Handling Correlations in Random Forests: which Impacts on Variable Importance and Model Interpretability?

Marie Chavent[1], Jerome Lacaille[2], Alex Mourer[1,2,3], Madalina Olteanu[4]

1- Inria bdx sud ouest, équipe ASTRAL - IMB, UMR CNRS 5251, U. bdx - France

2- Safran Aircraft Engines - Datalab - Villaroche - France

3- SAMM, EA 4543 - Université Pantheon Sorbonne - France

4. CEREMADE, UMR 7534 - Université Paris Dauphine PSL - France

Abstract. The present manuscript tackles the issues of model interpretability and variable importance in random forests, in the presence of correlated input variables. Variable importance criteria based on random permutations are known to be sensitive when input variables are correlated, and may lead for instance to unreliability in the importance ranking. In order to overcome some of the problems raised by correlation, an original variable importance measure is introduced. The proposed measure builds upon an algorithm which clusters the input variables based on their correlations, and summarises each such cluster by a synthetic variable. The effectiveness of the proposed criterion is illustrated through simulations in a regression context, and compared with several existing variable importance measures.

1 Introduction

Variable importance and model interpretability are essential for conducting comprehensive data analysis. Random Forests (RF) are an attractive technique for supervised learning because of their good empirical performances, yet they are often considered as black-boxes because of their lack of interpretability. Usually, for RF, interpretability may be assessed by quantifying variable importance. From a statistical point of view, computing variable importance aims at two goals: i) find the contribution of each input variable to the prediction, and ii) measure the dependency between the input variables and the output one. However, one should note that these objectives might be contradictory when variables are correlated. In fact, if two variables are strongly correlated, one of them may be discarded without degrading the performance of the model, while still being related to the output variable.

Two criteria are currently the benchmark for measuring variable importance for RF: the Mean Decreased Accuray (MDA) [1] which quantifies the decrease of accuracy when a given input variable is permuted, and the Mean Decreased Impurity (MDI) [1] which quantifies the decrease of impurity over all nodes in the forest split according to a given input variable. Both criteria are known to raise some issues in practice: the MDI usually overestimates the importance of non-discriminant variables, while the MDA may not detect important variables when dependency is present [2]. In particular, it has been proven by [2]

that, in the case of independent input variables, the MDA tends to estimate the proportion of output variance explained by each input variable, whereas in the case of dependent inputs, the MDA is ill-defined because of what is called the *sampling bias*. The latter has been studied empirically by [3], who argue that permutation-based methods tend to over-estimate the importance of correlated input variables.

Several alternatives have been proposed in the literature in order to overcome some of the known flaws of MDA. [4] introduced a conditional version of the MDA (CMDA), based on a conditional permutation scheme, where the considered input variable is permuted only within groups of observations, so as to preserve the correlation structure and make the criterion more reliable in the presence of correlated variables. Recently, [2] proposed the Sobol-MDA criterion, along with a theoretical analysis. The Sobol-MDA method is based on the total Sobol index and computes the degraded accuracy by ignoring the nodes containing the considered **the** input variable in each tree. Unlike the MDA, in the case of dependent input variables, the Sobol-MDA aims to estimate the proportion of output variance explained by each input variable.

Nevertheless, relying on any of these measures alone may be misleading. If, for example, two input variables are identical, they would be expected to explain the same amount of the output variance. However, this is not guaranteed in practice since each tree will randomly chose one of them. Thus, the amount of output variance explained by each of them will be different according to their being included in the model or not. Not only all of the above methods are hence subject to what may be called *selection bias*, but one should note also that they are mostly designed to meet objective i) and only aim at assessing the contribution of the input variables to the prediction.

This manuscript focuses mainly on objective ii) above, and on discovering relationships between the input variables and output one. It introduces a new criterion for assessing variable importance, based on the MDA computed on a set of synthetic variables and using RF. The synthetic variables are defined as summaries of the input variables, using the correlation structure of the input data and a clustering procedure, as proposed by [5]. The rest of the paper is organised as follows: Section 2 contains the details of the proposed methodology and enumerates its main steps, while Section 3 illustrates the proposed criterion and compares it with the existing literature, for a simulated data set.

2 Methodology

The proposed criterion is based on the following steps: first, the input variables are clustered according to their correlation structure; second, each cluster of input variables is summarised by a synthetic variable and more specifically by the first principal component computed within the cluster; third, a RF algorithm is trained on the synthetic variables and for each of them the associate MDA is computed; fourth, the importance of each of the original input variables is assessed using the MDA of the synthetic variables and the correlations between

synthetic and original variables. The first three steps were introduced in [5], while the last one represents the original contribution of the manuscript. Each of these four steps will be briefly described in the following paragraphs.

Variable clustering and computation of synthetic variables Let the data set $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq n}$, consisting of n observations with the input matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ of p variables, $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{x}^j \in \mathbb{R}^n$ indicates the j-th variable of $\mathbf{X}$.

Input variables are clustered together using a Hierarchical Clustering Analysis (HCA). The criterion used in the dendrogram construction is based on a homogeneity measure and aims at reducing information redundancy. Two clusters C_k and $C_{k'}$ are compared through a dissimilarity defined as

$$d(C_k, C_{k'}) = H(C_k) + H(C_{k'}) - H(C_k \cup C_{k'}),$$

where

$$H(C_k) = \sum_{\mathbf{x}^j \in C_k} r^2_{\mathbf{f}^k, \mathbf{x}^j} \text{ and } \mathbf{f}^k = \underset{\mathbf{u} \in \mathbb{R}^n}{argmax} \sum_{\mathbf{x}^j \in C_k} r^2_{\mathbf{u}, \mathbf{x}^j}.$$

In the equation above, $r^2_{\mathbf{u},\mathbf{x}^j}$ stands for the squared Pearson correlation between some synthetic feature $\mathbf{u}$ and the input variable $\mathbf{x}^j$, and one may notice, as shown in [5], that $\mathbf{f}^k$ is actually the first principal component associated to cluster C_k.

Random Forests on Synthetic Variables The above procedure establishes a hierarchy of dependency between input variables in terms of redundant information, but does not provide the *optimal* number of clusters, and implicitly the *optimal* number of synthetic variables $\mathbf{f}^k$ to use hereafter. In order to establish it, one may train a RF algorithm for each level of the dendrogram, hence for a number of clusters $K = 1, ..., p$. For each K and for each associate input variable partitioning, one will train a RF on the corresponding synthetic variables and compute the resulting Out-Of-Bag (OOB) error. The optimal number of clusters of variables K^* is chosen as the one leading to the minimum OOB error rate. [5] have shown that clustering variables not only enhances interpretability, but can also improve the predictive performance of the algorithm. Once the optimal number of synthetic features K^* has been selected, one may select the RF trained on those and compute the MDA associated to $(\mathbf{f}^k)_{k=1,...,K^*}$.

Synthetic MDA (SMDA) Variable clustering combined with RF trained on synthetic variables provides an appealing solution to avoid selection and sampling bias, at least to some extent. Indeed, homogeneity-based clustering should result into small correlations between synthetic variables, and may prevent both types of bias. Nevertheless, the above procedure allows to compute MDA measures for synthetic variables only and does not assess the importance of the original ones. To fill this gap, a new criterion for measuring the importance of the original variables through the synthetic ones is defined. The Synthetic-MDA (SMDA) conditionally to a variable partitioning into $K^\star$ clusters is defined as:

$$\text{SMDA}^{K^*}(\mathbf{x}^j) = \text{MDA}^{K^*}(\mathbf{f}^k) \times r^2_{\mathbf{f}^k, \mathbf{x}^j},$$

where $\mathbf{x}^j$ is a feature clustered in cluster C_k, and $\mathbf{f}^k$ is the synthetic variable summarising the same cluster, for $k = 1, ..., K^*$ and $j = 1, ..., p$. $\text{MDA}^{K^*}(\mathbf{f}^k)$ is computed as the average decrease of accuracy in the OOB sample, before and after permutation of the synthetic feature $\mathbf{f}^k$ over all trees:

$$\text{MDA}^{K^*}(\mathbf{f}^k) = \frac{1}{T}\sum_{t=1}^{T}\Big[R_m(D_{oob,t}^{\pi_{k,t}}, t) - R_m(D_{oob,t}, t)\Big],$$

with T the number of tree of the RF m, $R_m(D_{oob,t}, t)$ is the risk of the t-th tree computed on its OOB sample and $D_{oob,t}^{\pi_{k,t}}$ is the OOB sample of the t-th tree where the k-th variable has been permuted, i.e. OOB samples and permutations are different between trees. The risk can be defined differently depending on the context, and in the simulations below we consider the regression context with the mean squared error as associated risk. The clustering of variables reduces but does not completely remove the correlation in the data since synthetic variables are not necessarily orthogonal. Hence, there is a trade-off between interpretability and performance that can be adjusted via the number of clusters used to fit the RF m. Nonetheless, thanks to the SMDA formulation, the relationship between the input variables and the output variable should be better assessed since the importance is not diluted by the correlations, which helps to meet objective ii), unlike other measures which give a lower contribution to the correlated variables due to the selection bias.

3 Experimental Results

For the experimental section, the data was simulated in a regression context and using a linear model, as follows:

$$\mathbf{y} = \sum_{j=1}^{p_1}\beta_j \times \mathbf{x}^j + \frac{1}{2}\times\tilde{\mathbf{x}}^1 + \boldsymbol{\varepsilon}, \text{ where } \beta_j = \frac{p_1 - j + 1}{p_1}, j = 1, \ldots, p_1,$$

where $\mathbf{y} \in \mathbb{R}^n$, $\boldsymbol{\varepsilon}$ normally distributed with 0 mean and variance 0.5. The important variables are normally distributed s.t.:

1. $\mathbf{X}_{p_1} \in \mathbb{R}^{n\times p_1}$ are p_1 independent variables with 0 mean.
2. $\tilde{\mathbf{x}}^1 \in \tilde{\mathbf{X}}_{p_2}$, $\tilde{\mathbf{X}}_{p_2} \in \mathbb{R}^{n\times p_2}$ have 0 mean and pairwise correlation of 0.9.

Two additional set of noise variables normally distributed are added:

1. $\mathbf{Z}_{q_1} \in \mathbb{R}^{n\times q_1}$ are q_1 independent variables with 0 mean.
2. $\tilde{\mathbf{Z}}_{q_2} \in \mathbb{R}^{n\times q_2}$ have 0 mean and pairwise correlation of 0.9.

Hence, the input matrix $\mathbf{X}$ is composed of 4 groups of variables, independent of each other, such that $\mathbf{X} = [\mathbf{X}_{p_1}|\tilde{\mathbf{X}}_{p_2}|\mathbf{Z}_{q_1}|\tilde{\mathbf{Z}}_{q_2}]$, with $p = p_1 + p_2 + q_1 + q_2$.

Since a linear regression model without interaction is considered, the importance of a variable $\mathbf{x}^j$ in the *true* model may be defined via its correlation with

$\mathbf{y}$: $r^2_{\mathbf{y},\mathbf{x}^j}$. Indeed, since it is objective ii) that is being considered, the importance of a variable must not depend on the presence or the absence of other variables. (given our model without interaction). Note that the specific value of the coefficient $r^2_{\mathbf{y},\mathbf{x}^j}$ is not relevant, yet it is the ranking of variables it provides that is pertinent. Once the data has been simulated with samples of size $n = 500$, the RF in the proposed strategy were fitted using the `R`-package `ranger` [6] with default parameters and `T` $= 1000$.

In the first experiment, $p_1 = 3, p_2 = 3, q_1 = 2, q_2 = 0$. Table 1 compares the theoretical importance as defined by the squared correlation with the empirical variable importance as computed with the MDA, the CMDA, the Sobol-MDA and the SMDA over 100 independent samples of data. As one may see, the MDA underestimates the importance of the input variables correlated with $\tilde{\mathbf{x}}_1$. The CMDA fairly ranks the variables but the importance of the input variables correlated with $\tilde{\mathbf{x}}_1$ is almost 0, which may raise some issues if one seeks to discover relationships between input variables and $\mathbf{y}$. The Sobol-MDA estimates negative weights, in particular for $\tilde{\mathbf{x}}^1$, suggesting, wrongly, that the performance of the model could be improved by discarding $\tilde{\mathbf{x}}^1$. On the other hand, the SMDA provides a good estimate of the ranking, and the computed values appear as meaningful. For the SMDA procedure, over the 100 samples, the algorithm repeatedly selected $K^\star = 6$, and clustered the input variables in the $\tilde{\mathbf{X}}_{p_2}$ block together. Hence, one shortcoming of the proposed method is that all variables in this block have consequently the same importance.

Table 1: Results of simulations with $p_1 = 3, p_2 = 3, q_1 = 2, q_2 = 0$. Average importance over 100 samples is reported. Standard deviations are below 10^{-3}.

	$\mathbf{x}^1$	$\mathbf{x}^2$	$\mathbf{x}^3$	$\tilde{\mathbf{x}}^1$	$\tilde{\mathbf{x}}^2$	$\tilde{\mathbf{x}}^3$	$\mathbf{z}^1$	$\mathbf{z}^2$
MDA	0.690	0.269	0.063	0.128	0.061	0.074	0.000	-0.001
Sobol-MDA	0.482	0.179	0.027	-0.002	-0.014	-0.015	-0.011	-0.011
CMDA	1.076	0.423	0.095	0.011	0.000	0.001	-0.001	-0.001
SMDA	0.802	0.311	0.064	0.151	0.151	0.150	0.001	0.001
$r^2_{\mathbf{y},\mathbf{x}^j}$	0.56	0.24	0.06	0.14	0.11	0.11	0.00	0.00

In the second experiment, $p_1 = 10, q_1 = 25$, and p_2 and q_2 take different values as described in Table 2. The comparison of variable-importance measures has been extended here to MDI and MDICor. The latter is supposed to handle correlation, according to [6]. The theoretical importance was compared with the empirical importance as measured by the various criteria using the Spearman correlation. The percentage of important variables among the $p_1 + p_2$ firstly ranked input variables is also reported. The case $p_2 = 1, q_2 = 50$ should be the easiest one for all criteria, since all important variables are independent, the others being noise variables, and indeed most criteria are equivalent.

In the presence of correlated important variables and correlated noise ($p_2 = 50, q_2 = 50$), the MDA is negatively affected by the sampling bias which is stronger in the presence of many correlated variables in constrast to the MDI favourably biased because of the selection bias (and by definition it is not sampling-based). Overall, our method was significantly better. The number

of clusters chosen is almost equivalent in the three settings and is in average equal to 36 with a standard deviation of 5. In particular, depending on the simulations, seldom independent variables were put together and each group of correlated variables may have been separated into two clusters.

Table 2: Spearman correlation between the *true* and the estimated variable-importance, and percentage of important variables in the $p_1 + p_2$ firstly ranked input variables. $p_1 = 10, q_1 = 25$. Average values over 100 samples are reported. Standard deviations are provided in the brackets.

	$p_2 = 1; q_2 = 50$		$p_2 = 50; q_2 = 0$		$p_2 = 50; q_2 = 50$	
	Sp	%sel	Sp	%sel	Sp	%sel
MDICor	**0.41**(.089)	**86**(7.3)	0.64 (.085)	91(4.1)	0.68(.053)	87(6.1)
MDI	**0.42**(.045)	**87**(5.9)	0.13(.182)	61(3.2)	0.59(.072)	65(4.5)
MDA	**0.27**(.054)	**73**(5.7)	**0.75**(.026)	**98**(1.1)	0.69(.051)	88(4.1)
CMDA	0.21(.078)	**81**(6.4)	0.15(.152)	81(3.2)	0.16(.087)	43(5.2)
Sobol-MDA	**0.27**(.086)	**82**(7.1)	0.11(.101)	67(2.9)	0.25(.099)	31(4.5)
SMDA	**0.42**(.093)	**89**(6.8)	**0.77**(.031)	**98**(1.4)	**0.81**(.035)	**98**(0.9)

4 Conclusion

According to the results above, the performance of the proposed criterion is superior or similar with the existing measures for variable importance. Simulation results pointed out that handling correlations is important if one seeks to discover relationships between input variables and output variable. The proposed criterion may be generalised whether in terms of clustering method for the input variables (k-means), or in terms of algorithm for the supervised learning step. Furthermore, variable clustering can be done for mixed data [5] (mixture of categorical and numerical variables). A theoretical analysis of the sampling bias and several additional experiments have been also conducted, but are not presented here because of the reduced number of pages. The implementation of our method and the code to simulate data are publicly available[1].

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] C. Bénard, S. Da Veiga, and E. Scornet. Mda for random forests: inconsistency, and a practical solution via the sobol-mda. *arXiv:2102.13347*, 2021.

[3] G. Hooker and L. Mentch. Please stop permuting features: An explanation and alternatives. *arXiv:1905.03151*, 2019.

[4] C. Strobl, A-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):1–11, 2008.

[5] M. Chavent, R. Genuer, and J. Saracco. Combining clustering of variables and feature selection using random forests. *Communications in Statistics-Simulation and Computation*, 50(2):426–445, 2021.

[6] M. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv:1508.04409*, 2015.

[1] https://github.com/MourerAlex/SMDA