1. Is selection of markers different than expected??
   * Close-ness of two markers in the trees (>expected?)
   * Selection Asymmetry
   * Paired SF
   * Maxsubtree (randomForestSRC)

‘method="maxsubtree"’ This invokes a maximal subtree analysis. In this case, a matrix is returned where entries [i][i]are the normalized minimal depth of variable [i] relative to the root node (normalized wrt thesize of the tree) and entries [i][j] indicate the normalized minimal depth of a variable [j] wrt the maximal subtree for variable [i] (normalized wrt the size of [i]’s maximal subtree). Smaller[i][i] entries indicate predictive variables. Small [i][j] entries having small [i][i] entries are asign of an interaction between variable i and j (note: the user should scan rows, not columns, for small entries). See Ishwaran et al. (2010, 2011) for more details.

1. Are importance/decrease in variance values dependant one each other?
   * Breiman’s suggestion: <https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#inter>
     + Rank gini/variance decrease of predictors for each tree
     + For each pair, compute absolute difference of ranks and average over all trees 🡪 Dobserved (representing average “distance” between predictors)
     + Compute the value for this difference that would be expected when the pair is independent (unclear how) 🡪 Dexpected
     + X = Dexpected-Dobserved
     + X>0 🡪 correlated. X=0 🡪 independent. X<0 🡪 interacting
   * My idea: does decrease in impurity depend on prior split?
     + Compare importances of A after split on B with importances of A before/without split on B
     + Compare importances of B after split on A with importances of B before/without split on A
   * Split Asymmetry
   * Reduction in variance decrease for the second split asymmetry
2. Conditional permutation
   * ~~PPI (pairwise permutation importance) from Wright~~
   * ~~JVIMP (randomForestSRC)~~ = Maxsubtree
   * My idea
     + Same idea in Ishwaran 2007 with full theoretical explnation: <https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-1/issue-none/Variable-importance-in-binary-regression-trees-and-forests/10.1214/07-EJS039.full>
     + randomForestSRC, ‘method="vimp"’: This invokes a joint-VIMP approach. Two variables are paired and their paired VIMP calcu-lated (refered to as ’Paired’ importance). The VIMP for each separate variable is also calcu-lated. The sum of these two values is refered to as ’Additive’ importance. A large positive ornegative difference between ’Paired’ and ’Additive’ indicates an association worth pursuing ifthe univariate VIMP for each of the paired-variables is reasonably large. See Ishwaran (2007)for more details.
     + Permute A 🡪 performance decrease D(A)
     + Permute B 🡪 performance decrease D(B)
     + Permute both A&B 🡪 performance decrease D(AB)observed
     + D(AB)expected = D(A)+D(B) ~~or D(A)\*D(B)~~
     + X = - D(AB)observed - D(AB)expected
     + X=0 🡪 independent. X<0 🡪interaction. X>0 🡪 correlated or interaction.
     + Care: co-permute correlated markers.

Others

* MCC (maximal conditional chi-squared): <https://pubmed.ncbi.nlm.nih.gov/20130032/>
  + “maximal conditional chi-square (MCC) as a measure of association between a SNP and the trait conditional on other SNPs”
  + “To overcome the problems stated above, we introduce an alternative importance score using maximal conditional chi-square (MCC) statistic to assess the conditional significance of SNPs in GWAS. For example, with two SNPs (A and B), if SNP B confounds the effect of SNP A on the trait, failure to control the effects of SNP B may lead to inefficient tests for SNP A. Like the Mantel–Haenszel test, a test by stratifying by SNP B is an effective approach to adjusting for the confounding effect of SNP B.”
  + For categorical predictor and outcome

|  |  |  |  |
| --- | --- | --- | --- |
|  | Diseased | healthy |  |
| Allele A | 50 | 20 |  |
| Allele a | 3 | 70 |  |
|  |  |  |  |

* + Get only Chi-squared statistics for splits of B following A and take its maximum.

Simulations:

* Data types:
  + Categorical predictors and categorical outcome
  + Categorical predictors and continuous outcome
  + Continuous predictors and categorical outcome
  + Continuous predictors and continuous outcome
  + Mix of predictors and categorical outcome
  + Mix of predictors and continuous outcome
* Interactions:
  + AND
  + XOR
* One interaction, 1-5 other effects (i.e. noise)

*Real Datasets:*

* *well known “air pollution” data set [15]. The variables are daily readings of various air quality values measured from May 1,1973 to September 30, 1973 in the New York metropolitan area.Solar radiationin Langleys (Solar), average wind speed in miles per hour (Wind), maximumdaily temperature in Fahrenheit (Temp), month of the year (Month) and day ofmonth (Day) were all recorded. The outcome is the mean ozone value (Ozone),which was transformed by taking its cube-root (Ishwaran 2007)*