Threshold Values for the Gini Variable Importance
A Empirical Bayes Approach

Robert Dunne

Wednesday, April 6, 2022

Abstract

Random Forests (RF) are a very widely used modelling tool. Lundberg et al. (2019) concludes that no nonlinear model had a more widespread popularity, from health care to academia to industry, than random forests and decision trees. The bounds of the methodology are still being extended. Bayat et al. (2020) give an example with 80 million variables.

It is highly desirable that RF models be made more interpretable and a large part of that is a better understanding of the characteristics of the variable importance measures generated by the RF. Due to its speed and ease of calculation, we consider the mean decrease in node "impurity" (MDI) variable importance (VI) and address the question of setting a significance level.

The report is organized as follows:

- we first consider the question of multiple testing in the case of multiple measurements made on two groups (the standard microarray set up Efron (2008)). We show that some standard approaches for multiple testing can fail severely due to the correlation structure of the measurement and other modelling failures. We show that deriving the null distribution by permutation does not fix the problem. This point applies to determining the null distribution of variable importances as well as many other statistical tests;
- we show that variable correlation can either increase or decrease the (MDI) of variables in different settings. We also show that there is an additional problem with the permutation null due to the functional relationships between the statistics;
- We consider the empirical Bayes argument of Efron (2005) and model the VI as a mixture of two distribution, a null and a non-null distribution. We find that unlike the relatively well behaved case considered in Efron's papers, there are a number of issues here:
  - the distribution may be multi-modal, which creates modelling difficulties;
  - the null distribution is not of an obvious form, as it is not symmetric.
- we resolve these issues to derive a fast, plausible, empirical Bayes method for selecting significant variables while controlling the false discovery rate.

Contents

1 Introduction **3**

2 Methods **3**

2.1 Random Forests 3

2.2 Variable importance 4

2.3 GWAS and multiple testing 4

2.3.1 An empirical Bayes approach. 5

2.3.2 Machine Learning approaches to Genomic data 7

2.4 Other approaches to significance for variable importance 8

2.5 Empirical Bayes for Gini Variable Importance 9

3 Results **10**

3.1 Demonstration on synthetic data 10

3.1.1 Classification causes variable importance to smear over correlated variables . . 10

3.1.2 Regression inflates the importance of individually weak but correlated variables 10

3.2 Thousand genomes 11

4 Conclusion **12**

5 Software **13**

5.1 R implementation 13

5.2 Python implementation: 13

1 Introduction

Random Forests (RF) are a very widely used modelling tool. Lundberg et al. (2019) concludes that no nonlinear model had a more widespread popularity, from health care to academia to industry, than random forests and decision trees. The popularity is in part due to the ability of RF to process large volumes of data, as demonstrated by Bayat et al. (2020) who process 80 million variables.

However, a shortcoming of RF models are the interpretability of the results, specifically the variable importance measures to assess the association with the predicted label. Here a statistical assessment is needed to confidently set a threshold of which variable are label-associated.

There have been many attempts to do this (briefly reviewed in section 2.4). Generally they have used either a permutation based approach (with many variants) or a recursive feature elimination approach. We have gone to the literature on controlling the false discovery rate for multiple testing in high throughput genomic data. Here we find that Efron's "local fdr" approach, calculated from an empirical Bayes estimate of the null distribution.

We have adopted the same strategy in the case of high dimensional vectors of MDI importances. This turns a problem, the high dimensionality of the data, into an asset, we have enough data points to estimate the null distribution accurately.

2 Methods

2.1 Random Forests

Random Forest (RF) models have a number of desirable features:

- they give an out-of-bag (OOB) estimate of model accuracy, by using the samples not selected for the bootstrapped sample as an independent test set for each tree;
- they provide a measure of variable importance, for example by summing the number of times a variable is used to split a node, over all nodes and all trees. More complex measures of variable importance are also used;
- they inherit some of the qualities of tree models such as invariance to monotonic scaling of the data, and are applicable in the case of large p;
- in many applications they provide accurate predictions (Kaggle, 2019; Arik and Pfister, 2020).

Since the original implementation of random forests (Breiman, 2001), there have been a number of developments leading to a better understanding of the algorithm.

- random forests are resistant to overfitting. However they are not immune to it. Segal (2004) gives examples of situations where overfitting can occur;
- Strobl et al. (2007) demonstrates that the random forest algorithm is subject to a bias in variable selection via two mechanisms:
    1. differences in variable scale (for X_i continuous) and number of levels (for X_i discrete) introduces a bias. An uninformative variable with a large number of levels may be selected over a more informative variable with fewer levels;
    2. bagging introduces a bias, but sampling $(1 - 1/e) \approx 0.632$ of the data without replacement is an effective strategy for removing the problem.

Strobl et al. (2007) introduce changes to the random forest algorithm to fix 1) and 2). For correlated variables, Strobl et al. (2008) introduces "conditional variable importance." While these methodological advances address a number of problems with RF models, we still find a compelling use case for the standard RF model. This is particularly so with respect to genomic data,

2.2 Variable importance

Despite the desirability and utility of variable importance measures, there is an underlying difficulty with them. There is no theoretically defined variable importance metric in the sense of a parametric quantity that a variable importance estimator should try to estimate. See Grömping (2009) for a history of work with the concept of variable importance and also for a discussion of RF variable importance compared with a more explored variable importance context, that of linear models.

We note that there are two levels of variable importance:

- global, that is the variables importance in the model;
- local, the variables importance for single prediction.

Both are of importance and both can provided by a RF. Our concern here is specifically with the global RF Gini importance measure.

There are two other questions closely related to variable importance.

- variable selection. There is a vast literature on this in the context of linear regression. For RF, variable importance measures are often used as a precursor for a variable selection procedure, as we are doing here. Although variable importance is not well defined, variable selection is. If we know the number of variables in the generating model (as in a simulated example), we can count the number of true and false positives.
- dependency plots such as Individual Conditional Expectation (ICE) and Partial Dependence Plot (PDP) visualize the interactions of variables and may indicate how a variable is importance. Such plots are generally prohibitive in the context of RF with large p. There are several effort to develop methods to visualize variable effects in RF (Welling et al., 2016), and we have found them useful when there is a smaller subset of variables to investigate.

Shapley values, derived from games theory, has a number of desirable theoretical qualities (efficiency, symmetry, additivity, consistency). The importance of a variable in a prediction is based on the marginal utility of a variable over all possible subsets of the other variables. However, computing Shapley values is NP-hard. There are a several attempts to approximate it (see Lundberg and Lee (2017); Lundberg et al. (2019, 2020)) and it is the focus of a lot of work in interpretable ML at the present time. However, for genomics data, the computational burden is far too high.

2.3 GWAS and multiple testing

- describe GWAS data, wide data sets, comparing case and control
- give examples and size

$$\begin{array}{c} \overbrace{\qquad\qquad\qquad\qquad}^{\text{cases}} \quad \overbrace{\qquad\qquad\qquad\qquad}^{\text{controls}} \\ \begin{bmatrix} x_{1,1} & \dots & x_{1,i} & x_{1,i+1} & \dots & x_{1,n-1} & x_{1,n} \\ \vdots & & & \ddots & & & \vdots \\ \vdots & & & \ddots & & & \vdots \\ \vdots & & & \ddots & & & \vdots \\ \vdots & & & \ddots & & & \vdots \\ x_{p,1} & \dots & x_{p,i} & x_{p,i+1} & \dots & x_{p,n-1} & x_{p,n} \end{bmatrix} \end{array} \begin{bmatrix} t_1 & p_1 \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ t_p & p_p \end{bmatrix}$$

- Say we have data $\underset{n\times p}{X}$ with $p >> n$. We perform p simultaneous tests to get the test statistics $\{t_i\}_{i=1}^p$.
- Under normality assumptions, we can compute statistics z_j comparing patients with controls which satisfy, to a good approximation, $z_j \sim N(\delta_j, 1)$, where δ_j is the effect size for gene j : δ_j equals 0 for "null genes," genes that show the same genetic activity in patients and controls, while

$|\delta_j|$ is large for the kinds of genes being sought, namely, those having much different responses for patients versus controls. Inferences for the individual genes by themselves are immediate. For instance,

This is the position that the widely used approach (and code body) Ritchie et al. (2015) *limma* (Linear Models for Microarray Data) takes. It used a empirical Bayes argument to make an adjustment to the variance of each gene based on a model for the variance of all genes in the sample. This prevents genes with a small variance from having a greatly inflated t statistic and appearing significant.

It is also a common practice to permute the phenotype and derive a matrix of statistics of dimension p genes by M permutations. The "classic" approaches to determining the threshold for GWAS studies are given in Churchill and Doerge (1994) "Empirical Threshold Values for Quantitative Trait Mapping" where two methods for determining the threshold are given. They both depend on doing M permutations of the phenotype trait.

- the issues of multiple testing have long been understood. If you perform 20 tests at an $\alpha = 0.05$ then you expect one false positive. However in a high-throughput experiment this can mean that you expect 1000s of false positives. If there are only expect a relatively small number of genes to be involved in the process, the true positives can be drowned by the false positives;
- hence we do a multiple testing correction on the statistics t_i
  - we can control the probability of at least one false discovery (the family-wise error rate) via the Bonferroni correction. This is effective but at the cost of greatly reduced power. In the case of genomics problems the large number of tests reduces the power to essentially 0
  - Benjamini and Hochberg (1995) introduced the false discovery rate (FDR), or expected proportion of discoveries which are falsely rejected. This controlling the proportion of type I errors at a specified level but has greater power to detect true positives.
  - Storey and Tibshirani (2003) introduced the q-value
- Korthauer et al. (2019) gives an overview of many methods of controlling the false discovery rate and some optimal conditions for particular approaches to yield good results.

See Price et al. (2006) for a discussion of some of the issues in the linear regression approach to the analysis of GWAS data and see George et al. (2021) for a view of the current practice in the area, particularly attempts to extend the single locus analysis considered here to a multi-locus approach.

2.3.1 An empirical Bayes approach.

Efron (2010) suggests an empirical Bayes approach. The conceptual model underlying this is that the data is made up of two groups, a large group of data that will generate NULL values of some statistic and a smaller group that will generate non NULL values. The method involves taking the histogram of the data, as in figure 2 and modelling the NULL distribution so that the non-NULL can be identified by some criteria.

The method is ideally suited to the case/control experiments with genomic data, where it has become a standard method of analysis. The principal takeaways from Efrons approach are:

- despite the large "sample" sizes, the very poor fit that the $N(0, 1)$ Gaussian distribution has to the z values
- the permutation distribution is very similar to the $N(0, 1)$ Gaussian
- a problem like this, with
  - a large amount of data drawn from a NULL distribution (in this case, most of the z values)
  - a much smaller set of data drawn from a non-NULL distribution.

  is an ideal situation for an empirical Bayes estimate of the NULL distribution.

We consider a example discussed in Efron (2007, 2010)). A breast cancer data set from Hedenfalk et al. (2001) consisting of a matrix with 3226 rows corresponding to the measured genes and 15 columns

corresponding to the patients, divided between cancers with BRCA1 and BRCA2 mutations.

- let t_i be the set of statistics arising from the cases and controls. In this instance they are standard t-statistics, but they could be variance stabilised t-statistics or some other statistic.
- the method of Efron is working on the histogram of the data as is our method
- let $z_i = \Phi^{-1}(G_0(t_i))$, where Φ is the standard normal cdf, and G_0 is a putative null cdf for the t-values. G_0 can be a theoretical null distribution, or a permutation null (Efron, 2010)
- if the modelling assumptions are reasonable then $z_i \sim N(0,1)$, However, looking at figure 1 we see that the histogram of the data is far from the theoretical Gaussian. See Efron (2010) for a discussion of the many causes of this in practice. For both data sets we plot the histogram of $z_i = \Phi^{-1}(G_0(t_i))$, where Φ is the standard normal cdf, and G_0 is a putative null cdf for the t-values. The $N(0,1)$ distribution is given in red. If the modelling assumptions were accurate then the z_i should have a $N(0,1)$. However, figures 1 that they are far from the theoretical distribution.Efron gives a discussion of why this may occur:
  - failed assumptions;
  - correlations between cases, or between variables;
  - unobserved covariates.
- A surprise observation is that the permutation density is very similar to the theoretical density Taking the permutation distribution does not fix the problem (figure 1) as the permutation distribution is similar to the theoretical distribution.
- we model the observed distribution as a mixture $f(z) = p_o f_o(z) + p_1 f_1(z)$
  - where $f_o(z)$ is the null distribution of importance's
  - and $f_1(z)$ is the distribution of significant importance's.
- we will not go into the modelling process in depth here. It involves using the central mass of the distribution to fit a Gaussian, this is the null density, and then calculating a local fdr as the ratio of null density and the observed density of the tails $f_1(z)$.

We see the following results on this data set:

- 498 p-values < 0.05
- BH multiple testing correction reduces it to 2
- `limma` (Ritchie et al., 2015), with `eBayes` gives 23
- give the results for locfdr

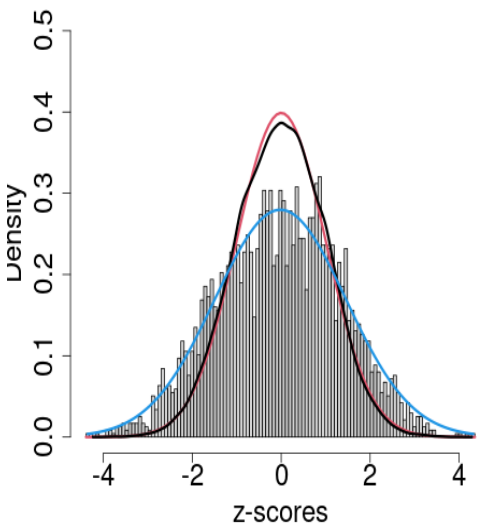


Figure 1: The histogram and the permutation null distribution for the Hedenfalk data set

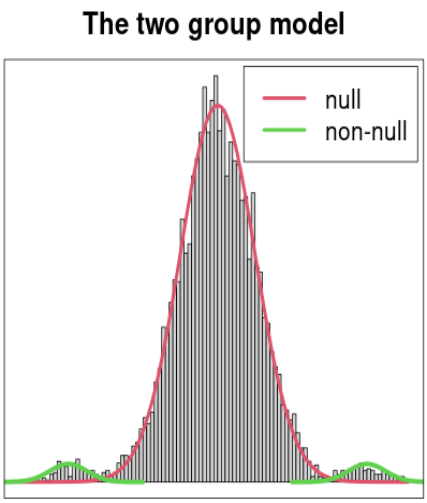


Figure 2: The two group model.

2.3.2 Machine Learning approaches to Genomic data

A distinguishing features of genomic data is the very large number of variables with respect to the relatively small number of observations. Just selecting a sample of data sets over the period from 2000 to now (2022) we see the following:

- 2001 – cDNA array 10×4000 Efron (2007)
- 2004 – effects of cigarette smoke on human airway epithelial cells, transcriptome 57×22283 (Spira et al., 2004)
- 2006 – colon cancer data set – Affymetrix HU133 A/B arrays 500×50000 probe sets (over 1M if analysed at probe level) (LaPointe et al., 2012)
- current – wheat varieties (Broadbent et al., 2019) 280×14614 plus replicates (Broadbent et al., 2019)
- current – (Clarke et al., 2017) UK BioBank datasets $50000 \times 5,000,000$. Human SNP data sets are typically bigger than transcriptome data sets because they are not tissue specific.

The data sets have grown over time but roughly p is about 2 orders of magnitude larger than n. The sizes are important. It is not just that genomic data sets are big, it is how they are big, and what we want from them also determines how we analyse them. Take the colon cancer data set mentioned above. If there are, say, $k = 100$ genes that are highly involved in early stage colon cancer, then $\rho = k/n = 0.1$ and $\delta = n/p = 0.01$. Consulting the phase diagram in Donoho and Stodden (2006); Donoho and Tanner (2009) we see that we are in the "difficult" region and many algorithmic methods of variable recovery will have a high probability of failing. Unfortunately there is no literature extending such phase diagrams to RF.

It is rare to see deep learning models used in genomics as variable selection is often a vital aspect of the analysis. In addition, the notable successes of DL have been in structured data, such as image and speech, where the DL has been designed to take advantage of the structure. There have not been similar successes in "tabular" data where we have variables by observations Rudin et al. (2021). The models outlined by Arik and Pfister (2020) and Cui et al. (2021) appear interesting (they either offer variable selection or are designed for genomic data). However, in none of these cases have we been able to successfully use the model, out of the box, on a genomic size data set. Also Lemhadri et al. (2021), which combines a lasso (variable selection) penalty on skip-layer connections with a deep leaning network, holds some promise.

Alipanahi et al. (2021) give an example using SNP data from the UK biobank Clarke et al. (2017). The aim is to predict glaucomatous optic nerve head features, which is normally done by a trained ophthalmologist. Using a labelled set of photographs they were able to predict the value of a useful diagnostic marker for glaucoma. There are two points we want to make about this study:

- The significant achievement of the analysis was the identification of SNP loci involved in glaucoma. That is, it resolved into a variable selection problem; However,
- while the authors describe the analysis as a "ML-based GWAS," the GWAS data was handled in a completely "classical" fashion The challenge that genomic data poses for deep learning methods is still open.

There are problems in genomics where random forests may be a good candidate technique and Gini variable importance may be more practical due to the large number of variables. In some of these area (like GWAS studies), all the variable will be on the same scale so the bias effect may be small. Michaelson et al. (2010) show that random forests have a better performance than some other multi-locus models (lasso and the elastic net) and single loci mapping methods (the standard).

See Bayat et al. (2020), where we introduced VariantSpark, a tailored Hadoop/Spark-based implementation of random forests, designed to handle data with extremely large number of variables per sample (O'Brien et al., 2018). In a test of the algorithm, VariantSpark was used to accurately predict ethnicity from whole genome sequencing profiles of 2,500 individuals and 80 million genomic

variants each (1000 Genomes Project). The current implementation of VariantSpark returns a Gini variable importance.

2.4 Other approaches to significance for variable importance

There are broadly two approaches to determining the significance level for RF variable importance, permutations of the response vector, and recursive feature elimination.

permutation methods

- we could permute the response vector a number of times calculate the p values for significance from the permuted importances. This is a standard procedure and is not specific to RF variable importances in any way. False discovery rates are often estimated by permutations, which are simulated from a null distribution. The validity of this procedure relies on the fact that the statistic for gene j is a function of only the data for gene j. Witten and Tibshirani (2008) comment that "[one can not] use permutations to estimate the null distribution for an arbitrary statistic that is a function of multiple genes." As the variable importances are functionally related in that the choice of one variable to split on influences (precludes) the choice of another, assumption is violated and a permutation approach is wrong. That this observation is relevant to RF variable importance is noted in Huynh-Thu et al. (2012), but seems widely ignored.
- Altmann et al. (2010) propose a methods that relies on permuting the response vector s times and thus getting a vector of s importances for each variable. The PIMP algorithm fits a probability distribution to the population of null importances. This is subject to the same problems as above.
- The Boruta algorithm (Kursa and Rudnicki, 2010) combines permutation and feature elimination. For the algorithm each variable has a shadow variable generated by permuting the original values. A RF is trained on the extended data set and the variable importance values are collected. For each real variable a test is performed comparing its importance with the maximum value of all the shadow variables. All unimportant variables and their shadow variables are removed and the previous steps are repeated until a stopping criteria is met.

recursive feature elimination

- stepwise selection based on accuracy (varSelRF) (Díaz-Uriarte and Alvarez de Andrés, 2006) A proportion of the least important variables are removed and a new RF is generated using the remaining variables. This is done recursively until a single variable is left. At each step the prediction performance is estimated based on the out-of-bag samples that were not used for model building. Degenhardt et al. (2019) modifies the original varSelRF to recompute the rankings at each step.

Vita (Janitza et al., 2016) is only applicable to permutation importances, but is similar in spirit to the algorithm proposed here in that it estimates the null distribution. Permutation importances can take negative values and it is argues that the variable taking negative values are null. By mirroring the negative permutation importances we can estimate the null distribution.

Degenhardt et al. (2019) provides a thorough review and comparison of methods of setting the cut-off value, but does not include impurity importance. The conclusions drawn are that Boruta and Vita are the two preferred methods. The size of the genomic data sets used up to a 1000 samples and 20000 variables, so small numbers of variables by genomic standards.

2.5 Empirical Bayes for Gini Variable Importance

The variables importances are the data for this modelling exercise. Consequently, the method depends on having a large number of variables and a large number of trees. It is not going to calculate variable importances for Fisher's Iris data, as we can't estimate a density with 4 points.

This makes it ideally suited for the GWAS case.

- We have shown in examples about that there is often a discrepancy between a theoretical null distribution or a permutation derived null and the observed data. Here we review a method, largely inspired by Efrons approach, for modelling distribution of Gini Variable Importance, with a view to setting a significance threshold.
- We have shown that there are significant problems with the null permutation distribution for VI. We have also shown that in the context of multiple testing for equality of means, an empirical Bayes method can be extremely useful. We now consider the potential of empirical Bayes in the context of Gini VI.
- we ref to the figures for the Thousand Genome example to illustrate the argument (see section 3.2)
- we log the data, the distributions are clearly very far from Gaussian, the distribution in multi-modal.
- we model the observed distribution is a mixture $f(z) = p_C f_c(z) + p_0 f_0(z) + p_1 f_1(z)$
  - $f_c(z)$ here may be multi-modal (see figure 8a)
  - The distribution we are interested in $p_0 f_0(z) + p_1 f_1(z)$ is indicated in figure 8a
  - we initially modelled $f(z)$ as a mixture of Gaussians. However a simpler and seemingly more robust procedure is to
    * count the number of times that each variable is used – this is "C"
    * progressively threshold this to get $f(z)|C > c$
    * using a model selection procedure, select $C^\star$ so that the distribution is unimodal. See figure 8b for $f(z)|C > C^\star$.

    The modes of $f(z)$ can be associated with particular counts, for example, the first mode of figure 8a is the importances of the variables that were only used once in the forest. In this case there is a significant mode for variables that were not used in any tree. This mode has been moved to $-\infty$ by logging the data.
  - and $f^\star(z) \approx p_0 f_0(z) + p_1 f_1(z)$ is the distribution of significant importances.
- Having extracted the desired distribution, we consider how to model it. It is clear that it is often far from Gaussian, with a a thicker tail to the right
- from this point, the object of the analysis is the histogram of $f^\star(z)$
- using a standard Poisson regression software, we fit a spline to the
- the model is $f = \pi_0 f_o + (1 - \pi_o) f_1$
- We explore a number of densities for modelling this density.
- There is no theoretical guidance on the distribution of variable importances in the null case. A number of densities were explored including Pareto, Burr, log-logistic, generalized lambda,..... In addition Cullen and Frey plots were used to understand the features of the fitted density.
- heavy tailed distributions – Moment generating function does not exist $\int_\infty^\infty \exp(tx) dF(x) = \infty$
- Jiang (2021) *Multiple Testing with Test Statistics Following Heavy-Tailed Distributions*
  - if f_0 is correct distribution then the p values will be uniform
  - $\pi_0 \approx 0.9$
- in our case $\pi_0 \approx 0.99$ and the data is noisy – f_0 is not a perfect approximation – so it does not work
- determine C such that to the left of C $f \approx f_o$ (a change point method related to penalized model selection Gauran et al. (2018))
- fit f_o to x s.t. $x < C$ where f_o is a skew-normal distribution
- fit a spline f_{spline} to the histogram using standard Poisson GLM i.e
  `glm(y ~ splines::ns(x, df = df), Poisson)`

- using the ratio f_0/f_{spline} determine the local false discovery rate

We consider a number of examples (see below) as we explore the possibilities of Empirical Bayes.

3 Results

3.1 Demonstration on synthetic data

We consider two simulations that show that correlation may increase a variables importance or decreas it

1. First a classification problem that shows that importance is smeared over a set of correlated variables, which then have a reduced variable importance, depending on the size of the set. That is, larger sets of correlated variables will have a greater reduction in importance;
2. the simulation from Strobl et al. (2008), a regression problem. This shows that variables with a weak relationship with the independent variable y but correlated with a significant variable may have an increased variable importance, compared to similar but uncorrelated variables.

We note that these two simulations are very different and that both effects could exist in a real data set. However, this is enough to show that the the effects of correlations between the variables may be hard to untangle in any one particular data set.

Both Gini and permutation importance are subject to the same effect. Conditional inference trees (Hothorn et al., 2006) and "conditional variable importance" (Strobl et al., 2008) also suffer from the same problem.

3.1.1 Classification causes variable importance to smear over correlated variables

Each "band" contains "blocks" of sizes $\{1, 2, 4, 8, 16, 32, 64\}$ and each block consists of correlated (identical variables). Each variable is $\in \{0, 1, 2\}$ and the dependent variable y is 1 if any of $X[, c(1, 2, 4, 8, 16, 32, 64)]$ is non-zero. So y contains a lot more 1's than 0's.

We have 4 bands (as shown in figure 3). We fit a standard RF and record the importances. Figure 4 shows the importance scores arranged by variable number and colored by band. We see that blocks with fewer variables have a higher importance score. This is because the importance is "smeared" over the correlated (identical in this case) variables.

Figure 5 shows the importances for the first 127 variables (i.e the true non-null variables). The plot shows several things:

- for a small number of correlated variables, the importance in increased. In this example it is only in evidence for the set of 2 correlated variables;
- for larger groups of correlated variables, the importance is smeared across the whole set and is decreased
- strangely, the permuted data set (where none of the variables is a true non-null) the same effect seems to be apparent

3.1.2 Regression inflates the importance of individually weak but correlated variables

Strobl et al. (2008) shows a simulation with the data generated according to a linear model with twelve predictor variables $y = \beta_1 x_1 + \ldots + \beta_{12} x_{12} + \epsilon$, with $\epsilon \sim N(0, 0.5)$. The predictor variables were sampled from a multivariate normal distribution $X_1, \ldots, X_{12} \sim N(0, \Sigma)$ where the covariance structure

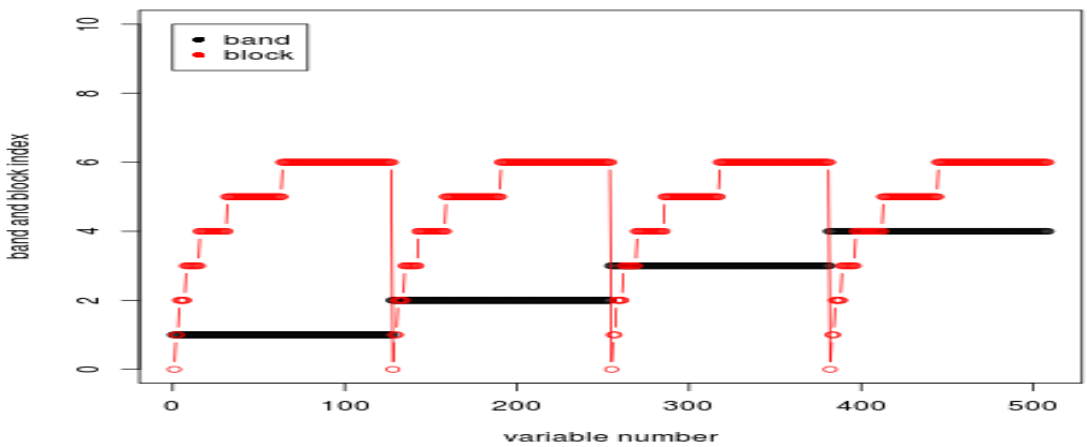


Figure 3: The data is structured into bands and blocks. The color and the y-axis show which band/block each variable belongs to, not the variables value. Each "band" contains "blocks" of sizes 1, 2, 4, 8, 16, 32, and 64. Each block consists of correlated (identical variables), where each variable is $\in \{0, 1, 2\}$. The dependent variable y is 1 if any of $X[, c(1, 2, 4, 8, 16, 32, 64)]$ is non-zero, so only band 1 has a relationship to the dependent variable.

Σ was chosen such that all variables have unit variance $\sigma_j, j = 1, \ldots, 12$ and the first four predictor variables are block-correlated with $\sigma_{j,j'} = 0.9$ for $j \neq j' \leq 4$, while the rest were independent with $\sigma_{j,j'} = 0$. The coefficients of the linear generating model were $\{5, 5, 2, 0, -5, -5, -2, 0, 0, 0, 0, 0\}$.

Strobl et al. (2008) uses a forest based on conditional inference tree (see Hothorn et al. (2006)). In addition Strobl et al.'s argument is based on the frequency with which a variable is selected for splitting, not the variable importance. We show that the simulation leads to the same results with a standard forest algorithm and with variable importance.

3.2 Thousand genomes

We obtained the 1000 Genomes Project data as VCF files from their FTP site. Each VCF file contains the variants for every individual across one chromosome. We use the phase 3 dataset which contains 2,504 individuals and 1M SNPs for chr22. The task was to predict ethnicity from the SNPs.

If we example the density of the log of the importances we see that it is multi-modal (figure 8a). An easier procedure than modelling this with a mixture distribution, is to take only the importances where the variable has be used greater then n times. This is quite problem dependent and with 10^5 trees and 10^6 variables it can be quite large. In this instance, neither the Hartigan's diptest for unimodality (Maechler, 2021) nor Cramer-von Mises test of goodness-of-fit to a Skew-Normal distribution give reasonable results. By inspection, the desired distribution is isolated when $n \geq 40$. This removed 0.87 of the variables, 0.43 of variables never appeared in the RF and many only appeared a small number of times. We shift the distribution along the x positive axis (and add the Machine double.eps). This gives the density shown in figure 8b.

The procedure was:

- examine the density (figure 8a). As it is multi-modal:
  - take the importances where the variable appears 40 or more times in the RF.
  - shift the distribution to the positive axis (and add the Machine double.eps). This gives the density shown in figure 8b.
- model this density as the sum of a Null and a non-Null density.

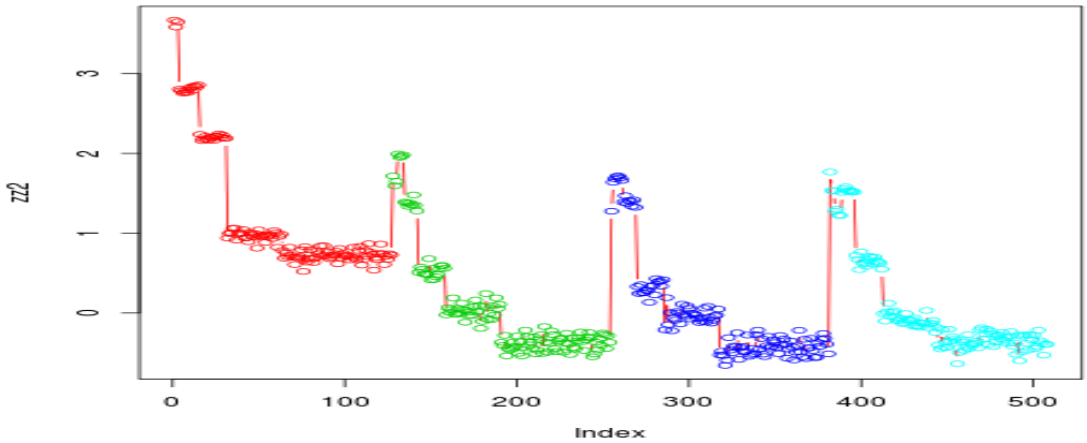


Figure 4: The log importance scores for simulation 2, arranged by variable number and colored by band.

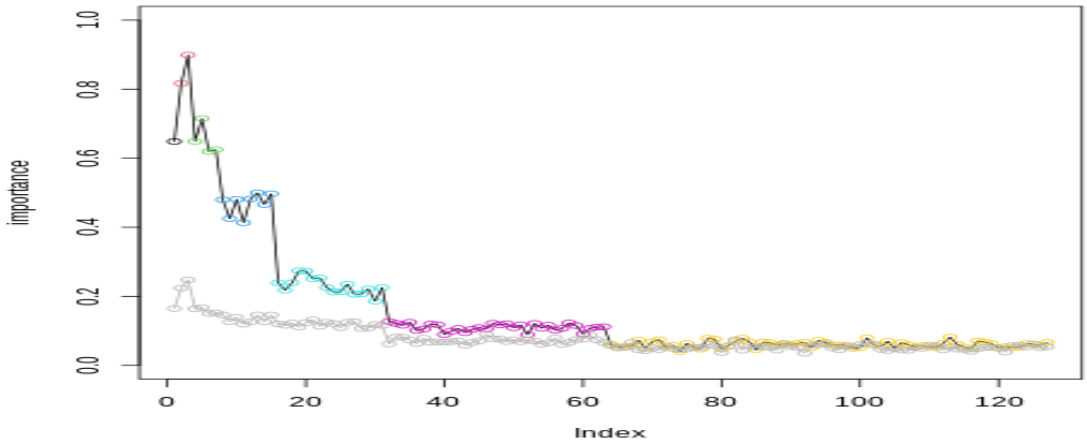


Figure 5: The importances for the first 127 variables (colored) and for a permuted data set.

  - taking the histogram, fit a spline to the observed bin counts modelled as a Poisson random variables, this is f
  - fit a Burr distribution to the data truncated at x=3, This is f_0
  - using f and f_0, estimate the local fdr.
- Figure 8e shows a plot of the fitted distribution and empirical distribution for the data. From this we can determine the number of significant SNPs.

4 Conclusion

A method is presented for calculating the significance threshold of RF impurity importance using an empirical Bayes approach. The method has a number of advantages:

- it is computationally fast, particularly when compared to recursive feature elimination or permutation methods
- it does not involve the issues that arise when calculating a false discovery rate using permutations.

It depends on a problem with sufficient varaiables and a sufficient number of trees in order to form useful estimates of the null distribution. However, the lower limit in terms of varaibles and trees has

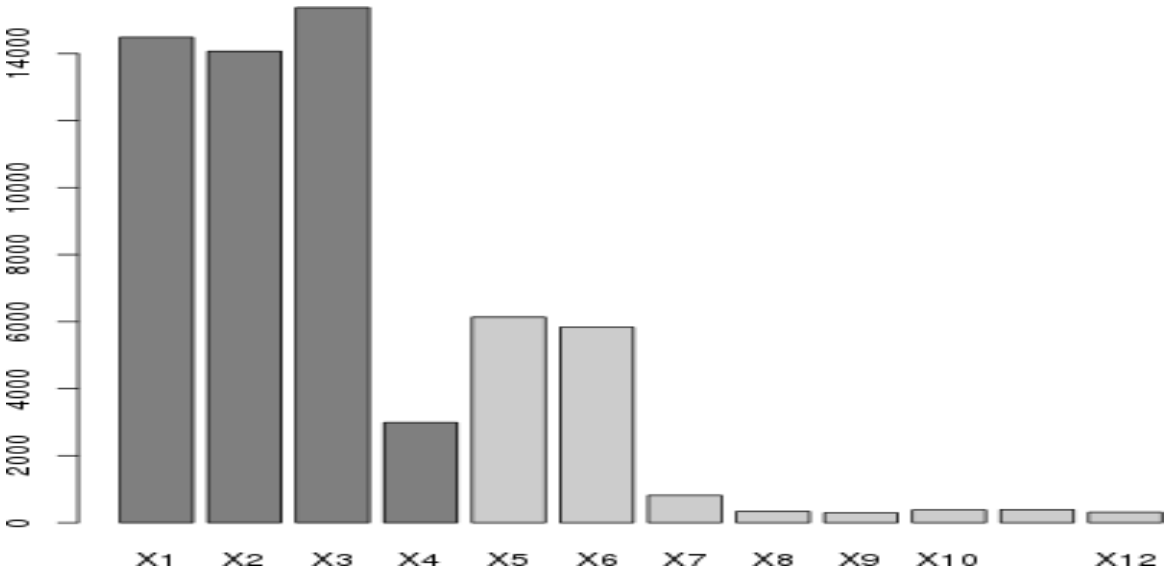


Figure 6: The Strobl et al. (2008) simulation using a standard tree algorithm (ranger). 6 shows the Gini impurity importance

not been determined.

5 Software

5.1 R implementation

```
devtools::install_github("parsifal9/RFlocalfdr", build_vignettes = TRUE, dependencies=TRUE)
library(RFlocalfdr)
vignette("RFlocalfdr")
```

5.2 Python implementation:

Along with the R implementation of this tool, we released a python version included in the VariantSpark repository. The python script has been tested using python version 3.8.12. The libraries used are the following: numpy version 1.21.2 for numeric transformations; pandas version 1.4.1 to create and manage data frames; patsy version 0.5.2 to create the cubic regression splines; scipy version 1.7.3 to fit the data using the last squares method, compute the cumulative distribution functions and percentile point functions; and statsmodels version 0.13.2 to fit a generalized linear model.

Although the script contains several functions, the end user is only required to use the run_it_importances. This function requires a data frame as input with a single column containing the logarithm of the RF importance for each variant. Currently the method returns a dictionary with the FDR values, the estimates for the fitted function, and the p-values for the statistically significant variants.

Link to the python implementation: `https://github.com/aehrc/VariantSpark/blob/3d9900e93f5a21768dfc0cf8733779ba147a3f48/python/varspark/pvalues_calculation.py` (Notice that this implementation is the current last commit on a branch of VariantSpark)

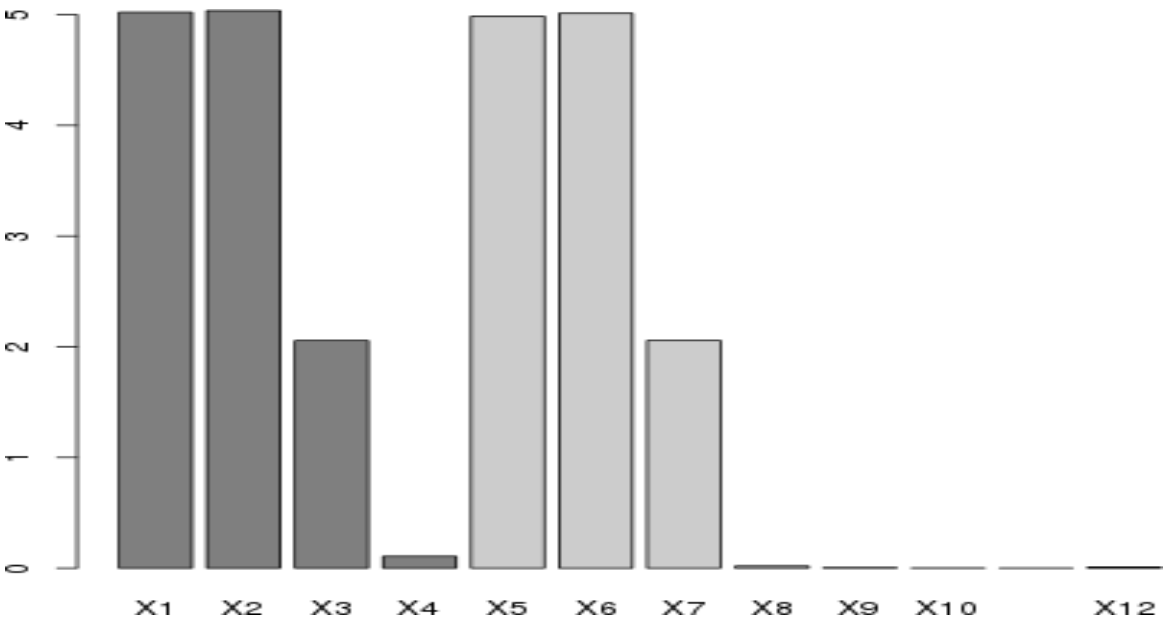
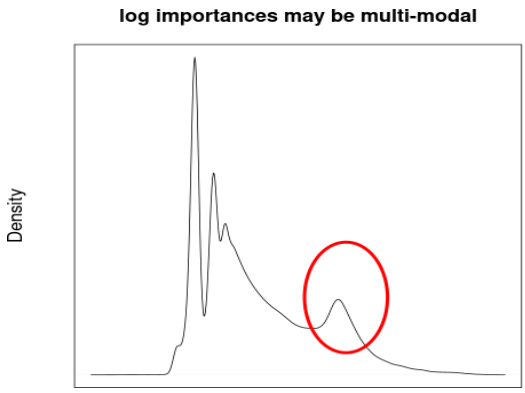
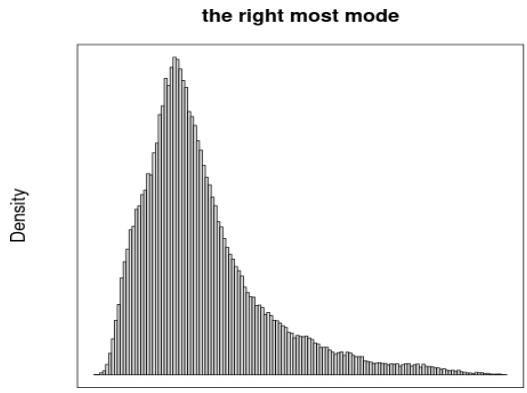


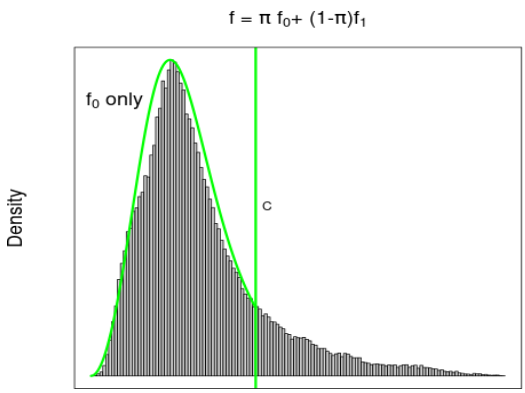
Figure 7: The absolute value of the coefficients from a linear model fitted to the Strobl et al. (2008) simulation data. The linear model is successful in recovering the generative model coefficients $\{5, 5, 2, 0, -5, -5, -2, 0, 0, 0, 0, 0\}$. The fact that variables 1-4 are correlated and 5-8 are uncorrelated has not greatly affected the fitted coefficients.



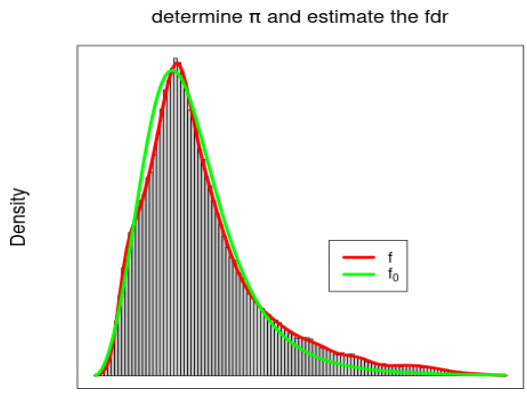
(a) The density of the log importances is multi-modal. The density $p_0 f_0(z) + p_1 f_1(z)$ is indicated in red.



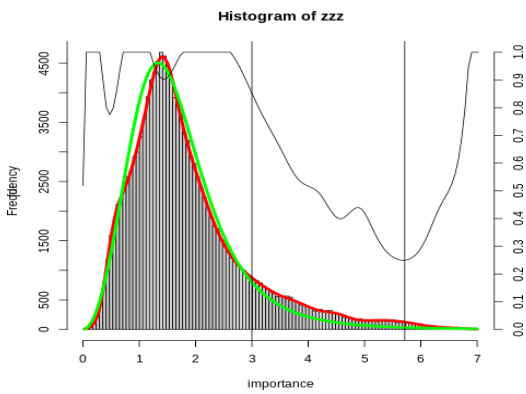
(b) If we take only variables that are used greater than 40 times in the forest, we isolate the desired distribution, which we can shift along the positive axis.



(c)



(d) Empirical versus theoretical CDF for a number of distributions fitted by ML.



(e) Empirical versus theoretical CDF for a small number of well fitting distributions.

References

Alipanahi, B., Hormozdiari, F., Behsaz, B., Cosentino, J., McCaw, Z. R., Schorsch, E., Sculley, D., Dorfman, E. H., Foster, P. J., Peng, L. H., Phene, S., Hammel, N., Carroll, A., Khawaja, A. P., and McLean, C. Y. (2021). Large-scale machine-learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology. *The American Journal of Human Genetics*, 108(7):1217–1230.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340.

Arik, S. O. and Pfister, T. (2020). TabNet: Attentive Interpretable Tabular Learning. *arXiv:1908.07442 [cs, stat]*.

Bayat, A., Szul, P., O'Brien, A. R., Dunne, R., Hosking, B., Jain, Y., Hosking, C., Luo, O. J., Twine, N., and Bauer, D. C. (2020). VariantSpark: Cloud-based machine learning for association study of complex phenotype and large-scale genomic data. *GigaScience*, 9(8).

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57(1):289-300.

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

Broadbent, J., Stockwell, S., Byrne, K., Bose, U., Ramm, K., Hyles, J., Trevaskis, B., Dillon, S., and Colgrave, M. (2019). OzWheat - A genetic diversity panel for classification and prediction of wheat traits. In *Human Proteome Organisation World Conference; Adelaide, Australia*. Human Proteome Organisation World Conference. `http://hdl.handle.net/102.100.100/421985?index=1`.

Churchill, G. A. and Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics*, 138(3):963–971.

Clarke, L., Fairley, S., Zheng-Bradley, X., Streeter, I., Perry, E., Lowy, E., Tassé, A.-M., and Flicek, P. (2017). The international Genome sample resource (IGSR): A worldwide collection of genome variation incorporating the 1000 Genomes Project data. *Nucleic Acids Research*, 45(D1):D854–D859.

Cui, T., El Mekkaoui, K., Reinvall, J., Havulinna, A. S., Marttinen, P., and Kaski, S. (2021). Gene-gene interaction detection with deep learning. *bioRxiv*.

Degenhardt, F., Seifert, S., and Szymczak, S. (2019). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, 20(2):492–503.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3.

Donoho, D. and Stodden, V. (2006). Breakdown point of model selection when the number of variables exceeds the number of observations. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1916–1921. IEEE.

Donoho, D. and Tanner, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293.

Efron, B. (2005). Local False Discovery Rates.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103.

Efron, B. (2008). Microarrays, Empirical Bayes and the Two-Groups Model. *Statistical Science*, 23(1):1–22.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge.

Gauran, I. I. M., Park, J., Lim, J., Park, D., Zylstra, J., Peterson, T., Kann, M., and Spouge, J. L. (2018). Empirical null estimation using zero-inflated discrete mixture distributions and its application to protein domain data. *Biometrics*, 74(2):458–471.

George, A. W., Verbyla, A., and Bowden, J. (2021). Eagle for better genome-wide association mapping. *G3 Genes—Genomes—Genetics*, 11(9):jkab204.

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, 63(4):308–319.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., Wilfond, B., Borg, A., Trent, J., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., and Sauter, G. (2001). Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, 344(8):539–548. [DOI:10.1056/NEJM200102223440801] [PubMed:11207349].

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674.

Huynh-Thu, V. A., Saeys, Y., Wehenkel, L., and Geurts, P. (2012). Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics*, 28(13).

Janitza, S., Celik, E., and Boulesteix, A.-L. (2016). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, pages 1–31.

Jiang, Z. (2021). Multiple testing with test statistics following heavy-tailed distributions. https://infoscience.epfl.ch/record/283692.

Kaggle (2019). Historical Data Science Trends on Kaggle. Accessed: 2019-04-20.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., Shukla, C., Alm, E. J., and Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20(1):118.

Kursa, M. B. and Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, 36(11):1–13.

LaPointe, L. C., Pedersen, S. K., Dunne, R., Brown, G. S., Pimlott, L., Gaur, S., McEvoy, A., Thomas, M., Wattchow, D., Molloy, P. L., and Young, G. P. (2012). Discovery and validation of molecular biomarkers for colorectal adenomas and cancer with application to blood testing. *PLoS ONE*, 7(1).

Lemhadri, I., Ruan, F., Abraham, L., and Tibshirani, R. (2021). LassoNet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2019). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv:1802.03888 [cs, stat]*.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774.

Maechler, M. (2021). *diptest: Hartigan's Dip Test Statistic for Unimodality - Corrected*. R package version 0.76-0.

Michaelson, J. J., Alberts, R., Schughart, K., and Beyer, A. (2010). Data-driven assessment of eQTL mapping methods. *BMC Genomics*, 11:502.

O'Brien, A., Szul, P., Dunne, R., Leo, P., Duncan, E. L., Twine, N. A., Li, S., Doecke, J., Ellis, N., and Bauer, D. C. (2018). Cursedforest- a random forest implementation for "big" and "wide" data. submitted.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2021). Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *arXiv:2103.11251 [cs, stat]*.

Segal, M. R. (2004). Machine learning benchmarks and random forest regression.

Spira, A. E., Beane, J., Pinto-Plata, V., Kadar, A., Liu, G., Shah, V., Celli, B., and Brody, J. S. (2004). Gene expression profiling of human lung tissue from smokers with severe emphysema. *Am J Respir Cell Mol Biol.*

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25. highly accessed.

Welling, S. H., Refsgaard, H. H. F., Brockhoff, P. B., and Clemmensen, L. H. (2016). Forest Floor Visualizations of Random Forests. *ArXiv e-prints.*

Witten, D. M. and Tibshirani, R. (2008). Testing significance of features by lassoed principal components. *Ann. Appl. Stat.*, 2(3):986–1012.