

Handling Correlations in Random Forests

Which Impacts on Variable Importance and Model Interpretability?

Marie Chavent¹, Jerome Lacaille², Alex Mourer^{1,2,3}, Madalina Olteanu⁴

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

Introduction

Motivations

- Context
 - High dimensional data.
 - Complex models.
 - Understand patterns of a phenomenon given variables.
- First Objectives
 - Explain an observed variable.
 - Discover relationships.
 - Accurate Variable Importance.
- Goal
 - Improve interpretability.
 - Variable importance in terms of link between the input variables and the output variable.

- Variable Importance (VI):
 - **Definition 1:** Find the contribution of each input variable to the prediction.
 - **Definition 2:** Measure the dependency between the input variables and the output one.
- Random Forest (RF):
 - Effective.
 - accurate model → accurate relation.
 - Easy + fast + no tuning.
 - Black Boxes.

- VI for RF:
 - Mean Decreased Accuracy (MDA); Breiman [2001].
 - Mean Decreased Impurity (MDI); Breiman [2001].
 - Conditional-MDA (CMDA); Strobl et al. [2008].
 - MDI-Corrected (MDICor); Wright and Ziegler [2015].
 - Sobol-MDA; B  nard et al. [2021].
- Goals:
 - Explain prediction → misleading.
- Problems:
 - Selection bias.

Methodology

Our solution in a nutshell



First four steps described in Chavent et al. [2021].

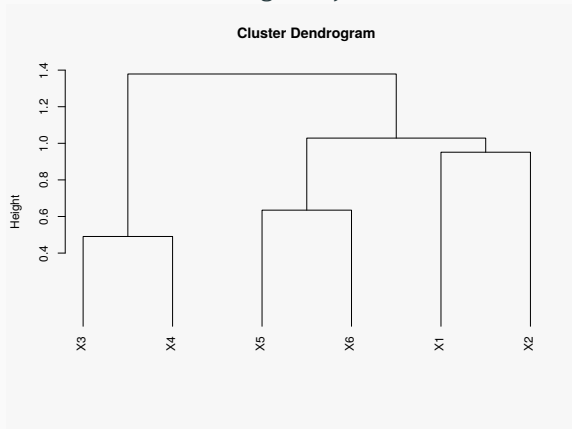
Cluster original variables



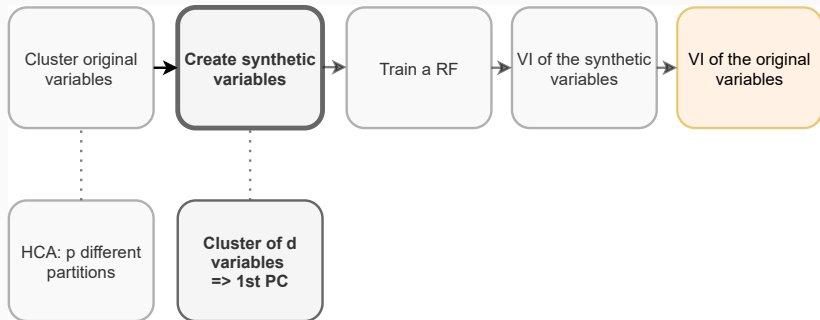
Cluster original variables

1. $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{1 \leq i \leq n}$.
2. $\mathbf{X} \in \mathbb{R}^{n \times p}$
3. $\mathbf{y} \in \mathbb{R}^n$.

Hierarchical Clustering Analysis (HCA):



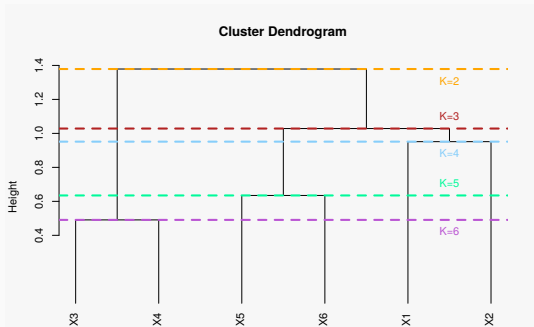
Create synthetic variables



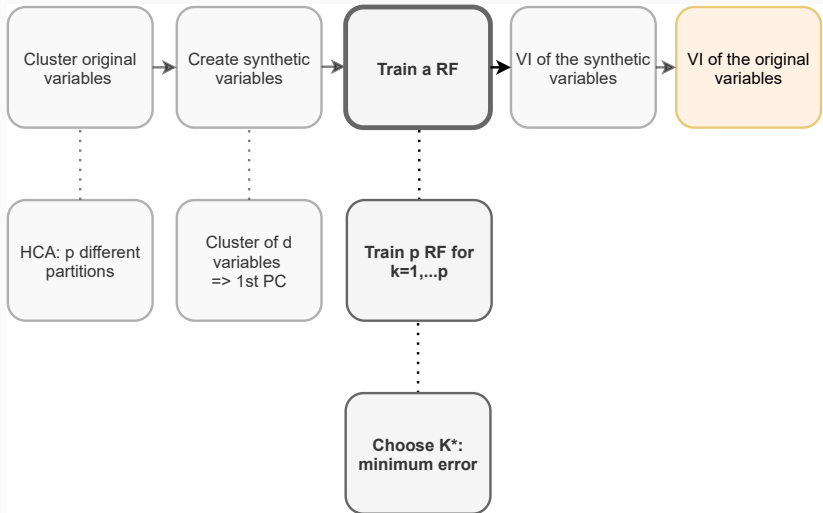
Create synthetic variables

$$\mathbf{f}^k = \underset{\mathbf{u} \in \mathbb{R}^n}{\operatorname{argmax}} \sum_{\mathbf{x}^i \in \mathcal{C}_k} \operatorname{cor}(\mathbf{u}, \mathbf{x}^i)^2.$$

- $\mathbf{x}^j \in \mathbb{R}^n$ the j -th variable of \mathbf{X} .
- $\mathbf{f}^k \in \mathbb{R}^n$ the k -th synthetic variable.
- \mathbf{f}^k is the 1st Principal Component (PC) of the variables in the cluster k noted \mathcal{C}_k .



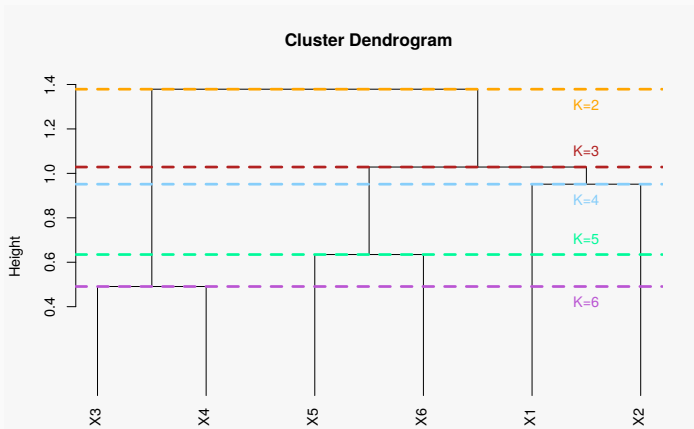
Train RF and choose K !



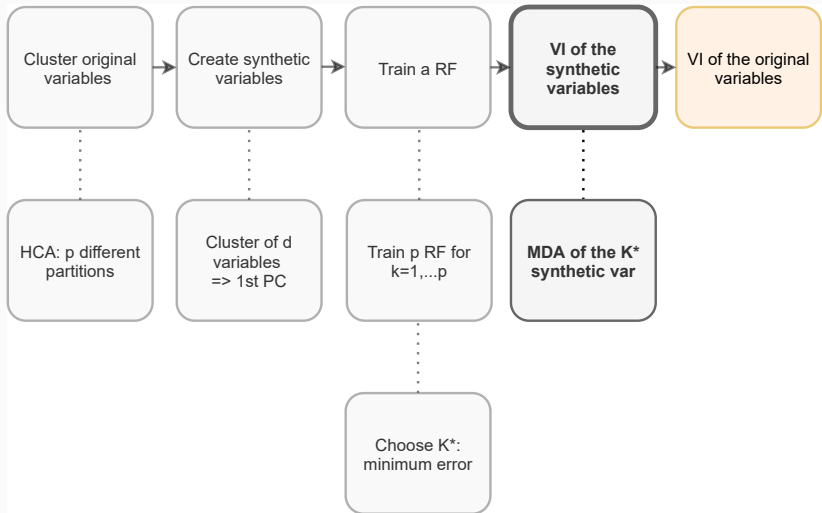
Train RF and choose K !

Procedure:

- for each partition in K clusters of the HCA, for $K = 1, \dots, p$ a RF is fitted with $\mathbf{f}^1, \dots, \mathbf{f}^K$.
- The optimal $K^* \rightarrow$ minimum out-of-the-bag (OOB) error rate.

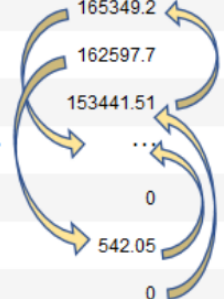


VI of the synthetic variables



Permutation importance

Illustration of the permutation procedure:



1	165349.2	136897.8	471784.1	192261.83	0
2	162597.7	151377.59	443898.53	191792.06	1
3	153441.51	101145.55	407934.54	191050.39	1
...
48	0	135426.92	0	42559.73	1
49	542.05	51743.15	0	35673.41	0
50	0	116983.8	45173.06	14681.4	1

Random Shuffle of the first feature

N

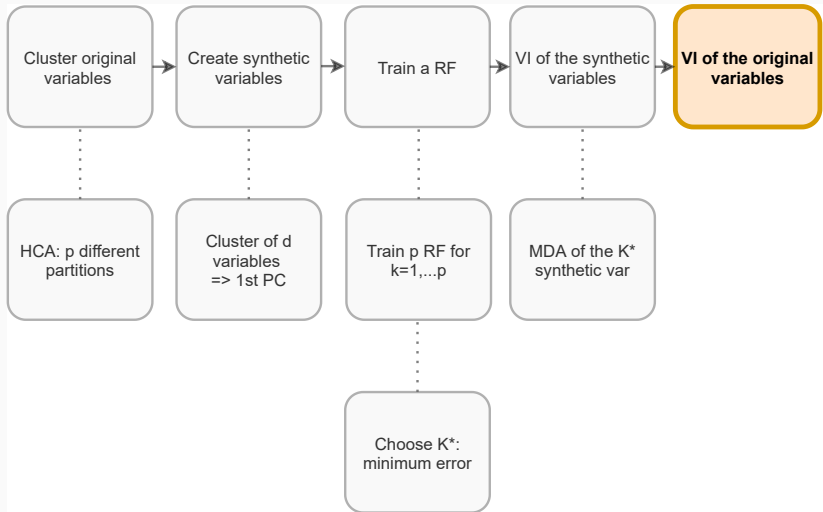
VI of the synthetic variables

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

- m a RF.
- T the number of tree in m .
- $R_m(D_{oob,t}, t)$ is the risk of the t -th tree computed on its OOB sample.
- $D_{oob,t}^{\pi_{k,t}}$ 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.

Our solution: VI of the original variables



The Synthetic-MDA (SMDA):

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

- \mathbf{x}^j is a variable clustered in cluster C_k .
- \mathbf{f}^k is the synthetic variable summarising the same cluster C_k .

Experiments

Simulated linear model

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

where $\mathbf{y} \in \mathbb{R}^n$, $\boldsymbol{\varepsilon}$ normally distributed with 0 mean and variance 0.5.

Important variables:

1. $\mathbf{X}_{p_1} \in \mathbb{R}^{n \times p_1}$ p_1 independent variables.
2. $\tilde{\mathbf{x}}^1 \in \tilde{\mathbf{X}}_{p_2}$, $\tilde{\mathbf{X}}_{p_2} \in \mathbb{R}^{n \times p_2}$: pairwise correlation of 0.9.

Noise variables:

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

Results 1

Table 1: Results of simulations¹ 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
$\text{cor}(\mathbf{y}, \mathbf{x}^j)^2$	0.56	0.24	0.06	0.14	0.11	0.11	0.00	0.00

¹using the *R*-package ranger [Wright and Ziegler, 2015] with default parameters and $T = 1000$.

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. $p_1 = 10, q_1 = 25$.

	$p_2 = 1; q_2 = 50$		$p_2 = 50; q_2 = 0$		$p_2 = 50; q_2 = 50$	
	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)

Conclusion

Conclusion

Summary:

- Use VI to discover relationships.
- VI with RF is a great tools but has limitations.
- Handling correlations allows solve issues.
- Cluster variables → additional gain in interpretability.

What to explore next?

- Interaction between features.

CODE IS AVAILABLE HERE: <https://github.com/MourerAlex/SM DA>

Questions?

CODE IS AVAILABLE HERE:

<https://github.com/MourerAlex/SMDA>

References

- Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- 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.
- M. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv:1508.04409*, 2015.
- C. B  nard, S. Da Veiga, and E. Scornet. Mda for random forests: inconsistency, and a practical solution via the sobol-mda. *arXiv:2102.13347*, 2021.

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.