Thresholding Gini Variable Importance with a single trained Random Forest: An Empirical Bayes Approach

Robert Dunne[1], Roc Reguant[2], Priya Ramarao-Milne[2], Piotr Szul[3], Letitia Sng[2], Mischa Lundberg[2,4], Natalie A. Twine[2,5], and Denis C. Bauer[2,5,6]

[1]*Data61, Commonwealth Scientific and Industrial Research Organisation, Sydney, Australia*
[2]*Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation, North Ryde, Australia*
[3]*Data61, Commonwealth Scientific and Industrial Research Organisation, Dutton Park, Australia*
[4]*Diamantina institute, The University of Queensland, St Lucia, Australia*
[5]*Macquarie University, Department of Biomedical Sciences, Faculty of Medicine and Health Science, Macquarie Park, Australia*
[6]*Macquarie University, Applied BioSciences, Faculty of Science and Engineering, Macquarie Park, Australia*

Wednesday, April 6, 2022

Abstract

Background: Random Forests (RF) are a widely used modelling tool, enabling feature-selection via a variable importance measure. For this, a threshold is required that separates label-associated features from false positives. In the absence of a good understanding of the characteristics of the variable importance measures, current approaches attempt to select features by training multiple random forest to generate statistical power via a permutation null, employ recursive feature elimination or a combination of both. However, for high-dimensional datasets, such as genome data with millions of variables, this is computationally infeasible.

Method: Here we present RFlocalfdr, a statistical approach to a threshold that identifies which features are significantly associated with the prediction label and reduces false positives. It builds on the empirical Bayes argument of Efron (2005) and models the variable importance as mixture of two distributions – null and non-null "genes."

Result: We demonstrate on synthetic data that RFlocalfdr has an equivalent performance to computationally more intensive approaches. But unlike these methods, RFlocalfdr successfully thresholds a 6 Million-feature dataset of 10,000 samples in real-time, being up to 100 times faster than other tools. RFlocalfdr can be applied to the output of any RF implementation returning variable importance and counts and we demonstrate its application to ranger and VariantSpark.

Conclusion: RFlocalfdr allows for robust feature selection by placing a confidence value on the predicted importance score. It does so without repeated fitting of the RF or the use of additional shadow variables and is thus usable for data sets with very large numbers of variables.

Contents

1 Introduction **3**

1.1 Random Forests 3

1.2 Feature selection for RF 4

1.3 Statistical methods for high-dimensional data 4

1.3.1 An empirical Bayes approach. 5

2 Methods **6**

2.1 Empirical Bayes for MDI Importance 7

2.1.1 task 1 7

2.1.2 task 2 8

3 Results **10**

3.1 Demonstration on a synthetic data set 10

3.2 Feature selection on the Thousand Genomes Dataset 11

3.3 Scaling to 10^6 datapoints 12

4 Conclusion **13**

1 Introduction

Random Forests (RF) are a widely used modelling tool. Lundberg et al. (2019) concludes that no nonlinear model has 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. In addition, in the genomics area especially, RF are widely used for the variable importance measures that they provide, which enables the selection of features that are associated with the prediction label. As such, RF can be seen as a step in the direction of interpretable machine learning.

However, the variable importance measure is not immune from false positives (examples in Section 3.1). A statistical assessment is needed to confidently identify which variable are label-associated.

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 (Grömping, 2009). The only variable importance measure that has a strong theoretical foundation is based on Shapley values, the calculation of which is NP-complete (Lundberg and Lee, 2017; Lundberg et al., 2019, 2020)). Likewise "conditional variable importance" (Strobl et al., 2008) addresses some of the deficiencies of Mean Decrease in Impurity (MDI) but for wide data genomic problems, the computational burden is far too high for practical use.

Several methodologies have been developed enabling the use of importance measures in separating signal from noise when selecting features associated with the label. Here we discuss Boruta (Kursa and Rudnicki, 2010), AIR (Nembrini et al., 2018), RFE and PIMP (Altmann et al., 2010). However, these tools are either too compute intensive, because several RF model need to be trained to gain statistical confidence, or lack the ability to fine tune the level of sensitivity.

Addressing these shortcomings, we introduce RFlocalfdr, a method for setting a significance level for MDI that does not involve any refitting of the RF model. In addition we avoid the use of "shadow variables" (Nembrini et al., 2018), which makes our approach applicable to extremely high dimensional data sets, such as genomic data, where the number of variables may be in the millions.

The paper is organized as follows:

- we introduce the RF model and discuss the characteristics of the MDI variable importance;
- we consider the characteristics of genomic data and consider other approaches to variable importance/selection, including the classical gene-by-gene analysis, Efron's empirical Bayes and machine learning approaches. We consider the use of a permutation null or recursive-feature-elimination (RFE) to set a threshold for RF variable importances;
- we introduce an empirical Bayes method of estimating the distribution of the null genes and setting a local false discovery rate. The method is called Random Forests-local-false-discovery-rate (RFlocalfdr) and R and python implementations are available;
- we compare RFlocalfdr and other widely used methods on several data sets.

1.1 Random Forests

RF models have a number of desirable features that make them particularly useful for genomic data with $p >> n$. The default settings are very robust and there is no call for a train/validation split of the data, which is very useful when we have small n. In addition, the out-of-bag (OOB) estimate of model accuracy is an acceptable alternative to a test set. It characteristics are still being explored but it appears that it is more likely to over- than under-estimate the true error (Janitza and Hornung, 2018).

In addition, and perhaps most importantly, they provide a measure of variable importance. In almost all genomic data applications this, rather than low prediction error, is the desired outcome. Strobl

et al. (2007, 2008) discusses a number of problems with both the MDI and permutation variable importances and 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, particularly with genomic data. The permutation variable importance measure is not feasible for genomic data and the "conditional variable importance" is significantly more computationally demanding.

1.2 Feature selection for RF

As reviewed by Degenhardt et al. (2019), there are broadly two approaches to determining the label-association for RF variable importance: permutations of the response vector, and recursive feature elimination.

Borrowing from the standard procedure for estimating false discovery rates, the response vector can be permuted a number of times to calculate the p values for significance from the permuted importances. However, 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 split variable influences (precludes) the choice of another, the 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 to be widely ignored.

The algorithm of Nembrini et al. (2018) involves taking one permutation of the class labels. At a node, variables are sampled from the original variables, plus a set of permuted variables. If the selected split is on one of the original variables, then that is used in the calculation of MDI as usual. If the split is on one of the permuted variable then the decrease in impurity is subtracted from that of the original, unpermuted, variable. Using the procedure of Janitza et al. (2016), these corrected MDI values are turned into p-values. Nembrini et al. (2018) refer to this method as actual-impurity-reduction (AIR).

Altmann et al. (2010) propose a method (Permutation IMPortance, PIMP) 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 and returns a p-value for each observed variable based on the permuted values for that variable.

The second widely used method is RFE. (Díaz-Uriarte and Alvarez de Andrés, 2006) introduce a stepwise selection based on accuracy (varSelRF). 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 OOB samples. Degenhardt et al. (2019) modifies the original varSelRF to recompute the rankings at each step.

Building on this, the Boruta algorithm (Kursa and Rudnicki, 2010) combines permutation and feature elimination. It is model agnostic in that any model that returns a vector of variable importances can be used with Boruta. 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.

1.3 Statistical methods for high-dimensional data

While offering possible approaches for feature selection, RFE, Bortua and AIR require multiple random forests to be trained. This can be computationally intensive and even prohibitive, depending on the dimensionality of the input data.

For example, genomic data is characterised by having a very large number of variables for a relatively small number of observations. Here, datasets can range from $(n \times p)$ 500 samples $\times$ 50,000 variables ((LaPointe et al., 2012) a colon cancer dataset) to 50,000 samples $\times$ 5,000,000 variables (UK BioBank (Clarke et al., 2017)).

So even for the smaller colon cancer data set if there are, say, $k = 100$ genes that are highly involved in early stage cancer, and we define $\rho = k/n$ and $\delta = n/p$ we have $\{\rho, \delta\} = \{0.1, 0.01\}$. Donoho and Stodden (2006); Donoho and Tanner (2009) puts this in the "difficult" region of $\{\rho, \delta\}$ phase space where many algorithmic methods of variable recovery have a high probability of failing. Unfortunately there is no literature, as yet, on extending such phase diagrams to RF.

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

Say we have a data set $\underset{n \times p}{X}$ with $p >> n$, as shown above. We perform p simultaneous tests to get the test statistics $\{t_i\}_{i=1}^{p}$. Under assumptions about the distribution of $\{t_i\}_{i=1}^{p}$, 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 activity in patients and controls, while $|\delta_j|$ is large for genes having much different responses for patients versus controls. Inferences for the individual genes gives the p-values, $\{p_i\}_{i=1}^{p}$.

This is the position that the widely used Linear Models for Microarray Data (LIMMA) (Ritchie et al., 2015) takes. It uses an 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 very 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, given such a matrix, are discussed in Churchill and Doerge (1994) "Empirical Threshold Values for Quantitative Trait Mapping."

For both approaches, the issue of multiple-testing needs to be evaluated and a widely used approach was introduced by Benjamini and Hochberg (1995) as the false discovery rate (FDR), or expected proportion of discoveries which are falsely rejected. This controls the proportion of type I errors at a specified level but has greater power to detect true positives than the simple Bonferroni correction (Korthauer et al., 2019). From this, Storey and Tibshirani (2003) introduced the q-value, the infimum over the p-values of the false discovery rate.

Building on this, Efron (2005; 2007; 2008; 2010) proposed a "local fdr" approach, which is able to control the false discovery rate for multiple testing using empirical Bayes estimate of the null distribution.

1.3.1 An empirical Bayes approach.

The conceptual model underlying this is that the data is composed 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 $N(0,1)$ Gaussian distribution often has a very poor fit 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 illustrate Efron's method with a small breast cancer data set discussed in Efron (2007, 2010). The data is from 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 comparison of cases and controls. In this instance they are standard t-statistics, but they could be variance stabilised t-statistics or some other statistic.
- 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. Interestingly the permutation density is very similar to the theoretical $N(0,1)$ density, so a non-parametric approach does not alleviate the problem. The method of Efron involves working on the histogram of the $\{z_i\}_{i=1}^p$.
- In figure 1 we plot the histogram of z_i. If the modelling assumptions were accurate then the z_i should have a $N(0,1)$, as shown in red. However, the histogram is far from the theoretical distribution. Efron gives a discussion the reasons this may occur, including:
  - failed assumptions;
  - correlations between cases, or between variables;
  - unobserved covariates.
- we model the observed distribution as a mixture $f(z) = p_o f_o(z) + (1 - p_0) f_1(z)$
  - where $f_o(z)$ is the null distribution of importance's
  - and $f_1(z)$ is the distribution of significant importance's.
- The modelling process involves using the central mass of the distribution to fit a Gaussian (the null distribution), and then calculating a local fdr as the ratio of a null density and the observed density of the tails $f_1(z)$. See Efron (2007, 2010) for more details.

We see the following results on this data set:

- 498 p-values < 0.05
- Benjamini and Hochberg (1995)'s multiple testing correction reduces it to 2
- `limma` (Ritchie et al., 2015), with `eBayes` gives 23
- `locfdr` gives 0.

Our conclusion in this instance would be that the experiment offers no evidence of genes that are different between the cases and controls and is perhaps, under-powered. Efron (2007) discusses this and gives other examples.

2 Methods

We have adopted the same strategy to enable false discovery evaluation for RF on high dimensional vectors of MDI importances. This turns a problem, the high dimensionality of the data, into an asset: we now have enough data points to estimate the null distribution accurately. In conjunction with RF implementations for high-dimensional data, e.g. VariantSpark Bayat et al. (2020), Random Forest can deliver on its promise of offering higher performance for GWAS-approaches than other multi-locus

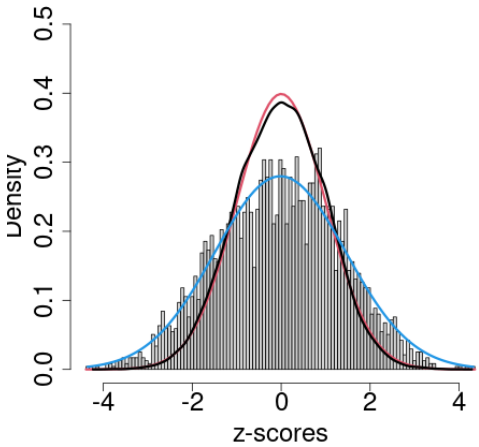


Figure 1: The histogram of z values for the Hedenfalk data set (Hedenfalk et al., 2001). The red curve shows the $N(0,1)$ distribution, the black curve shows the permutation null distribution. The blue curve is the empirical Bayes Gaussian fit to the data.

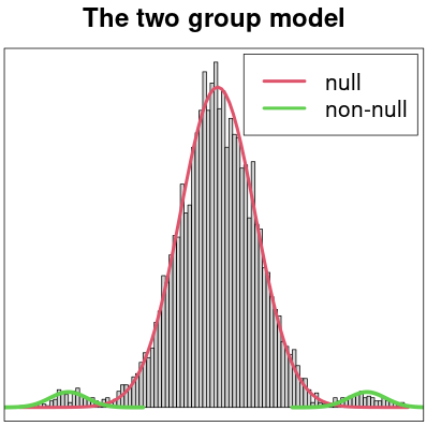


Figure 2: The two group model.

models (lasso and the elastic net, (Friedman et al., 2010)) and single loci mapping methods (Michaelson et al., 2010).

2.1 Empirical Bayes for MDI Importance

Efron (2005; 2007; 2008; 2010) approached the problem for case/control testing by introducing an empirical Bayes approach. Here we introduce a method, RFlocalfdr, inspired by Efrons approach, for modelling the distribution of MDI importance, with a view to setting a significance threshold. The variables importances are the data for this modelling exercise in that the method depends on having a large number of variables and a large number of trees. It is ideally suited for genomic data, both transcript and GWAS.

We use the vector of MDI values from the Thousand Genome example (discussed in section 3.2) to illustrate the method. We take the log of the MDI and consider the distribution shown in figure 3a. As is often the case, it is multi-model and we model it as a mixture,

$$f(z) = p_A f_A(z) + p_B f_B(z)$$

and

$$f_B = p_0 f_0(z) + (1 - p_0) f_1(z).$$

The distribution we are interested in $f_B(z) = p_0 f_0(z) + (1 - p_0) f_1(z)$ is indicated in red in figure 3a. We assume that this is unimodal and that it is a mixture of null genes and non-null genes. So we have two tasks. One is to separate $f_A(z)$ and $f_B(z)$, the other is similar to Efron's problem, to estimate the empirical null $f_0(z)$ and calculate the local fdr.

2.1.1 task 1

Due to the wide range of distributions between different data sets, modelling $p_A f_A(z) + p_B f_B(z)$ by mixtures is quite difficult. However we observe that the modes of $f_A(z)$ are associated with variables

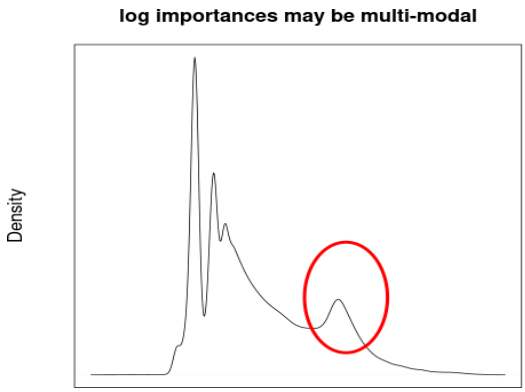
that were used a specific (and small number) of times in the RF. For example, the left most peak of f_A in figure 3a is composed of variables that were only used once in the RF. There will often be another peak (now at $-\infty$) of variables that were not used at any time in the RF. Building on this observation we:

- count the number of times that each variable is used – this is C
- progressively threshold C giving $f_c(z) = f(z)|C > c$
- for $f_c(z)$, we can fit a Skew-Normal distribution (Azzalini, 2022) and apply Hartigan's diptest for unimodality, or a goodness-of-fit test such as the Cramer-von Mises test. However none of these procedures selects a c value giving a satisfactory fit (see supplementary file 2 for some plots.) Our current procedure is
  - for the distribution $f_c(z)$, fit a Skew-Normal up to the qth quantile, $S_q(z)$
  - calculate the L_∞ norm $d_c = \max_z |f_c(z) - S_q(z)|$
- plot d_c versus c and pick the minimum value $c^\star$. Call the corresponding distribution $f_{c^\star}(z)$
- we shift $f_{c^\star}(z)$ along the z positive axis so the smallest value is 0. See figure 3b for the selected distribution for this data set.

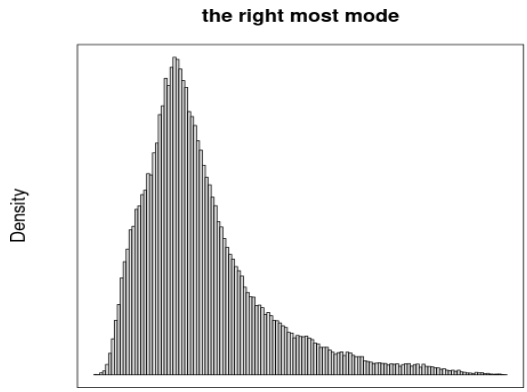
2.1.2 task 2

Starting with the histogram in figure 3b, we note that in this task our data is the histogram (midpoints of cells and counts). We:

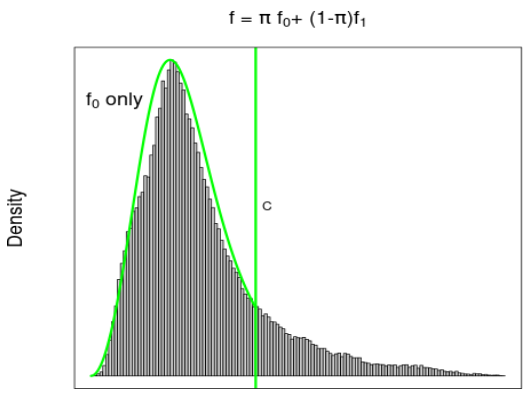
- fit a spline to the observed bin counts. This can be done using standard Poisson generalized linear modelling software. This is f, shown in green in figure 3d
- by assumption, there is a point q such that to the left of q $f_B \sim f_0(z)$. That is, there is a q such that there are only null values to the left of q. We determine q using a change point method related to penalized model selection (Gauran et al., 2018). See 3c for the estimated value of c
- estimate p_0
- fit f_0 using only the data $< q$. We use the formulation of Ashour and Abdel-hameed (2010) to fit a skew-Normal using only the truncated range. We fit the truncated skew-Normal by non-linear least squares (Elzhov et al., 2022).
- using f, p_0, f_0 we estimate the local $\text{fdr}(x) = \dfrac{p_0 f_0(x)}{f(x)}$. See figure 4 for the output of the R program RFlocalfdr showing the fdr curve for this example.



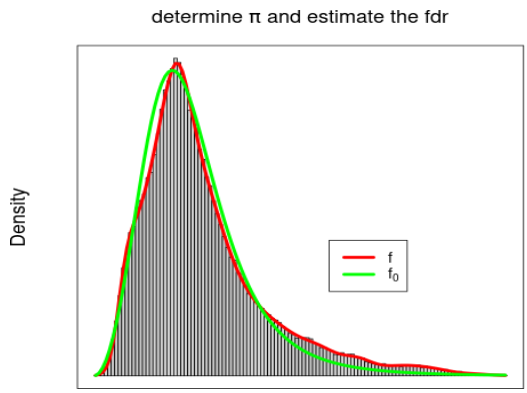
(a) The density of the log importances is multi-modal. The density $p_0 f_0(z) + p_1 f_1(z)$ is indicated in red.



(b) If we take only variables that are used greater than 30 times in the forest, we isolate the desired distribution, which we can shift along the positive axis.



(c) There is a value q such that to the left of q, f only depends on f_0.



(d) f (in red) and f_0 (in green).

Figure 3: The steps in estimating the local fdr for RF MDI importance values.

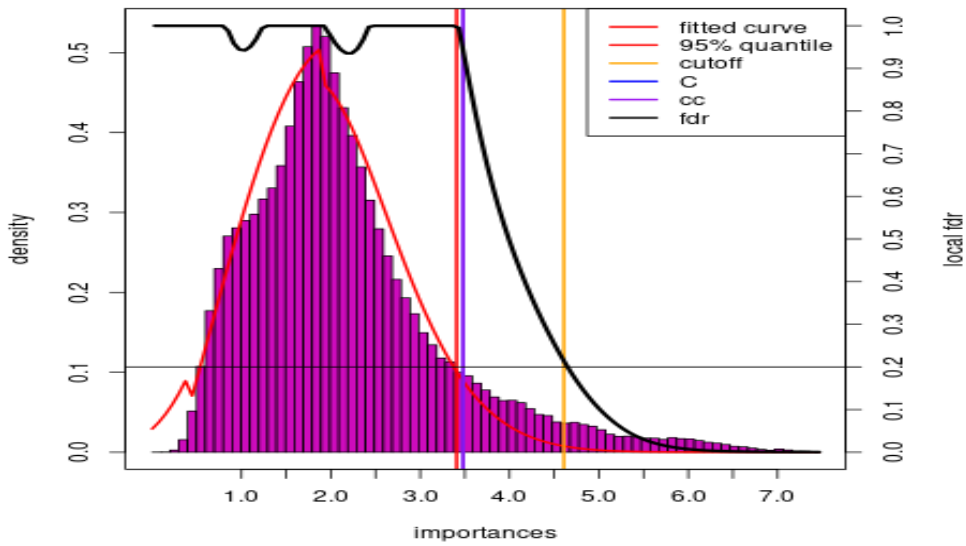


Figure 4: A plot produced by the code RFlocalfdr. The fdr curve is shown in black.

3 Results

3.1 Demonstration on a synthetic data set

We compare RFlocalfdr against the four published approaches, PIMP (Altmann et al., 2010), Bortua (Kursa and Rudnicki, 2010), RFE and AIR (Nembrini et al., 2018), on a simulated dataset with a strong correlation structure. This dataset consist of "bands", with "blocks" of $\{1, 2, 4, 8, 16, 32, 64\}$ of identical variables (see figure 5). The variables are $\in \{0, 1, 2\}$, a common encoding for genomic data where the numbers represent the number of copies of the minor allele. Only band 1 is used to calculate the y vector and y is 1 if any of $X[, c(1, 2, 4, 8, 16, 32, 64)]$ is non-zero, so y is unbalanced, containing more 1's than 0's. We have 50 bands and 300 observations so X is 300×6350 with 127 non-null variables. We fit a standard RF to this dataset and record the resulting MDI importance score.

As shown in figure 6, selecting variables above a single importance threshold can not achieve a perfect separation of the true positive variables from band one (red) and the false positives from the other bands. This issue is exacerbated by blocks with fewer variables having higher importance scores. This is because the importance is "smeared" over the correlated (identical in this case) variables. The effects of correlation on MDI scores is further discussed in Supplementary file 1.

It is hence necessary to employ a statistical approach for true positive selection and table 1 shows the results when applying the four feature selection methods. RFlocalfdr has the second highest AUC after AIR (0.73 vs 0.98), but unlike AIR it also has a high Sensitivity (0.99). Both Boruta and RFlocalfdr have a lower AUC but perform comparably on Sensitivity to RFlocalfdr by selecting only 2 and 22 true positives from the 128 true variables.

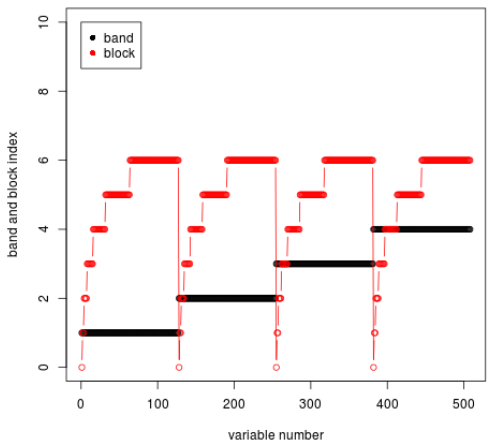


Figure 5: The data is structured into bands and blocks. The color and the y-axis show which band/block each variable belongs to, not the variables value. Each "band" contains "blocks" of sizes 1, 2, 4, 8, 16, 32, and 64. Each block consists of correlated (identical variables), where each variable is $\in \{0, 1, 2\}$. 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.

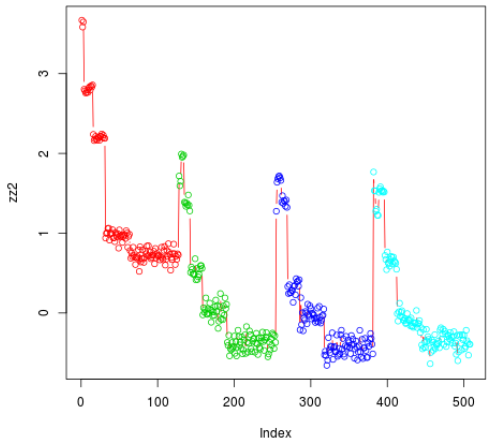


Figure 6: The log importance scores for the simulation, arranged by variable number and colored by band. Only the red variables are non-null, so we can see that it is not possible to threshold the importance scores so as to only recover the non-null variables.

Method	true positives	false positives	Sensitivity	AUC
RFlocalfdr	59	36	**0.99**	0.73
Boruta	2	2	**0.99**	0.50
RFE	22	2	**0.99**	0.58
AIR	176	273	0.96	**0.98**
PIMP	39	556	0.91	0.61

Table 1: Performance of variable selection for the simulated example by several methods. The best outcomes for each category are in bold face.

3.2 Feature selection on the Thousand Genomes Dataset

To demonstrate the applicability of RFlocalfdr on a real-world dataset, we applied it to the 1000 Genome Project phase 3 dataset (Auton et al., 2015). We obtained the 1000 Genomes Project data as VCF files from their FTP site. Each VCF file contains the genotypes for each of the Single Nucleotide Polymorphisms (SNPs) for every individual. There are 2,504 individuals and we limit the analysis to the 1 Million SNPs on chromosome 22. We train a RF to predict the ethnicity of each individual. This is the data set used to illustrate the method in section 2.1 and produce figure 3. A script detailing the analysis can be found in Supplementary file 1.

Using RFlocalfdr we can select 6,335 SNPs that are significantly associated with ethnicity at a FDR of 0.2. This removed 87% of the unassociated variables and results in the density shown in figure 3b.

The Boruta algorithm (Kursa and Rudnicki, 2010) returned a similar number (6,773) of significant SNPs with 82% overlap to RFlocalfdr (see figure 7). Unlike RFlocalfdr, which offers a continuous p-value scale for the adjustment of FDR, Boruta offers only two levels: at the more conservative "Confirmed" option it returns 1,443 SNPs all of which are included in our 6,335 list. The "Confirmed" SNPs have RFlocalfdr p-values that cluster at the low end. A one-way ANOVA shows the association between RFlocalfdr p-values and Boruta categories to be significant (p-value $< 2 \times 10^{-16}$).

As in the simulated example, AIR (Nembrini et al., 2018) selects an order of magnitude more significant SNPs (61,092) than RFlocalfdr and Boruta, potentially having a high sensitivity. However, with the p-value of all 61,092 SNPs receiving a score 0, AIR does not offer any option for tuning the specificity and the resulting false positives rate may be substantial (see Figure 7).

RFE returned 59 nested sets and selects the best set using the smallest prediction error. RFE appears to have focused on specificity and only returns 12 SNPs, which were also found by all other methods (see Figure 7).

We next evaluated the runtime. Since RFlocalfdr does not require multiple passes of the RF to be conducted for the statistical analysis, the runtime is essentially the runtime for the RF model. This can be run in parallel so the speed will depend on the number of cores. Using the same processing parameters, we found that RFlocalfdr was 25 times faster than the next best performing method, AIR, 57 times faster than RFE and 100 times faster than Boruta.

Method	SNPs returned	run time
RFlocalfdr	6,335	1
Boruta	6,773	100
RFE	12	57
AIR	61,092	25
PIMP	787502	100

Table 2: The number of SNPs returned by each methods and the run time. Run times are expressed in multiples of the run time of a single ranger (Wright and Ziegler, 2017) fit for the given hardware configuration. The runtime hence describes the number of "refits" that each method requires. The processing time outside of the RF fit is negligible, by comparison, in all cases.

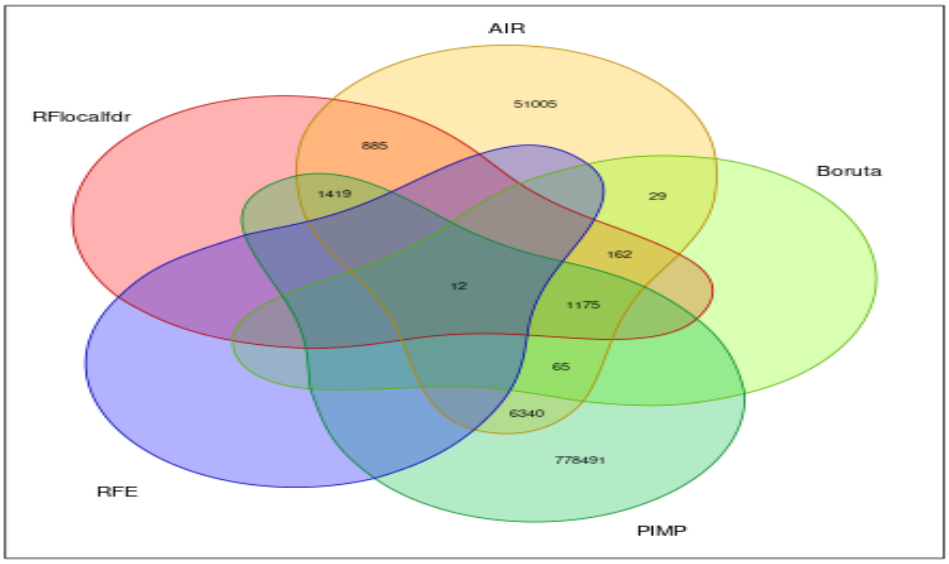


Figure 7: A Venn diagram of the overlaps in SNP for RFlocalfdr, Boruta, AIR, RFE and PIMP. AIR and PIMP are very much the outliers, with more than 10^4 unique SNPs (not found by the other methods) for each method.

3.3 Scaling to 10^6 datapoints

We consider one of the data sets presented in Bayat et al. (2020). This is the 64KX data set with 10,000 samples and 6 Million variables. It is based on the 1000 Genomes Project data but uses a synthetic y value. The data set is one of many considered in Bayat et al. (2020) and is used to explore the fitting of RF models on larger data.

Instead of the previously used ranger(Wright and Ziegler, 2017), we used VariantSpark (Bayat et al., 2020) to train the RF for this dataset. This was because ranger could not fit a model for this data

size on the hardware used for the trails in (Bayat et al., 2020)[1]. We also tested ReForest (a Spark based implementation of RF, see Lulli et al. (2017)), which was able to fit the model in about 15 hours, compared to VariantSpark, which took 5 hours.

Using the Gini importances and counts reported by VariantSpark, we can select associated genomic variants with a specified FDR using RFlocalfdr. RFlocalfdr selected 38 variants as significant, including the 5 variables used to generate the simulated y value, resulting in a 100% success rate with a 13% false discovery rate.

Given the runtimes reported in table 2, showing the comparitor methods taking between 25-100 times the RF run-time, fitting this dataset would require between 5 to 62 days to complete. Hence, only RFlocalfdr is capable of selecting variants associated with a trait in a feasible time on whole-genome data sets.

4 Conclusion

We present a method for calculating the significance threshold of RF impurity importance for detecting label-associated features using an empirical Bayes approach. Its accuracy is comparable to other published techniques, with the following advantages:

- **real-time**: it requires only one fit of the RF model so is computationally fast, particularly when compared to recursive feature elimination or permutation methods;
- **method-agnostic**: RFlocalfdr can be applied to any RF implementation that returns a MDI variable importance and counts of the variables used
- **sensitivity tuning**: it provides a range of p-values, which allows tailored sensitivity and specificity selections.

Specifically, RFlocalfdr offers the option to adjust the sensitivity/specificity trade-off, by adjusting the cutoff p-value of the selected SNPS. Of the other methods tested here, only RFE offers similar capability by selecting other nested sets with a larger or smaller number of elements. Boruta and AIR do not offer this flexibility as there is no criteria for subsetting the set of selected SNPs. Especially AIR stands out by reporting a vastly larger number of associated SNPS. Whether this is desirable or not would depend on the subsequent analysis and its required false discovery tolerance.

We have successfully processed the data from the example in section 3.3 using ranger (Wright and Ziegler, 2017), several hundred cores, and several Terra Byte of RAM. However, as data set sizes increase an implementation runtime that grows quickly with the data eventually becomes infeasible Bayat et al. (2020). This is particularly true for methods that require multiple refits. RFlocalfdr can be applied to any RF implementation that returns a MDI variable importance and counts of the variables used, such as VariantSpark.

We note that Efron's local fdr method is justified on its clearly better modelling of the histogram of statistics not on it computational speed. We have made no claim about the utility of RFlocalfdr, merely noting that it is comparable to other methods and computationally much faster. However, while the variable importances offer a number challenges that are not present in the case/control comparison that local fdr was designed for, it is conceivable that it may have better properties than other methods of variable selection.

More experience in its application would be required to ascertain this. Early applications of the method (Lundberg et al., 2022) have lead to promising results in that using RFlocalfdr, in tandem with other methods, (Bayat et al., 2021, 2020) captured more phenotypic variance in Alzheimers disease than standard GWAS analysis methods. In addition, there is evidence that these findings replicate in an independent data set.

[1] AWS EC2 (Elastic Compute Cloud) and EMR (Elastic Map Reduce). Eight compute nodes of class `r4.8xlarge` were used giving 254 CPUs and 2013 GB of RAM.

It is our expectation that RFlocalfdr with its direct and real-time capability of detecting trait-associated SNPs will greatly assist the analysis of GWAS data.

Software and data access

R implementation

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

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)

Supplementary information

- supplementary_file_1.Rmd
- supplementary_file_2.Rmd

References

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

Ashour, S. K. and Abdel-hameed, M. A. (2010). Approximate skew normal distribution. *Journal of Advanced Research*, 1(4):341–350.

Auton, A., Abecasis, G. R., Altshuler, D. M., and 298 other authors (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Azzalini, A. (2022). *The R package* `sn`*: The skew-normal and related distributions such as the skew-t and the SUN (version 2.0.2).* Università degli Studi di Padova, Italia. Home page: `http://azzalini.stat.unipd.it/SN/`.

Bayat, A., Hosking, B., Jain, Y., Hosking, C., Kodikara, M., Reti, D., Twine, N. A., and Bauer, D. C. (2021). Fast and accurate exhaustive higher-order epistasis search with BitEpi. *Scientific Reports*, 11(1):15923.

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.

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.

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.

Elzhov, T. V., Mullen, K. M., Spiess, A.-N., and Bolker, B. (2022). *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds*. R package version 1.2-2.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

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.

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].

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.

Janitza, S. and Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLOS ONE*, 13(8):e0201904.

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).

Lulli, A., Oneto, L., and Anguita, D. (2017). ReForeSt: Random Forests in Apache Spark. In Lintas, A., Rovetta, S., Verschure, P. F., and Villa, A. E., editors, *Artificial Neural Networks and Machine Learning – ICANN 2017*, Lecture Notes in Computer Science, pages 331–339, Cham. Springer International Publishing.

Lundberg, M., Sng, L. M., Szul, P., Dunne, R., Bayat, A., Burnham, S., Cuellar-Partida, G., Bauer, D. C., and Twine, N. A. (2022). Variantspark, a cloud-based random forest gwas platform, identifies novel loci and epistasis in alzheimer's disease. Submitted to Nature Aging.

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.

Nembrini, S., König, I. R., Wright, M. N., and Valencia, A. (2018). The revival of the Gini importance? *Bioinformatics*.

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.

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.

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

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.