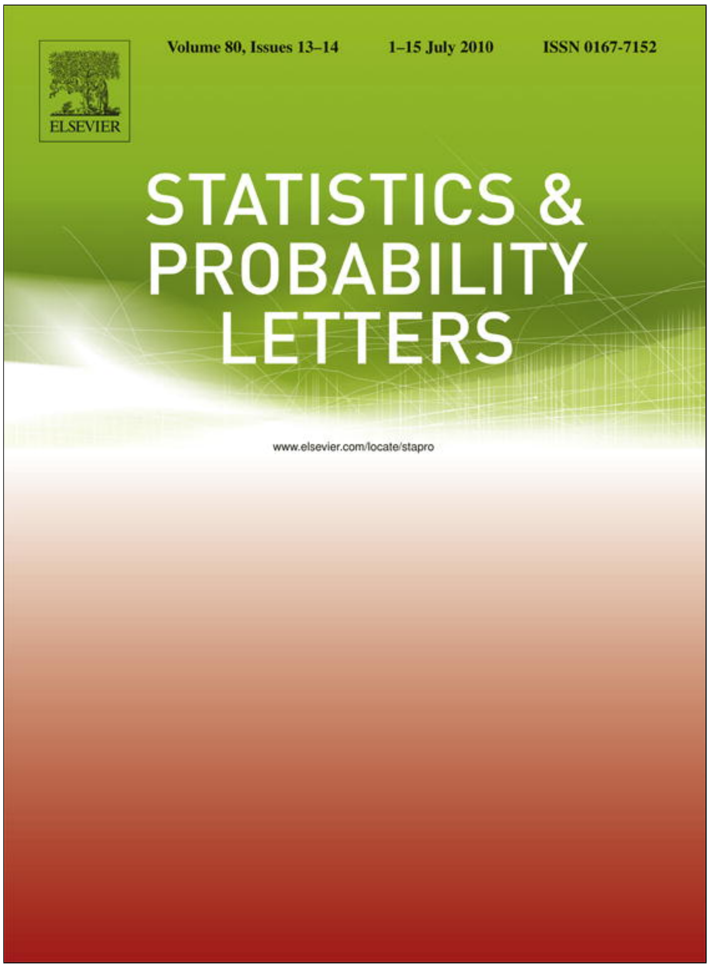
Volume 80, Issues 13–14
1–15 July 2010
ISSN 0167-7152
ELSEVIER
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

Statistics and Probability Letters 80 (2010) 1056–1064

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Consistency of random survival forests

Hemant Ishwaran [*], Udaya B. Kogalur

Cleveland Clinic, 9500 Euclid Avenue Cleveland, United States

ARTICLE INFO

Article history:
Received 17 July 2009
Accepted 22 February 2010
Available online 10 March 2010

Keywords:
Consistency
Ensemble
Factors
Kaplan–Meier
Random forests

ABSTRACT

We prove uniform consistency of Random Survival Forests (RSF), a newly introduced forest ensemble learner for analysis of right-censored survival data. Consistency is proven under general splitting rules, bootstrapping, and random selection of variables—that is, under true implementation of the methodology. Under this setting we show that the forest ensemble survival function converges uniformly to the true population survival function. To prove this result we make one key assumption regarding the feature space: we assume that all variables are factors. Doing so ensures that the feature space has finite cardinality and enables us to exploit counting process theory and the uniform consistency of the Kaplan–Meier survival function.

1. Introduction

Among the most exciting machine learning algorithms to have been proposed in the last decade is Random Forests (RF), an ensemble method introduced by Breiman (2001). RF is an all-purpose algorithm that can be applied in a wide variety of data settings. In regression settings (i.e. where the response is continuous) the method is referred to as RF-R. In classification problems, or multiclass problems, where the response is a class label, the method is referred to as RF-C. Recently, RF has also been extended to right-censoring survival settings, a method called random survival forests (RSF) (Ishwaran et al., 2008).

RF is considered an "ensemble learner". Ensemble learners are predictors formed by aggregating many elementary learners (base learners), each of which have been constructed from different realizations of the data. A widely used ensemble technique is bagging (Breiman, 1996). In bagging, the ensemble is formed by aggregating a base learner over independent bootstrap samples of the original data.

Although there are many variants of RF (Amit and Geman, 1997; Dietterich, 2000; Cutler and Zhao, 2001), the most popular, and the one we consider here, is that described by Breiman (2001) under the name Forest-RI (short for RF random input selection). In this version, RF can be viewed as an extension of bagging. Using independent bootstrap samples, a random tree is grown by splitting each tree node using a randomly selected subset of variables (features). The forest ensemble is constructed by aggregating over the random trees. The extra randomization introduced in the tree growing process is the crucial step distinguishing forests from bagging. Unlike bagging, each bootstrap tree is constructed using different variables, and not all variables are used. This is designed to encourage independence among trees, and unlike bagging, it not only reduces variance, but also bias. Results using benchmark data have shown that prediction error for RF is often substantially better than bagging and comparable to other state-of-the-art methods such as boosting (Freund and Shapire, 1996), and support vector machines (Cortes and Vapnik, 1995).

In his seminal 2001 paper (Breiman, 2001), Breiman discussed bounds on the generalization error for a forest as a trade-off involving number of variables randomly selected as candidates for splitting, and the correlation between trees. He showed as number of variables increases, strength of a tree (accuracy) increases, but at the price of increasing correlation

[*] Corresponding author.
E-mail address: hemant.ishwaran@gmail.com (H. Ishwaran).

0167-7152/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2010.02.020

among trees; which degrades overall performance. In Lin and Jeon (2006), lower bounds for the mean-squared error for a regression forest were derived under random splitting by drawing analogies between forests and nearest neighbor classifiers. Recently, Meinshausen (2006) proved consistency of RF-R for quantile regression, and Biau et al. (2008) proved consistency of RF-C under the assumption of random splitting.

In this paper, we prove consistency of RSF by showing that the forest ensemble survival function converges uniformly to the true population survival function. Because RSF is a new extension of RF to right-censored survival settings not much is known about its properties. Even consistency results for survival trees are sparse in the literature. For right-censored survival data, LeBlanc and Crowley (1993) showed survival tree cumulative hazard functions are consistent for smoothed cumulative hazard functions. The method of proof used convergence results for recursive-partitioned regression trees for uncensored data (Breiman et al., 1984).

We take a different approach and establish consistency by drawing upon counting process theory. We first prove uniform consistency of survival trees, and from this, by making use of bootstrap theory, we prove consistency of RSF (Section 3). These results apply to general tree splitting rules (not just random ones) and to true implementations of RSF. We make only one important assumption: that the feature space is a finite (but very large) discrete space and that all variables are factors (Section 3). In this regard we deviate from other proofs of forest consistency which assume that the feature space is continuous. A continuous space is a more general assumption than ours and we readily acknowledge this limitation (see the Discussion). At that same time, discrete variables are very often encountered in medical settings involving survival data. Section 2.3 discusses an example related to esophageal cancer. Furthermore, in Section 4 we investigate the extent to which an assumption of a discrete feature space limits our results. We show by way of example that embedding forests in a discrete setting is realistic in that one can analyze problems with continuous variables by treating them as factors having a large number of factor labels. For the interested user, we note that all computations in the paper were implemented using the freely available R-software package, `randomSurvivalForest` (Ishwaran and Kogalur, 2007, 2010).

2. Random survival forests

Let $(\mathbf{X}, T, \delta)$, $(\mathbf{X}_1, T_1, \delta_1), \ldots, (\mathbf{X}_n, T_n, \delta_n)$ be i.i.d. random elements such that $\mathbf{X}$, the feature, is a d-dimensional vector taking values in $\mathscr{X}$, a discrete space (to be described in Section 3). Here $T = \min(T^0, C)$ is the observed survival time and $\delta = I(T^o \leq C)$ is the binary $\{0, 1\}$ censoring value, where it is assumed that T^o, the true event time, is independent of C, the censoring time. An individual (case) i is said to be right-censored at time T_i if $\delta_i = 0$; otherwise, if $\delta_i = 1$, the individual is said to have experienced an event at T_i. It is assumed that $\mathbf{X}$ is independent of δ and that $(\mathbf{X}, T, \delta)$ has joint distribution $\mathbb{P}$. The marginal distribution for $\mathbf{X}$ is denoted by μ and defined via $\mu(A) = \mathbb{P}\{\mathbf{X} \in A\}$ for all subsets A of $\mathscr{X}$. It is assumed that $\mu(A) > 0$ for each $A \neq \emptyset$.

2.1. The RSF algorithm

The collection of values $\{(\mathbf{X}_i, T_i, \delta_i)\}_{1 \leq i \leq n}$ are referred to as the *learning data* and are used in the construction of the forest. We begin with a high-level description of the RSF algorithm. Specific details follow (Section 2.2).

1. Draw B independent Efron bootstrap samples (Efron, 1979) from the learning data and grow a binary recursive survival tree to each bootstrap sample.
2. When growing a survival tree, at each node of the tree randomly select p candidate variables to split on (use as many candidate variables as possible, up to $p - 1$, if there are less than p variables available within the node). Split the node by using the split that maximizes survival difference between daughter nodes (in the case of ties, a random tie breaking rule is used).
3. Grow a tree as near to saturation as possible (i.e. to full size) with the only constraint being that each terminal node should have no less than $d_0 > 0$ events.
4. Calculate the tree survival function. The forest ensemble is the averaged tree survival function.

2.2. Survival trees and forests: Details of the algorithm

The tree survival function calculated in Step 4 of the algorithm is the Kaplan–Meier (KM) estimator for the tree's terminal nodes. To be more precise, let $\mathscr{N}(\mathscr{T})$ denote the terminal nodes of a survival tree, $\mathscr{T}$. These are the extreme nodes of $\mathscr{T}$ reached when the tree can no longer be split to form new nodes (daughters). Let $h \in \mathscr{N}(\mathscr{T})$ be a terminal node. Define $Y_h^{(n)}(t) = \sum_{i=1}^{m(h)} I(T_{i,h} \geq t)$ to be the number of individuals in h observed to be at risk just prior to time t (i.e., the number of individuals who have neither experienced an event nor been censored prior to t). Let $N_h^{(n)}(t)$ be the counting process defined as the number of events in $[0, t]$ for all cases in h. Define the indicator process $J_h^{(n)}(t) = I(Y_h^{(n)}(t) > 0)$. The Nelson–Aalen estimator for cases within the terminal node h is

$$\hat{H}_h(t) = \int_0^t \frac{J_h^{(n)}(s)}{Y_h^{(n)}(s)} \mathrm{d}N_h^{(n)}(s),$$

where we adopt the convention that $J_h^{(n)}(s)/Y_h^{(n)}(s) = 0$ whenever $Y_h^{(n)}(s) = 0$. The KM estimator for cases within h is

$$\hat{S}_h(t) = \prod_{s \le t}(1 - d\hat{H}_h(s)) = \prod_{s \le t}\left(1 - \frac{dN_h^{(n)}(s)}{Y_h^{(n)}(s)}\right). \tag{1}$$

Each case i has a d-dimensional feature $\mathbf{x}_i \in \mathscr{X}$. To determine the survival function for i, drop $\mathbf{x}_i$ down the tree. Because of the binary nature of a survival tree, $\mathbf{x}_i$ will be assigned a unique terminal node $h' \in \mathscr{N}(\mathscr{T})$. The survival function for i is the KM estimator for $\mathbf{x}_i$'s terminal node:

$$\hat{S}(t|\mathbf{x}_i) = \hat{S}_{h'}(t), \quad \text{if } \mathbf{x}_i \in h'.$$

This defines the survival function for all cases and thus defines the survival function for the tree. To make this clear, we write the tree survival function as

$$\hat{S}(t|\mathbf{x}_i) = \sum_{h \in \mathscr{N}(\mathscr{T})} I(\mathbf{x}_i \in h)\hat{S}_h(t). \tag{2}$$

The forest ensemble survival function is the averaged tree survival function. If the forest is comprised of trees $\{\mathscr{T}_b\}_{1 \le b \le B}$, where $\hat{S}_b(t|\mathbf{x})$ is the survival function for $\mathscr{T}_b$ given $\mathbf{x}$, the ensemble survival function is

$$\frac{1}{B}\sum_{b=1}^{B}\hat{S}_b(t|\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\sum_{h \in \mathscr{N}(\mathscr{T}_b)} I(\mathbf{x}_i \in h)\hat{S}_h(t).$$

Note that in practice each tree in the forest is constructed from an independent bootstrap sample of the data (Step 1 of the algorithm), but for now we ignore this complication. We revisit this issue later in Section 3.3 when we consider bootstrap resampling.

2.3. Esophageal cancer

As motivation before presenting our theoretical results (Section 3), we present a previously published example that illustrates the use of RSF in a real application. In this example, RSF was applied to define a cancer stage grouping for esophageal cancer. Currently, staging of esophageal cancer is based solely on anatomic extent of disease using an orderly, progressive grouping of TNM cancer classifications. TNM stands for three anatomic features of esophageal cancer: depth of cancer invasion through the esophageal wall and adjacent tissues (T), presence of cancer-positive nodes along the esophagus (N), and presence of cancer metastases to distant sites (M). There are five subclassifications of T (Tis, T1, T2, T3, and T4), two of N (absence or presence of cancer-positive nodes), and two of M (absence or presence of distant metastases).

TNM classifications for esophageal cancer are known to be overly simplistic; in part because they reflect only anatomic extent of cancer. Other cancer characteristics known to affect prognosis include location of the cancer along the esophagus, cell type (squamous cell carcinoma vs. adenocarcinoma), and histologic grade; a crude reflector of biologic activity. It is also widely known that an increasing number of cancer-positive lymph nodes is associated with decreasing survival, and it is suspected this relationship is non-linear and may depend upon other factors, such as depth of cancer invasion, T.

In order to develop a more biologically plausible stage grouping, a RSF analysis was applied to a large group ($n = 4627$) of esophageal cancer patients (Ishwaran et al., 2009). Variables measured on each patient included TNM classifications, number of lymph nodes removed at surgery, number of cancer-positive lymph nodes, other non-TNM cancer characteristics, patient demographics, and additional variables to adjust for country and institution. All variables were discrete (this included patient age, which as customary was recorded in years). The primary outcome used in the analysis was time to death, measured from date of surgery.

Fig. 1 displays results from this analysis. Plotted in the figure is five-year predicted survival for lymph node-positive patients who were free of distant metastases. The curves plotted are averaged ensemble survival functions evaluated at five years with each curve being constructed from a forest using a different number of trees (5, 10, 50, and 250 trees, respectively). Each curve is stratified by depth of cancer invasion (T1, T2, T3, and T4; note that Tis plays no role in node-positive cancers) and by number of cancer-positive nodes. Each point on a curve is the average of the ensemble survival function, averaged over all patients for a given T, for a given number of cancer-positive nodes. As the number of trees increases, the curves begin to stabilize, and after about 250 trees one can see there is not much to be gained by using additional trees. This shows that forest ensembles can stabilize fairly quickly. The other interesting aspect of Fig. 1 is that RSF has clearly identified a non-linear relationship between survival and number of cancer-positive nodes, with survival decreasing rapidly as number of nodes increases. Furthermore, this relationship depends strongly upon T (the more deeply invasive the tumor, the less the effect). These results are consistent with the biology of esophageal cancer.

3. Properties of survival forests

3.1. Feature space

In establishing consistency of RSF we assume that each coordinate $1 \le j \le d$ of the d-dimensional feature $\mathbf{X}$ is a factor (discrete nominal variable) with $1 < L_j < \infty$ distinct labels. While this assumes that the feature space $\mathscr{X}$ has finite

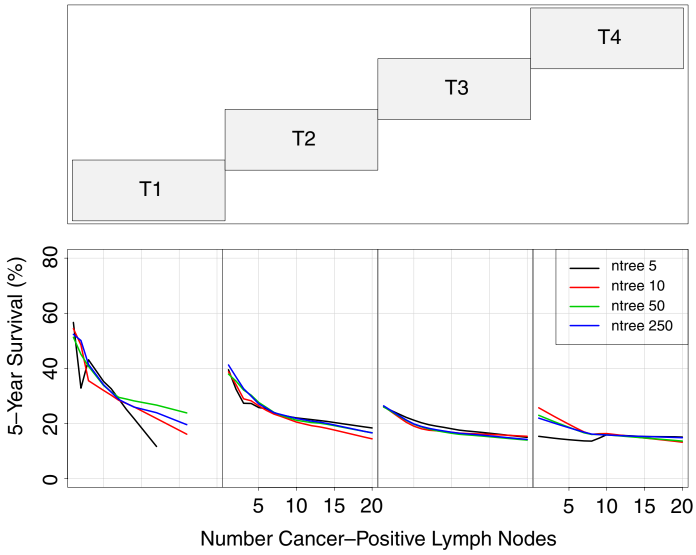

Fig. 1. RSF analysis of esophageal data. Five-year predicted survival for node-positive patients is plotted against number of cancer-positive nodes, stratified by depth of invasion (T1, T2, T3, and T4). Predicted survival is based on the first 5, 10, 50, and 250 trees, respectively from the forest.

cardinality, the actual size of $\mathscr{X}$ can be quite large, $L_1 \times \cdots \times L_d$, and moreover, the number of splits that a tree might make from such data can be even larger, depending on d and L_j.

To see this, note that a split on a factor in a tree results in data points moving left and right of the parent node such that the complementary pairings define the new daughter nodes. For example, if a factor has three labels, $\{A, B, C\}$, then there are three complementary pairings (daughters) as follows: $\{A\}$ and $\{B, C\}$; $\{B\}$ and $\{C, A\}$; and $\{C\}$ and $\{A, B\}$. In general, for a factor with L_j distinct labels, there are $2^{L_j-1} - 1$ distinct complementary pairs. Thus, the total number of splits evaluated when splitting the root node for a survival tree when all variables are factors can be as much as

$$\text{maximum number root-node splits} = \sum_{j=1}^{d} 2^{L_j-1} - d.$$

Following the root-node split, are splits on the resulting daughter nodes, and their daughter nodes, recursively, with each subsequent generation requiring a large number of evaluations. Each evaluation can result in a new tree, thus showing that number of trees (space of trees) associated with $\mathscr{X}$ can be extremely large.

3.2. Uniform consistency of survival trees

To show that the ensemble survival function converges to the true population survival function we first prove consistency of a single tree. Consistency of forests will be readily deduced from this (Section 3.3).

In the following, and throughout the paper, the true survival function, or population parameter, is assumed to be of the form

$$S(t|\mathbf{X}) := \mathbb{P}\{T^o > t|\mathbf{X}\} = \sum_{\mathbf{x}\in\mathscr{X}} I(\mathbf{X} = \mathbf{x}) \exp\left(-\int_0^t \alpha(s|\mathbf{x})\mathrm{d}s\right), \tag{3}$$

where $\alpha(\cdot|\mathbf{x})$ is the non-negative hazard function for the subpopulation $\mathbf{X} = \mathbf{x}$.

The following result proves uniform consistency of a survival tree, and is a consequence of the uniform consistency of the KM estimator. The proof takes advantage of the finiteness of $\mathscr{X}$ which turns the problem of proving tree consistency into a more manageable problem of establishing consistency for a single terminal node.

Theorem 1. *Let $t \in (0, \tau)$, where $\tau = \min\{\tau(\mathbf{x}) : \mathbf{x} \in \mathscr{X}\}$ and $\tau(\mathbf{x}) = \sup\{t : \int_0^t \alpha(s|\mathbf{x})\mathrm{d}s < \infty\}$. If $\mathbb{P}\{C > 0\} > 0$, and $\alpha(\cdot|\mathbf{x})$ is strictly positive over $[0, t]$ for at least one $\mathbf{x} \in \mathscr{X}$, then*

$$\sup_{s\in[0,t]} \int_{\mathscr{X}} |\hat{S}(s|\mathbf{x}) - S(s|\mathbf{x})|\mathrm{d}\mu(\mathbf{x}) \xrightarrow{\mathrm{p}} 0, \quad \text{as } n \to \infty,$$

where $\hat{S}(\cdot|\mathbf{x})$ is a tree survival function defined as in (2).

3.3. Uniform consistency of survival forests

We have so far implicitly assumed that trees are grown from the learning data, $\mathscr{L} = \{(\mathbf{X}_i, T_i, \delta_i)\}_{1\le i\le n}$. But in practice, trees are actually grown from independent bootstrap samples of the data. In order to prove consistency of a true implementation of RSF, we must extend our previous result to address bootstrap resampling.

Let $\mathscr{L}^* = \{(\mathbf{X}_i^*, T_i^*, \delta_i^*)\}_{1\le i\le n}$ denote an Efron bootstrap sample (Efron, 1979) of $\mathscr{L}$. Let $\mathscr{T}^*$ be the survival tree grown from $\mathscr{L}^*$ and let $\hat{S}^*(t|\mathbf{x})$ be the KM estimator for $\mathscr{T}^*$,

$$\hat{S}^*(t|\mathbf{x}) = \sum_{h\in\mathscr{N}(\mathscr{T}^*)} I(\mathbf{x}\in h)\hat{S}_h^*(t),$$

where $\hat{S}_h^*(t)$ is defined similar to (1), and the above sum is over $\mathscr{N}(\mathscr{T}^*)$, the set of terminal nodes of $\mathscr{T}^*$.

The ensemble survival function for a survival forest comprised of B survival trees is

$$\hat{S}_e(t|\mathbf{x}) := \frac{1}{B}\sum_{b=1}^{B}\hat{S}_b^*(t|\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\sum_{h\in\mathscr{N}(\mathscr{T}_b^*)} I(\mathbf{x}\in h)\hat{S}_h^*(t),$$

where $\hat{S}_b^*(t|\mathbf{x})$ is the survival function for the survival tree $\mathscr{T}_b^*$ grown using the bth bootstrap sample. We prove consistency of RSF by establishing consistency of $\hat{S}_b^*(t|\mathbf{x})$ for each b.

Theorem 2. *Let $\tau^* = \min(\tau, \sup(F))$, where $\sup(F)$ is the upper limit of the support of $F(s) = 1 - \mathbb{P}\{T^o > s\}\mathbb{P}\{C > s\}$. Then under the same conditions as in Theorem 1, for each $t \in (0, \tau^*)$:*

$$\sup_{s\in[0,t]}\int_{\mathscr{X}} |\hat{S}^*(s|\mathbf{x}) - S(s|\mathbf{x})|\mathrm{d}\mu(\mathbf{x}) = o_p^*(1) + o_p(1),$$

where o_p^* stands for o_p in bootstrap probability for almost all $\mathscr{L}$-sample sequences; i.e. with probability one under $\mathbb{P}^\infty$.

Uniform consistency of the ensemble, $\hat{S}_e(t|\mathbf{x})$, follows automatically from Theorem 2, because

$$\sup_{s\in[0,t]}\int_{\mathscr{X}} |\hat{S}_e(s|\mathbf{x}) - S(s|\mathbf{x})|\mathrm{d}\mu(\mathbf{x}) \le \frac{1}{B}\sum_{b=1}^{B}\sup_{s\in[0,t]}\int_{\mathscr{X}} |\hat{S}_b^*(s|\mathbf{x}) - S(s|\mathbf{x})|\mathrm{d}\mu(\mathbf{x}).$$

The right-hand side is $o_p^*(1) + o_p(1)$ by Theorem 2.

3.4. Uniform approximation by forests

Theorem 2 establishes consistency of a bootstrapped survival tree, and from this consistency of a survival forest follows. While this is a useful line of attack for establishing large sample properties of forests, it does not convey how in practice a forest might improve inference over a single tree. Indeed, in finite sample settings, a forest of trees can have a decided advantage when approximating the true survival function. Recall the esophageal cancer example of Section 2.3 where we saw first hand how averaging over trees improved prediction accuracy.

To provide theoretical support for superiority of forests, we make use of the following setting. Suppose that we are allowed to construct a binary survival tree $\mathscr{T}_b$ from a prechosen learning data set $\mathscr{L}_b = \{(\mathbf{X}_{b,i}, T_{b,i}, \delta_{b,i})\}_{1\le i\le n}$ in any manner we choose. The only constraint being that each terminal node of $\mathscr{T}_b$ must contain at least $d_0 = 1$ events. We are allowed to construct $b = 1, \dots, B$ such trees for any $B < \infty$ of our choosing. Let $S_b(t|\mathbf{x})$ be the KM tree survival function for $\mathscr{T}_b$, and let

$$S_e(t|\mathbf{x}) = \sum_{b=1}^{B} W_b S_b(t|\mathbf{x}) \tag{4}$$

be the ensemble survival function constructed from $\{\mathscr{T}_b\}_{1\le b\le B}$, where $\{W_b\}_{1\le b\le B}$ are non-negative forest weights that we are free to specify. The next theorem shows that one can always find an ensemble that uniformly approximates the true survival function (3). Trees do not possess this property.

Theorem 3. *If $n > d$, and $s \in [0, \tau)$, then for each $\varepsilon > 0$ there exists an ensemble survival function (4) for a survival forest comprised of $B = B(\varepsilon)$ survival trees, with each tree consisting of $d + 1$ terminal nodes, such that*

$$\int_0^s\int_{\mathscr{X}} (S_e(t|\mathbf{x}) - S(t|\mathbf{x}))^2\,\mu(\mathrm{d}\mathbf{x})\mathrm{d}t \le \varepsilon.$$

4. Treating a continuous problem as discrete

Our theory has been predicted on the assumption that all variables are factors, but in practice data with continuous variables are often encountered. Here we show that one can discretize continuous variables and treat them as factors

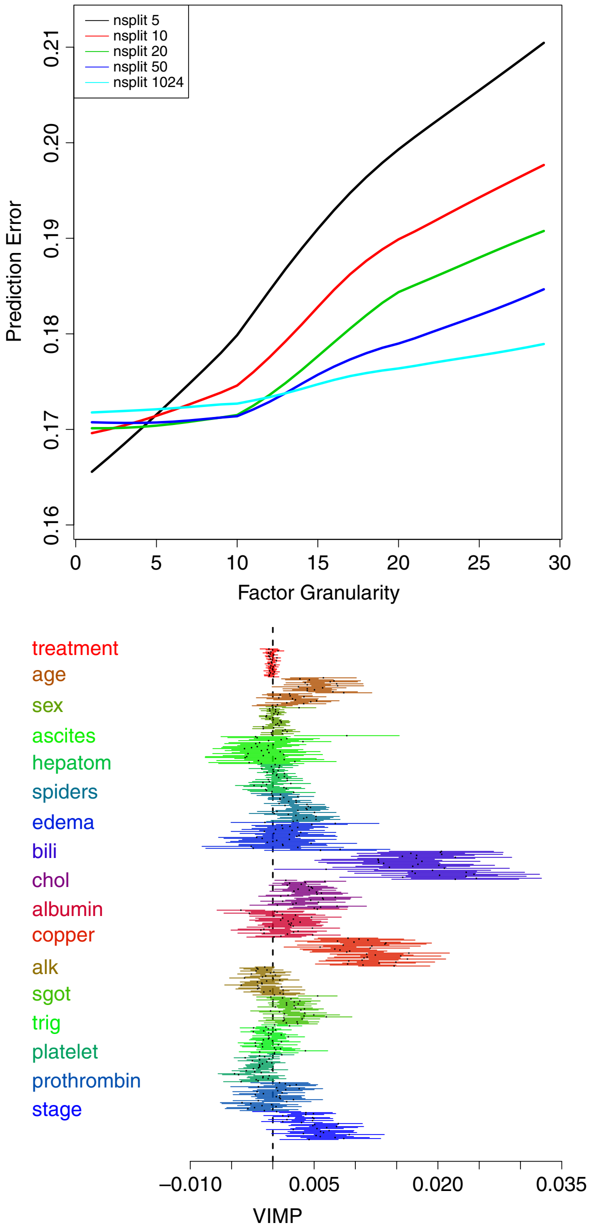

Fig. 2. RSF analysis of PBC data using 1000 trees with random log-rank splitting where variables, both nominal and continuous, were discretized to have a maximum number of labels (factor granularity). Top figure is out-of-bag prediction error versus factor granularity, stratified by number of random splits used for a node, nsplit. Bottom figure shows 68% bootstrap confidence region for variable importance (VIMP) from 1000 bootstrap samples using an nsplit value of 1024 for each factor granularity value in the top figure. Color coding is such that the same color has been used for a variable over the different granularity values (factor granularity for a variable increases going from top to bottom).

without unduly affecting prediction error and inference: thus showing that our theory extrapolates reasonably to general data settings.

For illustration we consider the primary biliary cirrhosis (PBC) data of Fleming and Harrington (1991). The data is from a randomized clinical trial studying the effectiveness of the drug D-penicillamine on PBC. The data set involves 312 individuals and contains 17 variables as well as censoring information and time until death for each individual. Of the 17 features, seven are discrete and 10 are continuous. Each of the 10 continuous variables were discretized and converted to a factor with L labels. We investigated different amounts of granularity: $L = 2, \dots, 30$.

For each level of granularity, L, we fit a survival forest of 1000 survival trees using log-rank splitting with node-adaptive random splits. Splits for nodes were implemented as follows. A maximum of "nsplit" complementary pairs were chosen randomly for each of the p randomly selected candidate variables within a node (if nsplit exceeded the number of cases in a node, then nsplit was set to the size of the node). Log-rank splitting was applied to the randomly selected complementary pairs and the node was split on that variable and complementary pair maximizing the log-rank test. Five different values for nsplit were tried: nsplit = 5, 10, 20, 50, 1024.

The top plot in Fig. 2 shows out-of-bag prediction error as a function of granularity and nsplit value. As granularity rises, prediction error increases—but this increase is reasonably slow and well contained with larger values of nsplit. This is quite remarkable because the total number of complementary pairs with a granularity level of $L = 30$ is on order 2^{30} (over 1 billion pairs) and yet our results show that using only a handful of randomly selected complementary pairs keeps prediction error in check.

Prediction error measures overall performance, but we should also consider how inference for a variable is affected by increasing granularity. To study this we looked at variable importance (VIMP). VIMP measures predictiveness of a variable, adjusting for all other variables (Ishwaran et al., 2008). Positive values of VIMP indicate predictiveness, and negative and zero values indicate noise. For each forest we dropped bootstrapped data down the forest and computed VIMP for each variable. This was repeated 1000 times independently. The bottom plot of Fig. 2 displays the 68% bootstrap confidence region from this distribution. The analysis was restricted to only those forests grown under an nsplit value of 1024 but was carried out for each level of granularity (color coding scheme used to depict granularity is described in the caption of the figure). Overall, one can see that the bootstrap confidence regions are relatively robust to the level of granularity.

5. Discussion

We proved uniform consistency of RSF under settings that mirror those seen in actual data applications. Our consistency result followed by showing consistency of a single bootstrapped survival tree. As we remarked, while this is a useful line of attack for establishing large sample properties, it does not tell us why RF work. Theorem 3 presented one argument, but a general theory explaining why RF works still remains elusive. We hope that our work will motivate others to study this problem.

The major assumption in our approach was the assumption of a discrete feature space. Other proofs of consistency have used continuous feature spaces, and these results are more general. At the same time, because applications of RF to survival settings is much less studied than regression and classification settings, our results are a useful addition to the literature on forests. It is also interesting to note that a discrete space assumption implies, with probability one, that each node of the tree will contain one unique $\mathbf{x}$ value. This is in line with the original Forest-RI method suggested by Breiman (2001).

Appendix. Proofs

Proof of Theorem 1. Independence of T^o and C ensures that

$$\mathbb{P}\{\delta = 1\} = \int_0^\infty \mathbb{P}\{T^o \le c\}\mathbb{P}\{C \in \mathrm{d}c\}.$$

The assumption $\mathbb{P}\{C > 0\} > 0$ implies that the censoring distribution has mass bounded away from the origin. Thus, we can find a $c > 0$ such that $\mathbb{P}\{C > c\} > 0$. The assumed form of the survival function (3) ensures that the distribution function for T^0 is continuous over $[0, t]$. Combining this with the assumption $\alpha(\cdot|\mathbf{x})$ is strictly positive for some $\mathbf{x}$, deduce that $\mathbb{P}\{\delta = 1\} > 0$. Recall that a survival tree is grown to full length with the proviso that a terminal node should have no less than $d_0 > 0$ events. Let $A \subseteq \mathscr{X}$ be any non-null set. Then by the law of large numbers, and by the assumed independence of $\mathbf{X}$ and δ,

$$\frac{1}{n}\sum_{i=1}^n I(\mathbf{X}_i \in A, \delta_i = 1) \overset{\text{a.s.}}{\to} \mathbb{P}\{\mathbf{X} \in A, \delta = 1\} = \mu\big(A\big)\mathbb{P}\{\delta = 1\} > 0.$$

Therefore,

$$I\left(\sum_{i=1}^n I(\mathbf{X}_i \in A, \delta_i = 1) \ge d_0\right) \overset{\text{a.s.}}{\to} 1. \tag{5}$$

Hence, almost surely, each terminal node in the tree corresponds to one and only one distinct element of $\mathscr{X}$ (by this we mean that the terminal node is associated with a set in the tree partition of $\mathscr{X}$, and that this set is a unique element $\mathbf{x} \in \mathscr{X}$ for some $\mathbf{x}$). If it were not, then we could find a terminal node corresponding to a set A with $|A| > 1$, such that any split on it yields daughters A_1 and A_2 with at least one daughter having fewer than d_0 deaths. But because the tree is grown to full size, this contradicts (5) which holds almost surely for any $A \neq \emptyset$. Thus, the tree almost surely splits on all possible values of $\mathscr{X}$ and has terminal nodes for each distinct $\mathbf{x} \in \mathscr{X}$. In other words,

$$\hat{S}(s|\mathbf{X}) = \sum_{\mathbf{x}\in\mathscr{X}} I(\mathbf{X} = \mathbf{x} = h)\hat{S}_h(s) + o_p(1), \quad \text{uniformly in } s. \tag{6}$$

We slightly abuse notation by using $h = \mathbf{x}$ to indicate the terminal node h containing $\mathbf{x}$ (and only $\mathbf{x}$). Note that the $o_p(1)$ term on the right-side is uniform in s because of the boundedness of $|\hat{S}_h|$.

Let $Y^{(n)}(s|\mathbf{x}) = \sum_{i=1}^n I(T_i \ge s, \mathbf{X}_i = \mathbf{x})$ be the number of cases with feature $\mathbf{x}$ who are at risk just prior to s. Define $J^{(n)}(s|\mathbf{x}) = I(Y^{(n)}(s|\mathbf{x}) > 0)$. Now if we can show that for each $\mathbf{x} \in \mathscr{X}$, as $n \to \infty$,

$$\int_0^t \frac{J^{(n)}(s|\mathbf{x})}{Y^{(n)}(s|\mathbf{x})}\alpha(s|\mathbf{x})\mathrm{d}s \overset{\mathrm{p}}{\to} 0, \tag{7}$$

and

$$\int_0^t (1 - J^{(n)}(s|\mathbf{x}))\alpha(s|\mathbf{x})\mathrm{d}s \overset{\mathrm{p}}{\to} 0, \tag{8}$$

then by Theorem IV.3.1 (Andersen et al., 1993), for each $h = \mathbf{x}$:

$$\sup_{s\in[0,t]} |\hat{S}_h(s) - S(s|\mathbf{x})| \overset{\mathrm{p}}{\to} 0, \quad \text{as } n \to \infty.$$

This would establish the result, because

$$\begin{aligned} \sup_{s\in[0,t]} \int_{\mathscr{X}} |\hat{S}(s|\mathbf{x}) - S(s|\mathbf{x})|\mathrm{d}\mu(\mathbf{x}) &\le \int_{\mathscr{X}} \sup_{s\in[0,t]} |\hat{S}(s|\mathbf{x}) - S(s|\mathbf{x})|\mathrm{d}\mu(\mathbf{x}) \\ &= \sum_{h=\mathbf{x}} \mu\{\mathbf{X} = \mathbf{x}\}\left(\sup_{s\in[0,t]} |\hat{S}_h(s) - S(s|\mathbf{x})|\right) + o_p(1). \end{aligned}$$

The $o_p(1)$ on the right-hand side holds due to (6).

Therefore, to complete the proof we only need to verify that conditions (7) and (8) hold. By the definition of τ, $\sup_{s\in[0,t]} \alpha(s|\mathbf{x}) < \infty$ and thus a sufficient condition for (7) and (8) is that

$$\inf_{s\in[0,t]} Y^{(n)}(s|\mathbf{x}) \overset{\mathrm{p}}{\to} \infty.$$

This condition holds by noting that for each $s \in [0, t]$,

$$\begin{aligned} n^{-1}Y^{(n)}(s|\mathbf{x}) &\ge \frac{1}{n}\sum_{i=1}^{n} I(T_i \ge t, \delta_i = 1, \mathbf{X}_i = \mathbf{x}) \\ &\overset{\text{a.s.}}{\to} \mu\{\mathbf{X} = \mathbf{x}\}\mathbb{P}\{T^o \ge t|\mathbf{x}\} > 0. \end{aligned}$$

Note that $\mathbb{P}\{T^o \ge t|\mathbf{x}\} > 0$ because of (3) and the definition of τ. □

Proof of Theorem 2. Let $(M^*_{n,1}, \ldots, M^*_{n,n})^T$ be a multinomial random vector from n trials in which each cell has probability $1/n$ of occurring (as customary we use a "$*$" to indicate that randomness comes from bootstrapping). For each non-null $A \subseteq \mathscr{X}$,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}^*_i \in A, \delta^*_i = 1) &\overset{d^*}{=} \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \in A, \delta_i = 1)M^*_{n,i} \\ &= \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \in A, \delta_i = 1) + \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \in A, \delta_i = 1)(M^*_{n,i} - 1) \\ &= \mathbb{P}\{\mathbf{X} \in A, \delta = 1\} + o^*_p(1), \quad \text{a.s.}, \end{aligned}$$

where the almost sure convergence is in $\mathbb{P}$-probability and the $o^*_p(1)$ term follows from

$$\begin{aligned} \mathbb{P}^*\left\{\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \in A, \delta_i = 1)(M^*_{n,i} - 1) \ge \varepsilon\right\} &\le \frac{1}{\varepsilon^2 n^2}\mathbb{E}^*\left(\sum_{i=1}^{n} I(\mathbf{X}_i \in A, \delta_i = 1)(M^*_{n,i} - 1)\right)^2 \\ &\le \frac{1}{\varepsilon^2 n^2}\sum_{i=1}^{n} \mathrm{Var}^*(M^*_{n,i}) + \frac{1}{\varepsilon^2 n^2}\sum_{i\ne j} |\mathrm{Cov}^*(M^*_{n,i}, M^*_{n,j})| \\ &= O(n^{-1}). \end{aligned}$$

In the proof of Theorem 1 it was shown that $\mathbb{P}\{\mathbf{X} \in A, \delta = 1\} > 0$. From this, and using similar arguments as in that proof, it follows that

$$\hat{S}^*(s|\mathbf{X}) = \sum_{\mathbf{x}\in\mathscr{X}} I(\mathbf{X} = \mathbf{x} = h)\hat{S}^*_h(s) + o^*_p(1),$$

where the $o^*_p(1)$ term is uniform in s because of the boundedness of $|\hat{S}^*_h|$. Notice that

$$\int_{\mathscr{X}} |\hat{S}^*(s|\mathbf{x}) - S(s|\mathbf{x})|\mathrm{d}\mu(\mathbf{x}) \le \int_{\mathscr{X}} |\hat{S}^*(s|\mathbf{x}) - \hat{S}(s|\mathbf{x})|\mathrm{d}\mu(\mathbf{x}) + \int_{\mathscr{X}} |\hat{S}(s|\mathbf{x}) - S(s|\mathbf{x})|\mathrm{d}\mu(\mathbf{x}).$$

The second term on the right side is $o_p(1)$ uniformly in s by Theorem 1. To deal with the first term, we use the representation (6) for $\hat{S}(s|\mathbf{X})$ given in the proof of Theorem 1, to obtain

$$\int_{\mathscr{X}} |\hat{S}^*(s|\mathbf{x}) - \hat{S}(s|\mathbf{x})| \mathrm{d}\mu(\mathbf{x}) = \sum_{h=\mathbf{x}} \mu\{\mathbf{X} = \mathbf{x}\} |\hat{S}_h^*(s) - \hat{S}_h(s)| + o_p^*(1) + o_p(1), \quad \text{uniformly in } s.$$

An Efron bootstrap sample of size n from $\mathscr{L}$ can be drawn equivalently using a two-stage process by first drawing a multinomial vector $(n_h^*)_h$ from n trials with each cell $h = \mathbf{x}$ having probability $\mu\{\mathbf{X} = \mathbf{x} = h\}$ and then drawing a bootstrap sample of size n_h^*, for each h, from $\mathscr{L}_h = \{(\mathbf{X}_i, T_i, \delta_i) : \mathbf{X}_i = \mathbf{x} = h\}$. It is not hard to show that $n_h^*/n_h \overset{\mathrm{p}^*}{\to} 1$, where $n_h = |\mathscr{L}_h| = \sum_{i=1}^n I(\mathbf{X}_i = \mathbf{x} = h) \overset{\mathrm{a.s.}}{\to} \infty$. Thus to complete the proof it suffices to show uniform convergence of $\hat{S}_h^*(s) - \hat{S}_h(s)$ for each $h = \mathbf{x}$, where $\hat{S}_h^*(s)$ is derived from a bootstrap sample of size n_h from $\mathscr{L}_h$. We use part (b) of Lemma 3 from Lo and Singh (1985) for this. This theorem applies if $s < \sup(F)$ under the random censorship model for a continuous survival distribution. All these conditions are met under our assumptions. Thus applying this Lemma, one can show that

$$\sup_{s\in[0,t]} |\hat{S}_h^*(s) - \hat{S}_h(s)| = O_p^*(n_h^{-1/2} \log n_h^{1/2}) = o_p^*(1). \quad \Box$$

Proof of Theorem 3. It suffices to show that for a given $\mathbf{x}$, we can find an ensemble $S_e(\cdot|\mathbf{x}')$ that is zero if $\mathbf{x}' \neq \mathbf{x}$, and for $\mathbf{x}' = \mathbf{x}$, uniformly approximates $S(\cdot|\mathbf{x})$ over $[0, s]$ (this suffices because we can always combine such ensembles to uniformly approximate $S_e(\cdot|\mathbf{x})$ for all $\mathbf{x}$). Choose $\mathscr{L}_b$ such that $\delta_{b,i} = 1$ for all b and i. By repeated splitting on the left, it is clear we can use d splits to construct a tree $\mathscr{T}_b$ having $d + 1$ terminal nodes such that the left-most daughter node corresponds to $\mathbf{x}$. Because $n > d$, we can assign at least one event to the left-most node. For concreteness, assume that the node contains exactly one event, with event time $T > 0$. Over the remaining d terminal nodes assign event times $T = 0$ for all cases. Thus $S_b(\cdot|\mathbf{x}') = 0$ if $\mathbf{x}' \neq \mathbf{x}$. On the other hand, if $\mathbf{x}' = \mathbf{x}$, then $S_b(t|\mathbf{x})$ is a step function with value 1 if $t < T$ and value 0 if $t \geq T$. Because $S(\cdot|\mathbf{x})$ is continuous (by definition (3)), it is uniformly continuous over the compact set $[0, s]$. A uniformly continuous, monotonically decreasing function over a compact set can be uniformly approximated by a linear combination of a finite number of step functions such as $S_b(\cdot|\mathbf{x})$. Thus one can construct a finite number of survival trees like $\mathscr{T}_b$, that when suitably weighted, uniformly approximates $S(\cdot|\mathbf{x})$ over $[0, s]$. $\Box$

References

Amit, Y., Geman, D., 1997. Shape quantization and recognition with randomized trees. Neural Computation 9, 1545–1588.
Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N., 1993. Statistical Methods Based on Counting Processes. Springer, New York.
Biau, G., Devroye, L., Lugosi, G., 2008. Consistency of random forests and other classifiers. Journal of Machine Learning Research 9, 2039–2057.
Breiman, L., 1996. Bagging predictors. Machine Learning 26, 123–140.
Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., Classification and Regression Trees, Belmont, California, 1984.
Cortes, C., Vapnik, V.N., 1995. Support-vector networks. Machine Learning 20, 273–297.
Cutler, A., Zhao, G., 2001. Pert — perfect random tree ensembles. Computing Science and Statistics 33, 490–497.
Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning 40, 139–157.
Efron, B., 1979. Bootstrap methods: another look at the jackknife. The Annals of Statistics 7, 1–26.
Fleming, T., Harrington, D., 1991. Counting Processes and Survival Analysis. Wiley, New York.
Freund, Y., Shapire, R.E., 1996. Experiments with a new boosting algorithm, in: Proc. of the 13th. Int. Conf. on Machine Learning. pp. 148–156.
Ishwaran, H., Blackstone, E.H., Hansen, C.A., Rice, T.W., 2009. A novel approach to cancer staging: application to esophageal cancer. Biostatistics 10, 603–620.
Ishwaran, H., Kogalur, U.B., 2007. Random survival forests for R. Rnews 7/2, 25–31.
Ishwaran, H., Kogalur, U.B., 2010. RandomSurvivalForest: Random Survival Forests. R package version 3.6.1. http://cran.r-project.org.
Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival forests. The Annals of Applied Statistics 2, 841–860.
LeBlanc, M., Crowley, J., 1993. Survival trees by goodness of split. Journal of the American Statistical Association 88, 457–467.
Lin, Y., Jeon, Y., 2006. Random forests and adaptive nearest neighbors. Journal of the American Statistical Association 101, 578–590.
Lo, S.-H., Singh, K., 1985. The product-limit estimator and the bootstrap: some asymptotic representations. Probability Theory and Related Fields 71, 455–465.
Meinshausen, N., 2006. Quantile regression forests. Journal of Machine Learning Research 7, 983–999.