ggRandomForests: Exploring Random Forest Survival

John Ehrlinger and **Jeevanantham Rajeswaran** and **Eugene H. Blackstone**
Cleveland Clinic

Abstract

Random forest (?) (RF) is a non-parametric statistical method requiring no distributional assumptions on covariate relation to the response. RF is a robust, nonlinear technique that optimizes predictive accuracy by fitting an ensemble of trees to stabilize model estimates. Random survival forests (RSF) (??) are an extension of Breiman's RF techniques allowing efficient non-parametric analysis of time to event data. The **randomForestSRC** package (?) is a unified treatment of Breiman's random forest for survival, regression and classification problems.

Predictive accuracy makes RF an attractive alternative to parametric models, though complexity and interpretability of the forest hinder wider application of the method. We introduce the **ggRandomForests** package, tools for visually understand random forest models grown in R (?) with the **randomForestSRC** package. The **ggRandomForests** package is structured to extract intermediate data objects from **randomForestSRC** objects and generate figures using the **ggplot2** (?) graphics package.

This document is structured as a tutorial for building random forest for survival with the **randomForestSRC** package and using the **ggRandomForests** package for investigating how the forest is constructed. We analyse the Primary Biliary Cirrhosis of the liver data from a clinical trial at the Mayo Clinic (?). We demonstrate random forest variable selection using Variable Importance (VIMP) (?) and Minimal Depth (?), a property derived from the construction of each tree within the forest. We will also demonstrate the use of variable dependence and partial dependence plots (?) to aid in the interpretation of RSF results. We then examine variable interactions between covariates using conditional variable dependence plots. Our aim is to demonstrate the strength of using Random Forest methods for both prediction and information retrieval, specifically in time to event data settings.

Keywords: random forest, survival, VIMP, minimal depth, R, **randomForestSRC**.

1. Introduction

Random forest (?) (RF) is a non-parametric statistical method which requires no distributional assumptions on covariate relation to the response. RF is a robust, nonlinear technique that optimizes predictive accuracy by fitting an ensemble of trees to stabilize model estimates. Random Survival Forest (RSF) (??) is an extension of Breiman's RF techniques to survival settings, allowing efficient non-parametric analysis of time to event data. The **randomForestSRC** package (?, `http://CRAN.R-project.org/package=randomForestSRC`) is a unified treatment of Breiman's random forest for survival, regression and classification problems.

Predictive accuracy make RF an attractive alternative to parametric models, though complexity and interpretability of the forest hinder wider application of the method. We introduce the **ggRandomForests** package (`http://CRAN.R-project.org/package=ggRandomForests`) for visually exploring random forest models. The **ggRandomForests** package is structured to extract intermediate data objects from **randomForestSRC** objects and generate figures using the **ggplot2** graphics package (**?**, `http://CRAN.R-project.org/package=ggplot2`).

Many of the figures created by the **ggRandomForests** package are also available directly from within the **randomForestSRC** package. However **ggRandomForests** offers the following advantages:

- Separation of data and figures: **ggRandomForests** contains functions that operate on either the `rfsrc` forest object directly, or on the output from **randomForestSRC** post processing functions (i.e., `plot.variable`, `var.select`) to generate intermediate **ggRandomForests** data objects. **ggRandomForests** functions are provide to further process these objects and plot results using the **ggplot2** graphics package. Alternatively, users can use these data objects for their own custom plotting or analysis operations.

- Each data object/figure is a single, self contained unit. This allows simple modification and manipulation of the data or `ggplot` objects to meet users specific needs and requirements.

- We chose to use the **ggplot2** package for our figures for flexibility in modifying the output. Each **ggRandomForests** plot function returns either a single `ggplot` object, or a `list` of `ggplot` objects, allowing the use of additional **ggplot2** functions to modify and customize the final figures.

This document is structured as a tutorial for using the **randomForestSRC** package for building and post-processing random survival forest models and using the **ggRandomForests** package for understanding how the forest is constructed. In this tutorial, we will build a random survival forest for the primary biliary cirrhosis (PBC) of the liver data set (**?**), available in the **randomForestSRC** package.

In Section 2 we introduce the `pbc` data set and summarize the proportional hazards analysis of this data from Chapter 4 of **?**. In Section **??**, we describe how to grow a random survival forest with the **randomForestSRC** package. Random forest is not a parsimonious method, but uses all variables available in the data set to construct the response predictor. We demonstrate random forest variable selection techniques (Section **??**) using Variable Importance (VIMP) (**?**) in Section **??** and Minimal Depth (**?**) in Section **??**. We then compare both methods with variables used in the **?** model.

Once we have an idea of which variables we are most interested in, we use dependence plots (**?**) (Section **??**) to understand how these variables are related to the response. Variable dependence (Section **??**) plots give us an idea of the overall trend of a variable/response relation, while partial dependence plots (Section **??**) show us the risk adjusted relation by averaging out the effects of other variables. Dependence plots often show strongly non-linear variable/response relations that are not easily obtained through parametric modeling.

We then graphically examine forest variable interactions with the use of variable and partial dependence conditioning plots (coplots) (**??**) (Section **??**) and close with concluding remarks in Section **??**.

2. Data summary: primary biliary cirrhosis (PBC) data set

The *primary biliary cirrhosis* of the liver (PBC) study consists of 424 PBC patients referred to Mayo Clinic between 1974 and 1984 who met eligibility criteria for a randomized placebo controlled trial of the drug D-penicillamine (DPCA). The data is described in (?, Chapter 0.2) and a partial likelihood model (Cox proportional hazards) is developed in Chapter 4.4. The `pbc` data set, included in the **randomForestSRC** package, contains 418 observations, of which 312 patients participated in the randomized trial (?, Appendix D).

```
R> data("pbc", package = "randomForestSRC")
```

For this analysis, we modify some of the data for better formatting of our results. Since the data contains about 12 years of follow up, we prefer using `years` instead of `days` to describe survival. We also convert the `age` variable to years, and the `treatment` variable to a factor containing levels of `c("DPCA", "placebo")`. The variable names, type and description are given in Table 1.

Variable name	Description	Type
years	Time (years)	numeric
status	Event (F = censor, T = death)	logical
treatment	Treament (DPCA, Placebo)	factor
age	Age (years)	numeric
sex	Female = T	logical
ascites	Presence of Asictes	logical
hepatom	Presence of Hepatomegaly	logical
spiders	Presence of Spiders	logical
edema	Edema (0, 0.5, 1)	factor
bili	Serum Bilirubin (mg/dl)	numeric
chol	Serum Cholesterol (mg/dl)	integer
albumin	Albumin (gm/dl)	numeric
copper	Urine Copper (ug/day)	integer
alk	Alkaline Phosphatase (U/liter)	numeric
sgot	SGOT (U/ml)	numeric
trig	Triglicerides (mg/dl)	integer
platelet	Platelets per cubic ml/1000	integer
prothrombin	Prothrombin time (sec)	numeric
stage	Histologic Stage	factor

Table 1: `pbc` data set variable dictionary.

2.1. Exploratory data analysis

It is good practice to view your data before beginning analysis. Exploratory Data Analysis (EDA) ? will help you to understand the data, and find outliers, missing values and other data anomalies within each variable before getting deep into the analysis. To this end, we use **ggplot2** figures with the `facet_wrap` function to create two sets of panel plots, one of