based on % decrease of median C-index with respect to best method

Name	Cox	RFsrc1	RFsrc2	RFsrc3	RIST3	RIST5	L_1-forest	L_1^*-forest
PBC	0.94	1.21	1.05	0.00	0.35	0.52	1.09	0.86
Veteran	0.00	0.37	1.72	4.22	0.54	0.17	0.92	1.26
GBC	1.81	0.36	0.00	0.05	0.15	0.09	0.90	0.07
CSL	0.00	4.40	4.17	4.09	4.64	4.46	4.48	4.25
NWTCO	0.57	8.64	7.28	3.83	2.48	2.52	0.00	0.09
WPBC	4.36	7.49	0.27	0.00	0.04	1.05	2.80	5.08
Average	1.28	3.74	2.41	2.03	1.37	1.47	1.70	1.94

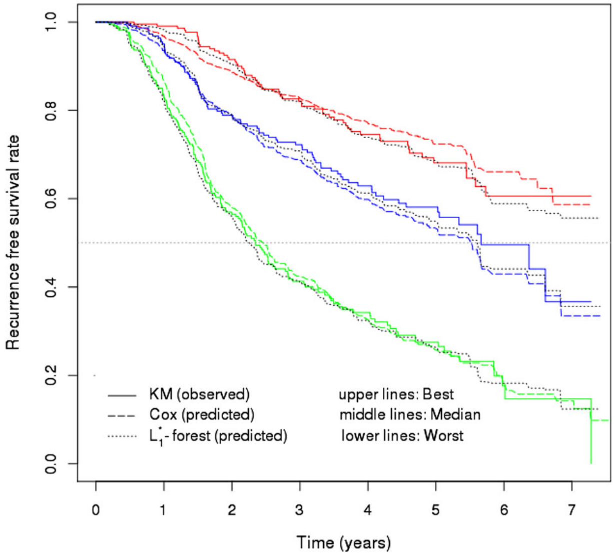

Fig. 6 Predicted and observed survival curves by prognosis group in GBC data

simplicity. After a careful analysis and using fractional polynomials, SR came up with a seemingly good Cox model based on the covariates $(X_1/50)^{-2}$, $(X_1/50)^{-0.5}$, $I(X_4 \geq 2)$, $\exp(-0.12X_5)$, and $(X_6+1)^{0.5}$; see Model III in Table 4 of SR. Following SR, we divide the sample into three groups of nearly equal sizes based on the prognostic index $\mathrm{PI}_i = x_i\hat{\beta}$, where $\hat{\beta}$ are the estimates of the parameters in the above Cox model. The 228 subjects with the lowest PI scores are assigned to the best prognosis group, the following 229 subjects are assigned to the median prognosis group, and the remaining 229 with the highest PI scores go into the worse prognosis group. Note however that SR used another simpler model to define the three groups. But that model had only 27 different covariates patterns which caused the group sizes to be imbalanced. Figure 4 of SR shows the Kaplan–Meier curves of each group. Fleming and Harrington (1991) present a similar plot (see their Figure 4.6.13 on page 195) in their analysis of the PBC data set. They also plotted the average estimated survival curves of each group in order to visually inspect the goodness-of-fit of their model. Figure 6 is such a plot where the average survival curves of the Cox model and of the L_1^*-forest are depicted. We can observe that the Cox model seems to fit the data fairly well in all groups. But strikingly, the L_1^*-forest seems to fit the data slightly better, even though the groups are derived from the Cox model. This simple example illustrates the fact that, sometimes, a good off-the-shelf method like a forest do as well as a carefully crafted parametric model. When confronted between two models with similar performance, the choice should often be the one which is easier to interpret. In this case, it would be the Cox model even though interpreting the effect of the transformed covariates is not straightforward. Thus, a forest can serve as a benchmark to evaluate if an easier to interpret parametric model fits well enough.

5 Discussion and concluding remarks

The log-rank test is commonly used as the splitting rule in the various implementations of survival forests within the CART paradigm, such as in Ishwaran et al. (2008) and Zhu and Kosorok (2012). However, the log-rank test is not designed to detect all possible differences between two survival curves. For instance, the log-rank test is inadequate to detect a difference between two groups when the hazard or survival functions cross each other in the two compared groups. Consequently, if the goal is to accurately estimate the conditional survival function, then using the log-rank test as splitting criterion may not be optimal. This was never thoroughly investigated for survival forests. At a time where more refinements and features are added to existing packages, it seemed that going back to "basics" was in order. In this paper, it was showed that forests built with a simple splitting rule, based on the integrated absolute difference between the two children nodes survival functions, are very competitive compared to forests built with the log-rank splitting rule. Indeed, these forests often got the best performance in the cases considered, either with simulated data or with real data sets. Hence, it is certainly worthwhile to consider the proposed methods as potential competitors. It would certainly be helpful if some well established, very comprehensive and useful package like `randomForestSRC` could incorporate L_1-type splitting rules.

The two splitting rules investigated are $L_1 = (n_L n_R)\Delta$ and $L_1^* = \sqrt{n_L n_R}\Delta$ where $\Delta = \int_t |\hat{S}_L(t) - \hat{S}_R(t)| dt$. The factors $(n_L n_R)$ and $\sqrt{n_L n_R}$ can be seen as penalization factors that favor splits with children nodes of nearly equal sizes. For instance, if we have two potential splits with the same value of Δ, then the one with the largest value of $n_L n_R$ should be favored because the two $\hat{S}$ in Δ are obtained from larger sample sizes. As an extreme example, assume that $n_L + n_R = 100$ and that two splits have an equal value of Δ. We would be more confident with a split with $n_L = 50$ and $n_R = 50$, than one with $n_L = 5$ and $n_R = 95$ because, in the second case, there is a lot of variability in the estimation of $\hat{S}_L$ which induces a larger variance for Δ. In fact, Appendix establishes that, under some simplifying assumptions, the factor $\sqrt{n_L n_R}$ is the one producing a test statistic which is asymptotically normally distributed. The factor $\sqrt{n_L n_R}$ favors more heavily splits with nearly equal size children nodes than the factor $n_L n_R$. But in practice, at a given node, we do not know if the best split is located more towards the center (with respect to the children nodes sizes) or not. So we do not know how much we should favor splits towards the center. Even if the factor $\sqrt{n_L n_R}$ has a theoretical justification, the factor $n_L n_R$ had globally a slightly better performance with the data generating processes considered in the simulation study. But with the real data sets, the factor $\sqrt{n_L n_R}$ had a slightly better performance according to the integrated Brier score and was slightly worse according to the C-index. In practice, one possibility would be to try forests built with both splitting rules and select the one according to the integrated Brier score under a cross-validation scheme. More research on this aspect could lead to more specific recommendations.

The scope of this work is limited to traditional forests built within the CART paradigm. Other paradigms including "unbiased" trees are also available; see Loh (2002, 2013) for the GUIDE approach and Hothorn et al. (2006b) for the ctree approach. For example, the GUIDE approach fits different types of proportional hazards regression models for censored data. Hence, it would be interesting to investigate the robustness of this method when the proportionality assumption is not met. Moreover, we saw in the additional simulations that the covariates' types can have an impact on the performance. Hence, developing forests built with unbiased trees using L_1 types criteria for variable and split selections might be interesting.

Acknowledgements The authors would like to thank the Associate Editor and two reviewers whose comments helped in preparing an improved version of this article. This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by Le Fonds québécois de la recherche sur la nature et les technologies (FQRNT).

Appendix

Hazard function formula for DGP 1

$$\begin{cases} 0.27t & x_1 \le 0.5, t \le 2 \\ 0.2(t-2)+5.4 & x_1 \le 0.5, t > 2 \\ 0.1t & x_1 > 0.5, t \le 6 \\ 5.5(t-6)+0.6 & x_1 > 0.5, t > 6. \end{cases}$$

Hazard function formula for DGP 2

$$\begin{cases} 0.27t & x_1 \le 0.5, x_2 \le 0.5, t \le 2 \\ 0.2(t-2)+5.4 & x_1 \le 0.5, x_2 \le 0.5, t > 2 \\ 0.27t & x_1 \le 0.5, x_2 > 0.5, t \le 2 \\ 5.5(t-2)+5.4 & x_1 \le 0.5, x_2 > 0.5, t > 2 \\ 0.1t & x_1 > 0.5, x_2 \le 0.5, t \le 6 \\ 0.2(t-6)+0.6 & x_1 > 0.5, x_2 \le 0.5, t > 6 \\ 0.1t & x_1 > 0.5, x_2 > 0.5, t \le 6 \\ 5.5(t-6)+0.6 & x_1 > 0.5, x_2 > 0.5, t > 6. \end{cases}$$

Simplification of the Lin and Xu (2010) statistic leading to the L_1^* splitting rule

For $i = L, R$, the left and right nodes, denote by $\hat{\sigma}_i^2$ the estimated variance of $\hat{S}_i$ from Greenwood's formula. To perform a formal test of the equality of the survival functions in the left and right nodes, Lin and Xu (2010) propose the statistic

$$\Delta^* = \frac{\Delta - \hat{E}(\Delta)}{\sqrt{\widehat{Var}(\Delta)}}$$

where

$$\hat{E}(\Delta) = \sum_{j|t_j<\tau} \{2/\pi(\hat{\sigma}_L^2(t_j) + \hat{\sigma}_R^2(t_j))\}^{1/2}(t_{j+1} - t_j)$$

and

$$\begin{aligned}\widehat{Var}(\Delta) = &\sum_{j|t_j<\tau} (t_{j+1} - t_j)^2(1 - 2/\pi)(\hat{\sigma}_L^2(t_j) + \hat{\sigma}_R^2(t_j))\\ &+ \sum_{j<j'|t_j,t_{j'}<\tau} (t_{j+1} - t_j)(t_{j'+1} - t_{j'})(1 - 2/\pi)\\ &\times\{(\hat{\sigma}_L^2(t_j) + \hat{\sigma}_R^2(t_j))(\hat{\sigma}_L^2(t_{j'}) + \hat{\sigma}_R^2(t_{j'}))\}^{1/2}\end{aligned}$$

are estimates of $E(\Delta)$ and $Var(\Delta)$. These estimates arise from a normal approximation for $\hat{S}_L(t) - \hat{S}_R(t)$, and the test statistic Δ^* is asymptotically normally distributed under the null hypothesis of equality of the two survival functions. To simplify this statistic in order to speed up computations for tree building, assume that all observations are from the same population with survival function $S(t)$, that is we are under the null hypothesis and there is no censoring. Then $Var(\hat{S}_i(t)) = S(t)(1 - S(t))/n_i$, for $i = L, R$. In that case,

$$\begin{aligned}\hat{E}(\Delta) &= \sqrt{2/\pi}\sqrt{(n_L + n_R)/(n_L n_R)} \sum_{j|t_j<\tau} (S(t_j)(1 - S(t_j)))^{1/2}(t_{j+1} - t_j)\\ &= c_1/\sqrt{n_L n_R}\end{aligned}$$

where c_1 is the same constant for all candidate splits. Similarly, $\widehat{Var}(\Delta) = c_2^2/(n_L n_R)$ where c_2 is the same constant for all candidate splits. Hence,

$$\frac{\Delta - \hat{E}(\Delta)}{\sqrt{\widehat{Var}(\Delta)}} = \frac{\sqrt{n_L n_R}\Delta}{c_2} - \frac{c_1}{c_2}.$$

But using this last expression is equivalent to using $\sqrt{n_L n_R}\Delta$ as the splitting criterion.

References

Ambler G, Benner A (2014) mfp: multivariable fractional polynomials. R package version 1.5.0. http://CRAN.R-project.org/package=mfp

Bache K, Lichman M (2013) UCI machine learning repository. http://archive.ics.uci.edu/ml

Bou-Hamad I, Larocque D, Ben-Ameur H (2011) A review of survival trees. Stat Surv 5:44–71

Boulesteix AL, Janitza S, Kruppa J, König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov 2(6):493–507

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth & Brooks, Monterey

Breslow NE, Chatterjee N (1999) Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. J R Stat Soc Ser C (Appl Stat) 48(4):457–468

Chen X, Ishwaran H (2012) Random forests for genomic data analysis. Genomics 99(6):323–329

Chen X, Ishwaran H (2013) Pathway hunting by random survival forests. Bioinformatics 29(1):99–105

Ciampi A, Thiffault J, Nakache JP, Asselain B (1986) Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. Comput Stat Data Anal 4(3):185–204

Ciampi A, Hogg SA, McKinney S, Thiffault J (1988) Recpam: a computer program for recursive partition and amalgamation for censored survival data and other situations frequently occurring in biostatistics. i. methods and program features. Comput Methods Progr Biomed 26(3):239–256

Cutler A, Zhao G (2001) Pert-perfect random tree ensembles. Comput Sci Stat 33:490–497

De Bin Riccardo, Sauerbrei Willi, Boulesteix Anne-Laure (2014) Investigating the prediction ability of survival models based on both clinical and omics data: two case studies. Stat Med 33(30):5310–5329

Fleming TR, Harrington DP (1991) Counting processes and survival analysis. Wiley, Hoboken

Gordon L, Olshen RA (1985) Tree-structured survival analysis. Cancer Treat Rep 69(10):1065

Graf E, Schmoor C, Sauerbrei W, Schumacher M (1999) Assessment and comparison of prognostic classification schemes for survival data. Stat Med 18(17–18):2529–2545

Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. JAMA 247(18):2543–2546

Hosmer DW Jr, Lemeshow S, May S (2011) Applied survival analysis: regression modeling of time to event data. Wiley, Chichester

Hothorn T, Lausen B (2003) On the exact distribution of maximally selected rank statistics. Comput Stat Data Anal 43(2):121–137

Hothorn T, Bühlmann P, Dudoit S, Molinaro A, Van Der Laan MJ (2006a) Survival ensembles. Biostatistics 7(3):355–373

Hothorn T, Hornik K, Zeileis A (2006b) Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat 15(3):651–674

Ishwaran H, Kogalur UB (2010) Consistency of random survival forests. Stat Probab Lett 80(13):1056–1064

Ishwaran H, Kogalur UB (2014) Random forests for survival, regression and classification (rf-src). R package version 1.5.5. http://cran.r-project.org/web/packages/randomForestSRC/

Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. Ann Appl Stat 2(3):841–860

Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS (2010) High-dimensional variable selection for survival data. J Am Stat Assoc 105(489):205–217

Ishwaran H, Kogalur UB, Chen X, Minn AJ (2011) Random survival forests for high-dimensional data. Stat Anal Data min 4(1):115–132

Kalbfleisch JD, Prentice RL (1980) The statistical analysis of failure time data. Wiley series in probability and mathematical statistics. Wiley, New York

Leblanc M, Crowley J (1993) Survival trees by goodness of split. J Am Stat Assoc 88(422):457–467

Lin X, Wang H (2004) A new testing approach for comparing the overall homogeneity of survival curves. Biom J 46(5):489–496

Lin X, Xu Q (2010) A new method for the comparison of survival distributions. Pharm Stat 9(1):67–76

Lin Y, Jeon Y (2006) Random forests and adaptive nearest neighbors. J Am Stat Assoc 101(474):578–590

Loh WY (2002) Regression trees with unbiased variable selection and interaction detection. Stat Sin 12(2):361–386

Loh WY (2013) Guide classification and regression trees user manual for version 15

Mogensen UB, Ishwaran H, Gerds TA (2012) Evaluating random forests for survival analysis using prediction error curves. J Stat Softw 50(11):1

R Core Team (2014) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.http://www.R-project.org/

Rokach L (2009) Taxonomy for characterizing ensemble methods in classification tasks: a review and annotated bibliography. Comput Stat Data Anal 53(12):4046–4072

Sauerbrei Willi, Royston Patrick (1999) Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J R Stat Soc Ser A (Stat Soc) 162(1):71–94

Scheike T, Martinussen T, Silver J (2009) timereg: timereg package for flexible regression models for survival data. R package version, pp 1–2

Schlichting P, Christensen E, Andersen PK, Fauerholdt L, Juhl E, Poulsen H, Tygstrup N (1983) Prognostic factors in cirrhosis identified by Cox's regression model. Hepatology 3(6):889–895

Schumacher M, Bastert G, Bojar H, Huebner K, Olschewski M, Sauerbrei W, Schmoor C, Beyerle C, Neumann RL, Rauschecker HF (1994) Randomized 2 × 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. J Clin Oncol 12(10):2086–2093

Segal MR (1988) Regression trees for censored data. Biometrics 44:35–47

Siroky DS (2009) Navigating random forests and related advances in algorithmic modeling. Stat Surv 3:147–163

Therneau TM (2014) A package for survival analysis in S. R package version 2.37-7. http://CRAN.R-project.org/package=survival

Verikas A, Gelzinis A, Bacauskiene M (2011) Mining data with random forests: a survey and results of new tests. Pattern Recogn 44(2):330–349

Zhu R, Kosorok MR (2012) Recursively imputed survival trees. J Am Stat Assoc 107(497):331–340