Les Cahiers du GERAD ISSN: 0711–2440

L_1 splitting rules in survival forests

H. Moradian, D. Larocque, F. Bellavance

G–2015–76

August 2015

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

L_1 splitting rules in survival forests

Hoora Moradian
Denis Larocque
François Bellavance

GERAD & Department of Decision Sciences,
HEC Montréal, Montréal (Québec) Canada,
H3T 2A7

denis.larocque@hec.ca
francois.bellavance@hec.ca

August 2015

Les Cahiers du GERAD
G–2015–76

Abstract: The log-rank test is commonly 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 better results, compared to forests built with the log-rank splitting rule.

Key Words: Survival data, right-censored data, ensemble methods, random forests, survival forests.

Acknowledgments: This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by the Fonds de recherche du Québec – Nature et technologies (FRQNT).

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 (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., 2008). 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 a number of covariates and either a categorical or a continuous outcome in a complete dataset. 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) as the splitting rule. According to Ishwaran et al. (2008), RSF is the only survival forest technique that adheres to all random forest principals 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 datasets 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. 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) and 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.

A simple motivating example is used to illustrate this fact. We first investigate the ability of the log-rank splitting rule to identify the only covariate related with the response. In this example, there is no censoring and ten covariates $X_1, \ldots, X_{10}$ are given. They are independently and identically distributed (iid) uniformly on the interval (0,1). Only the first one is related with the response. The nine others are noise covariates. In fact, there are only two survival patterns depicted in Figure 1. If $X_1 \leq 0.5$, then the true survival function is the smooth exponential one, while it is the one with the sharp drop when $X_1 > 0.5$. The sample size is 200 and the results are based on 500 simulation runs. For each sample, we built a tree with a single split by using two different splitting rules.

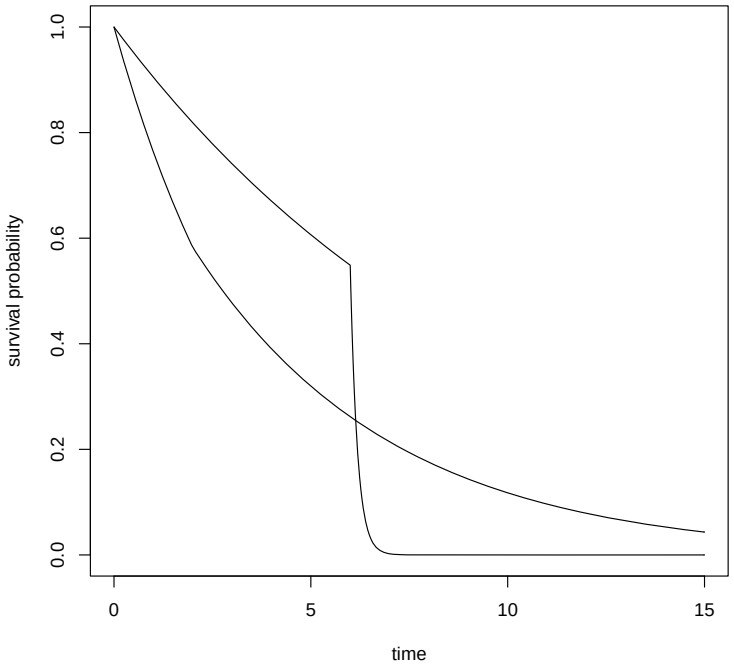

Figure 1: Crossing survival curves

The first splitting rule is the log-rank test. Strikingly, the log-rank splitting rule could only detect the right covariate 14.4% of the times, which is just a bit higher than the random selection probability of 10%. The second splitting rule is the one we are proposing and investigating in this paper and we call it L_1 splitting rule:

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

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

by Lin and Xu (2010). In the simple example, this splitting rule was able to detect the right covariate (X_1) 100% of times, a major improvement over the 14.4% achieved by the log-rank splitting rule.

These results are promising. However, in practice, we are not building a single tree with a single split, but a forest. Forests are well-known to be able to adapt themselves to many situations. It might be the case that even if the log-rank splitting rule is not able to find the right covariate in the first split, it could be able to detect the structure somehow deeper in the trees and the forest could still provide a good performance at the end. To investigate this possibility, we built forests with 100 trees and the default parameters in `randomForestSRC`. Hence, this time, a random subset of three covariates is selected in each node of a tree. The same tuning parameters were used to build forests with the L_1 splitting rule. 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:

$$\text{IAE} = \int_t |S(t) - \hat{S}(t)| dt \tag{2}$$

and

$$\text{ISE} = \int_t (S(t) - \hat{S}(t))^2 dt. \tag{3}$$

of $\hat{S}$ is interpolated at 1000 points in each method.

These criteria are evaluated using a test set of size 1000. The results are reported in Figure 2. We see that the L_1 forest is clearly better than a forest built with the log-rank splitting rule, according to both

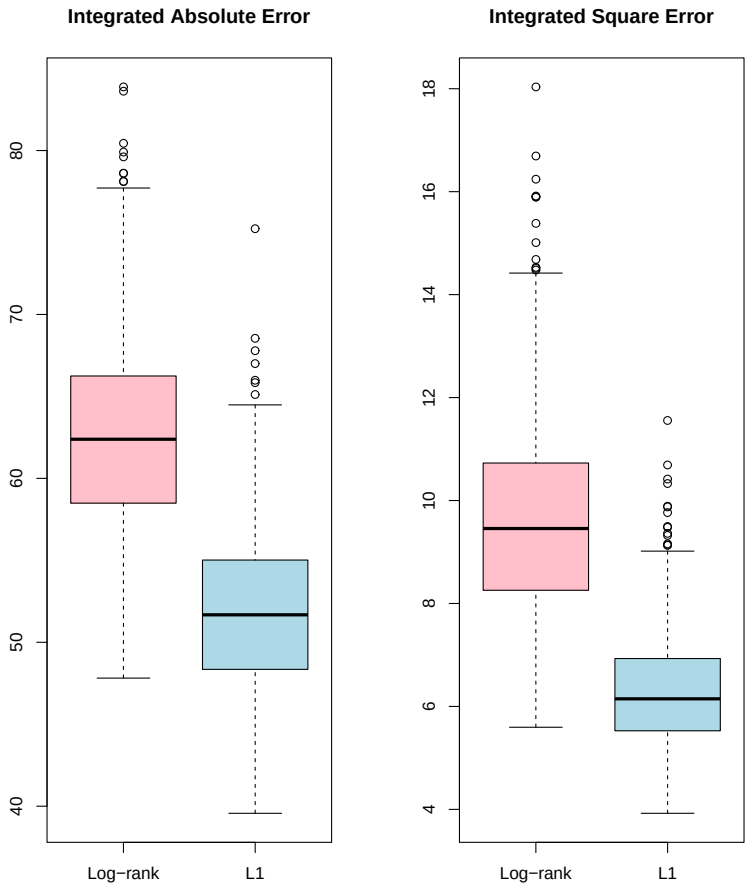

Figure 2: Comparison of IAE and ISE between forests built with the log-rank and the L_1 splitting rules

criteria. Hence, the forest paradigm is not able to fully compensate for the fact that the splitting rule is not appropriate. It is thus worthwhile to investigate further forests built with the seemingly more robust L_1 splitting rule.

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 Section 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 Description of the method

2.1 Data description

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 a covariates vector. 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 that τ_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.

2.2 Survival forest approach and splitting criterion

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, it was also illustrated with a simple example that the forest paradigm is not able to compensate for the fact that 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 the Appendix, to speed up computations. With this simplification, we consider

$$L_1^* = \sqrt{n_L n_R}\Delta = \sqrt{n_L n_R} \int_t |\hat{S}_L(t) - \hat{S}_R(t)| dt \tag{4}$$

where n_L and n_R are the left and right node sizes, as the splitting criterion. 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 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.

3 Simulation study

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

1. A forest where trees are build with the proposed L_1 splitting rule. It is denoted by L_1-forest.
2. A forest where trees are build with the proposed L_1^* splitting rule. It is denoted by L_1^*-forest.
3. A forest where trees are build with the log-rank splitting rule. It is denoted by LR-forest.
4. A Cox model. This is the first benchmark. It is denoted by Cox.
5. 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 Development Core Team, 2014). The `R` package `randomForestSRC` (Ishwaran and Kogalur, 2014) was used for the LR-forest. The `R` package `survival` (Therneau, 2014) was used for both the Cox model and the Kaplan-Meier estimator. In all three 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 terminal nodes is 3, the default value in `randomForestSRC`. Only the results for the IAE criterion, defined in (2) is reported. The results for the ISE (3) are similar and are available from the authors.

3.1 Simulation design

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. distribution. 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 the 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.

DGP 1. It is the one described in the Introduction section and illustrated in Figure 1. The model is a tree with two terminal nodes. Ten iid uniform covariates on the interval (0,1) are available, $X_1, \ldots, X_{10}$. The response is only related to X_1. The censoring times are uniformly distributed on the interval (0,α). The

hazard function is:

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

DGP 2. This DGP is slightly more complex than DGP 1. It is a tree with four terminal nodes. 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 $(0,\alpha)$. There are four survival patterns depicted in Figure 3. The hazard function is:

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

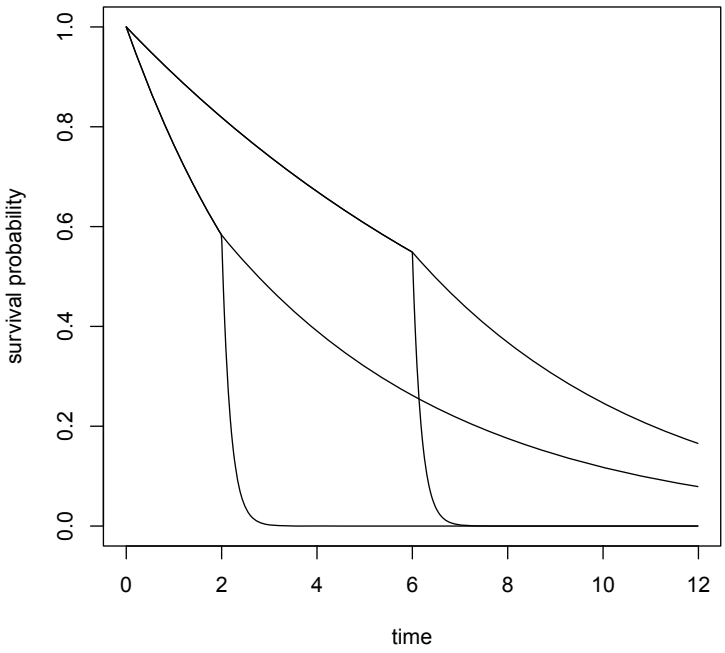

Figure 3: DGP 2

DGP 3. It is an altered version of scenario 2 from Section 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 $(0,\alpha)$.

DGP 4. It is adapted from scenario 3 in Section 4.1 of Zhu and Kosorok (2012). 25 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 $(0,\alpha)$.

DGP 5. This is a dependent censoring DGP. It is adapted from scenario 1 in Section 4.1 of Zhu and Kosorok (2012). 25 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

The results of the simulation study for the IAE criterion (2) are summarized in Figures 4 to 8. The first striking facts is that the L_1-type forests (L_1 and L_1^*) are performing better in terms of IAE than LR-forests for all DGPs and all proportion of censoring except for GDP 3. For DGP 3, the LR-forest is better for censoring proportions up to 20%, then the L_1^*-forest does better when the censoring proportion reaches 40%. The L_1-forest is better than the L_1^*-forest except for DGP 3. The L_1-type forests have usually a lower IAE compared to the two benchmarks (Cox and KM), except for a few instances with high censoring proportions.

4 Real data sets

In this section, we compare the performance of the same five methods used in the simulation study with six real datasets: 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 1.

Table 1: 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	194	76%	(Bache and Lichman, 2013) in `randomForestSRC` package (Ishwaran and Kogalur, 2014)
Veteran	6	137	7%	(Kalbfleisch and Prentice, 2002) in `randomForestSRC` package (Ishwaran and Kogalur, 2014)
NWTCO	4	4088	85%	(Breslow and Chatterjee, 1999) in `survival` package (Therneau, 2014)

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 `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 package `mfp` (Ambler and Benner, 2014). The data contains 686 observations and eight covariates. There is no missing data. The WPBC data is provided by package `randomForestSRC` (Ishwaran and Kogalur, 2014) and available in the UCI machine learning repository (Bache and Lichman, 2013). There are 198 observations in the data. However, four observations are deleted due to missing data. Thirty-two covariates are used in this example. The Veteran data (Kalbfleisch and Prentice, 2002) 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 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.

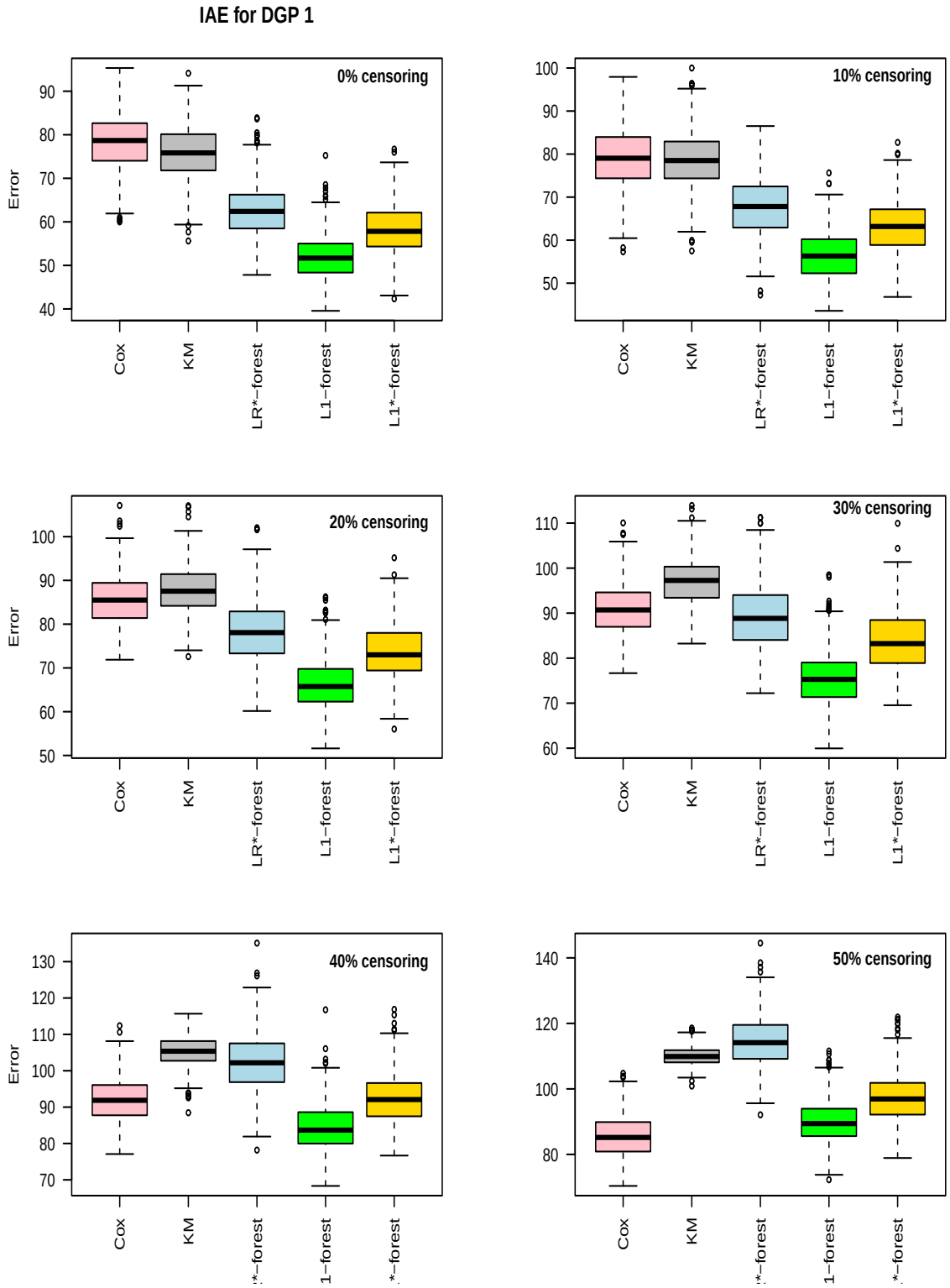

Figure 4: IAE of methods for DGP 1 across different censoring proportions

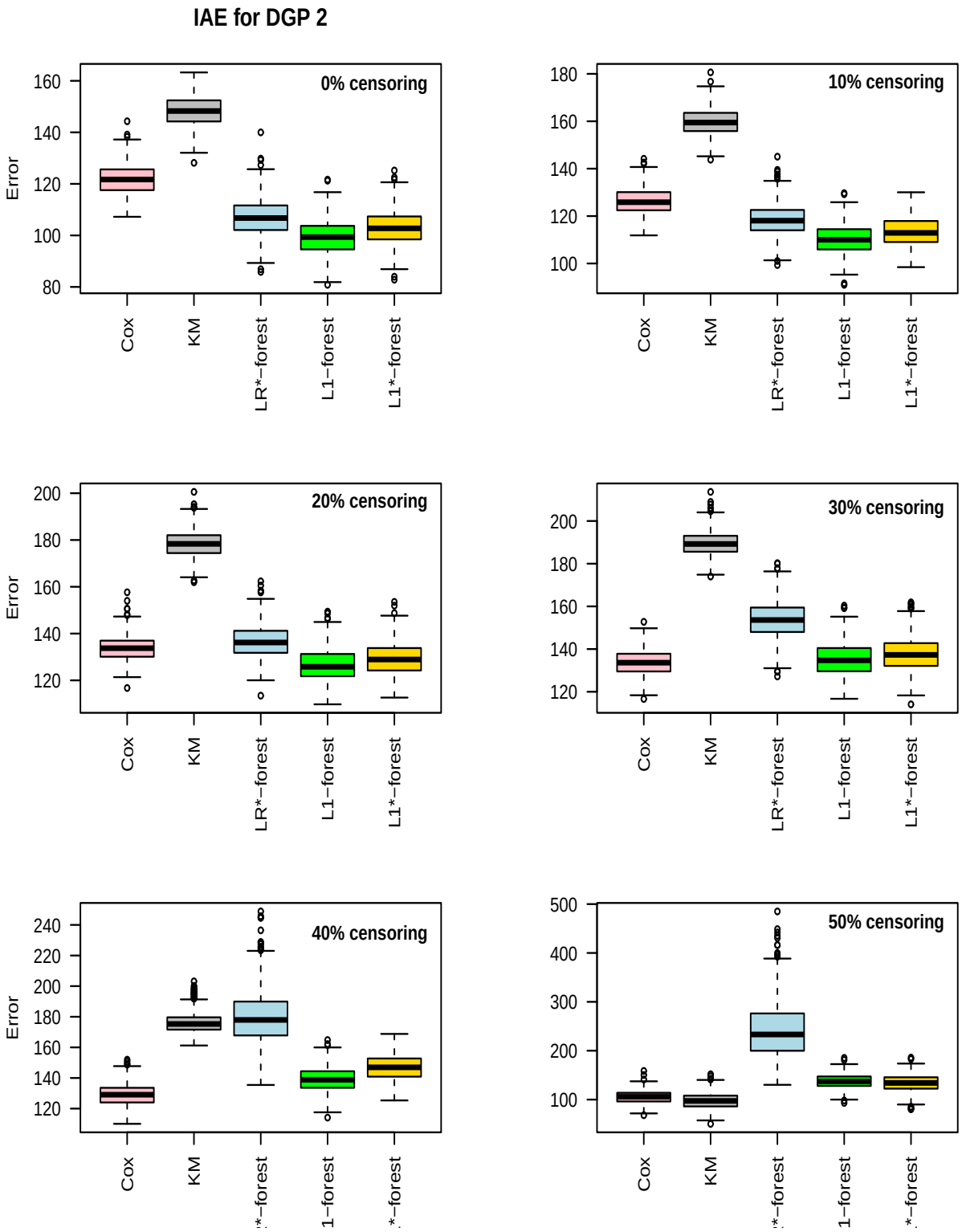

Figure 5: IAE of methods for DGP 2 across different censoring proportions

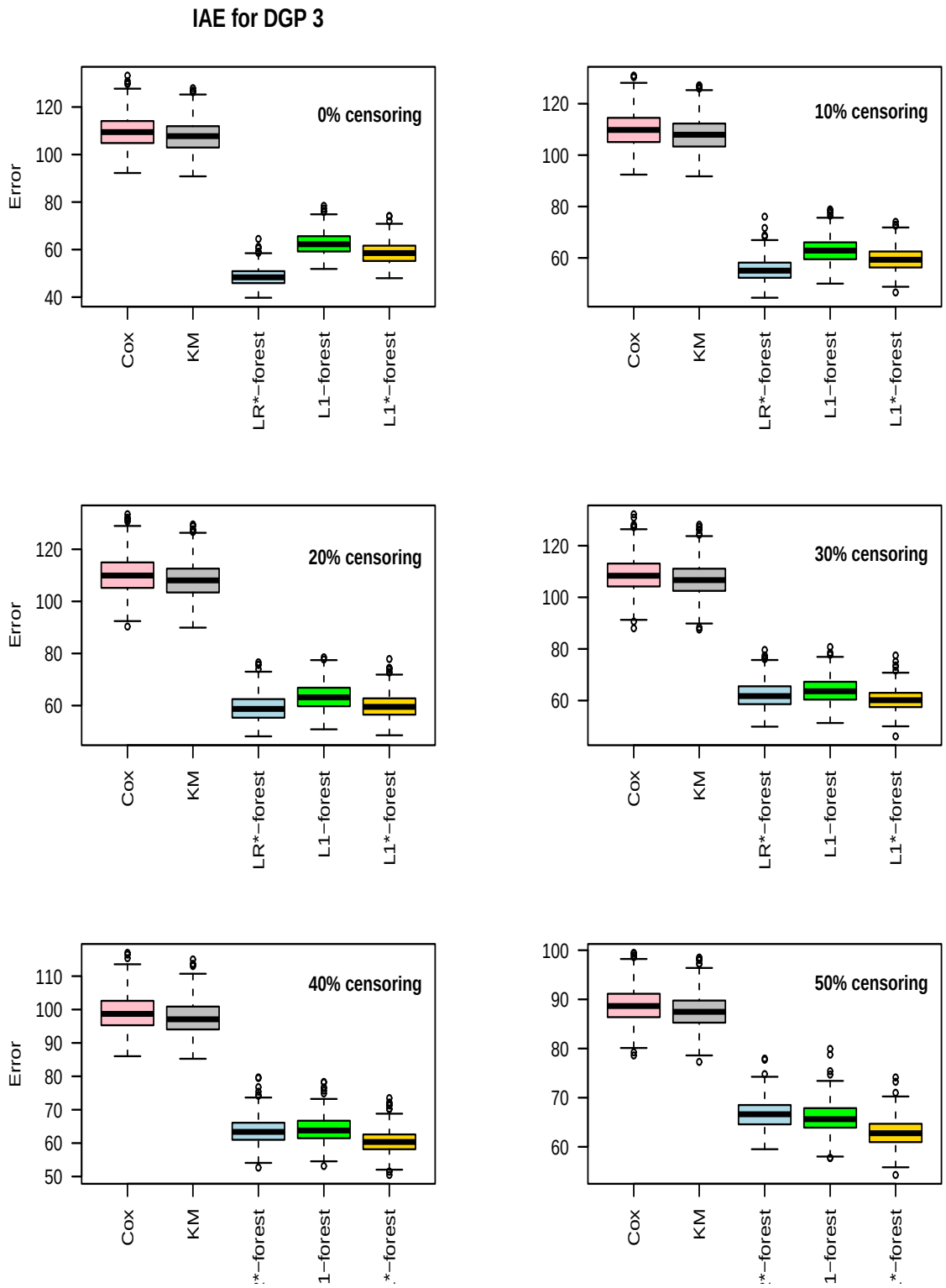

Figure 6: IAE of methods for DGP 3 across different censoring proportions

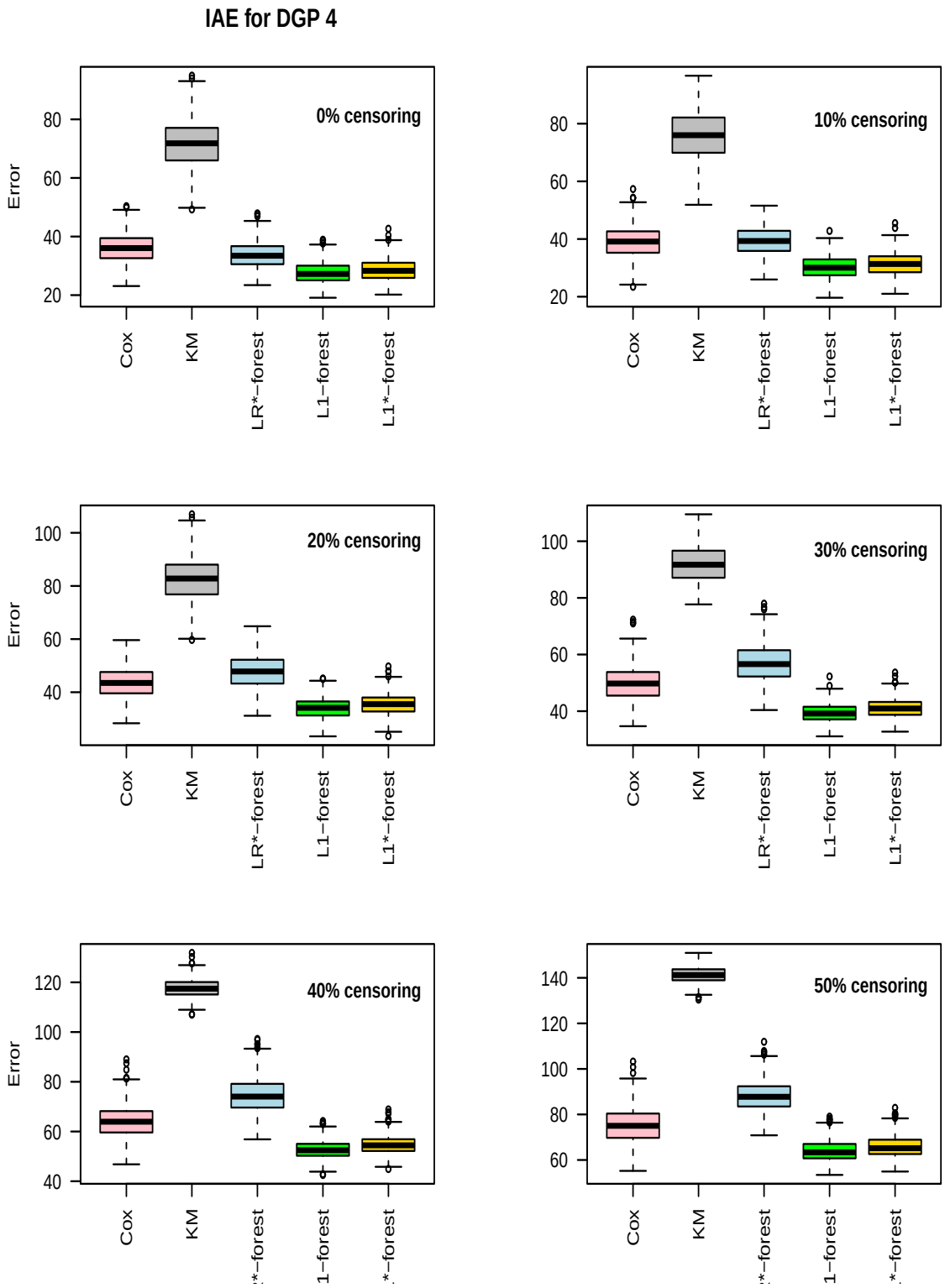

Figure 7: IAE of methods for DGP 4 across different censoring proportions

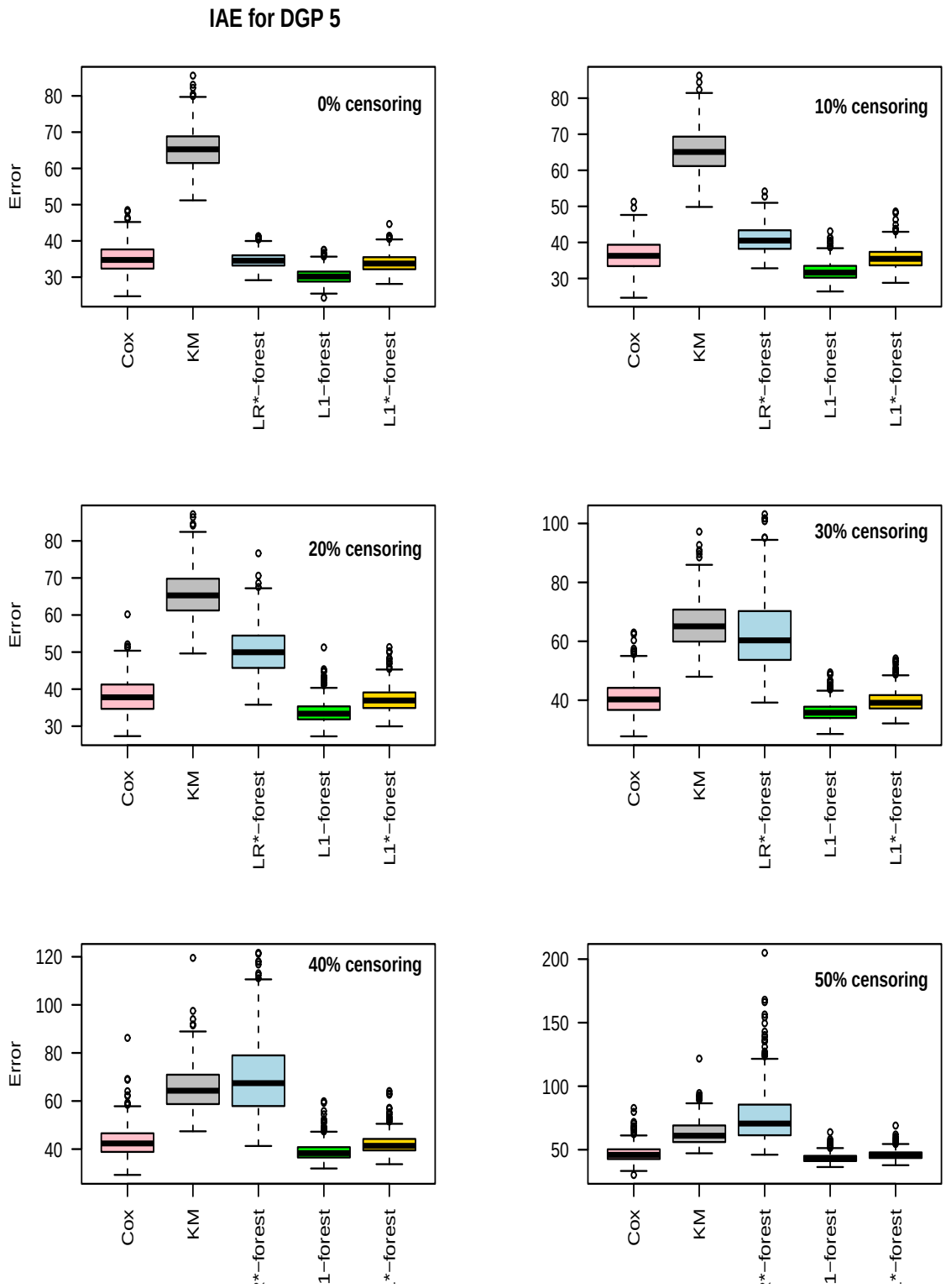

Figure 8: IAE of methods for DGP 5 across different censoring proportions

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 nodes is 3.

Since the true survival function is not known, we use the integrated Brier score (Graf et al., 1999) as the criterion for comparison. Let $\hat{S}(t|x)$ denote the estimated survival function, estimated by any model, at time t of a subject with covariate vector x. Let $\hat{G}$ denote the Kaplan-Meier estimate of the survival function. The Brier score at any time t is computed as

$$\mathrm{BS}(t) = \frac{1}{N}\sum_{i=1}^{n}\left((\hat{S}(t|x_i)^2 I(\tau_i \leq t \; and \; \delta_i = 1)\hat{G}^{-1}(\tau_i) + (1-\hat{S}(t|x_i)^2 I(\tau_i > t)\hat{G}^{-1}(t)\right)$$

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. Figure 9 illustrates the results of IBS across the 20 runs of 10-fold cross-validation for each data set. To avoid distorting the plots, the Kaplan-Meier method is not shown when its IBS is very high compared to the others.

Similarly to the results for the simulated data, the L_1-type forests perform better than the LR-forest. Their IBS are always lower than the one of the LR-forest, except for the Veteran data set. This time, however, the two L_1-type forests are closer. The L_1-forest is better than the L_1^*-forest for two data sets, the opposite is true for two others, and they are on par for two others. The Cox model has the best performance for the CSL and GBC data.

5 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 adequate. 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, gave better results compared to forests built with the log-rank splitting rule in almost all cases considered, either with simulated data or with real data sets. Hence, it is certainly worthwhile to consider it as a potential competitor. 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 $\Delta = \int_t |\hat{S}_R(t) - \hat{S}_L(t)|dt$. 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, the 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.

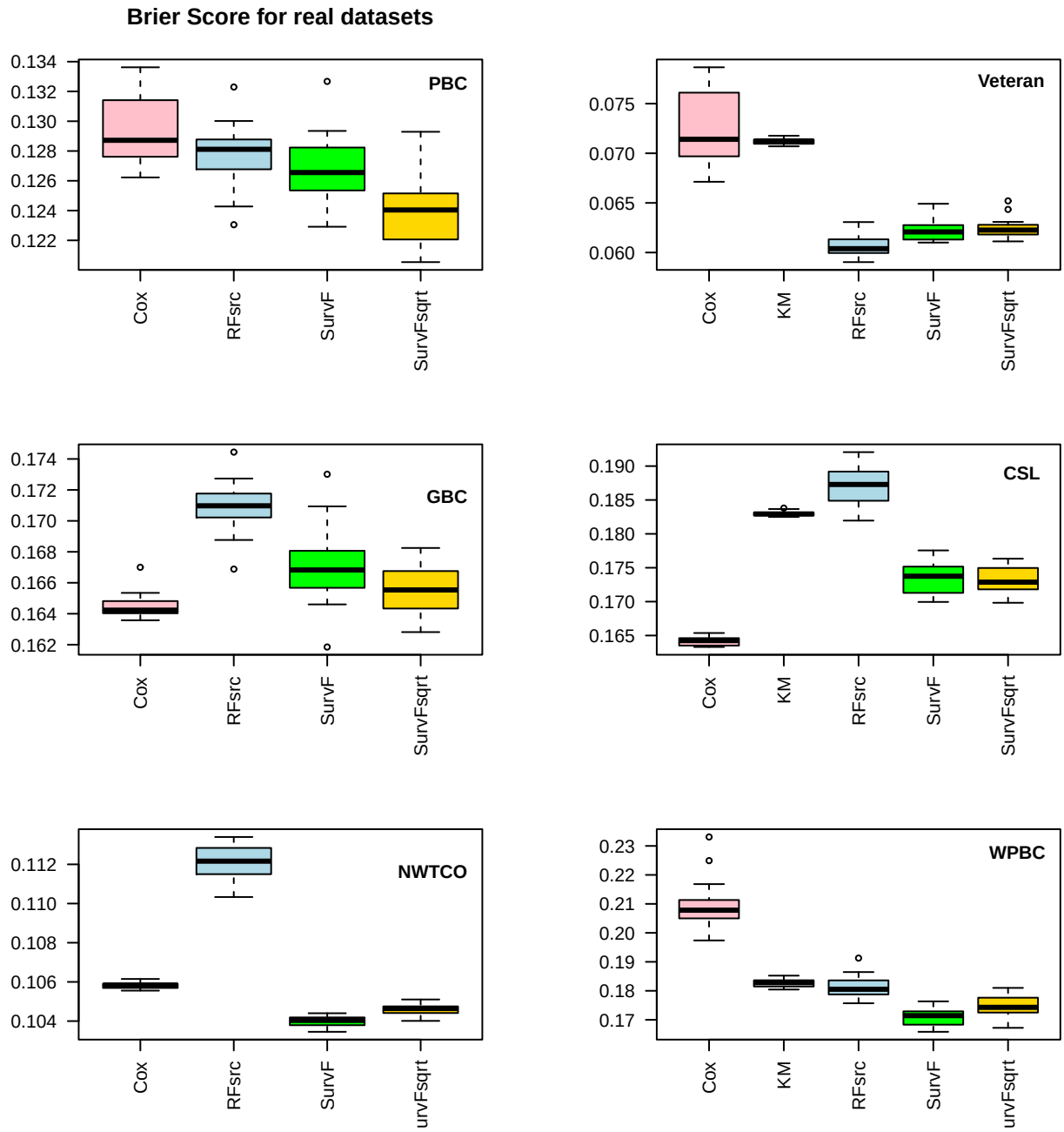

Figure 9: Brier score of methods for real data sets

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 and with the real data sets. 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 be interesting.

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) and Loh (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.

6 Appendix

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,

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

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

G. Ambler and A. Benner. mfp: Multivariable Fractional Polynomials, 2014. http://CRAN.R-project.org/package=mfp. R package version 1.5.0.

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

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

A.L. Boulesteix, S. Janitza, J. Kruppa, and I.R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(6):493–507, 2012.

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

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and regression trees. Wadsworth & Brooks. Monterey, CA, 1984.

N.E. Breslow and N. Chatterjee. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. Journal of the Royal Statistical Society: Series C (Applied Statistics), 48(4):457–468, 1999.

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

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

A. Ciampi, J. Thiffault, J.P. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition: A comparison of three methods of analysis for survival data with covariates. Computational Statistics & Data Analysis, 4(3):185–204, 1986.

A. Ciampi, S.A. Hogg, S. McKinney, and J. Thiffault. 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. Computer Methods and Programs in Biomedicine, 26(3):239–256, 1988.

T.R. Fleming and D.P. Harrington. Counting processes and survival analysis. John Wiley & Sons, Hoboken, NJ, USA, 1991.

L. Gordon and R.A. Olshen. Tree-structured survival analysis. Cancer Treatment Reports, 69(10):1065, 1985.

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

D.W. Hosmer Jr, S. Lemeshow, and S. May. Applied Survival Analysis: Regression Modeling of Time to Event Data, Wiley Series in Probability and Statistics, 2008.

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

T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15(3):651–674, 2006b.

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

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

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. The Annals of Applied Statistics, 2(3):841–860, 2008.

H. Ishwaran, U.B. Kogalur, E.Z. Gorodeski, A.J. Minn, and M.S. Lauer. High-dimensional variable selection for survival data. Journal of the American Statistical Association, 105(489):205–217, 2010.

H. Ishwaran, U.B. Kogalur, X. Chen, and A.J. Minn. Random survival forests for high-dimensional data. Statistical Analysis and Data Mining, 4(1):115–132, 2011.

J.D. Kalbfleisch and R.L. Prentice. The Statistical Analysis of Failure Time Data. Wiley Series in Probability and Mathematical Statistics, 2002.

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

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

W.Y. Loh. Regression trees with unbiased variable selection and interaction detection. Statistica Sinica, 12(2): 361–386, 2002.

W.Y. Loh. Guide classification and regression trees user manual for version 15. 2013.

L. Rokach. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography. Computational Statistics & Data Analysis, 53(12):4046–4072, 2009.

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

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

M. Schumacher, G. Bastert, H. Bojar, K. Huebner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R.L. Neumann, and H.F. Rauschecker. Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German Breast Cancer Study Group. Journal of Clinical Oncology, 12(10):2086–2093, 1994.

M.R. Segal. Regression trees for censored data. Biometrics, 44(1):35–47, 1988.

D.S. Siroky. Navigating random forests and related advances in algorithmic modeling. Statistics Surveys, 3:147–163, 2009.

T.M. Therneau. A Package for Survival Analysis in S, 2014. http://CRAN.R-project.org/package=survival. R package version 2.37-7.

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

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