Variable importance for random forests: a sensitivity analysis perspective

Clément Bénard

Thesis Advisors: Gérard Biau, Sébastien da Veiga, Erwan Scornet

Safran Tech, LPSM

April 2021

Interpretability

- State-of-the-art learning algorithms = black-boxes
  (e.g. random forests, neural networks)

Interpretability

- State-of-the-art learning algorithms = black-boxes (e.g. random forests, neural networks)
- Why? a high number of operations is performed to generate a prediction

Interpretability

- State-of-the-art learning algorithms = black-boxes (e.g. random forests, neural networks)
- Why? a high number of operations is performed to generate a prediction
- Consequence: impossible to grasp how inputs are combined to generate predictions

Interpretability

- State-of-the-art learning algorithms = black-boxes (e.g. random forests, neural networks)
- Why? a high number of operations is performed to generate a prediction
- Consequence: impossible to grasp how inputs are combined to generate predictions
- Strong limitation for applications with critical decisions at stake (e.g. healthcare, industrial processes)

Industrial processes

- Context
  Manufacturing process driven by controllable variables.

Production Line

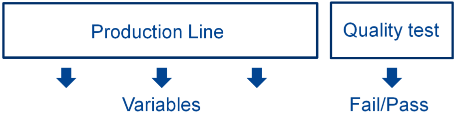
Variables

Quality test

Fail/Pass

Industrial processes

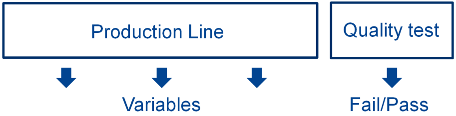
- Context
  Manufacturing process driven by controllable variables.



- Objective
  Identify production conditions generating defects: variable settings.

Industrial processes

- Context
  Manufacturing process driven by controllable variables.



- Objective
  Identify production conditions generating defects: variable settings.

- Method
  1. Fit a learning algorithm
  2. Use variable importance to detect influential variables
  3. Explore associated physical phenomenon with domain experts

Random forests

- Random forests are an efficient approach
    - State-of-the-art accuracy on a wide range of problems
    - Built-in variable importance algorithm: MDA (Breiman, 2001)

Random forests

- Random forests are an efficient approach
  - State-of-the-art accuracy on a wide range of problems
  - Built-in variable importance algorithm: MDA (Breiman, 2001)

- MDA has flaws
  - Poor understanding of the MDA: what is estimated ?
  - Empirical studies show that the MDA is biased for dependent inputs (Strobl et al., 2007; Gregorutti et al., 2017; Hooker and Mentch, 2019)

Random forests

- Random forests are an efficient approach
  - State-of-the-art accuracy on a wide range of problems
  - Built-in variable importance algorithm: MDA (Breiman, 2001)

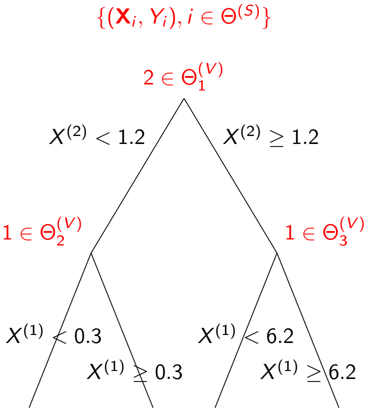
- MDA has flaws
  - Poor understanding of the MDA: what is estimated ?
  - Empirical studies show that the MDA is biased for dependent inputs (Strobl et al., 2007; Gregorutti et al., 2017; Hooker and Mentch, 2019)

- Our objective (Bénard et al., 2021)
  - Theoretical analysis of the MDA
    - First convergence result for the original MDA (Ishwaran, 2007; Zhu et al., 2015)
    - Theoretical understanding of MDA bias
  - Design of Sobol-MDA algorithm to fix the MDA flaws

Random forests

- Regression setting
  - input vector $\mathbf{X} = (X^{(1)}, \dots, X^{(p)}) \in \mathbb{R}^p$
  - output $Y \in \mathbb{R}$
  - dataset $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i), i = 1, \dots, n\}$, where $(\mathbf{X}_i, Y_i) \sim \mathbb{P}_{\mathbf{X},Y}$.

- Regression setting
  - input vector $\mathbf{X} = (X^{(1)}, \dots, X^{(p)}) \in \mathbb{R}^p$
  - output $Y \in \mathbb{R}$
  - dataset $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i), i = 1, \dots, n\}$, where $(\mathbf{X}_i, Y_i) \sim \mathbb{P}_{\mathbf{X},Y}$.
- Random forest algorithm
  - Aggregation of Θ-random trees $\Theta = (\Theta^{(S)}, \Theta^{(V)})$
  - M: number of trees
  - $m_{M,n}(\mathbf{X}, \Theta_M)$: the forest estimate at $\mathbf{X}$



MDA principle

MDA principle:

decrease of accuracy of the forest when a variable is noised up

MDA principle

MDA principle:

decrease of accuracy of the forest when a variable is noised up

1. fit a random forest with $\mathcal{D}_n$

MDA principle

MDA principle:

decrease of accuracy of the forest when a variable is noised up

1. fit a random forest with $\mathcal{D}_n$
2. compute the accuracy of the forest

MDA principle

MDA principle:

decrease of accuracy of the forest when a variable is noised up

1. fit a random forest with $\mathcal{D}_n$
2. compute the accuracy of the forest
3. permute randomly the values of a given input variable $X^{(j)}$:
   break the dependence between $X^{(j)}$ and Y

MDA principle

MDA principle:

decrease of accuracy of the forest when a variable is noised up

1. fit a random forest with $\mathcal{D}_n$
2. compute the accuracy of the forest
3. permute randomly the values of a given input variable $X^{(j)}$:
   break the dependence between $X^{(j)}$ and Y
4. compute the decrease of accuracy of the forest with the permuted data

MDA illustration

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

Table: Example of the permutation of a dataset $\mathcal{D}_n$ for $n = 5$.

MDA illustration

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

Table: Example of the permutation of a dataset $\mathcal{D}_n$ for $n = 5$.

MDA illustration

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	6.7	...	2.6	2.3
1.7	4.1	...	3.2	...	3.8	0.4
3.4	9.2	...	9.2	...	3.6	10.2
5.6	1.2	...	0.1	...	4.2	9.1
8.9	6.8	...	8.2	...	2.9	4.5

Table: Example of the permutation of a dataset $\mathcal{D}_n$ for $n = 5$.

MDA illustration

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	6.7	...	2.6	2.3
1.7	4.1	...	3.2	...	3.8	0.4
3.4	9.2	...	9.2	...	3.6	10.2
5.6	1.2	...	0.1	...	4.2	9.1
8.9	6.8	...	8.2	...	2.9	4.5

Table: Example of the permutation of a dataset $\mathcal{D}_n$ for $n = 5$.

Explained variance of Y = 16.4

Explained variance of Y = 13.7

$$\mathrm{MDA}(X^{(j)}) = 16.4 - 13.7 = 2.7$$

MDA illustration

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	6.7	...	2.6	2.3
1.7	4.1	...	3.2	...	3.8	0.4
3.4	9.2	...	9.2	...	3.6	10.2
5.6	1.2	...	0.1	...	4.2	9.1
8.9	6.8	...	8.2	...	2.9	4.5

Table: Example of the permutation of a dataset $\mathcal{D}_n$ for $n = 5$.

Question: Can I use $\mathcal{D}_n$ to both fit the forest and compute accuracy ?

No: overfitting and inflated accuracy.

How to handle this in practice?

MDA versions

The explained variance estimate of MDA algorithms differ across implementations

Train-Test MDA: train data to fit the forest, and test data for accuracy

The explained variance estimate of MDA algorithms differ across implementations

Train-Test MDA: train data to fit the forest, and test data for accuracy

Out-of-bag (OOB) samples: $\mathcal{D}_n$ is bootstrap prior to the construction of each tree, leaving aside a portion of $\mathcal{D}_n$, which is not involved in the tree growing and defines the "out-of-bag" sample.

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

Selected samples: $\Theta_\ell^{(S)} = \{1, 3, 4\}$

MDA versions

The explained variance estimate of MDA algorithms differ across implementations

Train-Test MDA: train data to fit the forest, and test data for accuracy

Out-of-bag (OOB) samples: $\mathcal{D}_n$ is bootstrap prior to the construction of each tree, leaving aside a portion of $\mathcal{D}_n$, which is not involved in the tree growing and defines the "out-of-bag" sample.

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

OOB samples: $\{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$

MDA versions

The explained variance estimate of MDA algorithms differ across implementations

Train-Test MDA: train data to fit the forest, and test data for accuracy

Out-of-bag (OOB) samples: $\mathcal{D}_n$ is bootstrap prior to the construction of each tree, leaving aside a portion of $\mathcal{D}_n$, which is not involved in the tree growing and defines the "out-of-bag" sample.

MDA Version	Package	Error	Data
Train-Test	`scikit-learn` `randomForestSRC`	Forest	Testing dataset
Breiman-Cutler	`randomForest` (normalized) `ranger` / `randomForestSRC`	Tree	OOB sample
Ishwaran-Kogalur	`randomForestSRC`	Forest	OOB sample

Table: Summary of the different MDA algorithms.

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree

$X^{(1)}$	$X^{(2)}$	$\ldots$	$X^{(j)}$	$\ldots$	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

- $i \in \{1, \dots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

Breiman-Cutler MDA

- $i \in \{1, \dots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathbf{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

$\mathbf{X}_i$

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	6.7	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	9.2	...	2.9	4.5

$\mathbf{X}_{i,\pi_{j\ell}}$

Breiman-Cutler MDA

- $i \in \{1, \dots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathbf{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

$$\widehat{\text{MDA}}_{M,n}^{(BC)}(X^{(j)}) = \frac{1}{M} \sum_{\ell=1}^{M} \frac{1}{N_{n,\ell}} \sum_{i=1}^{n} \Big[(Y_i - m_n(\mathbf{X}_{i,\pi_{j\ell}}, \Theta_\ell))^2 - (Y_i - m_n(\mathbf{X}_i, \Theta_\ell))^2 \Big] \mathbb{1}_{i \notin \Theta_\ell^{(S)}}$$

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathbf{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

$$\widehat{\mathrm{MDA}}_{M,n}^{(BC)}(X^{(j)}) = \frac{1}{M} \sum_{\ell=1}^{M} \frac{1}{N_{n,\ell}} \sum_{i=1}^{n} \Big[(Y_i - m_n(\mathbf{X}_{i,\pi_{j\ell}}, \Theta_\ell))^2 - (Y_i - m_n(\mathbf{X}_i, \Theta_\ell))^2\Big] \mathbb{1}_{i \notin \Theta_\ell^{(S)}}$$

Quadratic risk of the ℓ-th tree

Breiman-Cutler MDA

- $i \in \{1, \dots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^n \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathbf{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

$$\widehat{\mathrm{MDA}}_{M,n}^{(BC)}(X^{(j)}) = \frac{1}{M}\sum_{\ell=1}^{M}\frac{1}{N_{n,\ell}}\sum_{i=1}^{n}\Big[(Y_i - m_n(\mathbf{X}_{i,\pi_{j\ell}}, \Theta_\ell))^2 - (Y_i - m_n(\mathbf{X}_i, \Theta_\ell))^2\Big]\mathbb{1}_{i \notin \Theta_\ell^{(S)}}$$

Inflated quadratic risk of the ℓ-th tree where $X^{(j)}$ is permuted

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathbf{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

$$\widehat{\mathrm{MDA}}_{M,n}^{(BC)}(X^{(j)}) = \frac{1}{M} \sum_{\ell=1}^{M} \frac{1}{N_{n,\ell}} \sum_{i=1}^{n} \Big[(Y_i - m_n(\mathbf{X}_{i,\pi_{j\ell}}, \Theta_\ell))^2 - (Y_i - m_n(\mathbf{X}_i, \Theta_\ell))^2 \Big] \mathbb{1}_{i \notin \Theta_\ell^{(S)}}$$

Risks are computed over the OOB sample of each tree

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^n \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathbf{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

$$\widehat{\mathrm{MDA}}_{M,n}^{(BC)}(X^{(j)}) = \frac{1}{M} \sum_{\ell=1}^{M} \frac{1}{N_{n,\ell}} \sum_{i=1}^{n} \Big[(Y_i - m_n(\mathbf{X}_{i,\pi_{j\ell}}, \Theta_\ell))^2 - (Y_i - m_n(\mathbf{X}_i, \Theta_\ell))^2\Big] \mathbb{1}_{i \notin \Theta_\ell^{(S)}}$$

Average over all trees

Assumptions

(A1)

The response $Y \in \mathbb{R}$ *follows*

$$Y = m(\boldsymbol{X}) + \varepsilon$$

where

- $\boldsymbol{X} = (X^{(1)}, \dots, X^{(p)}) \in [0,1]^p$
- $\boldsymbol{X}$ *admits a density* f *such that* $c_1 < f(\boldsymbol{x}) < c_2$*, with constants* $c_1, c_2 > 0$
- m *is continuous*
- *the noise* ε *is sub-Gaussian and centered*

Assumptions

(A2): the theoretical tree is consistent
(always true with slight modifications of the forest algorithm)

Assumptions

(A2): the theoretical tree is consistent
(always true with slight modifications of the forest algorithm)

(A2)

The randomized theoretical CART tree built with the distribution of $(\boldsymbol{X}, Y)$ is consistent, that is, for all $\boldsymbol{x} \in [0,1]^p$, almost surely,

$$\lim_{k \to \infty} \Delta(m, A_k^{\star}(\boldsymbol{x}, \Theta)) = 0.$$

Assumptions

(A2): the theoretical tree is consistent
(always true with slight modifications of the forest algorithm)

(A2)

The randomized theoretical CART tree built with the distribution of $(\boldsymbol{X}, Y)$ *is consistent, that is, for all* $\boldsymbol{x} \in [0,1]^p$*, almost surely,*

$$\lim_{k \to \infty} \Delta(m, A_k^{\star}(\boldsymbol{x}, \Theta)) = 0.$$

(A3): tree partition is not too complex with respect to n

Assumptions

(A2): the theoretical tree is consistent
(always true with slight modifications of the forest algorithm)

(A2)

The randomized theoretical CART tree built with the distribution of $(\boldsymbol{X}, Y)$ is consistent, that is, for all $\boldsymbol{x} \in [0,1]^p$, almost surely,

$$\lim_{k \to \infty} \Delta(m, A_k^\star(\boldsymbol{x}, \Theta)) = 0.$$

(A3): tree partition is not too complex with respect to n

(A3)

The asymptotic regime of a_n, the size of the subsampling without replacement, and the number of terminal leaves t_n is such that $a_n \leq n-2$, $a_n/n < 1-\kappa$ for a fixed $\kappa > 0$, $\lim\limits_{n \to \infty} a_n = \infty$, $\lim\limits_{n \to \infty} t_n = \infty$, and $\lim\limits_{n \to \infty} t_n \frac{(\log(a_n))^9}{a_n} = 0$.

MDA Convergence

Theorem (Bénard et al. (2021))

If Assumptions (A1), (A2), and (A3) are satisfied, then, for all $M \in \mathbb{N}^\star$ *and* $j \in \{1, \ldots, p\}$ *we have*

$$\widehat{MDA}_{M,n}^{(BC)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[(m(\boldsymbol{X}) - m(\boldsymbol{X}_{\pi_j}))^2]$$

$\mathbf{X}_{\pi_j}$: $\mathbf{X}$ where the j-th component is replaced by an independent copy, i.e. $\mathbf{X}_{\pi_j} = (X^{(1)}, \ldots, X'^{(j)}, \ldots, X^{(p)})$

Limit interpretation?

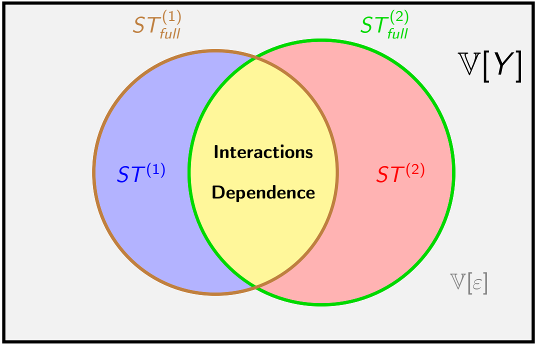


Figure: Standard and full total Sobol indices for $Y = m(X^{(1)}, X^{(2)}) + \varepsilon$.

Total Sobol index (Sobol, 1993)

$$ST^{(1)} = \frac{\mathbb{E}[\mathbb{V}(m(\mathbf{X})|\mathbf{X}^{(-1)})]}{\mathbb{V}(Y)}$$

Full total Sobol index (Mara et al., 2015; Benoumechiara, 2019)

$$ST^{(1)}_{full} = \frac{\mathbb{E}[\mathbb{V}(m(\mathbf{X}_{\pi_j})|\mathbf{X}^{(-1)})]}{\mathbb{V}(Y)}$$

MDA Decomposition

Proposition (Bénard et al. (2021))

If Assumptions (A1), (A2) and (A3) are satisfied, then for all $M \in \mathbb{N}^\star$ and $j \in \{1, \dots, p\}$ we have

$$\widehat{MDA}_{M,n}^{(BC)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST_{full}^{(j)} + MDA_3^{\star(j)}.$$

The term $\mathrm{MDA}_3^{\star(j)}$ is not an importance measure and is defined by

$$\mathrm{MDA}_3^{\star(j)} = \mathbb{E}[(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}] - \mathbb{E}[m(\mathbf{X}_{\pi_j})|\mathbf{X}^{(-j)}])^2].$$

MDA Decomposition

Proposition (Bénard et al. (2021))

If Assumptions (A1), (A2) and (A3) are satisfied, then for all $M \in \mathbb{N}^\star$ and $j \in \{1, \dots, p\}$ we have

$$(i) \quad \widehat{MDA}_{M,n}^{(TT)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST_{full}^{(j)} + MDA_3^{\star(j)}$$

$$(ii) \quad \widehat{MDA}_{M,n}^{(BC)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST_{full}^{(j)} + MDA_3^{\star(j)}.$$

If additionally $M \longrightarrow \infty$, then

$$(iii) \quad \widehat{MDA}_{M,n}^{(IK)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + MDA_3^{\star(j)}.$$

Independent inputs

If inputs X are independent: $\mathrm{MDA}_3^{\star(j)} = 0$ and $ST^{(j)} = ST^{(j)}_{full}$.

Corollary (Bénard et al. (2021))

If $\boldsymbol{X}$ *has independent components, and if Assumptions (A1)-(A3) are satisfied, for all* $M \in \mathbb{N}^\star$ *and* $j \in \{1, \dots, p\}$ *we have*

$$\widehat{MDA}^{(TT)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} 2\mathbb{V}[Y] \times ST^{(j)}$$

$$\widehat{MDA}^{(BC)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} 2\mathbb{V}[Y] \times ST^{(j)}.$$

If additionally $M \longrightarrow \infty$, *then*

$$\widehat{MDA}^{(IK)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)}.$$

This Corollary completes the result from (Gregorutti, 2015).

Additive regression function

If m **is additive**: $\mathrm{MDA}_3^{\star(j)} = 0$.

Corollary (Bénard et al. (2021))

If the regression function m is additive, and if Assumptions (A1)-(A3) are satisfied, for all $M \in \mathbb{N}^\star$ *and* $j \in \{1, \ldots, p\}$ *we have*

$$\widehat{MDA}_{M,n}^{(TT)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST_{full}^{(j)}$$

$$\widehat{MDA}_{M,n}^{(BC)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST_{full}^{(j)}.$$

If additionally $M \longrightarrow \infty$, *then*

$$\widehat{MDA}_{M,n}^{(IK)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)}.$$

MDA summary

- When inputs **X** are dependent and have interactions, the MDA is artificially inflated by the term MDA_3 and is therefore misleading.

MDA summary

- When inputs $\mathbf{X}$ are dependent and have interactions, the MDA is artificially inflated by the term MDA_3 and is therefore misleading.
- MDA versions have different theoretical counterparts, and thus different meanings: be careful when using forest packages !

MDA summary

- When inputs $\mathbf{X}$ are dependent and have interactions, the MDA is artificially inflated by the term MDA_3 and is therefore misleading.
- MDA versions have different theoretical counterparts, and thus different meanings: be careful when using forest packages !
- For variable selection, the total Sobol index is the relevant component

$$\mathbb{V}[Y] \times ST^{(j)} + \cancel{\mathbb{V}[Y] \times ST^{(j)}_{full}} + \cancel{\mathrm{MDA}_3^{\star(j)}}$$

MDA summary

- When inputs $\mathbf{X}$ are dependent and have interactions, the MDA is artificially inflated by the term MDA_3 and is therefore misleading.
- MDA versions have different theoretical counterparts, and thus different meanings: be careful when using forest packages !
- For variable selection, the total Sobol index is the relevant component

$$\mathbb{V}[Y] \times ST^{(j)} + \cancel{\mathbb{V}[Y] \times ST^{(j)}_{full}} + \cancel{\mathrm{MDA}_3^{\star(j)}}$$

- We develop the Sobol-MDA: a fast and consistent estimate of $ST^{(j)}$ for random forests

Sobol-MDA

Principle: **project** the partition of each tree along the j-th direction to remove $X^{(j)}$ from the prediction process.

Sobol-MDA

Principle: **project** the partition of each tree along the j-th direction to remove $X^{(j)}$ from the prediction process.

$$\widehat{\text{S-MDA}}_{M,n}(X^{(j)}) = \frac{1}{\hat{\sigma}_Y^2}\frac{1}{n}\sum_{i=1}^{n}\big[Y_i - m_{M,n}^{(-j,OOB)}(\mathbf{X}_i^{(-j)}, \Theta_M)\big]^2 - \big[Y_i - m_{M,n}^{(OOB)}(\mathbf{X}_i, \Theta_M)\big]^2$$

Principle: **project** the partition of each tree along the j-th direction to remove $X^{(j)}$ from the prediction process.

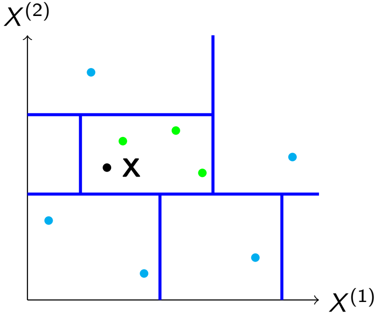


Figure: Example of the partition of $[0,1]^2$ by a random tree (left side) projected on the subspace span by $\mathbf{X}^{(-2)} = X^{(1)}$ (right side), for $p = 2$ and $j = 2$.

$$\widehat{\text{S-MDA}}_{M,n}(X^{(j)}) = \frac{1}{\hat{\sigma}_Y^2} \frac{1}{n} \sum_{i=1}^{n} \big[Y_i - m_{M,n}^{(-j,OOB)}(\mathbf{X}_i^{(-j)}, \Theta_M) \big]^2 - \big[Y_i - m_{M,n}^{(OOB)}(\mathbf{X}_i, \Theta_M) \big]^2$$

Sobol-MDA

Principle: **project** the partition of each tree along the j-th direction to remove $X^{(j)}$ from the prediction process.

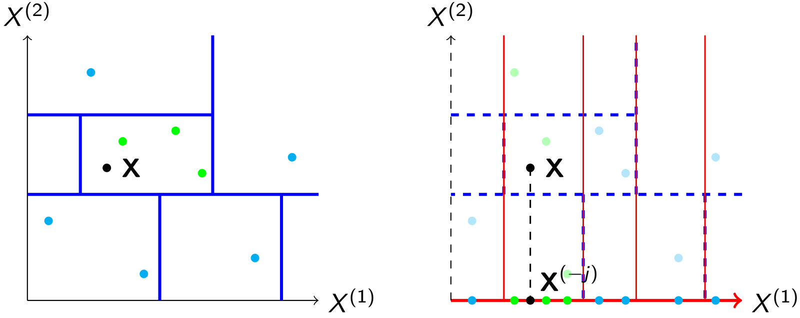


Figure: Example of the partition of $[0,1]^2$ by a random tree (left side) projected on the subspace span by $\mathbf{X}^{(-2)} = X^{(1)}$ (right side), for $p = 2$ and $j = 2$.

$$\widehat{\text{S-MDA}}_{M,n}(X^{(j)}) = \frac{1}{\hat{\sigma}_Y^2}\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - m_{M,n}^{(-j,OOB)}(\mathbf{X}_i^{(-j)}, \Theta_M)\right]^2 - \left[Y_i - m_{M,n}^{(OOB)}(\mathbf{X}_i, \Theta_M)\right]^2$$

Consistency of the Sobol-MDA

The Sobol-MDA recovers the appropriate theoretical counterpart for variable selection: the total Sobol index

Theorem (Bénard et al. (2021))

If Assumptions (A1), (A2'), and (A3') are satisfied, for all $M \in \mathbb{N}^\star$ and $j \in \{1, \dots, p\}$

$$\widehat{S\text{-}MDA}_{M,n}(X^{(j)}) \xrightarrow{p} ST^{(j)}.$$

Sobol-MDA Experiments

Settings (Archer and Kimes, 2008; Gregorutti et al., 2017)

- $p = 200$ input variables
- 5 independent groups of 40 variables
- each group is a Gaussian vector, strongly correlated

Sobol-MDA Experiments

Settings (Archer and Kimes, 2008; Gregorutti et al., 2017)

- $p = 200$ input variables
- 5 independent groups of 40 variables
- each group is a Gaussian vector, strongly correlated
- 1 variable from each group involved in m

$$m(\mathbf{X}) = 2X^{(1)} + X^{(41)} + X^{(81)} + X^{(121)} + X^{(161)}.$$

- independent Gaussian noise with $\mathbb{V}[\varepsilon] = 10\%\mathbb{V}[Y]$

$$Y = m(\mathbf{X}) + \varepsilon$$

Sobol-MDA Experiments

Settings (Archer and Kimes, 2008; Gregorutti et al., 2017)

- $p = 200$ input variables
- 5 independent groups of 40 variables
- each group is a Gaussian vector, strongly correlated
- 1 variable from each group involved in m

$$m(\mathbf{X}) = 2X^{(1)} + X^{(41)} + X^{(81)} + X^{(121)} + X^{(161)}.$$

- independent Gaussian noise with $\mathbb{V}[\varepsilon] = 10\%\mathbb{V}[Y]$

$$Y = m(\mathbf{X}) + \varepsilon$$

- $n = 1000$ observations
- $M = 300$ trees

Sobol-MDA Experiments

$\widehat{\text{S-MDA}}$		$\widehat{\text{BC-MDA}}/2\mathbb{V}[Y]$		$\widehat{\text{IK-MDA}}/\mathbb{V}[Y]$	
$\mathbf{X}^{(1)}$	0.035	$\mathbf{X}^{(1)}$	0.048	$\mathbf{X}^{(1)}$	0.056
$\mathbf{X}^{(161)}$	0.005	$\mathbf{X}^{(25)}$	0.010	$\mathbf{X}^{(5)}$	0.009
$\mathbf{X}^{(81)}$	0.004	$\mathbf{X}^{(31)}$	0.008	$\mathbf{X}^{(81)}$	0.007
$\mathbf{X}^{(121)}$	0.004	$\mathbf{X}^{(14)}$	0.008	$\mathbf{X}^{(41)}$	0.005
$\mathbf{X}^{(41)}$	0.002	$\mathbf{X}^{(40)}$	0.007	$\mathbf{X}^{(161)}$	0.005
$\mathbf{X}^{(179)}$	0.002	$\mathbf{X}^{(3)}$	0.007	$\mathbf{X}^{(15)}$	0.005
$\mathbf{X}^{(13)}$	0.001	$\mathbf{X}^{(17)}$	0.006	$\mathbf{X}^{(121)}$	0.005
$\mathbf{X}^{(25)}$	0.001	$\mathbf{X}^{(26)}$	0.006	$\mathbf{X}^{(7)}$	0.005
$\mathbf{X}^{(73)}$	0.001	$\mathbf{X}^{(41)}$	0.006	$\mathbf{X}^{(4)}$	0.004
$\mathbf{X}^{(155)}$	0.001	$\mathbf{X}^{(121)}$	0.006	$\mathbf{X}^{(28)}$	0.004

Table: Sobol-MDA, normalized BC-MDA, and normalized IK-MDA estimates with influential variables in blue.

Additional Experiments

Additional experiments are available in Bénard et al. (2021) (non-linear data with interactions and dependence)

- analytical example
- backward variable selection with real data

Conclusion

- Strong connections between the MDA and Sobol indices
- MDA does not target the appropriate quantity

Conclusion

- Strong connections between the MDA and Sobol indices
- MDA does not target the appropriate quantity
- Sobol-MDA fixes the flaws of original MDA
- R/C++ package SobolMDA, available online on Gitlab (https://gitlab.com/drti/sobolmda), and based on the package `ranger`
- Sobol-MDA can be associated with any black-box algorithm
  - fit a black box $\hat{f}$ on $\mathcal{D}_n$
  - generate a large sample $\mathcal{D}'_N$ with $\hat{f}$
  - run the Sobol-MDA with $\mathcal{D}'_N$

Conclusion

- Strong connections between the MDA and Sobol indices
- MDA does not target the appropriate quantity
- Sobol-MDA fixes the flaws of original MDA
- R/C++ package SobolMDA, available online on Gitlab (https://gitlab.com/drti/sobolmda), and based on the package `ranger`
- Sobol-MDA can be associated with any black-box algorithm
  - fit a black box $\hat{f}$ on $\mathcal{D}_n$
  - generate a large sample $\mathcal{D}'_N$ with $\hat{f}$
  - run the Sobol-MDA with $\mathcal{D}'_N$
- Perspectives: generalization to Shapley effects

Questions ?

K.J. Archer and R.V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52:2249–2260, 2008.

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

N. Benoumechiara. *Treatment of dependency in sensitivity analysis for industrial reliability*. PhD thesis, Sorbonne Université ; EDF R&D, 2019.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

B. Gregorutti. *Random forests and variable selection : analysis of the flight data recorders for aviation safety*. PhD thesis, Université Pierre et Marie Curie - Paris VI, 2015.

B. Gregorutti, B. Michel, and P. Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27:659–678, 2017.

G. Hooker and L. Mentch. Please stop permuting features: an explanation and alternatives. *arXiv preprint arXiv:1905.03151*, 2019.

H. Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

T. A Mara, S. Tarantola, and P. Annoni. Non-parametric methods for global sensitivity analysis of model output with dependent inputs. *Environmental Modelling & Software*, 72: 173–183, 2015.

I.M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1:407–414, 1993.

C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics*, 8:25, 2007.

R. Zhu, D. Zeng, and M. R. Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110:1770–1784, 2015.