Supplementary Material to “On what to permute in test-based approaches for variable importance measures in Random Forests”

December 7, 2018

Gaussianization of the AIR importances

The results in the following section (Fig. 7-15) show that the AIR importances under the null have zero mean and zero skewness, but with increasing variance and a different kind of kurtosis. Using usual parametric distributions like the normal, logistic or t seems non reliable. But we propose the use of a rank-based inverse normal transformation. We use a transformation method that allows the VIM to be “gaussianized”, i.e. transformed to a standard normal distribution. Assume the ranks corresponding to the VIMs are denoted by $R^1 < R^2 < \ldots < R^N$. Divide each $R^k$ by $N + 1$. This creates a distribution of scores lying between 0 and 1. If these scores are treated as percentile points under a normal distribution, it is possible to work backwards and determine the $z$ scores corresponding to the amount of area under the normal curve lying below these points. Here we use the common rank-based INT transformation proposed by Van der Waerden [1952] which entails creating a modified rank variable and then computing a new transformed value

$$AIR_{z}^k = \Phi^{-1} \left( \frac{R^k}{N + 1} \right)$$

where $R^k$ is the rank of $AIR^k$ and $N$ is the sample size and $\Phi^{-1}$ denotes the standard normal quantile (or probit) function. This transformation can be
applied to the null distribution and inference is obtained interpolating (or extrapolating) the AIR value of variable $X_i$ obtained on the original data using the standard pnorm function.

Figure S5 shows the Concordance Correlation Coefficient (CCC) between nonparametric $p$-values obtained with 4,999 iterations and $p$-values obtained from gaussianizing the AIR VIM for three datasets. Already after 100 iteration, CCC values are extremely high and close to one.

**Alternative Hypothesis**

Figures S1, S2 and S3 show VIMs and regression coefficient for the univariate regression as explained. Blue distribution is obtained permuting the outcome, while the red distribution is obtained from a wild bootstrap with a Rademacher distribution (Fig. S1), from a wild bootstrap with a Mammen distribution (Fig S2), and from a wild bootstrap with a Normal distribution (S3).

**xForest**

Instead of refitting a new forest for every permutation as in Hapfelmeier and Ulm [2013], one could extend the original dataset with $k$ row-wise permutations of and compute the AIR VIM and its corresponding null values inside the same forest. $k$ was set to 100 as in the other permutation-based methods.

Figure S6 shows that a certain amount of trees may be needed for the null distributions to be estimated with precision. This means that depending on the data structure a given number of trees is needed for the null distribution to have a large number of unique values, with small datasets needing a smaller number of trees. For small datasets convergence is obtained after 25 to 100 trees. For comparison and to avoid a suboptimal choice we used 5000 trees for all datasets, which is still 1/2 times the number of trees required by the permutation-based methods.
Figure S1: Euphydryas editha data: VIMs and regression coefficient for the univariate regression. Blue distribution is obtained permuting the outcome, while the red distribution is obtained from a wild bootstrap with a Rademacher distribution. The straight red line indicates the values obtained from the original data.

Distribution of the AIR VIMs

Mean variance kurtosis and skewness of the AIR importance for the null cases A-B-C explained in Nembrini et al. [2018], for regression, classification and survival.

Null Case A: Increasing minor allele frequency

Ten single nucleotide polymorphisms $X_1, \ldots, X_{10}$ with increasing minor allele frequency $MAF = \{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$ were simulated.

Null Case B: Increasing numbers of categories

Ten covariates $X_1, \ldots, X_{10}$ with increasing numbers of categories were gen-
Figure S2: Euphydryas editha data: VIMs and regression coefficient for the univariate regression. Blue distribution is obtained permuting the outcome, while the red distribution is obtained from a wild bootstrap with a Mammen distribution. The straight red line indicates the values obtained from the original data.
Figure S3: Euphydryas editha data: VIMs and regression coefficient for the univariate regression. Blue distribution is obtained permuting the outcome, while the red distribution is obtained from a wild bootstrap with a Normal distribution. The straight red line indicates the values obtained from the original data.
Figure S4: AIR VIM, Permutation VIM I and TDNI VIM density plots for uninformative variables in Scenario B. The blue density represents the VIM when the outcome is permuted, while the red density represents the VIM when the outcome is left unchanged. The red density for permutation VIM II is the permutation importance shuffling $X_i$ instead of the outcome $Y$. 
Figure S5: Concordance Correlation Coefficient between nonparametric $p$-values obtained with 4,999 iterations and $p$-values obtained from *gaussianizing* the AIR VIM for three datasets.
Figure S6: Percentage of unique values for the null distributions on the AIR VIM inside the xForest for all size datasets.
erated from a uniform distribution in a similar fashion as in Altmann et al. [2010], i.e., \( k = \{2, 3, 4, 5, 6, 7, 8, 10, 20, 30\} \), where \( k \) is the number of categories.

**Null Case C: Mixed type covariates**

Covariates were simulated according to the following scheme. \( B_{0.05}, B_{0.1}, B_{0.2}, B_{0.5} \) are binomial with increasing probabilities equal to \( p = \{0.05, 0.1, 0.2, 0.5\} \). \( O_5, O_{10} \) are ordered factors with 5 and 10 categories, \( N_5, N_8, N_{10} \) are nominal factors with 5, 8, and 10 categories, and \( C \) is a continuous covariate generated from a standard normal distribution. Nominal and ordered factors are generated so that each category has the same frequency.
Figure 7: Survival - Minor Allele Frequency. Mean, standard deviation, skewness and kurtosis.
Figure 8: Survival - Increasing Number of categories. Mean, standard deviation, skewness and kurtosis.
Figure 9: Survival - Mixed Type. Mean, standard deviation, skewness and kurtosis.
Figure 10: Classification - Minor Allele Frequency, standard deviation, skewness and kurtosis.
Figure 11: Classification - Increasing number of categories standard deviation, skewness and kurtosis.
Figure 12: Classification - Mixed Type Covariates, standard deviation, skewness and kurtosis.
Figure 13: Regression - Minor Allele Frequency, standard deviation, skewness and kurtosis.
Figure 14: Regression Increasing number of categories: standard deviation, skewness and kurtosis.
Figure 15: Regression - Mixed type covariates: standard deviation, skewness and kurtosis.
References


