MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA

Clément Bénard, Sébastien da Veiga, Erwan Scornet

MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA

Clément Bénard[1,2], Sébastien Da Veiga[1], and Erwan Scornet[3]

[1]Safran Tech, Digital Sciences & Technologies, 78114 Magny-Les-Hameaux, France
[2]Sorbonne Université, CNRS, LPSM, 75005 Paris, France
[3]Ecole Polytechnique, IP Paris, CMAP, 91128 Palaiseau, France

Abstract

Variable importance measures are the main tools to analyze the black-box mechanisms of random forests. Although the mean decrease accuracy (MDA) is widely accepted as the most efficient variable importance measure for random forests, little is known about its statistical properties. In fact, the exact MDA definition varies across the main random forest software. In this article, our objective is to rigorously analyze the behavior of the main MDA implementations. Consequently, we mathematically formalize the various implemented MDA algorithms, and then establish their limits when the sample size increases. In particular, we break down these limits in three components: the first one is related to Sobol indices, which are well-defined measures of a covariate contribution to the response variance, widely used in the sensitivity analysis field, as opposed to the third term, whose value increases with dependence within covariates. Thus, we theoretically demonstrate that the MDA does not target the right quantity when covariates are dependent, a fact that has already been noticed experimentally. To address this issue, we define a new importance measure for random forests, the Sobol-MDA, which fixes the flaws of the original MDA. We prove the consistency of the Sobol-MDA and show that the Sobol-MDA empirically outperforms its competitors on both simulated and real data. An open source implementation in `R` and `C++` is available online.

Keywords: MDA; Random forests; Sensitivity analysis; Sobol indices; Variable importance; Variable selection

1 Introduction

Random forests (Breiman, 2001) are a statistical learning algorithm, which aggregates a large number of trees to solve regression and classification problems, and achieve state-of-the-art performance on a wide range of problems. In particular, random forests exhibit a good behavior on high-dimensional or noisy data, without any parameter tuning, and are also well known for their robustness. However, they suffer from a major drawback: a given prediction is generated through a large number of operations, typically tens of thousands, which makes the interpretation of the prediction mechanism impossible. Because of this complexity, random

forests are often qualified as black boxes. More generally, the interpretability of learning algorithms is receiving an increasingly high interest since this black-box characteristic is a strong practical limitation. For example, applications involving critical decisions, typically healthcare, require predictions to be justified. The most popular way to interpret random forests is variable importance analysis: covariates are ranked by decreasing order of their importance in the algorithm prediction process. Thus, specific variable importance measures were developed along with random forests (Breiman, 2001, 2003a). However, we will see that they may not target the right variable ranking when covariates are dependent, and could therefore be improved. Firstly, we present the context and motivation of variable importance. Secondly, we review the existing variable importance measures for random forests, and then conduct a theoretical analysis of their limitations. Finally, we introduce the Sobol-MDA algorithm, a new importance measure for random forests, which outperforms the existing competitors as shown in the experiments.

2 Context and Objectives

2.1 Variable Importance for Random Forests.

There are essentially two importance measures for random forests: the mean decrease accuracy (MDA) (Breiman, 2001) and the mean decrease impurity (MDI) (Breiman, 2003a). The MDA measures the decrease of accuracy when the values of a given covariate are permuted, thus breaking its relation to the response variable and to the other covariates. On the other hand, the MDI sums the weighted decreases of impurity over all nodes that split on a given covariate, averaged over all trees in the forest. In both cases, a high value of the metric means that the covariate is used in many important operations of the prediction mechanism of the forest. Unfortunately, there is no precise and rigorous interpretation since these two definitions are purely empirical. Furthermore, in the last two decades, many empirical analysis have highlighted the flaws of the MDI (Strobl et al., 2007). Although Li et al. (2019), Zhou and Hooker (2019), and Loecher (2020) recently improved the MDI to partially remove its bias, Scornet (2020) demonstrated that the MDI is consistent only under a strong and restrictive assumption: the regression function is additive and the covariates are independent. Otherwise, the MDI is ill-defined. Overall, the MDA is widely considered as the most efficient variable importance measure for random forests (Strobl et al., 2007; Ishwaran, 2007; Genuer et al., 2010; Boulesteix et al., 2012), and we therefore focus on the MDA. Although it is extensively used in practice, little is known about its statistical properties. To our knowledge, only Ishwaran (2007) and Zhu et al. (2015) provide theoretical analyses of modified versions of the MDA, but the asymptotic behavior of the original MDA algorithm (Breiman, 2001) is unknown: Ishwaran (2007) considers Breiman's forests but simplifies the MDA procedure, whereas Zhu et al. (2015) considers the original MDA but assumes the independence of the covariates and an exponential concentration inequality on the random forest estimate, the latter being proved only for purely random forests (which do not use the data set to build the tree partitions). On the practical side, many empirical analyses provide evidence that when covariates are dependent, the MDA may fail to detect some relevant covariates (Archer and Kimes, 2008; Strobl et al., 2008; Nicodemus and Malley, 2009; Genuer et al., 2010; Auret and Aldrich, 2011; Toloşi and Lengauer, 2011; Gregorutti et al., 2017; Hooker and Mentch, 2019;

Mentch and Zhou, 2020). Several proposals (Mentch and Hooker, 2016; Candes et al., 2016; Williamson et al., 2020) were recently made to overcome this issue. Mentch and Hooker (2016) prove the asymptotic normality of random forests, which enables to detect if the predictions of a forest built without a given covariate are significantly different from the ones of the original forest with all covariates. Alternatively, Candes et al. (2016) introduce model-X knockoffs, which rely on conditional randomization tests, where the relation between a covariate and the response variable is broken without modifying the joint distribution of the covariates. Finally, Williamson et al. (2020) propose to measure the decrease of accuracy between the original procedure and a new run without a given covariate. However, these methods have a much higher computational cost, as many model retrains are involved, and are in particular intractable in high dimension. Furthermore, it is critical to assess that the properties of a variable importance measure are in line with the final objective of the conducted analysis. In the following subsection, we review the possible goals of variable importance, and then introduce sensitivity analysis to deepen the theoretical understanding of the MDA.

2.2 Sensitivity Analysis

The analysis of variable importance is not an end in itself, the goal is essentially to perform variable selection, with usually two final aims (Genuer et al., 2010): (i) find a small number of covariates with a maximized accuracy, or (ii) detect and rank all influential covariates to focus on for further exploration with domain experts. Depending on which of these two objectives is of interest, different strategies should be used as the following example shows: if two influential covariates are strongly correlated, one must be discarded in the first case, while the two must be kept in the second case. Indeed, if two covariates convey the same statistical information, only one should be selected if the goal is to maximize the predictive accuracy with a small number of covariates, i.e., objective (i). On the other hand, these two covariates may be acquired differently and represent distinct physical quantities. Therefore, they may have different interpretations for domain experts, and both should be kept for objective (ii).

Sensitivity analysis is the study of uncertainties in a system. The main goal is to apportion the uncertainty of a system response to the uncertainty of the different covariates. Iooss and Lemaître (2015) and Ghanem et al. (2017) provide detailed reviews of global sensitivity analysis (GSA). In particular, GSA introduces well-defined importance measures of covariate contributions to the response variance: Sobol indices (Sobol, 1993; Saltelli, 2002) and Shapley effects (Shapley, 1953; Owen, 2014; Iooss and Prieur, 2017). These metrics are widely used to analyze computer code experiments, especially for the design of industrial systems. However, the literature about variable importance in the fields of statistical learning and machine learning rarely mentions sensitivity analysis. The reason of this hiatus is clear: until quite recently, GSA was focused on independent covariates, whereas the machine learning community essentially works with dependent ones. In the last years, Gregorutti (2015) first established a link between GSA and the MDA: in the case of independent covariates, the theoretical counterpart of the MDA is the unnormalized total Sobol index, i.e., twice the amount of explained variance lost when a given covariate is removed from the model, which is the expected quantity for both objectives (i) and (ii) in this independent setting. Accordingly, the algorithm from Williamson et al. (2020) also estimates the total Sobol index when the accuracy metric is the explained variance, even when covariates are dependent, and although this connection is

not explicitly mentioned. Additionally, Owen (2014) reintroduced Shapley effects, originally proposed in game theory (Shapley, 1953). Shapley effects exhibit very interesting properties as they equitably allocate the mutual contribution due to dependence and interactions to individual covariates, and are now widely used by the machine learning community to interpret both tree ensembles and neural networks. SHAP values (Lundberg and Lee, 2017) also adapt Shapley effects for local interpretation of model predictions, and Lundberg et al. (2018) provide a fast algorithm for tree ensembles. Finally, we refer to Antoniadis et al. (2020) for a review of random forests and sensitivity analysis.

2.3 Article Outline

In Section 3, we review and clarify the different MDA algorithms implemented in the main random forest software: several definitions coexist, and we first formalize them mathematically. Then, we conduct an asymptotic analysis to draw connections between the MDA and sensitivity analysis in the general case with dependent covariates. We thus demonstrate that all MDA versions are indeed inappropriate for the two possible objectives of variable importance analysis. To our knowledge, this is the first asymptotic result on Breiman's MDA, which sheds light on the empirical limitations observed in practice. Next, for objective (ii), it is widely accepted that Shapley effects are an efficient importance measure as they equitably handle interactions and dependence. On the other hand, when one is using variable importance to select a small number of covariates while maximizing predictive accuracy, i.e. objective (i), the total Sobol index is clearly the relevant measure to eliminate the less influential covariates, as also suggested by Williamson et al. (2020). Therefore, we focus on objective (i) in Section 4, and propose the Sobol-MDA, an augmented version of the MDA which consistently estimates the total Sobol index even when covariates are dependent. The Sobol-MDA outperforms the existing competitors on both simulated and real data, and is proved to be consistent. An implementation in `R` and `C++` of the Sobol-MDA is available at `https://gitlab.com/drti/sobolmda`, and is based on `ranger` (Wright and Ziegler, 2017), a fast implementation of random forests. Notice that proofs, additional details, and experiments are to be found in the Supplementary Material.

3 MDA Theoretical Limitations

3.1 MDA Definitions

The MDA was originally proposed by Breiman in his seminal article (Breiman, 2001), and works as follows. The values of a specific covariate are permuted to break its relation to the response variable. Then, the predictive accuracy is computed for this perturbed dataset. The difference between this degraded accuracy and the original one gives the importance of the covariate: a high decrease of accuracy means that the considered covariate has a strong influence on the prediction mechanism. However, a review of the literature on random forests and their software implementations reveals that there is no consensus on the exact mathematical formulation of the MDA. We focus on the most popular random forest algorithms:

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

Table 1: Summary of the different MDA characteristics.

- the `R` package `randomForests` (Liaw and Wiener, 2002) based on the original `Fortran` code from Breiman and Cutler
- the fast `R/C++` implementation `ranger` (Wright and Ziegler, 2017)
- the most widely used `python` machine learning library `scikit-learn` (Pedregosa et al., 2011) (`RandomForestClassifier`/`RandomForestRegressor`)
- the `R` package `randomForestSRC` (Ishwaran and Kogalur, 2020) which implements survival forests in addition to the original algorithm.

To give an order of magnitude, the typical number of users of each of these packages during the year 2020 is about half a million. A close inspection of their code exhibits that essentially three distinct definitions of the MDA are widely used. References and details about the MDA implementation in the package codes are provided in the Supplementary Material. The differences between the three MDA versions are twofold: the MDA can be computed based on the tree error or the whole forest error, and via a test set or out-of-bag samples, as summarized in Table 1. We first introduce the required notations, and then mathematically formalize these different MDA definitions. We define a standard regression setting with the following Assumption 1, as well as the random forest notations below.

Assumption 1 *The response variable $Y \in \mathbb{R}$ follows $Y = m(X)+\varepsilon$, where the covariate vector $X = (X^{(1)}, \ldots, X^{(p)}) \in [0,1]^p$ admits a density over $[0,1]^p$ bounded from above and below by strictly positive constants, m is continuous, and the noise ε is sub-Gaussian, independent of X, and centered. A sample $\mathscr{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of n independent random variables distributed as (X, Y) is available.*

The random CART estimate $m_n(x, \Theta)$ is trained with $\mathscr{D}_n$ and Θ, where Θ is used to generate the bootstrap sampling and the split randomization, and $x \in [0,1]^p$ is a new observation. The component of Θ used to resample the data is denoted $\Theta^{(S)} \subset \{1, \ldots, n\}$. The random forest estimate $m_{M,n}(x, \Theta_{(M)})$ aggregates M Θ-random CART, each of which is randomized by a component of $\Theta_{(M)} = (\Theta_1, \ldots, \Theta_M)$. In the sequel, we consider a fixed index $j \in \{1, \ldots, p\}$. Next, we define X_{i,π_j} as the vector X_i where the j-th component is permuted between observations. Similarly, X_{π_j} is the vector X where the j-th component is replaced by an independent copy of $X^{(j)}$. Finally, we also introduce $X^{(-j)}$, as the random vector X without the j-th component. Now, we can detail the three MDA definitions, summarized in Table 1.

The most simple approach is taken by `scikit-learn` where the forest is fit with a training sample and the accuracy decrease is estimated with an independent testing sample $\mathscr{D}'_n = \{(X'_1, Y'_1), \ldots, (X'_n, Y'_n)\}$. Throughout the article, we call the generalization error of the forest the expected squared error for a new observation, usually estimated with an independent sample. Thus, forest predictions are run for both the test set and its permuted version, and the corresponding mean squared errors are subtracted to give the generalization error increase, denoted the Train-Test MDA.

Definition 1 (Train/Test MDA) *The Train/Test MDA is defined by*

$$\widehat{MDA}^{(TT)}_{M,n}(X^{(j)}) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y'_i - m_{M,n}(X'_{i,\pi_j}, \Theta_{(M)})\right\}^2 - \left\{Y'_i - m_{M,n}(X'_i, \Theta_{(M)})\right\}^2.$$

This algorithm is the only MDA version implemented in `scikit-learn`, and is one possibility in `randomForestSRC`. Note that the Train/Test-MDA is straightforward to implement with any random forest package by simply running predictions.

In practice, splitting the sample in two parts for training and testing often hurts the accuracy of the procedure, then decreasing the accuracy of the MDA estimate. Since the data is bootstrapped prior to the construction of each tree, a portion of the sample is left out, which is called the out-of-bag sample and can be used to measure accuracy. Despite the lack of mathematical formulation in the original MDA formulation of Breiman (Breiman, 2001), it seems clear that for each tree, the generalization error is estimated using its out-of-bag sample and the permuted version. Then, the two errors are subtracted and this difference is averaged across all trees to give the Breiman-Cutler MDA.

Definition 2 (Breiman-Cutler MDA) *If $X_{i,\pi_{j\ell}}$ is the i-th permuted out-of-bag sample for the ℓ-th tree and for $i \in \{1, \ldots, n\} \setminus \Theta^{(S)}_\ell$, then the Breiman-Cutler MDA (BC-MDA) (Breiman, 2001) is defined by*

$$\widehat{MDA}^{(BC)}_{M,n}(X^{(j)}) = \frac{1}{M}\sum_{\ell=1}^{M}\frac{1}{N_{n,\ell}}\sum_{i=1}^{n}\left[\{Y_i - m_n(X_{i,\pi_{j\ell}}, \Theta_\ell)\}^2 - \{Y_i - m_n(X_i, \Theta_\ell)\}^2\right]\mathbb{1}_{i\notin\Theta^{(S)}_\ell},$$

where $N_{n,\ell} = \sum_{i=1}^{n}\mathbb{1}_{i\neq\Theta^{(S)}_\ell}$ is the size of the out-of-bag sample of the ℓ-th tree.

Among the four main random forest implementations introduced above, only `ranger` and `randomForestSRC` exactly follow this original definition. In `randomForests`, the final quantity is normalized by the standard deviation of the generalization error differences. However, this procedure is questionable (Díaz-Uriarte and De Andres, 2006; Strobl and Zeileis, 2008): a non-influential covariate would constantly have a standard deviation close to zero, potentially leading to a high normalized MDA.

More importantly, observe that Breiman's MDA definition is in fact a Monte-Carlo estimate of a random tree decrease of accuracy when a covariate is noised up. Since we are interested in the covariate influence in the entire forest, and not only in a single tree, it seems natural to extend the out-of-bag procedure to estimate the forest risk (Ishwaran, 2007; Ishwaran et al., 2008) as implemented in `randomForestSRC`: for each observation X_i, we retrieve the random

set $\Lambda_{n,i}$ of trees which do not involve X_i in their construction because of the resampling step, formally defined by

$$\Lambda_{n,i} = \{\ell \in \{1, \dots, M\} : i \notin \Theta_\ell^{(S)}\}.$$

We can take advantage of such batch of trees to define the out-of-bag random forest estimate by averaging the tree predictions considering only trees that belong to $\Lambda_{n,i}$, i.e., for $i \in \{1, \dots, n\}$,

$$m_{M,n}^{(OOB)}(X_i, \Theta_{(M)}) = \frac{1}{|\Lambda_{n,i}|} \sum_{\ell \in \Lambda_{n,i}} m_n(X_i, \Theta_\ell) \mathbb{1}_{|\Lambda_{n,i}|>0}.$$

It is therefore possible to estimate the random forest error using $\mathscr{D}_n$ alone. Recall that for each Θ_ℓ-random tree, we randomly permute the j-th component of the out-of-bag dataset to define $X_{i,\pi_{j\ell}}$, and we stress that the permutation is independent for each tree. Then, we define the permuted out-of-bag forest estimate as

$$m_{M,n,\pi_j}^{(OOB)}(X_i, \Theta_{(M)}) = \frac{1}{|\Lambda_{n,i}|} \sum_{\ell \in \Lambda_{n,i}} m_n(X_{i,\pi_{j\ell}}, \Theta_\ell) \mathbb{1}_{|\Lambda_{n,i}|>0}.$$

These estimates enable to compute both the out-of-bag error of the forest and the inflated out-of-bag forest error when a covariate is noised up. Finally, the difference between these two errors forms the Ishwaran-Kogalur MDA. From an algorithmic point of view, also notice that the only difference with Breiman's definition is the mechanism to aggregate tree predictions and compute the errors.

Definition 3 (Ishwaran-Kogalur MDA) *The Ishwaran-Kogalur MDA (IK-MDA) (Ishwaran, 2007; Ishwaran et al., 2008) is defined by*

$$\widehat{MDA}_{M,n}^{(IK)}(X^{(j)}) = \frac{1}{N_{M,n}} \sum_{i=1}^{n} \{Y_i - m_{M,n,\pi_j}^{(OOB)}(X_i, \Theta_{(M)})\}^2 - \{Y_i - m_{M,n}^{(OOB)}(X_i, \Theta_{(M)})\}^2,$$

where $N_{M,n} = \sum_{i=1}^{n} \mathbb{1}_{|\Lambda_{n,i}|>0}$ *is the number of points which are not used in all tree constructions.*

An asymptotic analysis of these three MDA versions, summarized in Table 1, reveals that they do not share the same theoretical counterpart. Consequently, they have different meanings and generate different variable rankings, from which divergent conclusions can be drawn. However, these MDA versions are used interchangeably in practice. The convergence of the MDA is established in the next subsection, and then the different theoretical counterparts are analyzed in the following subsection.

3.2 MDA Inconsistency

The out-of-bag estimate is involved in both the Breiman-Cutler MDA and Ishwaran-Kogalur MDA, but is also used in practice to provide a fast estimate of the random forest error. We begin our asymptotic analysis by a result on the efficiency of the out-of-bag estimate, stated in Proposition 1 below, which shows that the out-of-bag error consistently estimates

the generalization error of the forest. This result will be later used to establish the convergence of the Ishwaran-Kogalur MDA. The only difference between the implemented algorithms and our theoretical results, is that the resampling in the forest growing is done without replacement to alleviate the mathematical analysis (Scornet et al., 2015; Mentch and Hooker, 2016; Wager and Athey, 2018). We define a_n the number of subsampled training observations used to build each tree.

Proposition 1 *If Assumption 1 is satisfied, for a fixed sample size n and $i \in \{1, \dots, n\}$, we have*

$$\left| \mathbb{E}\left[\left\{ m^{(OOB)}_{M,a_n,n}(X_i, \Theta_{(M)}) - m(X_i) \right\}^2 \right] - \mathbb{E}\left[\left\{ m_{M,a_n,n-1}(X, \Theta_{(M)}) - m(X) \right\}^2 \right] \right| = O\Big(\frac{1}{M}\Big).$$

First observe that, by construction of the set of trees $\Lambda_{n,i}$, the out-of-bag estimate aggregates a smaller number of trees than in the standard forest: $\mathbb{E}[|\Lambda_{n,i}|] = (1 - a_n/n)M$ trees in average. Therefore, the errors of the out-of-bag and standard forest estimates are different quantities. To our knowledge, this is the first result which states the convergence of the out-of-bag error towards the forest error for any fixed sample size, with a fast rate of $1/M$. This suggests that growing a large number of trees in the forest, which is computationally possible and what is done in practice, ensures that the out-of-bag estimate provides a good approximation of the forest error.

Next, the convergence of the three versions of the MDA holds under the following Assumption 2 of the consistency of a theoretical randomized CART. Since we are interested in the random forest interpretation through the MDA, it seems natural to conduct our analysis assuming that each tree of the forest is an efficient learner, i.e., consistent. To formalize such an assumption, we first define the variation of the regression function within a cell $A \subset [0,1]^p$ by

$$\Delta(m, A) = \sup_{x, x' \in A} |m(x) - m(x')|,$$

and secondly, we introduce $A^\star_k(x, \Theta)$ the cell of the theoretical CART of depth k (randomized with Θ) in which the observation $x \in [0,1]^p$ falls.

Assumption 2 *The randomized theoretical CART tree built with the distribution of (X, Y) is consistent, that is, for all $x \in [0,1]^p$, almost surely,*

$$\lim_{k \to \infty} \Delta(m, A^\star_k(x, \Theta)) = 0.$$

At first glance, Assumption 2 seems quite obscure since it involves the theoretical CART. However, Scornet et al. (2015) show that Assumption 2 holds if the regression function is additive. Because the original CART (Breiman et al., 1984) is a greedy algorithm, Assumption 2 may not always be satisfied when the regression function m has interaction terms. However, it holds if the CART algorithm is slightly modified to avoid splits to be close to the edges of cells, and the split randomization is slightly increased to have a positive probability to split in all directions at all nodes (Meinshausen, 2006; Wager and Athey, 2018). Indeed in that case, all cells become infinitely small as the tree depth k increases, and therefore Assumption

2 holds by continuity of m. Such modifications of CART have a negligible impact in practice on the random forest estimate since the cut threshold and the split randomization increase can be chosen arbitrarily small. Notice that such asymptotic regime is specifically analyzed in the next section.

As specified above, a_n is the number of training observations subsampled without replacement to build each tree, and we define t_n as the final number of terminal leaves in every tree. Notice that we can specify a_n in $m_{M,a_n,n}(x, \Theta_{(M)})$ or $m_{a_n,n}(x, \Theta)$ when needed, but we omit it in general to avoid cumbersome notations. In order to properly define the MDA procedures, the out-of-bag sample needs to be at least of size 2 to enable permutations, i.e., $a_n \leq n-2$. Finally, we need the following Assumption 3 on the asymptotic regime of the empirical forest as stated in Scornet et al. (2015), which essentially controls the number of terminal leaves with respect to the sample size n to enforce the random forest consistency.

Assumption 3 *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 $\kappa > 0$, $\lim_{n\to\infty} a_n = \infty$, $\lim_{n\to\infty} t_n = \infty$, and $\lim_{n\to\infty} t_n \frac{(\log(a_n))^9}{a_n} = 0$.*

In the case of the Ishwaran-Kogalur MDA, the number of trees has to tend to infinity with the sample size to ensure convergence. To lighten notations, we drop the dependence of M_n to n.

Assumption 4 *The number of trees grows to infinity with the sample size n: $M \underset{n\to\infty}{\longrightarrow} \infty$.*

Theorem 1 *If Assumptions 1, 2, and 3 are satisfied, then, for all $M \in \mathbb{N}^\star$ and $j \in \{1, \ldots, p\}$ we have*

$$(i) \quad \widehat{MDA}^{(TT)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[\{m(X) - m(X_{\pi_j})\}^2]$$
$$(ii) \quad \widehat{MDA}^{(BC)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[\{m(X) - m(X_{\pi_j})\}^2].$$

If Assumption 4 is additionally satisfied, then

$$(iii) \quad \widehat{MDA}^{(IK)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[\{m(X) - \mathbb{E}[m(X_{\pi_j})|X^{(-j)}]\}^2].$$

Theorem 1 reveals that the theoretical MDA counterparts are not identical across the different MDA definitions. Thus, covariates are ranked according to different criteria when the Breiman-Cutler MDA or Ishwaran-Kogalur MDA is used. We deepen this discussion in the following subsection.

3.3 MDA Analysis

The theoretical counterparts of the MDA established in Theorem 1 are hard to interpret since X_{π_j} has a different distribution from the original covariate vector X whenever components of X are dependent. These different MDA versions are widely used in practice to assess the variable importance of random forests, but the relevance of such analyses completely relies on the ranking criteria $\mathbb{E}[\{m(X) - m(X_{\pi_j})\}^2]$ or $\mathbb{E}[\{m(X) - \mathbb{E}[m(X_{\pi_j})|X^{(-j)}]\}^2]$, according to

Theorem 1. It is possible to deepen the discussion, observing that X and X_{π_j} are independent conditionally on $X^{(-j)}$ by construction. It enables to break down the MDA limit using Sobol indices that are well-defined quantity to measure the contribution of a covariate to the response variance.

Definition 4 (Total Sobol Index) *The total Sobol index of covariate $X^{(j)}$ (Sobol, 1993; Saltelli, 2002) gives the proportion of explained response variance lost when $X^{(j)}$ is removed from the model, that is*

$$ST^{(j)} = \frac{\mathbb{E}[\mathbb{V}(m(X) \mid X^{(-j)})]}{\mathbb{V}(Y)}.$$

Notice that $ST^{(j)}$ is also called the independent total Sobol index in Kucherenko et al. (2012) and Benoumechiara (2019).

We also introduce a new sensitivity index: the total Sobol index computed for the input vector X_{π_j}. We call it the marginal total Sobol index, since the distribution of X_{π_j} is the product of the marginal distributions of $X^{(j)}$ and $X^{(-j)}$. It can take high values even when $X^{(j)}$ is strongly correlated with other covariates, as opposed to the original total Sobol index. We derive the main properties of this new sensitivity index below, proved in the Supplementary Material.

Definition 5 (Marginal Total Sobol Index) *The marginal total Sobol index of covariate $X^{(j)}$ is defined by*

$$ST_{mg}^{(j)} = \frac{\mathbb{E}[\mathbb{V}(m(X_{\pi_j}) \mid X^{(-j)})]}{\mathbb{V}(Y)}.$$

Property 1 (Marginal Total Sobol Index) *If Assumption 1 is satisfied, the marginal total Sobol index $ST_{mg}^{(j)}$ satisfies the following properties.*

(a) $ST_{mg}^{(j)} = 0 \iff ST^{(j)} = 0$.

(b) If the components of X are independent, then we have $ST_{mg}^{(j)} = ST^{(j)}$.

(c) If m is additive, i.e. $m(X) = \sum_k m_k(X^{(k)})$, then we have $ST_{mg}^{(j)} = \mathbb{V}[m_j(X^{(j)})]/\mathbb{V}[Y]$, and $ST_{mg}^{(j)} \geq ST^{(j)}$.

Notice that the last property states that $ST_{mg}^{(j)} \geq ST^{(j)}$ for additive regression functions, which may also hold in the general case with interactions. However, such extension is out of the scope of the article at the moment. It is now possible to break down the MDA limits as the sum of positive terms using total Sobol indices and the following quantity $\text{MDA}_3^{\star(j)}$, further discussed below and defined as

$$\text{MDA}_3^{\star(j)} = \mathbb{E}[(\mathbb{E}[m(X) \mid X^{(-j)}] - \mathbb{E}[m(X_{\pi_j}) \mid X^{(-j)}])^2].$$

Proposition 2 *If Assumptions 1, 2 and 3 are satisfied, then for all* $M \in \mathbb{N}^\star$ *and* $j \in \{1, \dots, p\}$ *we have*

$$(i) \quad \widehat{MDA}^{(TT)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST^{(j)}_{mg} + MDA^{\star(j)}_3$$
$$(ii) \quad \widehat{MDA}^{(BC)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST^{(j)}_{mg} + MDA^{\star(j)}_3.$$

If Assumption 4 is additionally satisfied, then

$$(iii) \quad \widehat{MDA}^{(IK)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + MDA^{\star(j)}_3.$$

Importantly, each term of the decompositions of Proposition 2 is positive, and can be interpreted alone. We denote $\text{MDA}^{\star(j)}_1 = \mathbb{V}[Y] \times ST^{(j)}$ and $\text{MDA}^{\star(j)}_2 = \mathbb{V}[Y] \times ST^{(j)}_{mg}$.

$\text{MDA}^{\star(j)}_1$ is the non-normalized total Sobol index that has a straightforward interpretation: the amount of explained output variance lost when $X^{(j)}$ is removed from the model. This quantity is really the information one is looking for when computing the MDA for objective (i).

$\text{MDA}^{\star(j)}_2$ is the non-normalized marginal total Sobol index. Its interpretation is more difficult. Intuitively, in the case of $\text{MDA}^{\star(j)}_1$, contributions due to the dependence between $X^{(j)}$ and $X^{(-j)}$ are excluded because of the conditioning on $X^{(-j)}$. For $\text{MDA}^{\star(j)}_2$, this dependence is ignored, and therefore such removal does not take place. For example, if $X^{(j)}$ has a strong influence on the regression function but is highly correlated with other covariates, then $\text{MDA}^{\star(j)}_1$ is small, whereas $\text{MDA}^{\star(j)}_2$ is high. For objective (i), one wants to keep only one covariate of a group of highly influential and correlated inputs, and therefore $ST^{(j)}_{mg}$ can be a misleading component.

$\text{MDA}^{\star(j)}_3$ is not a known measure of importance, and seems to have no clear interpretation: it measures how the permutation shifts the average of m over the j-th covariate, and thus characterizes the structure of m and the dependence of X combined. $\text{MDA}^{\star(j)}_3$ is null if covariates are independent. The value of $\text{MDA}^{\star(j)}_3$ increases with dependence, and this effect can be amplified by interactions between covariates.

Overall, all MDA definitions are misleading with respect to both objectives (i) and (ii) since they include $\text{MDA}^{\star(j)}_3$ in their theoretical counterparts. In the Supplementary Material, we provide an analytical example to show how the MDA can fail to detect relevant covariates when the data has both dependence and interactions. From a practical perspective, it is only possible to conclude in general that the Breiman-Cutler MDA or Ishwaran-Kogalur MDA should be used rather than the Train/Test-MDA. Indeed, on the one hand we only have access to one finite sample $\mathscr{D}_n$ in practice, which has to be split in two parts to use the Train/Test-MDA, hurting the forest accuracy. On the other hand, it is possible to grow many trees at a reasonable linear computational cost, and Proposition 1 ensures that the out-of-bag estimate is efficient in this case. With additional assumptions on the data distribution, the Breiman-Cutler MDA and the Ishwaran-Kogalur MDA recover meaningful theoretical counterparts.

Corollary 1 *If covariates are independent, and if Assumptions 1-3 are satisfied, for all* $M \in \mathbb{N}^\star$ *and* $j \in \{1, \dots, p\}$ *we have*

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

In addition, if Assumption 4 is satisfied,

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

Thus, Corollary 1 states that when covariates are independent, all MDA versions estimate the same quantity, the unnormalized total Sobol index (up to a factor 2), as stated in Gregorutti (2015). However, since the Train/Test-MDA is based on a portion of the training sample, the Breiman-Cutler MDA on the accuracy of a single tree, and the Ishwaran-Kogalur MDA on the accuracy of the forest, the Ishwaran-Kogalur MDA appears to be a more efficient estimate than the two others in this independent setting. Also notice that in the case of independent covariates, the total Sobol index is a relevant measure for both objectives (i) and (ii). Interestingly, when covariates are dependent but without interactions, all MDA versions then estimate the marginal total Sobol index, as stated in the following Corollary.

Corollary 2 *If the regression function m is additive, and if Assumptions 1-3 are satisfied, for all $M \in \mathbb{N}^\star$ and $j \in \{1, \ldots, p\}$ we have*

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

In addition, if Assumption 4 is satisfied,

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

In this correlated and additive setting, the MDA versions now estimate the marginal total Sobol index, which takes the simple form stated in Property 1-(c), but is difficult to estimate with a finite sample because of dependence. The MDA is thus quite relevant for objective (ii): while contributions due to the dependence between covariates are removed in the total Sobol index, it is not the case here. Also notice that covariates with no influence in the regression function are excluded. If we further assume that the regression function is linear, the MDA limits can be written with the linear coefficients and the input variances as stated in Gregorutti et al. (2015); Hooker and Mentch (2019), and also left as an exercise in Chapter 15 of Friedman et al. (2001).

Remark 1 (Distribution Support) *Our asymptotic analysis relies on Assumption 1, which states that the support of the covariate distribution X is a hypercube. Without such geometrical assumption, the support of X_{π_j} may differ from the support of X in the dependent case. It means that the random forest estimate may be applied on regions with no training samples, resulting in inconsistent forest and MDA estimates, and then in a low predictive accuracy (Hooker and Mentch, 2019). This is an additional source of confusion of the MDA when inputs are dependent, induced by the permutation trick.*

4 Sobol-MDA

4.1 Objectives

When covariates are dependent, the MDA fails to estimate the total Sobol index, which is our true objective to solve problem (i), as shown in Section 3. Therefore, we introduce an

improved MDA procedure for random forests: the Sobol-MDA, that consistently estimates the total Sobol index even when covariates are dependent and have interactions. The Sobol-MDA is able to identify the less relevant covariates, as the total Sobol index is the proportion of response explained variance lost when a given covariate is removed from the model. Therefore, a recursive feature elimination procedure based on the Sobol-MDA is highly efficient for our objective (i) of selecting a small number of covariates while maximizing predictive accuracy. Notice that training a random forest without the covariate of interest would also enable to get an estimate of the total Sobol index, and is the approach taken by Williamson et al. (2020). However, the Sobol-MDA only requires to perform forest predictions, which is computationally faster than the forest growing, and scales with the dimension p as opposed to this brute force approach from Williamson et al. (2020). Similarly, Mentch and Hooker (2016) detect influential covariates with hypothesis tests based on the asymptotic normality of random forests and a model retrain without the considered covariate. However, this approach is only valid when the subsampling size a_n is about $\sqrt{n}$, which considerably reduces the accuracy of tree ensembles compared to Breiman's algorithm, and therefore the ability to identify influential covariates. It is also possible to estimate total Sobol indices with existing algorithms which are not specific to random forests. Indeed, this type of methods only requires a black-box estimate to generate predictions from given values of the covariates. Initially, Mara et al. (2015) introduce Monte-Carlo algorithms for the estimation of total Sobol indices in a dependent setting. The first step of the method is to generate a sample from the conditional distributions of the covariates. However, in our setting defined in Assumption 1, we do not have access to these conditional distributions, and their estimation is a difficult problem when only a limited sample $\mathscr{D}_n$ is available. Consequently, the approach of Mara et al. (2015) is not really appropriate for our setting. Notice that the promising approach from Candes et al. (2016) to detect relevant covariates also requires to sample from the conditional distributions of the covariates, and is therefore not adapted to our problem as well.

In the following subsection, we introduce the Sobol-MDA algorithm. Next, we prove the algorithm consistency. In the last two subsections, we show the good empirical behavior of the proposed algorithm through experiments on both simulated and real data, especially when used in a recursive feature elimination procedure.

4.2 Sobol-MDA Algorithm

The key feature of the original MDA procedures is to permute the values of the j-th covariate to break its relation to the response, and then compute the degraded accuracy of the forest. Observe that this is strictly equivalent to drop the original dataset down each tree of the forest, but when a sample hits a split involving covariate j, it is randomly sent to the left or right side with a probability equal to the proportion of points in each child node. This fact highlights that the goal of the MDA is simply to perturb the tree prediction process to cancel out the splits on covariate j. Besides, notice that this point of view on the MDA procedure (using the original dataset and noisy trees) is introduced by Ishwaran (2007) to conduct a theoretical analysis of a modified version of the MDA. Here, our Sobol-MDA algorithm builds on the same principle of ignoring splits on covariate j, such that the noisy CART tree predicts $\mathbb{E}[m(X)|X^{(-j)}]$ (instead of $m(X)$ for the original CART). It enables to recover the proper theoretical counterpart: the unnormalized total Sobol index, i.e., $\mathbb{E}[\mathbb{V}(m(X)|X^{(-j)})]$. To achieve this, we leave aside the

permutation trick, and use another approach to cancel out a given covariate j in the tree prediction process: the partition of the covariate space obtained with the terminal leaves of the original tree is projected along the j-th direction, as shown in Figure 1, and the outputs of the cells of this new projected partition are recomputed with the training data. From an algorithmic point of view, this procedure is quite straightforward as we will see below, and enables to get rid of covariate $X^{(j)}$ in the tree estimate. Then, it is possible to compute the accuracy of the associated out-of-bag projected forest estimate, subtract it from the original accuracy, and normalize the obtained difference by $\mathbb{V}[Y]$ to obtain the Sobol-MDA for covariate $X^{(j)}$.

Interestingly, to compute SHAP values for tree ensembles, Lundberg et al. (2018) also introduce an algorithm to modify the CART predictions to estimate $\mathbb{E}[m(X)|X^{(-j)}]$. More precisely, they propose the following recursive algorithm: the observation x is dropped down the tree, but when a split on covariate j is hit, x is sent to both the left and right children nodes. Then, x falls in multiple terminal cells of the tree. The final prediction is the weighted average of the cell outputs, where the weight associated to a terminal leave A is given by an estimate of $\mathbb{P}(X \in A|X^{(-j)} = x^{(-j)})$: the product of the empirical probabilities to choose the side that leads to A at each split on covariate j in the path of the original tree. At first sight, their approach seems suited to estimate total Sobol indices, but unfortunately, the weights are properly estimated by such procedure only if the covariates are independent. Therefore, as highlighted in Aas et al. (2019), this algorithm gives biased predictions in a correlated setting.

We improve over Lundberg et al. (2018) with the Projected-CART algorithm, formalized in Algorithm 1 in the Supplementary Material: both training and out-of-bag samples are dropped down the tree and sent on both right and left children nodes when a split on covariate j is met. Again, each observation may belong to multiple cells at each level of the tree. For each out-of-bag sample, the associated prediction is the output average over all training observations that belong to the same collection of terminal leaves. In other words, we compute the intersection of these terminal leaves to select the training observations belonging to every cell of this collection to estimate the prediction. This intersection gives the projected cell. Overall, this mechanism is equivalent to projecting the tree partition on the subspace span by $X^{(-j)}$, as illustrated in Figure 1 for $p = 2$ and $j = 2$. Recall that $A_n(X, \Theta)$ is the cell of the original tree partition where X falls, whereas the associated cell of the projected partition is denoted $A_n^{(-j)}(X^{(-j)}, \Theta)$. Formally, we respectively denote the associated projected tree and projected out-of-bag forest estimates as $m_n^{(-j)}(X^{(-j)}, \Theta)$ and $m_{M,n}^{(-j,OOB)}(X_i^{(-j)}, \Theta_{(M)})$, respectively defined by

$$m_n^{(-j)}(X^{(-j)}, \Theta) = \frac{\sum_{i=1}^{a_n} Y_i \mathbb{1}_{X_i \in A_n^{(-j)}(X^{(-j)}, \Theta)}}{\sum_{i=1}^{a_n} \mathbb{1}_{X_i \in A_n^{(-j)}(X^{(-j)}, \Theta)}},$$

and for $i \in \{1, \dots, n\}$,

$$m_{M,n}^{(-j,OOB)}(X_i^{(-j)}, \Theta_{(M)}) = \frac{1}{|\Lambda_{n,i}|} \sum_{\ell \in \Lambda_{n,i}} m_n^{(-j)}(X_i^{(-j)}, \Theta_\ell) \mathbb{1}_{|\Lambda_{n,i}|>0}.$$

The Projected-CART algorithm provides two sources of improvements over Lundberg et al. (2018): first, the training data points are dropped down the modified tree to recompute the cell outputs, and thus $\mathbb{E}[m(X)|X^{(-j)} \in A]$ is directly estimated in each cell. Secondly, the

projected partition is finer than in the original tree, which mitigates masking effects (when an influential covariate is not often selected in the tree splits because of other highly correlated covariates).

Finally, the Sobol-MDA estimate is given by the normalized difference of the squared error of the out-of-bag projected forest with the out-of-bag error of the original forest. Formally, we define the Sobol-MDA as

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

where $\hat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the standard variance estimate of the response Y. An implementation in `R` and `C++` of the Sobol-MDA is available at `https://gitlab.com/drti/sobolmda` and is based on `ranger` (Wright and Ziegler, 2017), a fast implementation of random forests. Given an initial random forest, the Sobol-MDA algorithm has a computational complexity of $O(Mn\log^3(n))$, which is in particular independent of the dimension p, and quasi-linear with the sample size n. On the other hand, the brute force approach from Williamson et al. (2020) has a complexity of $O(Mp^2n\log^2(n))$, which is quadratic with the dimension p and therefore intractable in high-dimensional settings, as opposed to the Sobol-MDA. Additional details are provided in the Supplementary Material.

Remark 2 (Empty Cells) *Some cells of the projected partition may contain no training samples. Consequently, the prediction for a new query point falling in such cells is undefined. In practice, the Projected-CART algorithm uses the following strategy to avoid empty cells. Recall that each level of the tree defines a partition of the input space (if a terminal leave occurs before the final tree level, it is copied down the tree at each level), and that a projected partition can thus be associated to each tree level. When a new observation is dropped down the tree, if it falls in an empty cell of the projected partition at a given tree level, the prediction is computed using the previous level. Notice that empty cells cannot occur in the partitions associated to the root and the first level of the tree by construction. Therefore, this mechanism enforces that the projected tree estimate is well defined over the full covariate space.*

4.3 Sobol-MDA Consistency

The original MDA versions do not converge towards the total Sobol index, which is the relevant quantity for our objective (i), as stated in Proposition 2. On the other hand, the Sobol-MDA is consistent as stated below. Before introducing this convergence result, we need to introduce additional assumptions. Indeed, in Section 3, we show the convergence of the different MDA versions provided that the forest is an efficient estimate, i.e. consistent. To enforce the consistency of random forests, we used Assumption 2 which controls the variation of the regression function in each cell of the theoretical tree: $\Delta(m, A_k^\star(x,\Theta)) \xrightarrow{a.s.} 0$. Because the covariates may be dependent, Assumption 2 does not imply the same property for the projected partition. Therefore, we cannot directly build on the consistency result from Scornet et al. (2015) to prove the consistency of the Sobol-MDA. Thus, we take another route and define a new Assumption 5 which brings two modifications to the random forest algorithm.

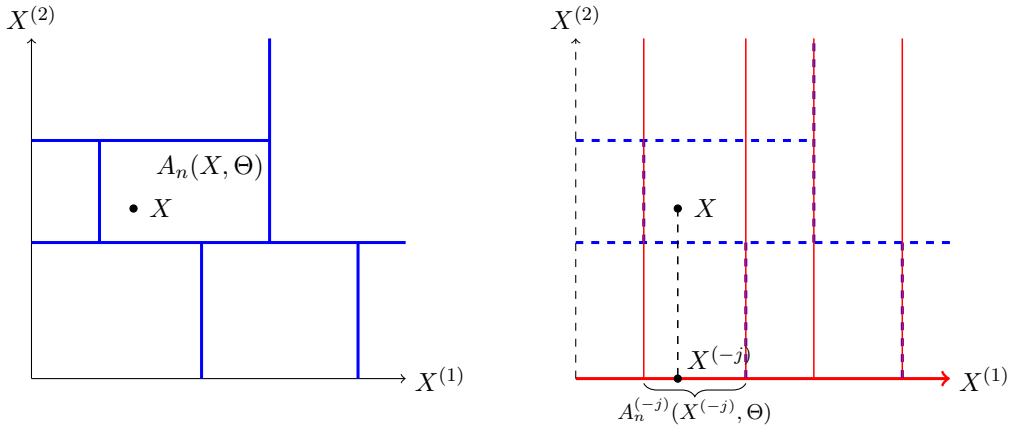

Figure 1: Example of the partition of $[0,1]^2$ by a random CART tree (left side) projected on the subspace span by $X^{(-2)} = X^{(1)}$ (right side). Here, $p = 2$ and $j = 2$.

Assumption 5 *A node split is constrained to generate child nodes with at least a small fraction $\gamma > 0$ of the parent node observations. Secondly, the split selection is slightly modified: at each tree node, the number* `mtry` *of covariates drawn to optimize the split is set to* `mtry` $= 1$ *with a small probability $\delta > 0$. Otherwise, with probability $1 - \delta$, the default value of* `mtry` *is used.*

Importantly, since γ and δ can be chosen arbitrarily small, the modifications of Assumption 5 are mild. Besides, notice that this assumption follows Meinshausen (2006) and Wager and Athey (2018): we slightly modify the random forest algorithm to enforce empirical cells to become infinitely small as the sample size increases. The projected forest inherits this property and an asymptotic analysis from Györfi et al. (2006) gives the consistency of the Sobol-MDA, provided that the complexity of tree partitions is appropriately controlled. If an original tree has t_n terminal leaves, the associated projected partition may have a higher number of terminal leaves, at most 2^{t_n}. Thus, we introduce Assumption 6, which slightly modifies Assumption 3 with a more restrictive regime for the number of terminal leaves t_n in the original trees.

Assumption 6 *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 $\kappa > 0$, $\lim_{n\to\infty} a_n = \infty$, $\lim_{n\to\infty} t_n = \infty$, and $\lim_{n\to\infty} 2^{t_n} \frac{(\log(a_n))^9}{a_n} = 0$.*

The Projected-CART algorithm ignores the splits based on the j-th covariate, and the associated out-of-bag projected forest consistently estimates $\mathbb{E}[m(X)|X^{(-j)}]$ under Assumptions 1, 5, and 6, which leads to the consistency of the Sobol-MDA as stated in the theorem below.

Theorem 2 *If Assumptions 1, 5, and 6 are satisfied, for all $M \in \mathbb{N}^\star$ and $j \in \{1, \ldots, p\}$*

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

Theorem 2 shows that the proposed Sobol-MDA algorithm consistently estimates the total Sobol index. Therefore, the Sobol-MDA targets the appropriate quantity for objective (i), of selecting a small number of covariates while maximizing accuracy, whereas original MDA versions target a bias quantity, as stated in Proposition 2.

	BC-MDA*	$\widehat{\text{BC-MDA}}$	IK-MDA*	$\widehat{\text{IK-MDA}}$	ST*	$\widehat{\text{S-MDA}}$	$\widehat{\psi_{n,j}}$	$\widehat{\text{S-MDA}}_{Ldg}$
$X^{(3)}$	0.47	0.37 (0.03)	0.47	0.43 (0.02)	**0.47**	0.45 (0.03)	0.42 (0.06)	0.43 (0.03)
$X^{(4)}$	0.21	0.10 (0.02)	0.37	0.14 (0.01)	**0.10**	0.08 (0.01)	0.06 (0.04)	0.13 (0.01)
$X^{(5)}$	0.21	0.09 (0.01)	0.37	0.13 (0.01)	**0.10**	0.08 (0.01)	0.06 (0.04)	0.13 (0.01)
$X^{(1)}$	0.64	0.24 (0.02)	1.0	0.29 (0.02)	**0.07**	0.05 (0.01)	0.03 (0.04)	0.22 (0.02)
$X^{(2)}$	0.64	0.24 (0.02)	1.0	0.28 (0.02)	**0.07**	0.05 (0.01)	0.03 (0.04)	0.23 (0.01)

Table 2: BC-MDA (normalized by $2\mathbb{V}[Y]$), IK-MDA (normalized by $\mathbb{V}[Y]$), Williamson et al. (2020) ($\widehat{\psi_{n,j}}$), and Sobol-MDA estimates for Example 1 (standard deviations over 10 repetitions in brackets). Theoretical counterparts are defined in Proposition 2.

4.4 Experiments with Simulated Data

We conduct three batches of experiments. First, we use the analytical example of the Supplementary Material, and show empirically that the Sobol-MDA leads to the accurate importance variable ranking, while original MDA versions do not. Next, we simulate a typical setting where several groups of covariates are strongly correlated and only few covariates are involved in the regression function. In such difficult setting, the Sobol-MDA identifies the relevant covariates, as opposed to its competitors. Finally, we apply the RFE on real data to show the performance improvement of the Sobol-MDA for variable selection.

We first consider the analytical example of the Supplementary Material, where the data has both dependence and interactions. In this example, the covariates are distributed as a Gaussian vector with $p = 5$, and the regression function is given by

$$m(X) = \alpha X^{(1)} X^{(2)} \mathbb{1}_{X^{(3)}>0} + \beta X^{(4)} X^{(5)} \mathbb{1}_{X^{(3)}<0}.$$

Here, we set $\alpha = 1.5$, $\beta = 1$, $\mathbb{V}[X^{(j)}] = 1$ for all covariates $j \in \{1, \dots, 5\}$. The correlation coefficients are set to $\rho_{1,2} = 0.9$ and $\rho_{4,5} = 0.6$, and other covariance terms are null. Finally, we define the model response as $Y = m(X) + \varepsilon$, where ε is an independent centered Gaussian noise whose variance verifies $\mathbb{V}[\varepsilon]/\mathbb{V}[Y] = 10\%$. Then, we run the following experiment: first, we generate a sample $\mathscr{D}_n$ of size $n = 3000$ and distributed as the Gaussian vector X. Next, a random forest of $M = 300$ trees is fit with $\mathscr{D}_n$ and we compute the Breiman-Cutler MDA, Ishwaran-Kogalur MDAthe algorithm from Williamson et al. (2020) denoted by $\widehat{\psi_{n,j}}$, and the Sobol-MDA. To enable comparisons, the Breiman-Cutler MDA is normalized by $2\mathbb{V}[Y]$, and the Ishwaran-Kogalur MDA by $\mathbb{V}[Y]$, as suggested by Proposition 2. To show the improvement of our Projected-CART algorithm, we also compute the Sobol-MDA using the algorithm from Lundberg et al. (2018), denoted $\widehat{\text{S-MDA}}_{Ldg}$. All results are reported in Table 2, along with the theoretical counterparts of the estimates. Notice that the associated standard deviations are reported in brackets, and that the covariates are ranked by decreasing values of the theoretical total Sobol index since it is the value of interest: $X^{(3)}$, then $X^{(4)}$ and $X^{(5)}$, and finally $X^{(1)}$ and $X^{(2)}$. Thus, only the Sobol-MDA computed with the Projected-CART algorithm and Williamson et al. (2020) rank the covariates in the same appropriate order than the total Sobol index. In particular, $X^{(4)}$ and $X^{(5)}$ have a higher total Sobol index than covariates 1 and 2 because of the stronger correlation between $X^{(1)}$ and $X^{(2)}$ than between $X^{(4)}$ and $X^{(5)}$.

For all the other importance measures, $X^{(1)}$ and $X^{(2)}$ are more important than $X^{(4)}$ and $X^{(5)}$. For the original MDA, this is due to the higher coefficient $\alpha = 1.5 > \beta = 1$, to the term $\text{MDA}_2^{\star(j)}$, and especially to $\text{MDA}_3^{\star(j)}$ which increases with correlation. Since the explained variance of the random forest is 82% in this experiment, all estimates have a negative bias. The bias of the Breiman-Cutler MDA and Ishwaran-Kogalur MDA dramatically increases with correlation. Indeed, a strong correlation between covariates leaves some regions of the input space free of training data. However, the out-of-bag permuted sample may fall in these regions, regions for which the forest has to extrapolate, resulting in a low predictive accuracy. This phenomenon combined with the $\text{MDA}_3^{\star(j)}$ component explains the high bias of the Breiman-Cutler MDA and Ishwaran-Kogalur MDA for correlated covariates. Also observe that since $X^{(3)}$ is independent of the other covariates, the bias is small for both MDA versions, and it is smaller for the Ishwaran-Kogalur MDA than the Breiman-Cutler MDA as the forest estimate is more accurate than a single tree. Finally, the Sobol-MDA computed with the algorithm of (Lundberg et al., 2018) is biased as suggested by (Aas et al., 2019), and the bias also seems to increase with correlation.

We then consider the following problem inspired by Archer and Kimes (2008); Gregorutti et al. (2017) and related to gene expressions. The goal is to identify relevant covariates among several groups of many strongly correlated covariates. More precisely, we define X, a random vector of dimension $p = 200$, composed of 5 independent groups of 40 covariates. Each group is a centered gaussian random vector where two distinct components have a correlation of 0.8 and the variance of each component is 1. The regression function m only involves one covariate from each group, and is simply defined by

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

Finally, we define the model response as $Y = m(X) + \varepsilon$, where ε is an independent gaussian noise ($\mathbb{V}[\varepsilon]/\mathbb{V}[Y] = 10\%$). Next, a sample of size $n = 1000$ is generated based on the distribution of X, and a random forest of $M = 300$ trees is fit. Thus, Tables 3 and 4 show that the Sobol-MDA identifies the five relevant covariates, whereas the Breiman-Cutler MDA, Ishwaran-Kogalur MDA, and Williamson et al. (2020) identify some noisy covariates among the top five. In this additive and correlated example, Corollary 2 states that all MDA algorithms have an appropriate theoretical counterpart to identify the five relevant covariates involved in the regression function, because these five covariates are mutually independent. However, in this finite sample setting, the original MDA versions give a high importance to the covariates of the first group because of their correlation with the most influential covariate $X^{(1)}$. Since the Ishwaran-Kogalur MDA is based on the forest error, it outperforms the Breiman-Cutler MDA, which relies on the tree error. Quite surprisingly, Williamson et al. (2020) is the worst performing algorithm although it uses a brute force approach by retraining the forest without a given covariate to consistently estimate its total Sobol index, the appropriate theoretical counterpart. In fact, the multiple layers of data splitting involved in Williamson et al. (2020) generate a high variance of the associated estimate, whereas the Breiman-Cutler MDA, Ishwaran-Kogalur MDA, and Sobol-MDA operate with a given dataset and a given initial forest structure to compute the decrease of accuracy, resulting in finer estimates and a higher performance to detect irrelevant covariates.

$\widehat{\text{S-MDA}}$		$\widehat{\text{BC-MDA}}/2\mathbb{V}[Y]$		$\widehat{\text{IK-MDA}}/\mathbb{V}[Y]$		$\widehat{\psi_{n,j}}$	
$X^{(1)}$	0.035	$X^{(1)}$	0.048	$X^{(1)}$	0.056	$X^{(1)}$	0.042
$X^{(161)}$	0.005	$X^{(25)}$	0.010	$X^{(5)}$	0.009	$X^{(119)}$	0.031
$X^{(81)}$	0.004	$X^{(31)}$	0.008	$X^{(81)}$	0.007	$X^{(155)}$	0.029
$X^{(121)}$	0.004	$X^{(14)}$	0.008	$X^{(41)}$	0.005	$X^{(24)}$	0.029
$X^{(41)}$	0.002	$X^{(40)}$	0.007	$X^{(161)}$	0.005	$X^{(54)}$	0.029
$X^{(179)}$	0.002	$X^{(3)}$	0.007	$X^{(15)}$	0.005	$X^{(72)}$	0.028
$X^{(13)}$	0.001	$X^{(17)}$	0.006	$X^{(121)}$	0.005	$X^{(103)}$	0.028
$X^{(25)}$	0.001	$X^{(26)}$	0.006	$X^{(7)}$	0.005	$X^{(124)}$	0.027
$X^{(73)}$	0.001	$X^{(41)}$	0.006	$X^{(4)}$	0.004	$X^{(60)}$	0.027
$X^{(155)}$	0.001	$X^{(121)}$	0.006	$X^{(28)}$	0.004	$X^{(185)}$	0.027

Table 3: Normalized BC-MDA, normalized IK-MDA, and Sobol-MDA estimates (influential covariates in blue) for Example 2.

$\widehat{\text{S-MDA}}$	$\widehat{\text{BC-MDA}}$	$\widehat{\text{IK-MDA}}$	$\widehat{\psi_{n,j}}$
0.90	0	0.33	0

Table 4: Probability to recover the 5 relevant covariates in Example 2 as the top 5 most important covariates ranked using the BC-MDA, IK-MDA, Sobol-MDA, and Williamson et al. (2020).

4.5 Experiments for Variable Selection with Real Data

The Recursive Feature Elimination algorithm (RFE) is originally introduced by Guyon et al. (2002) to perform variable selection with SVM. Gregorutti et al. (2017) apply the recursive feature elimination algorithm to random forests with the MDA as importance measure. The principle is to discard the less relevant covariates one by one, and is summarized in Algorithm 2 in the Supplementary Material. Thus, the recursive feature elimination algorithm is a relevant strategy for our objective (i) of building a model with a high accuracy and a small number of covariates. At each step of the algorithm, the goal is to detect the less relevant covariates based on the trained model. Since the total Sobol index measures the proportion of explained response variance lost when a given covariate is removed, the optimal strategy is therefore to discard the covariate with the smallest total Sobol index. As the Sobol-MDA directly estimates the total Sobol index whereas existing MDA all have additional noisy terms, using the Sobol-MDA improves the performance of the procedure, as shown in the following experiments.

The recursive feature elimination algorithm is illustrated with the "Ozone" data (Dua and Graff, 2017) and the high-dimensional dataset "HIV" as suggested in Williamson et al. (2020). The algorithm is run four times, respectively using the Breiman-Cutler MDA, Williamson et al. (2020), Ishwaran-Kogalur MDA, and the Sobol-MDA as importance measures to iteratively discard the less relevant covariate. At each step of the recursive feature elimination algorithm, the explained variance of the forest is retrieved. Following Gregorutti et al. (2017), we do not use the out-of-bag error since it gives optimistically bias results, but use instead a 10-

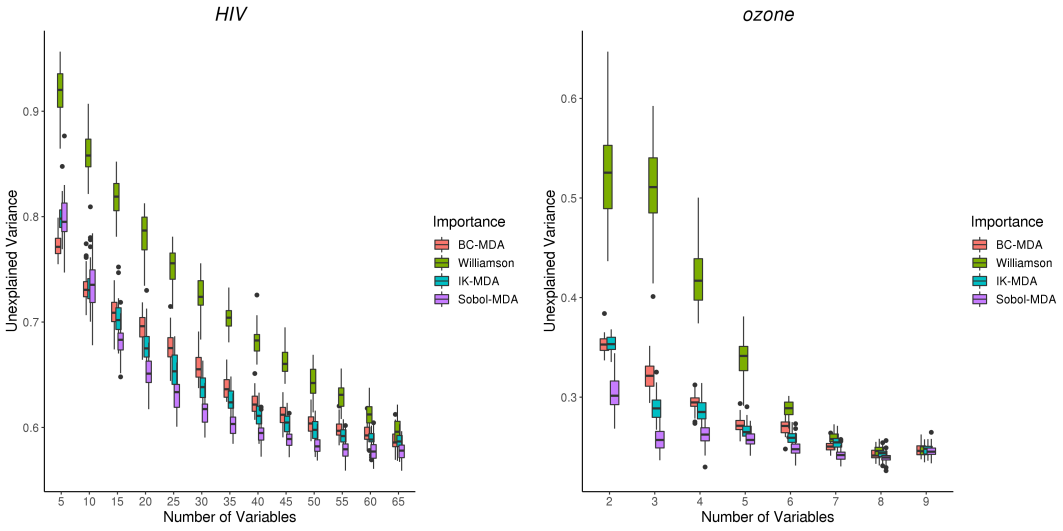

Figure 2: Random forest error versus the number of covariates for the "HIV" and "Ozone" datasets at each step of the RFE, using different importance measures.

fold cross-validation: the forest and the associated importance measure are computed with 9 folds, and the error is estimated with the 10-th fold. For each dataset, the cross-validation is repeated 40 times to get the result uncertainties, displayed as boxplots in the figures. Thus, Figure 2 highlights that the Sobol-MDA leads to a more efficient variable selection than the Breiman-Cutler MDA, Williamson et al. (2020), and the Ishwaran-Kogalur MDA for the "HIV" and "Ozone" datasets. We refer to the Supplementary Material for additional experiments. Notice that the Ishwaran-Kogalur MDA performs better than the Breiman-Cutler MDA, as expected from their theoretical counterparts stated in Proposition 2. Finally the algorithm from Williamson et al. (2020) is the worst performing approach because of the data splitting procedure, as explained in the previous subsection.

5 Conclusion

Variable importance is the main approach to analyze the black-box mechanisms of random forests, and the MDA is the most widely used importance measure. However, many empirical studies have shown that when covariates are dependent, the MDA fails to detect influential covariates. We conducted a theoretical analysis to understand this undesirable behavior. First, a close inspection of the literature and the main random forest software show that different definitions coexist: the Train-Test MDA, the Breiman-Cutler MDA, and the Ishwaran-Kogalur MDA. An asymptotic analysis shows that these different MDA versions do not converge towards the appropriate theoretical quantity when covariates are dependent, and are thus misleading for both objectives (i) and (ii) of variable importance. Therefore, we propose an augmented MDA algorithm: the Sobol-MDA, which consistently estimates the total Sobol index, i.e. the appropriate theoretical counterpart which tells how much explained variance of the response is lost when a given covariate is removed from the model, at an efficient com-

putational cost. We run many experiments to show the good empirical performance of the Sobol-MDA, especially the performance improvement over competitors for variable selection through the recursive feature elimination algorithm. An implementation in `R` and `C++` of the Sobol-MDA is available at `https://gitlab.com/drti/sobolmda`.

Acknowledgement

We thank the referees and the editors for their relevant suggestions to improve the article.

References

K. Aas, M. Jullum, and A. Løland. Explaining individual predictions when features are dependent: more accurate approximations to Shapley values. *arXiv preprint arXiv:1903.10464*, 2019.

A. Antoniadis, S. Lambert-Lacroix, and J.-M. Poggi. Random forests for global sensitivity analysis: a selective review. *Reliability Engineering & System Safety*, 206:107–312, 2020.

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

L. Auret and C. Aldrich. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, 105:157–170, 2011.

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

A.-L. Boulesteix, S. Janitza, J. Kruppa, and I.R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2:493–507, 2012.

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

L. Breiman. Setting up, using, and understanding random forests v3.1. Technical report, UC Berkeley, Department of Statistics, 2003a.

L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman & Hall/CRC, Boca Raton, 1984.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: model-X knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016.

R. Díaz-Uriarte and S.A. De Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:1–13, 2006.

D. Dua and C. Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics New York, 2001.

R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31:2225–2236, 2010.

R. Ghanem, D. Higdon, and H. Owhadi. *Handbook of Uncertainty Quantification*. Springer, New York, 2017.

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. Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis*, 90:15–35, 2015.

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

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46:389–422, 2002.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer, New York, 2006.

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

B. Iooss and C. Prieur. Shapley effects for sensitivity analysis with correlated inputs: comparisons with Sobol'indices, numerical estimation and applications. *arXiv preprint arXiv:1707.01334*, 2017.

Bertrand Iooss and Paul Lemaître. A review on global sensitivity analysis methods. In *Uncertainty Management in Simulation-Optimization of Complex Systems*, pages 101–122. Springer, Boston, 2015.

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

H. Ishwaran and U.B. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2020. URL https://cran.r-project.org/package=randomForestSRC. R package version 2.9.3.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2:841–860, 2008.

S. Kucherenko, S. Tarantola, and P. Annoni. Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications*, 183:937–946, 2012.

X. Li, Y. Wang, S. Basu, K. Kumbier, and B. Yu. A debiased MDI feature importance measure for random forests. In *Advances in Neural Information Processing Systems*, volume 32, pages 8049–8059, New York, 2019. Curran Associates, Inc.

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2:18–22, 2002.

Markus Loecher. Unbiased variable importance for random forests. *Communications in Statistics-Theory and Methods*, pages 1–13, 2020.

S.M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, New York, 2017. Curran Associates, Inc.

S.M. Lundberg, G.G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

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.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 2006.

L. Mentch and G. Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17:841–881, 2016.

Lucas Mentch and Siyu Zhou. Getting better from worse: augmented bagging and a cautionary tale of variable importance. *arXiv preprint arXiv:2003.03629*, 2020.

K.K. Nicodemus and J.D. Malley. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, 25:1884–1890, 2009.

A.B. Owen. Sobol'indices and Shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2:245–251, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145:280–297, 2002.

E. Scornet. Trees, forests, and impurity-based variable importance. *arXiv preprint arXiv:2001.04295*, 2020.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *The Annals of Statistics*, 43:1716–1741, 2015.

L.S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2:307–317, 1953.

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

C. Strobl and A. Zeileis. Danger: High power!–exploring the statistical properties of a test for random forest variable importance. In *Proceedings of the 18th International Conference on Computational Statistics*, Porto, Portugal, 2008.

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.

C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, 2008.

L. Toloşi and T. Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27:1986–1994, 2011.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113:1228–1242, 2018.

B.D. Williamson, P.B. Gilbert, N.R. Simon, and M. Carone. A unified approach for inference on algorithm-agnostic variable importance. *arXiv preprint arXiv:2004.03683*, 2020.

M.N. Wright and A. Ziegler. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77:1–17, 2017.

Z. Zhou and G. Hooker. Unbiased measurement of feature importance in tree-based methods. *arXiv preprint arXiv:1903.05179*, 2019.

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

Supplementary Material for "MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA"

1 Analytical Example for the MDA

To illustrate the behavior of the MDA, we take a simple example and analytically derive the MDA limit and its three associated components $\text{MDA}_1^{\star(j)}$, $\text{MDA}_2^{\star(j)}$, and $\text{MDA}_3^{\star(j)}$. This example shows how the MDA is misleading when input variables are dependent. We consider the Breiman-Cutler MDA, denoted by MDA to lighten notations. The TT-MDA or Ishwaran-Kogalur MDA lead to identical conclusions.

The input $\mathbf{X}$ is a Gaussian vector of dimension $p = 5$. Its covariance matrix is defined by $\mathbb{V}[X^{(j)}] = \sigma_j^2$ for $j \in \{1, \ldots, 5\}$, and all covariance terms are null except $\text{Cov}[X^{(1)}, X^{(2)}] = \rho_{1,2}\sigma_1\sigma_2$ and $\text{Cov}[X^{(4)}, X^{(5)}] = \rho_{4,5}\sigma_4\sigma_5$. The regression function m is given by

$$m(\mathbf{X}) = \alpha X^{(1)} X^{(2)} \mathbb{1}_{X^{(3)}>0} + \beta X^{(4)} X^{(5)} \mathbb{1}_{X^{(3)}<0}.$$

Notice that m has a simple form to enable an easy interpretation of the importance measures, but that interaction terms are required to highlight the different behaviors of the three MDA components in a correlated setting. Simple calculations give the analytical expression $\text{MDA}^{\star(1)}$ of the MDA limit for $X^{(1)}$ as

$$\text{MDA}^{\star(1)} = \underbrace{\frac{1}{2}(\alpha\sigma_1\sigma_2)^2(1-\rho_{1,2}^2)}_{\text{MDA}_1^{\star(1)}} + \underbrace{\frac{1}{2}(\alpha\sigma_1\sigma_2)^2}_{\text{MDA}_2^{\star(1)}} + \underbrace{\frac{3}{2}\rho_{1,2}^2(\alpha\sigma_1\sigma_2)^2}_{\text{MDA}_3^{\star(1)}}.$$

First, observe that $\text{MDA}_1^{\star(1)}$ decreases with the correlation between $X^{(1)}$ and $X^{(2)}$. Indeed, $\text{MDA}_1^{\star(1)}$ is the total Sobol index and when these two variables are strongly dependent, the additional information provided by $X^{(1)}$ alone is small. In the extreme case, $\rho_{1,2} = 1$ implies that $\text{MDA}_1^{\star(1)} = 0$, i.e., $X^{(1)}$ can be removed from the model without hurting the model accuracy since all its information is contained in $X^{(2)}$. On the other hand, $\text{MDA}_2^{\star(1)}$ does not rely on the dependence between $X^{(1)}$ and $X^{(2)}$. Indeed, recall that in the case of $\text{MDA}_1^{\star(1)}$, contributions due to the dependence between $X^{(1)}$ and $X^{(2)}$ are excluded because of the conditioning on $X^{(2)}$. For $\text{MDA}_2^{\star(1)}$, this dependence is ignored, and therefore such removal does not take place. Therefore, it is clear that the MDA mixes two terms with opposite meanings. Finally, the third term $\text{MDA}_3^{\star(1)}$ measures how the permutation of $X^{(1)}$ shifts the mean value of the regression function averaged over $X^{(1)}$, which is not a quantity of interest to rank variables. However, in a high correlation setting $\left(\rho_{1,2} > \frac{\sqrt{2}}{2}\right)$, we have $\text{MDA}_3^{\star(1)} > \text{MDA}_1^{\star(1)} + \text{MDA}_2^{\star(1)}$, which means that the meaningless third term is the main contribution of the MDA value of variable $X^{(1)}$. Besides, symmetrically for the other input variables, we have $\text{MDA}^{\star(1)} = \text{MDA}^{\star(2)}$, and the

same formula for $X^{(4)}$ and $X^{(5)}$ with the appropriate parameters. MDA formulas for variables $3, 4$, and 5 are to be found in the last section of the Supplementary Material.

As stated in the introduction, one of the main objective of variable importance analysis is usually to select a small number of variables while maximizing the model accuracy. In our example, we show how the MDA fails for this purpose. Let say we want to remove the less relevant input variable in a setting where the two vectors $\mathbf{X}^{(1,2)}$ and $\mathbf{X}^{(4,5)}$ are interchangeable ($\alpha\sigma_1\sigma_2 = \beta\sigma_4\sigma_5$), except that their dependence strengths differ and satisfy $\rho_{1,2} < \rho_{4,5}$. Since the correlation between variables 4 and 5 is higher than between variables 1 and 2, we should remove $X^{(4)}$ or $X^{(5)}$ to minimize the information loss, as suggested by the total Sobol index ranking

$$ST^{(4)} = ST^{(5)} < ST^{(1)} = ST^{(2)} < ST^{(3)}.$$

However, in such setting we have

$$\mathrm{MDA}^{\star(1)} = \mathrm{MDA}^{\star(2)} < \mathrm{MDA}^{\star(3)} < \mathrm{MDA}^{\star(4)} = \mathrm{MDA}^{\star(5)},$$

that would lead to discard $X^{(1)}$ or $X^{(2)}$, which is suboptimal—see the last section of the Supplementary Material for computation details. On the other hand, using only $\mathrm{MDA}_1^{\star(j)}$ or $\mathrm{MDA}_1^{\star(j)} + \mathrm{MDA}_2^{\star(j)}$ as importance measures gives the accurate variable selection. The term $\mathrm{MDA}_3^{\star(j)}$ artificially increases the MDA value because of correlation, and is thus misleading for both objectives (i) and (ii).

2 Algorithms

2.1 Ishwaran-Kogalur MDA by Blocks

The Ishwaran-Kogalur MDA is implemented in `randomForestSRC`. This package also provides the possibility to define the Ishwaran-Kogalur MDA by blocks: the trees of the forest are divided in a fixed number of blocks. The Ishwaran-Kogalur MDA is estimated for each block and then averaged. Thus, the Breiman-Cutler MDA can be seen as a specific case where the number of blocks is the number of trees M, and each block contains only one tree. On the theoretical side, if the number of blocks is fixed and Assumption 4 is satisfied, the number of trees in each block grows to infinity, and therefore Theorem 1-(iii) still holds.

2.2 Sobol-MDA Computational Complexity

Recall that the computational complexity of the brute force approach of Williamson et al. (2020), where a forest is retrained without each input variable, is $O(Mp^2n\log^2(n))$, which is quadratic with the dimension p and therefore intractable in high-dimensional settings.

On the other hand, the original MDA procedure has an average complexity of $O(Mpn\log(n))$: to run a balanced tree prediction for a given data point, it is dropped down the $\log(n)$ levels of the tree, which makes a complexity of $O(n\log(n))$ for the full out-of-bag sample, repeated for the M trees of the forest and the p variables. In the Sobol-MDA procedure, the complexity

analysis is similar, except that when a point is dropped down the tree, it can be sent to both the left and right children nodes, generating multiple operations at a given tree level and then an additional multiplicative factor of $\log(n)$. However, it is not necessary to run the Projected-CART algorithm for each of the p covariates. Indeed, when a given observation is dropped down the tree, it meets at most $\log(n)$ different variables in the original tree path. Therefore, the Projected-CART prediction has to be computed only for $\log(n)$ covariates for each observation. Thus, the Sobol-MDA algorithm has a computational complexity of $O(Mn\log^3(n))$, which is in particular independent of the dimension p, and quasi-linear with the sample size n.

2.3 Projected-CART

We provide below Algorithm 1 for an implementation of the projected random forests.

Algorithm 1 Projected-CART

Input: A Θ-random CART built with $\mathscr{D}_n$, and a variable index $j \in \{1, \ldots, p\}$. (Note that if a terminal leave occurs before the final tree level, it is copied at each level down the tree.)
Initialize both in-bag and OOB samples at the root node of the tree;
for all tree levels:
    for all level nodes:
        if the splitting variable is not j:
            send each data point to the right or left children node according to the node split;
        if the splitting variable is j:
            send the node sample to both the right and left children node ignoring the split;
    for all data points:
        retrieve the collection of nodes where the data point falls at the current tree level;
    for all OOB data points:
        retrieve the set of in-bag points which fall in the same node collection;
        if all nodes in the considered node collection are terminal:
            compute the output average of the in-bag points;
            set this average as the prediction for the considered OOB observation;
        if no in-bag points fall in the same node collection:
            retrieve the corresponding in-bag data points at the previous tree level;
            set the output average of these in-bag points as the prediction for the considered OOB observation;
return predictions;

2.4 Recursive Feature Elimination

Figures 3 and 4 provide additional experiments to show that the Sobol-MDA leads to a more efficient variable selection than the Breiman-Cutler MDA, Williamson et al. (2020), and the Ishwaran-Kogalur MDA. Notice that Algorithm 2 recalls the RFE procedure. The "Prostate"

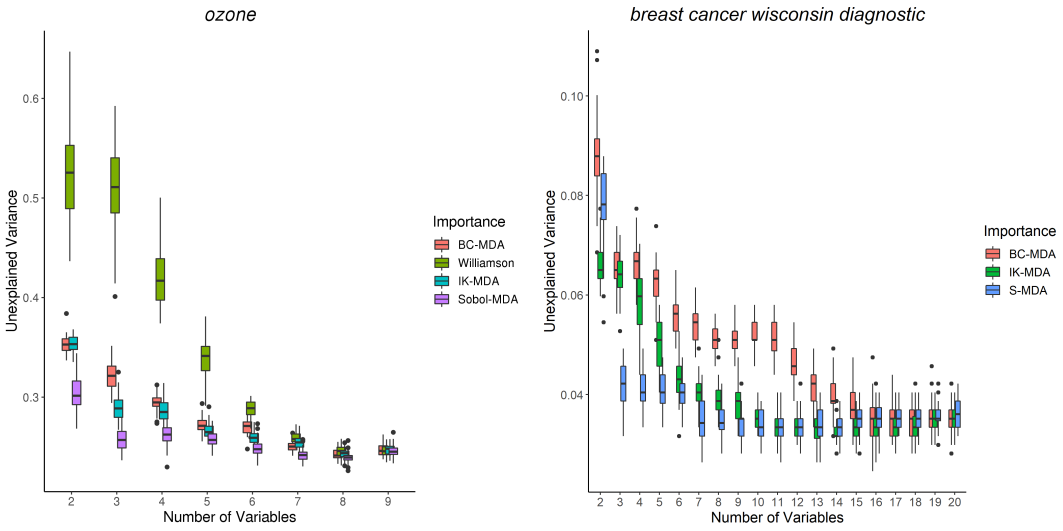

Figure 3: Random forest error versus the number of variables for the "Ozone" and "Breast Cancer Wisconsin Diagnostic" datasets at each step of the RFE, using different importance measures: BC-MDA, Williamson et al. (2020), IK-MDA, and Sobol-MDA.

dataset in Figure 4 is an example where the Sobol-MDA does not significantly improve over the original MDA.

Algorithm 2 Recursive Feature Elimination

1: for j in $1, \ldots, p$:
2: train a random forest
3: compute the MDA for all variables
4: remove the variable with the smallest MDA
5: return the ordered list of removed variables

3 Proof of the MDA Consistency

3.1 Assumptions and Theorem 1

We recall Assumptions 1, 2, 3, 4, Proposition 1, and Theorem 1 for the sake of clarity.

Assumption 1 *The response $Y \in \mathbb{R}$ follows*

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

where $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(p)}) \in [0,1]^p$ admits a density over $[0,1]^p$ bounded from above and below by strictly positive constants, m is continuous, and the noise ε is sub-Gaussian, independent of $\boldsymbol{X}$, and centered. A sample $\mathscr{D}_n = \{(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)\}$ of n independent random variables distributed as $(\boldsymbol{X}, Y)$ is available.

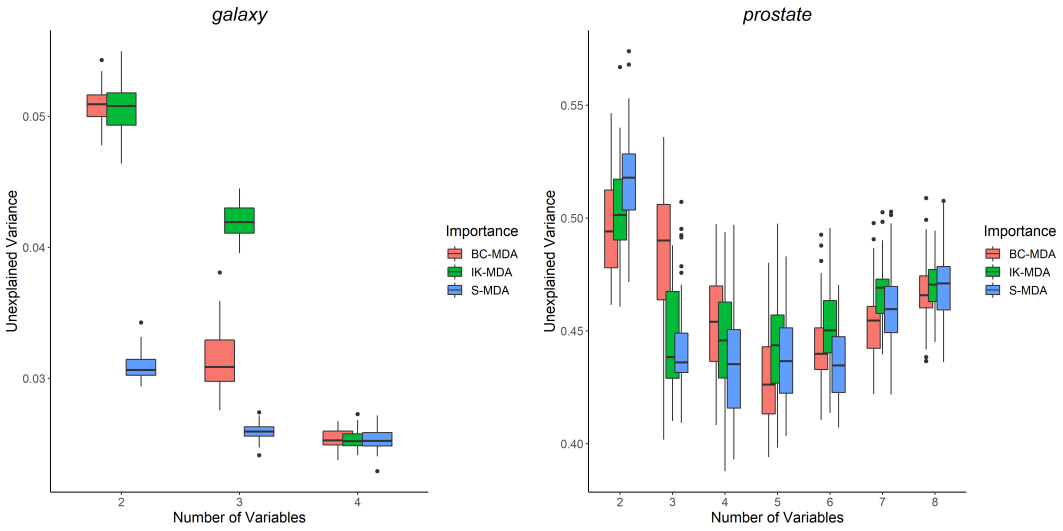

Figure 4: Random forest error versus the number of variables for the "Galaxy" and "Prostate" datasets at each step of the RFE, using different importance measures: BC-MDA, IK-MDA, and Sobol-MDA.

Assumption 2 *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,*

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

Assumption 3 *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* $\kappa > 0$*,* $\lim_{n\to\infty} a_n = \infty$*,* $\lim_{n\to\infty} t_n = \infty$*, and* $\lim_{n\to\infty} t_n \frac{(\log(a_n))^9}{a_n} = 0$*.*

Assumption 4 *The number of trees grows to infinity with the sample size* n*:* $M \underset{n\to\infty}{\longrightarrow} \infty$*.*

Proposition 1 *If Assumption 1 is satisfied, for a fixed* n *and* $i \in \{1, \ldots, n\}$*, we have*

$$\left| \mathbb{E}\left[\left(m^{(OOB)}_{M,a_n,n}(\boldsymbol{X}_i, \Theta_M) - m(\boldsymbol{X}_i) \right)^2 \right] - \mathbb{E}\left[\left(m_{M,a_n,n-1}(\boldsymbol{X}, \Theta_M) - m(\boldsymbol{X}) \right)^2 \right] \right| = O\left(\frac{1}{M}\right).$$

Theorem 1 *If Assumptions 1, 2, and 3 are satisfied, then, for all* $M \in \mathbb{N}^\star$ *and* $j \in \{1, \ldots, p\}$ *we have*

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

If Assumption 4 is additionally satisfied, then

$$(iii) \quad \widehat{MDA}^{(IK)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[(m(\boldsymbol{X}) - \mathbb{E}[m(X_{\pi_j})|\boldsymbol{X}^{(-j)}])^2].$$

3.2 Proof of Theorem 1-(i)

Assumptions 1, 2 and 3 are sufficient to slightly extend the $\mathbb{L}^2$-consistency of random forests from Scornet et al. (2015, Theorem 1) to the case where inputs are dependent, and also when the prediction is performed for the permuted sample (i.e, for a query point with a different distribution than the training data). Then, the TT-MDA consistency follows using a standard asymptotic analysis.

Lemma 1 *If Assumptions 1, 2, and 3 are satisfied, for $M \in \mathbb{N}^\star$ we have*

$$\lim_{n\to\infty} \mathbb{E}[(m_{M,n}(\boldsymbol{X}, \Theta_M) - m(\boldsymbol{X}))^2] = 0,$$

and for all $j \in \{1, \ldots, p\}$

$$\lim_{n\to\infty} \mathbb{E}[(m_{M,n}(X_{\pi_j}, \Theta_M) - m(X_{\pi_j}))^2] = 0.$$

of Theorem 1-(i). We assume that 1, 2, and 3 are satisfied, and fix $j \in \{1, \ldots, p\}$ and $M \in \mathbb{N}^\star$. Firstly, according to Lemma 1, we have

$$\lim_{n\to\infty} \mathbb{E}[(m_{M,n}(\mathbf{X}, \Theta_M) - m(\mathbf{X}))^2] = 0, \tag{3.1}$$

and

$$\lim_{n\to\infty} \mathbb{E}[(m_{M,n}(X_{\pi_j}, \Theta_M) - m(X_{\pi_j}))^2] = 0. \tag{3.2}$$

Next, we can break down the Train/Test-MDA as follows

$$\begin{aligned}
\widehat{\mathrm{MDA}}^{(TT)}_{M,n}(X^{(j)}) &= \frac{1}{n}\sum_{i=1}^{n} \big(Y'_i - m_{M,n}(X'_{i,\pi_j}, \Theta_M)\big)^2 - \big(Y'_i - m_{M,n}(\mathbf{X}'_i, \Theta_M)\big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \big(m(\mathbf{X}'_i) + \varepsilon'_i - m_{M,n}(X'_{i,\pi_j}, \Theta_M)\big)^2 - \big(m(\mathbf{X}'_i) + \varepsilon'_i - m_{M,n}(\mathbf{X}'_i, \Theta_M)\big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \big([m(\mathbf{X}'_i) - m(X'_{i,\pi_j})] + [m(X'_{i,\pi_j}) - m_{M,n}(X'_{i,\pi_j}, \Theta_M)] + \varepsilon'_i\big)^2 \\
&\qquad - \big(m(\mathbf{X}'_i) - m_{M,n}(\mathbf{X}'_i, \Theta_M) + \varepsilon'_i\big)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} [m(\mathbf{X}'_i) - m(X'_{i,\pi_j})]^2 + [m(X'_{i,\pi_j}) - m_{M,n}(X'_{i,\pi_j}, \Theta_M)]^2 + \varepsilon'^2_i \\
&\qquad + 2[m(\mathbf{X}'_i) - m(X'_{i,\pi_j})][m(X'_{i,\pi_j}) - m_{M,n}(X'_{i,\pi_j}, \Theta_M)] \\
&\qquad + 2\varepsilon'_i[m(\mathbf{X}'_i) - m(X'_{i,\pi_j})] + 2\varepsilon'_i[m(X'_{i,\pi_j}) - m_{M,n}(X'_{i,\pi_j}, \Theta_M)] \\
&\qquad - [m(\mathbf{X}'_i) - m_{M,n}(\mathbf{X}'_i, \Theta_M)]^2 - \varepsilon'^2_i - 2\varepsilon'_i[m(\mathbf{X}'_i) - m_{M,n}(\mathbf{X}'_i, \Theta_M)].
\end{aligned}$$

Then, we use the triangle inequality and the previous expression to get the following bound

$$\begin{aligned}
&\mathbb{E}\big[\big|\widehat{\mathrm{MDA}}^{(TT)}_{M,n}(X^{(j)}) - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2]\big|\big] \\
&\qquad \leq \mathbb{E}\big[\big|\frac{1}{n}\sum_{i=1}^{n}[m(\mathbf{X}'_i) - m(X'_{i,\pi_j})]^2 - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2]\big|\big] && (3.3)\\
&\qquad + \mathbb{E}\big[\frac{1}{n}\sum_{i=1}^{n}[m(X'_{i,\pi_j}) - m_{M,n}(X'_{i,\pi_j}, \Theta_M)]^2\big] && (3.4)\\
&\qquad + \mathbb{E}\big[\big|\frac{2}{n}\sum_{i=1}^{n}[m(\mathbf{X}'_i) - m(X'_{i,\pi_j})][m(X'_{i,\pi_j}) - m_{M,n}(X'_{i,\pi_j}, \Theta_M)]\big|\big] && (3.5)\\
&\qquad + \mathbb{E}\big[\big|\frac{2}{n}\sum_{i=1}^{n}\varepsilon'_i[m(\mathbf{X}'_i) - m(X'_{i,\pi_j})]\big|\big] && (3.6)\\
&\qquad + \mathbb{E}\big[\big|\frac{2}{n}\sum_{i=1}^{n}\varepsilon'_i[m(X'_{i,\pi_j}) - m_{M,n}(X'_{i,\pi_j}, \Theta_M)]\big|\big] && (3.7)\\
&\qquad + \mathbb{E}\big[\frac{1}{n}\sum_{i=1}^{n}[m(\mathbf{X}'_i) - m_{M,n}(\mathbf{X}'_i, \Theta_M)]^2\big] && (3.8)\\
&\qquad + \mathbb{E}\big[\big|\frac{2}{n}\sum_{i=1}^{n}\varepsilon'_i[m(\mathbf{X}'_i) - m_{M,n}(\mathbf{X}'_i, \Theta_M)]\big|\big]. && (3.9)
\end{aligned}$$

Now, let us consider all the terms on the right hand side one by one.

The first and fourth terms (3.3) and (3.6) do not depend on the forest estimate, but it is not possible to simply apply the law of large numbers since the permutation introduces dependence within samples. For both terms, we prove $\mathbb{L}^2$-convergence, which implies the $\mathbb{L}^1$-convergence we are looking for. For the first term (3.3), we define $\Delta_{n,1}$ as

$$\Delta_{n,1} = \frac{1}{n}\sum_{i=1}^{n}[m(\mathbf{X}'_i) - m(X'_{i,\pi_j})]^2 - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2].$$

Clearly, we have $\mathbb{E}[\Delta_{n,1}] = 0$. Its variance writes

$$\begin{aligned}
\mathbb{V}[\Delta_{n,1}] = \frac{1}{n^2}\mathbb{E}\big[\sum_{i,k=1}^{n}&([m(\mathbf{X}_i) - m(X_{i,\pi_j})]^2 - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2]) \\
&\times ([m(\mathbf{X}_k) - m(X_{k,\pi_j})]^2 - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2])\big].
\end{aligned}$$

Because of the permutation, each element of the sum is dependent on only two other terms. Therefore, only $3n$ terms of the double sum are not null, and because m is bounded (continuous on a compact), we get

$$\mathbb{V}[\Delta_{n,1}] \leq \frac{3}{n} \times 64||m||^4_\infty.$$

Thus, $\lim_{n\to\infty} \mathbb{V}[\Delta_{n,1}] = 0$, which proves $\mathbb{L}^2$-convergence of $\Delta_{n,1}$ towards $\mathbb{E}[\Delta_{n,1}] = 0$. We can handle the fourth term (3.6) in the same way. For the second term (3.4), by symmetry,

$$\mathbb{E}\big[\frac{1}{n}\sum_{i=1}^{n}[m(X'_{i,\pi_j}) - m_{M,n}(X'_{i,\pi_j}, \Theta_M)]^2\big] = \mathbb{E}[(m(X_{\pi_j}) - m_{M,n}(X_{\pi_j}, \Theta_M))^2],$$

which tends to zero according to (3.2). The sixth term (3.8) is handled similarly using (3.1). Since m is bounded, we can bound the third term (3.5)

$$\mathbb{E}\big[\big|\frac{2}{n}\sum_{i=1}^{n}[m(\mathbf{X}'_i) - m(X'_{i,\pi_j})][m(X'_{i,\pi_j}) - m_{M,n}(X'_{i,\pi_j}, \Theta_M)]\big|\big] \\ \leq 4\|m\|_\infty \mathbb{E}[|m(X_{\pi_j}) - m_{M,n}(X_{\pi_j}, \Theta_M)|],$$

and since $\mathbb{L}^2$ convergence implies $\mathbb{L}^1$ convergence, we use (3.2) to obtain the convergence towards 0 of this third term (3.5). For the fifth term (3.7) we first apply the triangle inequality, and by symmetry we get

$$\mathbb{E}\big[\big|\frac{2}{n}\sum_{i=1}^{n}\varepsilon'_i[m(X'_{i,\pi_j}) - m_{M,n}(X'_{i,\pi_j}, \Theta_M)]\big|\big] \leq 2\mathbb{E}[|\varepsilon'(m(X_{\pi_j}) - m_{M,n}(X_{\pi_j}, \Theta_M))|] \\ \leq 2\mathbb{E}[|\varepsilon'|]\mathbb{E}[|m(X_{\pi_j}) - m_{M,n}(X_{\pi_j}, \Theta_M)|],$$

which tends to zero according to (3.2). Similarly, the last term (3.9) is handled with (3.1). Gathering all previous convergence results on (3.3)-(3.9), we have for all M, for all $j \in \{1, \ldots, p\}$,

$$\widehat{\mathrm{MDA}}^{(TT)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2].$$

□

of Lemma 1. We assume that Assumptions 1, 2, and 3 are satisfied, and fix $j \in \{1, \ldots, p\}$ and $M \in \mathbb{N}^\star$. We first introduce the infinite forest estimate $m_n(\mathbf{x})$ defined as $m_n(\mathbf{x}) = \mathbb{E}_\Theta[m_n(\mathbf{x}, \Theta)]$ where $m_n(\mathbf{x}, \Theta)$ is the randomized CART estimate.

Theorem 1 from Scornet et al. (2015) states the $\mathbb{L}^2$-consistency of infinite random forests. It relies on Assumption 3 for the asymptotic regime of a_n and t_n, and on a modified version of 1, where the regression function is additive and $\mathbf{X}$ is uniformly distributed over $[0, 1]^p$. Here, we extend this result to any continuous regression function and any positive distribution for $\mathbf{X}$ with support on the unit cube. First, the extension to the case where $\mathbf{X}$ has any distribution bounded from above and below by positive constants can be easily obtained by several technical adaptations as already highlighted in Scornet (2020). Secondly, notice that the additive structure of the regression function is only required in Scornet et al. (2015) to show the consistency of a theoretical randomized CART. Therefore we can drop the additivity assumption and replace it by Assumption 2. Overall, we can extend Theorem 1 from Scornet et al. (2015): provided that Assumptions 1, 2, and 3 are satisfied, we have

$$\lim_{n\to\infty} \mathbb{E}[(m_n(\mathbf{X}) - m(\mathbf{X}))^2] = 0. \tag{3.10}$$

Next, this result needs to be extended when the query point $\mathbf{X}$ is replaced by X_{π_j}. From Assumption 1, $\mathbf{X}$ admits a density f_X over $[0,1]^p$. By construction, the random vector X_{π_j} is the vector $\mathbf{X}$ where the j-th component is replaced by an independent copy of $X^{(j)}$. Therefore X_{π_j} admits a density f_{π_j}, which is the product of the densities of $X^{(j)}$ and $\mathbf{X}^{(-j)}$, i.e., for $\mathbf{x} \in [0,1]^p$,

$$f_{\pi_j}(\mathbf{x}) = \int_{[0,1]^{p-1}} f_X(\mathbf{x}) d\mathbf{x}^{(-j)} \times \int_{[0,1]} f_X(\mathbf{x}) d\mathbf{x}^{(j)}. \tag{3.11}$$

From Assumption 1, f_X is bounded from above and below by positive constants. Thus, it exists $c_1, c_2 > 0$ such that for all $\mathbf{x} \in [0,1]^p$,

$$c_1 \leq f_X(\mathbf{x}) \leq c_2. \tag{3.12}$$

Combining (3.12) and (3.11), we obtain that for all $\mathbf{x} \in [0,1]^p$, $c_1^2 \leq f_{\pi_j}(\mathbf{x}) \leq c_2^2$, and consequently,

$$\sup_{\mathbf{x} \in [0,1]^p} \frac{f_{\pi_j}(\mathbf{x})}{f_X(\mathbf{x})} \leq \frac{c_2^2}{c_1}.$$

Now, we write

$$\begin{aligned}
\mathbb{E}[(m_n(X_{\pi_j}) - m(X_{\pi_j}))^2 | \mathscr{D}_n] &= \int_{[0,1]^p} (m_n(\mathbf{x}) - m(\mathbf{x}))^2 f_{\pi_j}(\mathbf{x}) d\mathbf{x} \\
&= \int_{[0,1]^p} (m_n(\mathbf{x}) - m(\mathbf{x}))^2 f_X(\mathbf{x}) \frac{f_{\pi_j}(\mathbf{x})}{f_X(\mathbf{x})} d\mathbf{x} \\
&\leq \frac{c_2^2}{c_1} \int_{[0,1]^p} (m_n(\mathbf{x}) - m(\mathbf{x}))^2 f_X(\mathbf{x}) d\mathbf{x} \\
&\leq \frac{c_2^2}{c_1} \mathbb{E}[(m_n(\mathbf{X}) - m(\mathbf{X}))^2 | \mathscr{D}_n].
\end{aligned}$$

Taking expectations on both sides and using (3.10), we finally obtain

$$\lim_{n \to \infty} \mathbb{E}[(m_n(X_{\pi_j}) - m(X_{\pi_j}))^2] = 0. \tag{3.13}$$

Equations (3.10) and (3.13) state that infinite forests evaluated at $\mathbf{X}$ or X_{π_j} are $\mathbb{L}^2$ consistent. The first of these two results can be extended to get the consistency of a single randomized CART $m_n(\mathbf{X}, \Theta)$, as shown in Scornet et al. (2015) by an easy adaptation of the infinite forest case. Formally, we obtain

$$\lim_{n \to \infty} \mathbb{E}[(m_n(\mathbf{X}, \Theta) - m(\mathbf{X}))^2] = 0. \tag{3.14}$$

The exact same reasoning as for the infinite forest above applies to get the extension to X_{π_j}, and thus, we have

$$\lim_{n \to \infty} \mathbb{E}[(m_n(X_{\pi_j}, \Theta) - m(X_{\pi_j}))^2] = 0. \tag{3.15}$$

Now, we expand the final quantity of interest $\mathbb{E}[(m_{M,n}(\mathbf{X},\Theta_M)-m(\mathbf{X}))^2]$ (and its counterpart for X_{π_j}):

$$\begin{aligned}
&\mathbb{E}[(m_{M,n}(\mathbf{X},\Theta_M)-m(\mathbf{X}))^2]\\
&\quad=\mathbb{E}\big[\big(\frac{1}{M}\sum_{\ell=1}^{M} m_n(\mathbf{X},\Theta_\ell)-m(\mathbf{X})\big)^2\big]\\
&\quad=\mathbb{E}\big[\mathbb{E}\big[\big(\frac{1}{M}\sum_{\ell=1}^{M} m_n(\mathbf{X},\Theta_\ell)-m(\mathbf{X})\big)^2\big|\mathbf{X},\mathscr{D}_n\big]\big]\\
&\quad=\frac{1}{M^2}\mathbb{E}\big[\mathbb{E}\big[\sum_{\ell,\ell'=1}^{M}[m_n(\mathbf{X},\Theta_\ell)-m(\mathbf{X})][m_n(\mathbf{X},\Theta_{\ell'})-m(\mathbf{X})]\big|\mathbf{X},\mathscr{D}_n\big]\big]\\
&\quad=\frac{1}{M^2}\mathbb{E}\big[\mathbb{E}\big[\sum_{\ell=1}^{M}\big(m_n(\mathbf{X},\Theta)-m(\mathbf{X})\big)^2\big|\mathbf{X},\mathscr{D}_n\big]\big]\\
&\qquad+\frac{1}{M^2}\mathbb{E}\big[\mathbb{E}\big[\sum_{\ell\neq\ell'}[m_n(\mathbf{X},\Theta_\ell)-m(\mathbf{X})][m_n(\mathbf{X},\Theta_{\ell'})-m(\mathbf{X})]\big|\mathbf{X},\mathscr{D}_n\big]\big].
\end{aligned}$$

Conditional on $(\mathbf{X},\mathscr{D}_n)$, the random variables $m_n(\mathbf{X},\Theta_\ell)$ for $\ell=1,\ldots,M$ are iid. Hence

$$\begin{aligned}
&\mathbb{E}[(m_{M,n}(\mathbf{X},\Theta_M)-m(\mathbf{X}))^2]\\
&\quad=\frac{1}{M}\mathbb{E}\big[\mathbb{E}\big[\big(m_n(\mathbf{X},\Theta)-m(\mathbf{X})\big)^2\big|\mathbf{X},\mathscr{D}_n\big]\big]\\
&\qquad+\frac{1}{M^2}\mathbb{E}\big[\sum_{\ell\neq\ell'}\big(\mathbb{E}[m_n(\mathbf{X},\Theta_\ell)|\mathbf{X},\mathscr{D}_n]-m(\mathbf{X})\big)\big(\mathbb{E}[m_n(\mathbf{X},\Theta_{\ell'})|\mathbf{X},\mathscr{D}_n]-m(\mathbf{X})\big)\big]\\
&\quad=\frac{1}{M}\mathbb{E}\big[\big(m_n(\mathbf{X},\Theta)-m(\mathbf{X})\big)^2\big]+\big(1-\frac{1}{M}\big)\mathbb{E}\big[\big(m_n(\mathbf{X})-m(\mathbf{X})\big)^2\big]. \qquad (3.16)
\end{aligned}$$

Using (3.10) and (3.14), we obtain the final result

$$\lim_{n\to\infty}\mathbb{E}[(m_{M,n}(\mathbf{X},\Theta_M)-m(\mathbf{X}))^2]=0,$$

which also holds for X_{π_j}, using (3.13) and (3.15):

$$\lim_{n\to\infty}\mathbb{E}[(m_{M,n}(X_{\pi_j},\Theta_M)-m(X_{\pi_j}))^2]=0.$$

□

3.3 Proof of Theorem 1-(ii)

Theorem 1-(i) can be quite easily adapted to the BC-MDA (ii).

of Theorem 1-(ii). We assume that Assumptions 1-3 are satisfied, and fix $j\in\{1,\ldots,p\}$ and $M\in\mathbb{N}^\star$. Recall that the Breiman-Cutler MDA is formally defined by

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

where $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \notin \Theta_\ell^{(S)}}$ is the size of the out-of-bag sample of the ℓ-th tree.

Since a_n observations are subsampled without replacement prior to the construction of each tree, all out-of-bag samples have the same constant size of $N_{n,\ell} = n - a_n$. Using the triangle inequality, we have

$$\begin{aligned}
&\mathbb{E}\big[\big|\widehat{\mathrm{MDA}}_{M,n}^{(BC)}(X^{(j)}) - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2]\big|\big] \\
&\qquad \leq \frac{1}{M}\sum_{\ell=1}^{M} \frac{1}{n-a_n}\mathbb{E}\big[\big|\sum_{i=1}^{n}[(Y_i - m_n(\mathbf{X}_{i,\pi_{j\ell}}, \Theta_\ell))^2 - (Y_i - m_n(\mathbf{X}_i, \Theta_\ell))^2 \\
&\qquad\qquad - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2]]\mathbb{1}_{i \notin \Theta_\ell^{(S)}}\big|\big],
\end{aligned}$$

and by symmetry, this boils down to

$$\begin{aligned}
&\mathbb{E}\big[\big|\widehat{\mathrm{MDA}}_{M,n}^{(BC)}(X^{(j)}) - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2]\big|\big] \\
&\qquad \leq \frac{1}{n-a_n}\mathbb{E}\big[\big|\sum_{i=1}^{n}[(Y_i - m_n(\mathbf{X}_{i,\pi_{j1}}, \Theta_1))^2 - (Y_i - m_n(\mathbf{X}_i, \Theta_1))^2 \\
&\qquad\qquad - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2]]\mathbb{1}_{i \notin \Theta_1^{(S)}}\big|\big].
\end{aligned}$$

Next, we expand the sum in the right hand side and obtain a similar decomposition as the one in the proof of Theorem 1-(i),

$$\begin{aligned}
&\frac{1}{n-a_n}\sum_{i=1}^{n}[(Y_i - m_n(\mathbf{X}_{i,\pi_{j1}}, \Theta_1))^2 - (Y_i - m_n(\mathbf{X}_i, \Theta_1))^2]\mathbb{1}_{i \notin \Theta_1^{(S)}} \\
&\qquad = \frac{1}{n-a_n}\sum_{i=1}^{n}[([m(\mathbf{X}_i) - m(\mathbf{X}_{i,\pi_{j1}})] + [m(\mathbf{X}_{i,\pi_{j1}}) - m_n(\mathbf{X}_{i,\pi_{j1}}, \Theta_1)] + \varepsilon_i)^2 \\
&\qquad\qquad - ([m(\mathbf{X}_i) - m_n(\mathbf{X}_i, \Theta_1)] + \varepsilon_i)^2]\mathbb{1}_{i \notin \Theta_1^{(S)}} \\
&\qquad = \frac{1}{n-a_n}\sum_{i=1}^{n}[m(\mathbf{X}_i) - m(\mathbf{X}_{i,\pi_{j1}})]^2\mathbb{1}_{i \notin \Theta_1^{(S)}} \\
&\qquad\qquad + [m(\mathbf{X}_{i,\pi_{j1}}) - m_n(\mathbf{X}_{i,\pi_{j1}}, \Theta_1)]^2\mathbb{1}_{i \notin \Theta_1^{(S)}} + \varepsilon_i^2\mathbb{1}_{i \notin \Theta_1^{(S)}} \\
&\qquad\qquad + 2[m(\mathbf{X}_i) - m(\mathbf{X}_{i,\pi_{j1}})][m(\mathbf{X}_{i,\pi_{j1}}) - m_n(\mathbf{X}_{i,\pi_{j1}}, \Theta_1)]\mathbb{1}_{i \notin \Theta_1^{(S)}} \\
&\qquad\qquad + 2\varepsilon_i[m(\mathbf{X}_i) - m(\mathbf{X}_{i,\pi_{j1}})]\mathbb{1}_{i \notin \Theta_1^{(S)}} \\
&\qquad\qquad + 2\varepsilon_i[m(\mathbf{X}_{i,\pi_{j1}}) - m_n(\mathbf{X}_{i,\pi_{j1}}, \Theta_1)]\mathbb{1}_{i \notin \Theta_1^{(S)}} \\
&\qquad\qquad - [m(\mathbf{X}_i) - m_n(\mathbf{X}_i, \Theta_1)]^2\mathbb{1}_{i \notin \Theta_1^{(S)}} - \varepsilon_i^2\mathbb{1}_{i \notin \Theta_1^{(S)}} \\
&\qquad\qquad - 2\varepsilon_i[m(\mathbf{X}_i) - m_n(\mathbf{X}_i, \Theta_1)]\mathbb{1}_{i \notin \Theta_1^{(S)}}.
\end{aligned}$$

Thus, we have the following bound

$$\begin{aligned}
&\mathbb{E}\big[\big|\widehat{\mathrm{MDA}}^{(BC)}_{M,n}(X^{(j)}) - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2]\big|\big] \\
&\quad \le \mathbb{E}\big[\big|\frac{1}{n-a_n}\sum_{i=1}^{n}([m(\mathbf{X}_i) - m(\mathbf{X}_{i,\pi_{j1}})]^2 - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2])\mathbb{1}_{i\notin\Theta_1^{(S)}}\big|\big] && (3.17)\\
&\quad + \mathbb{E}\big[\frac{1}{n-a_n}\sum_{i=1}^{n}[m(\mathbf{X}_{i,\pi_{j1}}) - m_n(\mathbf{X}_{i,\pi_{j1}},\Theta_1)]^2\mathbb{1}_{i\notin\Theta_1^{(S)}}\big] && (3.18)\\
&\quad + \mathbb{E}\big[\big|\frac{2}{n-a_n}\sum_{i=1}^{n}[m(\mathbf{X}_i) - m(\mathbf{X}_{i,\pi_{j1}})][m(\mathbf{X}_{i,\pi_{j1}}) - m_n(\mathbf{X}_{i,\pi_{j1}},\Theta_1)]\mathbb{1}_{i\notin\Theta_1^{(S)}}\big|\big] && (3.19)\\
&\quad + \mathbb{E}\big[\big|\frac{2}{n-a_n}\sum_{i=1}^{n}\varepsilon_i[m(\mathbf{X}_i) - m(\mathbf{X}_{i,\pi_{j1}})]\mathbb{1}_{i\notin\Theta_1^{(S)}}\big|\big] && (3.20)\\
&\quad + \mathbb{E}\big[\big|\frac{2}{n-a_n}\sum_{i=1}^{n}\varepsilon_i[m(\mathbf{X}_{i,\pi_{j1}}) - m_n(\mathbf{X}_{i,\pi_{j1}},\Theta_1)]\mathbb{1}_{i\notin\Theta_1^{(S)}}\big|\big] && (3.21)\\
&\quad + \mathbb{E}\big[\frac{1}{n-a_n}\sum_{i=1}^{n}[m(\mathbf{X}_i) - m_n(\mathbf{X}_i,\Theta_1)]^2\mathbb{1}_{i\notin\Theta_1^{(S)}}\big] && (3.22)\\
&\quad + \mathbb{E}\big[\big|\frac{2}{n-a_n}\sum_{i=1}^{n}\varepsilon_i[m(\mathbf{X}_i) - m_n(\mathbf{X}_i,\Theta_1)]\mathbb{1}_{i\notin\Theta_1^{(S)}}\big|\big]. && (3.23)
\end{aligned}$$

Now, let us consider all the terms on the right hand side one by one.

For the first term (3.17), we define $\Delta_{n,1}$ as

$$\Delta_{n,1} = \sum_{i=1}^{n}\frac{1}{n-a_n}([m(\mathbf{X}_i) - m(\mathbf{X}_{i,\pi_{j1}})]^2 - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2])\mathbb{1}_{i\notin\Theta_1^{(S)}}.$$

Its expectation is

$$\begin{aligned}
\mathbb{E}[\Delta_{n,1}] =& \mathbb{E}[\frac{n}{n-a_n}([m(\mathbf{X}_1) - m(\mathbf{X}_{1,\pi_{j1}})]^2 - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2])\mathbb{1}_{1\notin\Theta_1^{(S)}}] \\
&= \frac{n}{n-a_n}\mathbb{E}[(m(\mathbf{X}_1) - m(\mathbf{X}_{1,\pi_{j1}}))^2 - \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2]]\mathbb{P}(1\notin\Theta_1^{(S)}) \\
&= 0.
\end{aligned}$$

Next, observe that each term of the sum in $\Delta_{n,1}$ is dependent on two other terms because of the permutation of the j-th component, then we have $\mathbb{V}[\Delta_{n,1}] = O(1/(n-a_n))$. By Assumption 3, $a_n/n < 1-\kappa$ with a fixed $\kappa > 0$, thus $\mathbb{V}[\Delta_{n,1}] = O(1/n)$. Since $\mathbb{E}[\Delta_{n,1}] = 0$ and $\lim_{n\to\infty}\mathbb{V}[\Delta_{n,1}] = 0$, $\Delta_{n,1}$ converges towards 0 in $\mathbb{L}^2$, which implies $\mathbb{L}^1$-convergence. We

can handle the fourth term (3.20) in the same way. For the second term (3.18),

$$\begin{aligned}
&\mathbb{E}\big[\frac{1}{n-a_n}\sum_{i=1}^{n}[m(\mathbf{X}_{i,\pi_{j1}})-m_n(\mathbf{X}_{i,\pi_{j1}},\Theta_1)]^2\mathbb{1}_{i\notin\Theta_1^{(S)}}\big]\\
&\quad=\sum_{i=1}^{n}\mathbb{E}\big[[m(\mathbf{X}_{i,\pi_{j1}})-m_n(\mathbf{X}_{i,\pi_{j1}},\Theta_1)]^2\big|i\notin\Theta_1^{(S)}\big]\frac{\mathbb{P}(i\notin\Theta_1^{(S)})}{n-a_n}\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[[m(\mathbf{X}_{i,\pi_{j1}})-m_n(\mathbf{X}_{i,\pi_{j1}},\Theta_1)]^2\big|i\notin\Theta_1^{(S)}\big]
\end{aligned}$$

where the last equality results from $\mathbb{P}(i\notin\Theta_1^{(S)})=(n-a_n)/n$. The conditioning event $\{i\notin\Theta_1^{(S)}\}$ means that the observation of index i belongs to the out-of-bag sample. Thus, it is strictly equivalent to consider the tree trained with the sample $\mathscr{D}_n\setminus(\mathbf{X}_i,Y_i)$ of size $n-1$ with a subsampling size a_n. Furthermore, we can replace the query point $\mathbf{X}_{i,\pi_{j1}}$ by X_{π_j} because these two random vectors are iid and both independent of the training data of $m_{a_n,n-1}$. Then,

$$\begin{aligned}
&\mathbb{E}\big[\frac{1}{n-a_n}\sum_{i=1}^{n}[m(\mathbf{X}_{i,\pi_{j1}})-m_n(\mathbf{X}_{i,\pi_{j1}},\Theta_1)]^2\mathbb{1}_{i\notin\Theta_1^{(S)}}\big]\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[[m(X_{\pi_j})-m_{a_n,n-1}(X_{\pi_j},\Theta)]^2\big]\\
&\quad=\mathbb{E}[(m(X_{\pi_j})-m_{a_n,n-1}(X_{\pi_j},\Theta))^2],
\end{aligned}$$

which tends to zero according to the second statement in Lemma 1 for $M=1$. The sixth term (3.22) is handled similarly using the first part of Lemma 1. Since m is bounded, we can bound the third term (3.19)

$$\begin{aligned}
&\mathbb{E}\big[\big|\frac{2}{n-a_n}\sum_{i=1}^{n}[m(\mathbf{X}_i)-m(\mathbf{X}_{i,\pi_{j1}})][m(\mathbf{X}_{i,\pi_{j1}})-m_n(\mathbf{X}_{i,\pi_{j1}},\Theta_1)]\mathbb{1}_{i\notin\Theta_1^{(S)}}\big|\big]\\
&\quad\le\frac{4||m||_\infty}{n-a_n}\mathbb{E}\big[\sum_{i=1}^{n}\big|m(\mathbf{X}_{i,\pi_{j1}})-m_n(\mathbf{X}_{i,\pi_{j1}},\Theta_1)\big|\times\mathbb{1}_{i\notin\Theta_1^{(S)}}\big]\\
&\quad\le\frac{4||m||_\infty}{n}\sum_{i=1}^{n}\mathbb{E}\big[\big|m(\mathbf{X}_{i,\pi_{j1}})-m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)\big|\big]\\
&\quad\le4||m||_\infty\mathbb{E}\big[\big|m(X_{\pi_j})-m_{a_n,n-1}(X_{\pi_j},\Theta)\big|\big],
\end{aligned}$$

which tends to zero according to Lemma 1 (with $M=1$). Similarly, for the fifth term (3.21), we have

$$\begin{aligned}
&\mathbb{E}\big[\big|\frac{2}{n-a_n}\sum_{i=1}^{n}\varepsilon_i[m(\mathbf{X}_{i,\pi_{j1}})-m_n(\mathbf{X}_{i,\pi_{j1}},\Theta_1)]\mathbb{1}_{i\notin\Theta_1^{(S)}}\big|\big]\\
&\qquad\le2\mathbb{E}[|\varepsilon|]\mathbb{E}\big[\big|m(X_{\pi_j})-m_{a_n,n-1}(X_{\pi_j},\Theta)\big|\big],
\end{aligned}$$

and the convergence towards 0 is again given by Lemma 1. The last term (3.23) is handled in the same way. Gathering all previous convergence results on (3.17)-(3.23), we have for all M,

for all $j \in \{1, \dots, p\}$,

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

□

3.4 Proof of Theorems 1-(iii) and Proposition 1

The obstacle in the asymptotic analysis of the IK-MDA arises from the randomness of $\Lambda_{n,i}$, which can even be empty. However, the quadratic risk of the OOB estimate can be bounded using the risk of the standard forest, as stated in the following Lemma.

Lemma 2 *If Assumption 1 is satisfied, for all $M \in \mathbb{N}^\star$ and $i \in \{1, \dots, n\}$, we have*

$$\mathbb{E}\big[\big(m^{(OOB)}_{M,a_n,n}(\boldsymbol{X}_i, \Theta_M) - m(\boldsymbol{X}_i)\big)^2 \mathbb{1}_{|\Lambda_{n,i}|>0}\big] \leq \frac{2}{1 - a_n/n} \mathbb{E}\big[\big(m_{M,a_n,n-1}(\boldsymbol{X}, \Theta_M) - m(\boldsymbol{X})\big)^2\big].$$

We can draw interesting insights from Lemma 2. First by construction, the OOB estimate aggregates a smaller number of trees than in the standard forest: $\mathbb{E}[|\Lambda_{n,i}|] = (1 - a_n/n)M$ trees in average. Therefore the risk of the standard forest is inflated by the coefficient $2/(1 - a_n/n) > 2$ to bound the OOB risk. Since the risk of the OOB estimate is bounded by the risk of the standard forest, the $\mathbb{L}^2$-consistency of random forests can be extended to the OOB estimate.

Lemma 3 *If Assumptions 1, 2, and 3 are satisfied, for all $i \in \{1, \dots, n\}$ and $M \in \mathbb{N}^\star$ we have*

$$\lim_{n \to \infty} \mathbb{E}[(m^{(OOB)}_{M,n}(\boldsymbol{X}_i, \Theta_M) - m(\boldsymbol{X}_i))^2 \mathbb{1}_{|\Lambda_{n,i}|>0}] = 0,$$

and if Assumption 4 is additionally satisfied, for all $j \in \{1, \dots, p\}$

$$\lim_{n \to \infty} \mathbb{E}[(m^{(OOB)}_{M,n,\pi_j}(\boldsymbol{X}_i, \Theta_M) - \mathbb{E}[m(X_{i,\pi_j})|\boldsymbol{X}_i^{(-j)}])^2 \mathbb{1}_{|\Lambda_{n,i}|>0}] = 0.$$

To prove Lemma 2 and 3, we need the following Lemma 4, proved at the end of the section.

Lemma 4 *If $\delta_{M,n}$ and $\gamma_{M,n}$ are defined as*

$$\delta_{M,n} = M^2 \mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2} \Big| 1, 2 \in \Lambda_{n,i}\Big] \mathbb{P}(1, 2 \in \Lambda_{n,i})$$
$$\gamma_{M,n} = M^2 \mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2} \Big| 1 \in \Lambda_{n,i}\Big] \mathbb{P}(1 \in \Lambda_{n,i}),$$

for all $M \in \mathbb{N} \setminus \{0, 1\}$, we have

$$\delta_{M,n} \leq 1$$
$$\delta_{M,n} \leq \gamma_{M,n} \leq \frac{2}{1 - \frac{a_n}{n}},$$

and for a fixed sample size n,

$$1 - \delta_{M,n} = O\Big(\frac{1}{M}\Big).$$

Then, we can deduce the consistency of the IK-MDA.

of Theorem 1-(iii). We assume that Assumptions 1-4 are satisfied, and fix $j \in \{1, \dots, p\}$. Recall that Ishwaran-Kogalur MDA is defined as

$$\widehat{\mathrm{MDA}}^{(IK)}_{M,n}(X^{(j)}) = \frac{1}{N_{M,n}} \sum_{i=1}^{n} (Y_i - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i, \Theta_M))^2 - (Y_i - m^{(OOB)}_{M,n}(\mathbf{X}_i, \Theta_M))^2,$$

where $N_{M,n} = \sum_{i=1}^{n} \mathbb{1}_{|\Lambda_{n,i}|>0}$ is the number of points which do not belong to all trees, and

$$m^{(OOB)}_{M,n}(\mathbf{X}_i, \Theta_M) = \frac{1}{|\Lambda_{n,i}|} \sum_{\ell \in \Lambda_{n,i}} m_n(\mathbf{X}_i, \Theta_\ell) \mathbb{1}_{|\Lambda_{n,i}|>0},$$

$$m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i, \Theta_M) = \frac{1}{|\Lambda_{n,i}|} \sum_{\ell \in \Lambda_{n,i}} m_n(\mathbf{X}_{i,\pi_{j\ell}}, \Theta_\ell) \mathbb{1}_{|\Lambda_{n,i}|>0}.$$

To lighten derivations, we define $MDA^{\star}_{IK} = \mathbb{E}[(m(\mathbf{X}) - \mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2]$. We expand the following expression,

$$\begin{aligned}
&\mathbb{E}\big[\big|\widehat{\mathrm{MDA}}^{(IK)}_{M,n}(X^{(j)}) - MDA^{\star}_{IK}\big|\big] \\
&= \mathbb{E}\Big[\Big|\frac{1}{N_{M,n}} \sum_{i=1}^{n} \big[(Y_i - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i, \Theta_M))^2 - (Y_i - m^{(OOB)}_{M,n}(\mathbf{X}_i, \Theta_M))^2 - MDA^{\star}_{IK}\big] \mathbb{1}_{|\Lambda_{n,i}|>0}\Big|\Big].
\end{aligned}$$

Observe that $N_{M,n}$ is bounded between n and $n - a_n$, and consequently

$$\begin{aligned}
&\mathbb{E}\big[\big|\widehat{\mathrm{MDA}}^{(IK)}_{M,n}(X^{(j)}) - MDA^{\star}_{IK}\big|\big] \\
&\leq \mathbb{E}\Big[\Big|\frac{1}{n - a_n} \sum_{i=1}^{n} \big[(Y_i - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i, \Theta_M))^2 - (Y_i - m^{(OOB)}_{M,n}(\mathbf{X}_i, \Theta_M))^2 - MDA^{\star}_{IK}\big] \mathbb{1}_{|\Lambda_{n,i}|>0}\Big|\Big].
\end{aligned}$$

Then, we follow the proof of Theorem 1-(i) and (ii) with a similar decomposition of the sum of the above expression

$$\begin{aligned}
&\sum_{i=1}^{n} [(Y_i - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i, \Theta_M))^2 - (Y_i - m^{(OOB)}_{M,n}(\mathbf{X}_i, \Theta_M))^2 - MDA^{\star}_{IK}] \mathbb{1}_{|\Delta_{n,i}|>0} \\
&\quad = \sum_{i=1}^{n} [([m(\mathbf{X}_i) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}^{(-j)}_i]] + [\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}^{(-j)}_i] - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i, \Theta_M)] + \varepsilon_i)^2 \\
&\qquad\quad - ([m(\mathbf{X}_i) - m^{(OOB)}_{M,n}(\mathbf{X}_i, \Theta_M)] + \varepsilon_i)^2 - MDA^{\star}_{IK}] \mathbb{1}_{|\Delta_{n,i}|>0} \\
&\quad = \sum_{i=1}^{n} ([m(\mathbf{X}_i) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}^{(-j)}_i]]^2 - MDA^{\star}_{IK}) \mathbb{1}_{|\Delta_{n,i}|>0} \\
&\qquad\quad + [\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}^{(-j)}_i] - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i, \Theta_M)]^2 \mathbb{1}_{|\Delta_{n,i}|>0} + \varepsilon_i^2 \mathbb{1}_{|\Delta_{n,i}|>0} \\
&\qquad\quad + 2[m(\mathbf{X}_i) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}^{(-j)}_i]][\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}^{(-j)}_i] - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i, \Theta_M)] \mathbb{1}_{|\Delta_{n,i}|>0} \\
&\qquad\quad + 2\varepsilon_i [m(\mathbf{X}_i) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}^{(-j)}_i]] \mathbb{1}_{|\Delta_{n,i}|>0} \\
&\qquad\quad + 2\varepsilon_i [\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}^{(-j)}_i] - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i, \Theta_M)] \mathbb{1}_{|\Delta_{n,i}|>0} \\
&\qquad\quad - [m(\mathbf{X}_i) - m^{(OOB)}_{M,n}(\mathbf{X}_i, \Theta_M)]^2 \mathbb{1}_{|\Delta_{n,i}|>0} - \varepsilon_i^2 \mathbb{1}_{|\Delta_{n,i}|>0} \\
&\qquad\quad - 2\varepsilon_i [m(\mathbf{X}_i) - m^{(OOB)}_{M,n}(\mathbf{X}_i, \Theta_M)] \mathbb{1}_{|\Delta_{n,i}|>0}.
\end{aligned}$$

We then obtain the following bound

$$
\begin{aligned}
&\mathbb{E}\big[\big|\widehat{\mathrm{MDA}}^{(IK)}_{M,n}(X^{(j)}) - MDA^{\star}_{IK}\big|\big] \\
&\quad\leq \mathbb{E}\Big[\Big|\frac{1}{n-a_n}\sum_{i=1}^{n}([m(\mathbf{X}_i) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]]^2 - MDA^{\star}_{IK})\mathbb{1}_{|\Lambda_{n,i}|>0}\Big|\Big] &(3.24)\\
&\quad+\mathbb{E}\Big[\frac{1}{n-a_n}\sum_{i=1}^{n}[\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}] - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i,\Theta_M)]^2\mathbb{1}_{|\Lambda_{n,i}|>0}\Big] &(3.25)\\
&\quad+\mathbb{E}\Big[\Big|\frac{2}{n-a_n}\sum_{i=1}^{n}[m(\mathbf{X}_i) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]] \\
&\qquad\qquad\times[\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}] - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i,\Theta_M)]\mathbb{1}_{|\Lambda_{n,i}|>0}\Big|\Big] &(3.26)\\
&\quad+\mathbb{E}\Big[\Big|\frac{2}{n-a_n}\sum_{i=1}^{n}\varepsilon_i[m(\mathbf{X}_i) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]]\mathbb{1}_{|\Lambda_{n,i}|>0}\Big|\Big] &(3.27)\\
&\quad+\mathbb{E}\Big[\Big|\frac{2}{n-a_n}\sum_{i=1}^{n}\varepsilon_i[\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}] - m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i,\Theta_M)]\mathbb{1}_{|\Lambda_{n,i}|>0}\Big|\Big] &(3.28)\\
&\quad+\mathbb{E}\Big[\frac{1}{n-a_n}\sum_{i=1}^{n}[m(\mathbf{X}_i) - m^{(OOB)}_{M,n}(\mathbf{X}_i,\Theta_M)]^2\mathbb{1}_{|\Lambda_{n,i}|>0}\Big] &(3.29)\\
&\quad+\mathbb{E}\Big[\Big|\frac{2}{n-a_n}\sum_{i=1}^{n}\varepsilon_i[m(\mathbf{X}_i) - m^{(OOB)}_{M,n}(\mathbf{X}_i,\Theta_M)]\mathbb{1}_{|\Lambda_{n,i}|>0}\Big|\Big]. &(3.30)
\end{aligned}
$$

Now, let us consider all the terms on the right hand side one by one. For the first term (3.24), we can rewrite

$$
\begin{aligned}
&\frac{1}{n-a_n}\sum_{i=1}^{n}([m(\mathbf{X}_i) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]]^2 - MDA^{\star}_{IK})\mathbb{1}_{|\Lambda_{n,i}|>0} \\
&\qquad=\frac{n}{n-a_n}\frac{1}{n}\sum_{i=1}^{n}([m(\mathbf{X}_i) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]]^2 - MDA^{\star}_{IK})\mathbb{1}_{|\Lambda_{n,i}|>0},
\end{aligned}
$$

and the multiplicative term in front $n/(n-a_n)$ is upper bounded by $1/\kappa > 0$ by Assumption 3. Next, we can apply the strong law of large numbers to show that the sum converges almost surely towards

$$
\begin{aligned}
&\mathbb{E}\big[([m(\mathbf{X}_1) - \mathbb{E}[m(\mathbf{X}_{1,\pi_j})|\mathbf{X}_1^{(-j)}]]^2 - MDA^{\star}_{IK})\mathbb{1}_{|\Lambda_{n,1}|>0}\big] \\
&\quad= \mathbb{E}\big[([m(\mathbf{X}_1) - \mathbb{E}[m(\mathbf{X}_{1,\pi_j})|\mathbf{X}_1^{(-j)}]]^2 - MDA^{\star}_{IK})\big]\mathbb{P}(|\Lambda_{n,1}| > 0) \\
&\quad= 0.
\end{aligned}
$$

Since almost sure convergence implies $\mathbb{L}^1$-convergence, the first term (3.24) converges towards 0. The fourth term (3.27) is handled similarly with the strong law of large number since the

noise is centered and independent of $\mathscr{D}_n$. The second term

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{n-a_n}\sum_{i=1}^{n}[\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m_{M,n,\pi_j}^{(OOB)}(\mathbf{X}_i,\Theta_M)]^2\mathbb{1}_{|\Lambda_{n,i}|>0}\Big]\\ =\frac{n}{n-a_n}\mathbb{E}\Big[(\mathbb{E}[m(\mathbf{X}_{1,\pi_j})|\mathbf{X}_1^{(-j)}]-m_{M,n,\pi_j}^{(OOB)}(\mathbf{X}_1,\Theta_M))^2\mathbb{1}_{|\Lambda_{n,1}|>0}\Big],\end{aligned}$$

converges towards 0 from the second part of Lemma 3 and because $n/(n-a_n)<1/\kappa$. The sixth term (3.29) is handled identically using the first part of Lemma 3. For the third term (3.26), since m is bounded (continuous on a compact), we have

$$\begin{aligned}\mathbb{E}\Big[\Big|\frac{2}{n-a_n}\sum_{i=1}^{n}&[m(\mathbf{X}_i)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]]\\ &\times[\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m_{M,n,\pi_j}^{(OOB)}(\mathbf{X}_i,\Theta_M)]\mathbb{1}_{|\Lambda_{n,i}|>0}\Big|\Big]\\ \leq\frac{4n||m||_\infty}{n-a_n}&\mathbb{E}\Big[\Big|\mathbb{E}[m(\mathbf{X}_{1,\pi_j})|\mathbf{X}_1^{(-j)}]-m_{M,n,\pi_j}^{(OOB)}(\mathbf{X}_1,\Theta_M)\Big|\mathbb{1}_{|\Lambda_{n,1}|>0}\Big],\end{aligned}$$

which converges towards 0 by Lemma 3. Similarly, for the fifth (3.28) and seventh (3.30) terms, we have the following bound

$$\begin{aligned}\mathbb{E}\Big[\Big|\frac{2}{n-a_n}\sum_{i=1}^{n}&\varepsilon_i[\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m_{M,n,\pi_j}^{(OOB)}(\mathbf{X}_i,\Theta_M)]\mathbb{1}_{|\Lambda_{n,i}|>0}\Big|\Big]\\ \leq&\frac{2n}{n-a_n}E[|\varepsilon|]\mathbb{E}\Big[\Big|\mathbb{E}[m(\mathbf{X}_{1,\pi_j})|\mathbf{X}_1^{(-j)}]-m_{M,n,\pi_j}^{(OOB)}(\mathbf{X}_1,\Theta_M)\Big|\mathbb{1}_{|\Lambda_{n,1}|>0}\Big],\end{aligned}$$

and we conclude using Lemma 3 again. Overall, we have

$$\widehat{\mathrm{MDA}}_{M,n}^{(IK)}(X^{(j)})\xrightarrow{\mathbb{L}^1}\mathbb{E}[(m(\mathbf{X})-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2].$$

□

of Lemma 2. We assume that Assumption 1 is satisfied, and consider $i\in\{1,\dots,n\}$ and $M\in\mathbb{N}^\star$. To prove the first part of Lemma 2, we begin with and expansion of the OOB estimate

$$\begin{aligned}\mathbb{E}\Big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\big||\Lambda_{n,i}|>0\Big]\\ =\mathbb{E}\Big[\Big(\frac{1}{|\Lambda_{n,i}|}\sum_{\ell\in\Lambda_{n,i}}m_n(\mathbf{X}_i,\Theta_\ell)\mathbb{1}_{|\Lambda_{n,i}|>0}-m(\mathbf{X}_i)\Big)^2\big||\Lambda_{n,i}|>0\Big]\\ =\mathbb{E}\Big[\Big(\frac{1}{|\Lambda_{n,i}|}\sum_{\ell=1}^{M}[m_n(\mathbf{X}_i,\Theta_\ell)-m(\mathbf{X}_i)]\mathbb{1}_{\ell\in\Lambda_{n,i}}\Big)^2\big||\Lambda_{n,i}|>0\Big].\end{aligned}$$

Now, we expand the square with a double sum,

$$\begin{aligned}
&\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\big||\Lambda_{n,i}|>0\big]\\
&\quad=\sum_{\ell,\ell'=1}^{M}\mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}[m_n(\mathbf{X}_i,\Theta_\ell)-m(\mathbf{X}_i)][m_n(\mathbf{X}_i,\Theta_{\ell'})-m(\mathbf{X}_i)]\mathbb{1}_{\ell,\ell'\in\Lambda_{n,i}}\big||\Lambda_{n,i}|>0\Big]\\
&\quad=\sum_{\ell,\ell'=1}^{M}\mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}[m_n(\mathbf{X}_i,\Theta_\ell)-m(\mathbf{X}_i)][m_n(\mathbf{X}_i,\Theta_{\ell'})-m(\mathbf{X}_i)]\big|\ell,\ell'\in\Lambda_{n,i}\Big]\\
&\qquad\qquad\times\mathbb{P}\big(\ell,\ell'\in\Lambda_{n,i}\big||\Lambda_{n,i}|>0\big).
\end{aligned}$$

Observe that conditionally on $\{\ell,\ell'\in\Lambda_{n,i}\}$, $\Lambda_{n,i}$ only depends on $\{\Theta_k,k\in\{1,\dots,M\}\setminus\{\ell,\ell'\}\}$. This means that $\Lambda_{n,i}$ and $[m_n(\mathbf{X}_i,\Theta_\ell)-m(\mathbf{X}_i)][m_n(\mathbf{X}_i,\Theta_{\ell'})-m(\mathbf{X}_i)]$ are independent conditionally on $\{\ell,\ell'\in\Lambda_{n,i}\}$. We can then write

$$\begin{aligned}
&\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\big||\Lambda_{n,i}|>0\big]\mathbb{P}(|\Lambda_{n,i}|>0)\\
&\quad=\sum_{\ell,\ell'=1}^{M}\mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}\big|\ell,\ell'\in\Lambda_{n,i}\Big]\mathbb{P}\big(\ell,\ell'\in\Lambda_{n,i}\big||\Lambda_{n,i}|>0\big)\mathbb{P}(|\Lambda_{n,i}|>0)\\
&\qquad\qquad\times\mathbb{E}\big[[m_n(\mathbf{X}_i,\Theta_\ell)-m(\mathbf{X}_i)][m_n(\mathbf{X}_i,\Theta_{\ell'})-m(\mathbf{X}_i)]\big|\ell,\ell'\in\Lambda_{n,i}\big].\\
&\quad=\sum_{\ell,\ell'=1}^{M}\mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}\big|\ell,\ell'\in\Lambda_{n,i}\Big]\mathbb{P}\big(\ell,\ell'\in\Lambda_{n,i}\big)\\
&\qquad\qquad\times\mathbb{E}\big[[m_n(\mathbf{X}_i,\Theta_\ell)-m(\mathbf{X}_i)][m_n(\mathbf{X}_i,\Theta_{\ell'})-m(\mathbf{X}_i)]\big|\ell,\ell'\in\Lambda_{n,i}\big].
\end{aligned}$$

Since $|\Lambda_{n,i}|$ is a binomial distribution, $\mathbb{E}\big[\frac{1}{|\Lambda_{n,i}|^2}\big|\ell,\ell'\in\Lambda_{n,i}\big]\mathbb{P}(\ell,\ell'\in\Lambda_{n,i})$ takes the same value for each pair of distinct ℓ,ℓ' and any sample $i\in\{1,\dots,n\}$. Similarly for the case $\ell=\ell'$, $\mathbb{E}\big[\frac{1}{|\Lambda_{n,i}|^2}\big|\ell\in\Lambda_{n,i}\big]\mathbb{P}(\ell\in\Lambda_{n,i})$ is constant when ℓ varies. Therefore, we introduce

$$\delta_{M,n}=M^2\mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}\big|\ell,\ell'\in\Lambda_{n,i}\Big]\mathbb{P}(\ell,\ell'\in\Lambda_{n,i}),$$

and

$$\gamma_{M,n}=M^2\mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}\big|\ell\in\Lambda_{n,i}\Big]\mathbb{P}(\ell\in\Lambda_{n,i}).$$

Then, we have

$$\begin{aligned}
&\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\big||\Lambda_{n,i}|>0\big]\mathbb{P}(|\Lambda_{n,i}|>0)\\
&\qquad=\delta_{M,n}\frac{1}{M^2}\sum_{\ell,\ell'=1}^{M}\mathbb{E}\big[[m_n(\mathbf{X}_i,\Theta_\ell)-m(\mathbf{X}_i)][m_n(\mathbf{X}_i,\Theta_{\ell'})-m(\mathbf{X}_i)]\big|\ell,\ell'\in\Lambda_{n,i}\big]\\
&\qquad\quad+(\gamma_{M,n}-\delta_{M,n})\frac{1}{M^2}\sum_{\ell=1}^{M}\mathbb{E}\big[(m_n(\mathbf{X}_i,\Theta_\ell)-m(\mathbf{X}_i))^2\big|\ell\in\Lambda_{n,i}\big].
\end{aligned}$$

Recall that $m_n(\mathbf{X}_i, \Theta_\ell)$ is the randomized CART estimate, built with $\mathscr{D}_n$ and Θ_ℓ, where the component $\Theta_\ell^{(S)}$ is used to subsample a_n data points. When conditioned on $\{\ell \in \Lambda_{n,i}\}$ (i.e. $i \notin \Theta_\ell^{(S)}$), $m_n(\mathbf{X}_i, \Theta_\ell)$ can be seen as the CART estimate built with $\mathscr{D}_n \setminus \{(\mathbf{X}_i, Y_i)\}$ and with the subsample size a_n, i.e., $m_{a_n,n-1}(\mathbf{X}_i, \Theta_\ell)$. Therefore, we have for all pairs ℓ, ℓ',

$$\begin{aligned}
&\mathbb{E}\big[[m_n(\mathbf{X}_i, \Theta_\ell) - m(\mathbf{X}_i)][m_n(\mathbf{X}_i, \Theta_{\ell'}) - m(\mathbf{X}_i)]\big|\ell, \ell' \in \Lambda_{n,i}\big] \\
&\qquad = \mathbb{E}\big[[m_{a_n,n-1}(\mathbf{X}_i, \Theta_\ell) - m(\mathbf{X}_i)][m_{a_n,n-1}(\mathbf{X}_i, \Theta_{\ell'}) - m(\mathbf{X}_i)]\big] \\
&\qquad = \mathbb{E}\big[[m_{a_n,n-1}(\mathbf{X}, \Theta_\ell) - m(\mathbf{X})][m_{a_n,n-1}(\mathbf{X}, \Theta_{\ell'}) - m(\mathbf{X})]\big], \qquad (3.31)
\end{aligned}$$

where the last equality holds because $\mathbf{X}_i$ and $\mathbf{X}$ are identically distributed and both independent of the training data of $m_{a_n,n-1}$. Then, this last equality is plugged in the previous result to obtain

$$\begin{aligned}
&\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i, \Theta_M) - m(\mathbf{X}_i)\big)^2\big||\Lambda_{n,i}| > 0\big]\mathbb{P}(|\Lambda_{n,i}| > 0) \\
&\quad = \delta_{M,n}\frac{1}{M^2}\sum_{\ell,\ell'=1}^{M}\mathbb{E}\big[[m_{a_n,n-1}(\mathbf{X}, \Theta_\ell) - m(\mathbf{X})][m_{a_n,n-1}(\mathbf{X}, \Theta_{\ell'}) - m(\mathbf{X})]\big] \\
&\qquad + (\gamma_{M,n} - \delta_{M,n})\frac{1}{M^2}\sum_{\ell=1}^{M}\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}, \Theta_\ell) - m(\mathbf{X}))^2\big]. \qquad (3.32)
\end{aligned}$$

Next, we factorize the right hand side

$$\begin{aligned}
&\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i, \Theta_M) - m(\mathbf{X}_i)\big)^2\big||\Lambda_{n,i}| > 0\big]\mathbb{P}(|\Lambda_{n,i}| > 0) \\
&\quad = \delta_{M,n}\mathbb{E}\Big[\Big(\frac{1}{M}\sum_{\ell=1}^{M} m_{a_n,n-1}(\mathbf{X}, \Theta_\ell) - m(\mathbf{X})\Big)^2\Big] \\
&\qquad + (\gamma_{M,n} - \delta_{M,n})\frac{1}{M}\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}, \Theta) - m(\mathbf{X}))^2\big] \\
&\quad = \delta_{M,n}\mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X}, \Theta_M) - m(\mathbf{X})\big)^2\big] \\
&\qquad + (\gamma_{M,n} - \delta_{M,n})\frac{1}{M}\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}, \Theta) - m(\mathbf{X}))^2\big], \qquad (3.33)
\end{aligned}$$

where $m_{M,a_n,n-1}(\mathbf{X}, \Theta_M)$ is the standard random forest estimate, built with a dataset of size $n-1$ and the subsample size a_n. Using the decomposition (3.16) of the risk of the finite forest, we have

$$\frac{1}{M}\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}, \Theta) - m(\mathbf{X}))^2\big] \leq \mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X}, \Theta_M) - m(\mathbf{X})\big)^2\big].$$

Additionally, from Lemma 4, $\gamma_{M,n} - \delta_{M,n} > 0$. We combine the last two inequalities with the previous result and obtain

$$\begin{aligned}
&\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i, \Theta_M) - m(\mathbf{X}_i)\big)^2\big||\Lambda_{n,i}| > 0\big]\mathbb{P}(|\Lambda_{n,i}| > 0) \\
&\qquad \leq \delta_{M,n}\mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X}, \Theta_M) - m(\mathbf{X})\big)^2\big] \\
&\qquad\quad + (\gamma_{M,n} - \delta_{M,n})\mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X}, \Theta_M) - m(\mathbf{X})\big)^2\big] \\
&\qquad \leq \gamma_{M,n}\mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X}, \Theta_M) - m(\mathbf{X})\big)^2\big],
\end{aligned}$$

and using again Lemma 4, we finally get

$$\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\mathbb{1}_{|\Lambda_{n,i}|>0}\big]\leq\frac{2}{1-a_n/n}\mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X},\Theta_M)-m(\mathbf{X})\big)^2\big].$$

□

of Proposition 1. We need to bound the difference between the risks of the OOB estimate and the standard forest. To do so, we go back to equation (3.33)

$$\begin{aligned}\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\big||\Lambda_{n,i}|>0\big]\mathbb{P}(|\Lambda_{n,i}|>0)&\\ =\delta_{M,n}\mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X},\Theta_M)-m(\mathbf{X})\big)^2\big]&\\ +(\gamma_{M,n}-\delta_{M,n})\frac{1}{M}\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X},\Theta)-m(\mathbf{X}))^2\big],&\end{aligned}$$

and rewrite it

$$\begin{aligned}\Big|\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\mathbb{1}_{|\Lambda_{n,i}|>0}\big]-\mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X},\Theta_M)-m(\mathbf{X})\big)^2\big]\Big|&\\ \leq\big|\delta_{M,n}-1\big|\mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X},\Theta_M)-m(\mathbf{X})\big)^2\big]&\\ +(\gamma_{M,n}-\delta_{M,n})\frac{1}{M}\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X},\Theta)-m(\mathbf{X}))^2\big].&\end{aligned}$$

According to Lemma 4, $\delta_{M,n}-1=O(1/M)$ and $\gamma_{M,n}-\delta_{M,n}$ is bounded. Therefore, for a fixed sample size n, we have

$$\Big|\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\mathbb{1}_{|\Lambda_{n,i}|>0}\big]-\mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X},\Theta_M)-m(\mathbf{X})\big)^2\big]\Big|=O\Big(\frac{1}{M}\Big).\tag{3.34}$$

Finally, recall that $\mathbb{P}(|\Lambda_{n,i}|>0)$ is the probability that the i-th observation does not belong to all trees (in this case the OOB forest estimate is properly defined). A simple calculation gives that $\mathbb{P}(|\Lambda_{n,i}|>0)=1-(a_n/n)^M$, which converges towards 1 exponentially fast as M grows. Then, we have

$$\begin{aligned}&\Big|\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\big]-\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\mathbb{1}_{|\Lambda_{n,i}|>0}\big]\Big|\\ &\quad=\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\mathbb{1}_{|\Lambda_{n,i}|=0}\big]\\ &\quad=\mathbb{E}\big[m(\mathbf{X}_i)^2\big]\mathbb{P}(|\Lambda_{n,i}|=0)\\ &\quad\leq||m||_\infty^2(a_n/n)^M.\end{aligned}\tag{3.35}$$

From Assumption 3, $a_n/n<1$, and combining the bound (3.35) with the previous result (3.34), we conclude that

$$\Big|\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i,\Theta_M)-m(\mathbf{X}_i)\big)^2\big]-\mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X},\Theta_M)-m(\mathbf{X})\big)^2\big]\Big|=O\Big(\frac{1}{M}\Big).$$

□

of Lemma 3. We first assume that Assumptions 1, 2, 3, and 4 are satisfied, and we consider $i \in \{1, \dots, n\}$. Using Lemma 2, we have

$$\mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i, \Theta_M) - m(\mathbf{X}_i)\big)^2 \mathbb{1}_{|\Lambda_{n,i}|>0}\big] \leq \frac{2}{1 - a_n/n} \mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X}, \Theta_M) - m(\mathbf{X})\big)^2\big]. \tag{3.36}$$

According to Assumption 3, $1 - a_n/n > \kappa$ where κ is fixed positive constant. Thus, we can directly apply Lemma 1 to obtain

$$\lim_{n\to\infty} \mathbb{E}\big[\big(m_{M,a_n,n-1}(\mathbf{X}, \Theta_M) - m(\mathbf{X})\big)^2\big] = 0,$$

and then

$$\lim_{n\to\infty} \mathbb{E}\big[\big(m_{M,n}^{(OOB)}(\mathbf{X}_i, \Theta_M) - m(\mathbf{X}_i)\big)^2 \mathbb{1}_{|\Lambda_{n,i}|>0}\big] = 0.$$

Next, we extend this result to the permuted case, i.e., $\mathbf{X}_i$ is replaced by X_{i,π_j}. Following the same proof as in Lemma 2, we derive the following decomposition, similarly to equation (3.32)

$$\begin{aligned}
&\mathbb{E}\big[\big(m_{M,n,\pi_j}^{(OOB)}(\mathbf{X}_i, \Theta_M) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]\big)^2 \big| |\Lambda_{n,i}| > 0\big] \mathbb{P}(|\Lambda_{n,i}| > 0) \\
&\qquad = \delta_{M,n} \frac{1}{M^2} \sum_{\ell \neq \ell'} \mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j\ell}}, \Theta_\ell) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]) \\
&\qquad\qquad \times (m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j\ell'}}, \Theta_{\ell'}) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])\big] \\
&\qquad\quad + \gamma_{M,n} \frac{1}{M^2} \sum_{\ell=1}^{M} \mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j\ell}}, \Theta) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])^2\big].
\end{aligned}$$

By symmetry, we have

$$\begin{aligned}
&\mathbb{E}\big[\big(m_{M,n,\pi_j}^{(OOB)}(\mathbf{X}_i, \Theta_M) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]\big)^2 \big| |\Lambda_{n,i}| > 0\big] \mathbb{P}(|\Lambda_{n,i}| > 0) \\
&\qquad = \delta_{M,n} \frac{M-1}{M} \mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}}, \Theta_1) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]) \\
&\qquad\qquad \times (m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j2}}, \Theta_2) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])\big] \\
&\qquad\quad + \gamma_{M,n} \frac{1}{M} \mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{\pi_j}, \Theta) - \mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2\big].
\end{aligned}$$

In the first term of the right hand side, we need to deal with the specific case where $\pi_{j1} = \pi_{j2}$, which implies that $\mathbf{X}_{i,\pi_{j1}} = \mathbf{X}_{i,\pi_{j2}}$ since they have the same j-th permuted component:

$$\begin{aligned}
&\mathbb{E}\big[\big(m_{M,n,\pi_j}^{(OOB)}(\mathbf{X}_i, \Theta_M) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]\big)^2 \big| |\Lambda_{n,i}| > 0\big] \mathbb{P}(|\Lambda_{n,i}| > 0) \\
&\quad = \delta_{M,n} \frac{M-1}{M} \mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}}, \Theta_1) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]) \\
&\qquad\qquad \times (m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j2}}, \Theta_2) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])|\pi_{j1} \neq \pi_{j2}\big] \mathbb{P}(\pi_{j1} \neq \pi_{j2}) \\
&\qquad + \delta_{M,n} \frac{M-1}{M} \mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}}, \Theta_1) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]) \\
&\qquad\qquad \times (m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j2}}, \Theta_2) - \mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])|\pi_{j1} = \pi_{j2}\big] \mathbb{P}(\pi_{j1} = \pi_{j2}) \\
&\qquad + \gamma_{M,n} \frac{1}{M} \mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{\pi_j}, \Theta) - \mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2\big],
\end{aligned}$$

which can be simplified using Cauchy-Schwartz inequality for the second term as

$$\begin{aligned}
&\mathbb{E}\big[\big(m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i,\Theta_M)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]\big)^2\big||\Lambda_{n,i}|>0\big]\mathbb{P}(|\Lambda_{n,i}|>0) \qquad (3.37)\\
&\quad\le \delta_{M,n}\frac{M-1}{M}\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])\\
&\qquad\qquad\times(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j2}},\Theta_2)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])|\pi_{j1}\neq\pi_{j2}\big]\mathbb{P}(\pi_{j1}\neq\pi_{j2})\\
&\qquad+\Big(\frac{\gamma_{M,n}}{M}+\delta_{M,n}\frac{M-1}{M}\mathbb{P}(\pi_{j1}=\pi_{j2})\Big)\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{\pi_j},\Theta)-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2\big].
\end{aligned}$$

Now, we focus on the first term of the right hand side. We have

$$\begin{aligned}
&\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])\\
&\qquad\times(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j2}},\Theta_2)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])|\pi_{j1}\neq\pi_{j2}\big]\\
&=\mathbb{E}\big[[m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)-m(\mathbf{X}_{i,\pi_{j1}})-(\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m(\mathbf{X}_{i,\pi_{j1}}))]\\
&\qquad\times[m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j2}},\Theta_2)-m(\mathbf{X}_{i,\pi_{j2}})-(\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m(\mathbf{X}_{i,\pi_{j2}}))]|\pi_{j1}\neq\pi_{j2}\big]\\
&=\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)-m(\mathbf{X}_{i,\pi_{j1}}))(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j2}},\Theta_2)-m(\mathbf{X}_{i,\pi_{j2}}))|\pi_{j1}\neq\pi_{j2}\big]\\
&\qquad-2\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)-m(\mathbf{X}_{i,\pi_{j1}}))(\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m(\mathbf{X}_{i,\pi_{j2}}))|\pi_{j1}\neq\pi_{j2}\big]\\
&\qquad+\mathbb{E}\big[(\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m(\mathbf{X}_{i,\pi_{j1}}))(\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m(\mathbf{X}_{i,\pi_{j2}}))|\pi_{j1}\neq\pi_{j2}\big].
\end{aligned}$$

For the second term, the two multiplied terms are independent conditional on $\mathbf{X}_i^{(-j)}$ and $\pi_{j1}\neq\pi_{j2}$, then

$$\begin{aligned}
&\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)-m(\mathbf{X}_{i,\pi_{j1}}))(\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m(\mathbf{X}_{i,\pi_{j2}}))\big|\pi_{j1}\neq\pi_{j2}\big]\\
&\quad=\mathbb{E}\big[\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)-m(\mathbf{X}_{i,\pi_{j1}}))(\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m(\mathbf{X}_{i,\pi_{j2}}))\big|\mathbf{X}_i^{(-j)},\pi_{j1}\neq\pi_{j2}\big]\big]\\
&\quad=\mathbb{E}\big[\mathbb{E}\big[m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)-m(\mathbf{X}_{i,\pi_{j1}})\big|\mathbf{X}_i^{(-j)}\big]\mathbb{E}\big[\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]-m(\mathbf{X}_{i,\pi_{j2}}))\big|\mathbf{X}_i^{(-j)}\big]\big]\\
&\quad=0.
\end{aligned}$$

Similarly, the third term is also null. Finally, we apply Cauchy-Schwartz inequality to the first term to obtain

$$\begin{aligned}
&\delta_{M,n}\frac{M-1}{M}\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])\\
&\qquad\qquad\times(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j2}},\Theta_2)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}])|\pi_{j1}\neq\pi_{j2}\big]\\
&\qquad\le\delta_{M,n}\mathbb{E}\big[(m_{a_n,n-1}(\mathbf{X}_{i,\pi_{j1}},\Theta_1)-m(\mathbf{X}_{i,\pi_{j1}}))^2\big]\\
&\qquad\le\delta_{M,n}\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-m(X_{\pi_j}))^2\big],
\end{aligned}$$

where the last inequality holds because $\mathbf{X}_{i,\pi_{j1}}$ is independent of the sample used to train $m_{a_n,n-1}$ and have the same distribution as X_{π_j}. Overall, using this last inequality with the decomposition (3.37), we obtain the following bound

$$\begin{aligned}
&\mathbb{E}\big[\big(m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i,\Theta_M)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]\big)^2\big||\Lambda_{n,i}|>0\big]\mathbb{P}(|\Lambda_{n,i}|>0)\\
&\qquad\le\delta_{M,n}\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-m(X_{\pi_j}))^2\big]\\
&\qquad\quad+\Big(\frac{\gamma_{M,n}}{M}+\delta_{M,n}\frac{M-1}{M}\mathbb{P}(\pi_{j1}=\pi_{j2})\Big)\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2\big].
\end{aligned}$$

Furthermore, using Lemma 4, the bound can be simplified to get

$$\begin{aligned}
&\mathbb{E}\big[\big(m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i,\Theta_M)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]\big)^2\big||\Lambda_{n,i}|>0\big]\mathbb{P}(|\Lambda_{n,i}|>0)\\
&\quad\leq\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-m(X_{\pi_j}))^2\big]\\
&\qquad+\Big(\frac{2}{1-a_n/n}\frac{1}{M}+\mathbb{P}(\pi_{j1}=\pi_{j2})\Big)\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2\big].
\end{aligned}$$

Next, we break down the expectation of the second term

$$\begin{aligned}
&\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2\big]\\
&\quad=\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-m(X_{\pi_j})+(m(X_{\pi_j})-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]))^2\big]\\
&\quad=\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-m(X_{\pi_j}))^2\big]+\mathbb{E}\big[(m(X_{\pi_j})-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2\big]\\
&\qquad+2\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-m(X_{\pi_j}))(m(X_{\pi_j})-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])\big].
\end{aligned}$$

Since m is bounded, we get

$$\begin{aligned}
&\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2\big]\\
&\qquad\leq\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta)-m(X_{\pi_j}))^2\big]+4||m||^2_\infty\\
&\qquad\quad+4||m||_\infty\mathbb{E}\big[|m_{a_n,n-1}(X_{\pi_j},\Theta)-m(X_{\pi_j})|\big].
\end{aligned}$$

Finally we obtain the following bound

$$\begin{aligned}
&\mathbb{E}\big[\big(m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i,\Theta_M)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]\big)^2\big||\Lambda_{n,i}|>0\big]\mathbb{P}(|\Lambda_{n,i}|>0)\\
&\quad\leq\Big(1+\frac{2}{1-a_n/n}\frac{1}{M}+\mathbb{P}(\pi_{j1}=\pi_{j2})\Big)\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta_\ell)-m(X_{\pi_j}))^2\big]\\
&\qquad+\Big(\frac{2}{1-a_n/n}\frac{1}{M}+\mathbb{P}(\pi_{j1}=\pi_{j2})\Big)4||m||_\infty\mathbb{E}\big[|m_{a_n,n-1}(X_{\pi_j},\Theta)-m(X_{\pi_j})|\big]\\
&\qquad+4||m||^2_\infty\Big(\frac{2}{1-a_n/n}\frac{1}{M}+\mathbb{P}(\pi_{j1}=\pi_{j2})\Big).
\end{aligned}$$

The second part of Lemma 1 for $M=1$ gives that

$$\lim_{n\to\infty}\mathbb{E}\big[(m_{a_n,n-1}(X_{\pi_j},\Theta_\ell)-m(X_{\pi_j}))^2\big]=0,$$

and since $\mathbb{L}^2$-convergence implies $\mathbb{L}^1$-convergence, we also have

$$\lim_{n\to\infty}\mathbb{E}\big[|m_{a_n,n-1}(X_{\pi_j},\Theta_\ell)-m(X_{\pi_j})|\big]=0.$$

It is clear that $\mathbb{P}(\pi_{j1}=\pi_{j2})<1/(n-a_n)$, and then $\lim_{n\to\infty}\mathbb{P}(\pi_{j1}=\pi_{j2})=0$, since $1-a_n/n>\kappa>0$ by Assumption 3. Additionally, according to Assumption 4, $M\underset{n\to\infty}{\longrightarrow}\infty$, therefore

$$\lim_{n\to\infty}\frac{2}{1-a_n/n}\frac{1}{M}+\mathbb{P}(\pi_{j1}=\pi_{j2})=0.$$

Overall, we have

$$\lim_{n\to\infty}\mathbb{E}\big[\big(m^{(OOB)}_{M,n,\pi_j}(\mathbf{X}_i,\Theta_M)-\mathbb{E}[m(X_{i,\pi_j})|\mathbf{X}_i^{(-j)}]\big)^2\mathbb{1}_{|\Lambda_{n,i}|>0}\big]=0.$$

□

of Lemma 4. We consider $M \in \mathbb{N} \setminus \{0,1\}$, $i \in \{1,\dots,n\}$, and define

$$\begin{aligned}\delta_{M,n} =& M^2 \mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}\big|1,2 \in \Lambda_{n,i}\Big]\mathbb{P}(1,2 \in \Lambda_{n,i}) \\ =& M^2 \mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}\big|M-1, M \in \Lambda_{n,i}\Big]\mathbb{P}(M-1, M \in \Lambda_{n,i}).\end{aligned}$$

Recall that by definition, $|\Lambda_{n,i}| = \sum_{\ell=1}^{M} \mathbb{1}_{i \notin \Theta_\ell^{(S)}}$. Since Θ_ℓ are iid, $|\Lambda_{n,i}|$ is a binomial random variable. Then, we have

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}\big|M, M-1 \in \Lambda_{n,i}\Big] =& \mathbb{E}\Big[\frac{1}{(2+\sum_{\ell=1}^{M-2} \mathbb{1}_{i \notin \Theta_\ell^{(S)}})^2}\Big] \\ =& \sum_{k=0}^{M-2} \frac{1}{(k+2)^2}\binom{M-2}{k}\big(1-\frac{a_n}{n}\big)^k\big(\frac{a_n}{n}\big)^{M-2-k}.\end{aligned}$$

On the other hand,

$$\mathbb{P}(M-1, M \in \Lambda_{n,i}) = \big(1-\frac{a_n}{n}\big)^2.$$

Combining the previous two equations, we get

$$\begin{aligned}\delta_{M,n} =& M^2\big(1-\frac{a_n}{n}\big)^2 \sum_{k=0}^{M-2} \frac{1}{(k+2)^2}\binom{M-2}{k}\big(1-\frac{a_n}{n}\big)^k\big(\frac{a_n}{n}\big)^{M-2-k} \\ =& M^2 \sum_{k=0}^{M-2} \frac{1}{(k+2)^2}\frac{(M-2)!}{k!(M-(k+2))!}\big(1-\frac{a_n}{n}\big)^{k+2}\big(\frac{a_n}{n}\big)^{M-(k+2)} \\ =& M^2 \sum_{k=0}^{M-2} \frac{k+1}{(k+2)M(M-1)}\frac{M!}{(k+2)!(M-(k+2))!}\big(1-\frac{a_n}{n}\big)^{k+2}\big(\frac{a_n}{n}\big)^{M-(k+2)} \\ =& \frac{M}{M-1}\sum_{k=0}^{M-2} \frac{k+1}{k+2}\binom{M}{k+2}\big(1-\frac{a_n}{n}\big)^{k+2}\big(\frac{a_n}{n}\big)^{M-(k+2)}\end{aligned}$$

We reindex the sum with $k \leftarrow k+2$ and get

$$\begin{aligned}\delta_{M,n} =& \frac{M}{M-1}\sum_{k=2}^{M} \frac{k-1}{k}\binom{M}{k}\big(1-\frac{a_n}{n}\big)^k\big(\frac{a_n}{n}\big)^{M-k} \\ =& \frac{M}{M-1}\sum_{k=1}^{M} \Big(1-\frac{1}{k}\Big)\binom{M}{k}\big(1-\frac{a_n}{n}\big)^k\big(\frac{a_n}{n}\big)^{M-k}. \qquad (3.38)\end{aligned}$$

Next, we bound $\delta_{M,n}$,

$$\begin{aligned}\delta_{M,n} \leq & \frac{M}{M-1}\sum_{k=1}^{M}\Big(1-\frac{1}{M}\Big)\binom{M}{k}(1-\tfrac{a_n}{n})^k(\tfrac{a_n}{n})^{M-k} \\ \leq & \sum_{k=0}^{M}\binom{M}{k}(1-\tfrac{a_n}{n})^k(\tfrac{a_n}{n})^{M-k}-(\tfrac{a_n}{n})^M \\ \leq & 1-(\tfrac{a_n}{n})^M \\ \leq & 1.\end{aligned} \tag{3.39}$$

Similarly for the second inequality, we define

$$\begin{aligned}\gamma_{M,n} =& M^2\mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}\Big|1\in\Lambda_{n,i}\Big]\mathbb{P}(1\in\Lambda_{n,i}) \\ =& M^2\mathbb{E}\Big[\frac{1}{|\Lambda_{n,i}|^2}\Big|M\in\Lambda_{n,i}\Big]\mathbb{P}(M\in\Lambda_{n,i}),\end{aligned}$$

and get

$$\begin{aligned}\gamma_{M,n} =& M^2(1-\tfrac{a_n}{n})\sum_{k=0}^{M-1}\frac{1}{(k+1)^2}\binom{M-1}{k}(1-\tfrac{a_n}{n})^k(\tfrac{a_n}{n})^{M-1-k} \\ =& M\sum_{k=0}^{M-1}\frac{1}{k+1}\binom{M}{k+1}(1-\tfrac{a_n}{n})^{k+1}(\tfrac{a_n}{n})^{M-(k+1)} \\ =& M\sum_{k=1}^{M}\frac{1}{k}\binom{M}{k}(1-\tfrac{a_n}{n})^k(\tfrac{a_n}{n})^{M-k} \\ =& M\mathbb{E}\Big[\frac{1}{Z}\mathbb{1}_{Z\geq 1}\Big],\end{aligned}$$

where Z is a binomial random variable with M trials and parameter $1-\frac{a_n}{n}$. Lemma 4.1 from Györfi et al. (2006) states that

$$\mathbb{E}\Big[\frac{1}{Z}\mathbb{1}_{Z\geq 1}\Big] \leq \frac{2}{(M+1)(1-\frac{a_n}{n})}, \tag{3.40}$$

which implies that

$$\gamma_{M,n} \leq \frac{2M}{(M+1)(1-\frac{a_n}{n})} \leq \frac{2}{1-\frac{a_n}{n}}.$$

On the other hand,

$$\begin{aligned}\gamma_{M,n} =& M\sum_{k=1}^{M}\frac{1}{k}\binom{M}{k}(1-\tfrac{a_n}{n})^k(\tfrac{a_n}{n})^{M-k} \\ \geq & M\sum_{k=1}^{M}\frac{1}{M}\binom{M}{k}(1-\tfrac{a_n}{n})^k(\tfrac{a_n}{n})^{M-k} \\ \geq & 1-(\tfrac{a_n}{n})^M \\ \geq & \delta_{M,n},\end{aligned}$$

where the last inequality uses (3.39).

To prove the last statement of Lemma 4, we go back to equation (3.38):

$$\begin{aligned}
\delta_{M,n} =& \frac{M}{M-1}\sum_{k=1}^{M}\big(1-\frac{1}{k}\big)\binom{M}{k}\big(1-\frac{a_n}{n}\big)^k\big(\frac{a_n}{n}\big)^{M-k} \\
=& \frac{M}{M-1}\Big[\sum_{k=1}^{M}\binom{M}{k}\big(1-\frac{a_n}{n}\big)^k\big(\frac{a_n}{n}\big)^{M-k} - \sum_{k=1}^{M}\frac{1}{k}\binom{M}{k}\big(1-\frac{a_n}{n}\big)^k\big(\frac{a_n}{n}\big)^{M-k}\Big] \\
=& \frac{M}{M-1}\Big[1-\big(\frac{a_n}{n}\big)^M - \mathbb{E}\Big[\frac{1}{Z}\mathbb{1}_{Z\geq 1}\Big]\Big] \\
\geq& \frac{M}{M-1}\Big[1-\big(\frac{a_n}{n}\big)^M - \frac{2}{(M+1)(1-\frac{a_n}{n})}\Big],
\end{aligned}$$

where we use inequality (3.40) for the last statement. Overall, using also inequality (3.39), we have

$$0 \geq M(\delta_{M,n}-1) \geq \frac{M}{M-1}\Big[1-M\big(\frac{a_n}{n}\big)^M - \frac{2M}{(M+1)(1-\frac{a_n}{n})}\Big]$$

The right hand side is an increasing function of M and converges towards $-\frac{1+a_n/n}{1-a_n/n}$ as $M\to\infty$. Additionally, the right hand side is always defined since $1-a_n/n>\kappa>0$ from Assumption 3. Therefore, for a fixed sample size n, $M(\delta_{M,n}-1)$ is a bounded sequence. Finally,

$$\delta_{M,n}-1 = O\Big(\frac{1}{M}\Big).$$

□

3.5 Proof of Proposition 2

Proposition 2 *If Assumptions 1, 2 and 3 are satisfied, then for all $M\in\mathbb{N}^\star$ and $j\in\{1,\dots,p\}$ we have*

$$\begin{aligned}
&(i)\quad \widehat{MDA}^{(TT)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y]\times ST^{(j)} + \mathbb{V}[Y]\times ST^{(j)}_{mg} + MDA^{\star(j)}_3 \\
&(ii)\quad \widehat{MDA}^{(BC)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y]\times ST^{(j)} + \mathbb{V}[Y]\times ST^{(j)}_{mg} + MDA^{\star(j)}_3.
\end{aligned}$$

If Assumption 4 is additionally satisfied, then

$$(iii)\quad \widehat{MDA}^{(IK)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y]\times ST^{(j)} + MDA^{\star(j)}_3.$$

of Proposition 2. We assume that Assumptions 1, 2, and 3 are satisfied, and fix $j\in\{1,\dots,p\}$ and $M\in\mathbb{N}^\star$. Then, using Theorem 1-(i), we have

$$\widehat{\mathrm{MDA}}^{(TT)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[(m(\mathbf{X})-m(X_{\pi_j}))^2].$$

First, we rewrite the MDA limit as

$$\begin{aligned}
&\mathbb{E}[(m(\mathbf{X})-m(X_{\pi_j}))^2] \\
&\quad=\mathbb{E}[\mathbb{E}[(m(\mathbf{X})-m(X_{\pi_j}))^2|\mathbf{X}^{(-j)}]] \\
&\quad=\mathbb{E}\big[\mathbb{E}\big[\big((m(\mathbf{X})-\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}])-(m(X_{\pi_j})-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]) \\
&\qquad+(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}]-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])\big)^2|\mathbf{X}^{(-j)}\big]\big].
\end{aligned}$$

Now, observing that these three terms are independent conditionally on $\mathbf{X}^{(-j)}$, we can expand the MDA limit as follows

$$\begin{aligned}
&\mathbb{E}[(m(\mathbf{X})-m(X_{\pi_j}))^2] \\
&\quad=\mathbb{E}\big[\mathbb{E}\big[(m(\mathbf{X})-\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}])^2|\mathbf{X}^{(-j)}\big]+\mathbb{E}[(m(X_{\pi_j})-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2|\mathbf{X}^{(-j)}\big] \\
&\qquad+(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}]-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2\big] \\
&\quad=\mathbb{E}[\mathbb{V}[m(\mathbf{X})|\mathbf{X}^{(-j)}]]+\mathbb{E}[\mathbb{V}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]] \\
&\qquad+\mathbb{E}[(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}]-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2] \\
&\quad=\mathbb{V}[Y]\times ST^{(j)}+\mathbb{V}[Y]\times ST^{(j)}_{mg}+\mathbb{E}[(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}]-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2].
\end{aligned}$$

Theorem 1-(ii) gives the same theoretical counterpart for BC-MDA, and thus the same decomposition applies

$$\widehat{\mathrm{MDA}}^{(BC)}_{M,n}(X^{(j)})\xrightarrow{\mathbb{L}^1}\mathbb{V}[Y]\times ST^{(j)}+\mathbb{V}[Y]\times ST^{(j)}_{mg}+\mathbb{E}[(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}]-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2].$$

Now, we additionally assume that Assumption 4 is satisfied, i.e., the number of trees grows to infinity with n. Then, using Theorem 1-(iii) we have

$$\widehat{\mathrm{MDA}}^{(IK)}_{M,n}(X^{(j)})\xrightarrow{\mathbb{L}^1}\mathbb{E}[(m(\mathbf{X})-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2].$$

We decompose the theoretical counterpart as in the first case,

$$\begin{aligned}
&\mathbb{E}[(m(\mathbf{X})-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2] \\
&\quad=\mathbb{E}[(m(\mathbf{X})-\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}]-(\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]-\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}]))^2] \\
&\quad=\mathbb{E}[(m(\mathbf{X})-\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}])^2]+\mathbb{E}[(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}]-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]))^2] \\
&\quad=\mathbb{V}[Y]\times ST^{(j)}+\mathbb{E}[(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}]-\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}])^2].
\end{aligned}$$

□

3.6 Proof of Corollary 2

Corollary 1 *If covariates are independent, and if Assumptions 1-3 are satisfied, for all $M\in\mathbb{N}^\star$ and $j\in\{1,\dots,p\}$ we have*

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

In addition, if Assumption 4 is satisfied,

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

Corollary 2 *If the regression function m is additive, and if Assumptions 1-3 are satisfied, for all $M \in \mathbb{N}^\star$ and $j \in \{1, \dots, p\}$ we have*

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

In addition, if Assumption 4 is satisfied,

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

of Corollary 2. We assume that Assumptions 1, 2, and 3 are satisfied, and fix $j \in \{1, \dots, p\}$ and $M \in \mathbb{N}^\star$. Then, using Theorem 1-(i), we have

$$\widehat{\text{MDA}}^{(TT)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2].$$

Since the regression function is assumed additive, we can write m as

$$m(\mathbf{x}) = \sum_{k=1}^{p} m_k(x^{(k)}).$$

Then, the MDA limit writes

$$\begin{aligned}
\mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2] &= \mathbb{E}[\{m_j(X^{(j)}) - m_j(X'^{(j)})\}^2] \\
&= \mathbb{E}[\{(m_j(X^{(j)}) - \mathbb{E}[m_j(X^{(j)})]) - (m_j(X'^{(j)}) - \mathbb{E}[m_j(X^{(j)})])\}^2] \\
&= 2\mathbb{V}[m_j(X^{(j)})],
\end{aligned}$$

where $X'^{(j)}$ is an independent copy of $X^{(j)}$ by definition of X_{π_j}.

On the other hand, we have

$$\begin{aligned}
\mathbb{V}[Y] \times ST^{(j)}_{mg} &= \mathbb{E}[\mathbb{V}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]] \\
&= \mathbb{E}[\{m(X_{\pi_j}) - \mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]\}^2] \\
&= \mathbb{E}[\{m_j(X'^{(j)}) + \sum_{k\neq j}^{p} m_k(X^{(k)}) - \mathbb{E}[m_j(X'^{(j)}) + \sum_{k\neq j}^{p} m_k(X^{(k)})|\mathbf{X}^{(-j)}]\}^2] \\
&= \mathbb{E}[\{m_j(X'^{(j)}) - \mathbb{E}[m_j(X'^{(j)})]\}^2] \\
&= \mathbb{V}[m_j(X^{(j)})] \\
&= 1/2\mathbb{E}[(m(\mathbf{X}) - m(X_{\pi_j}))^2],
\end{aligned}$$

which gives the result of Corollary 2 for the Train-Test MDA.

The proof for the Breiman-Cutler MDA is identical. For the Iswharan-Kogalur MDA, we assume that Assumption (A4) is additionally satisfied, and Theorem 1 gives that

$$\widehat{\mathrm{MDA}}^{(IK)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[\{m(\mathbf{X}) - \mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]\}^2].$$

Again, we can simplify the MDA limit in the additive setting, and we get

$$\begin{aligned}\mathbb{E}[\{m(\mathbf{X}) - \mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]\}^2] &= \mathbb{E}[\{m_j(X^{(j)}) - \mathbb{E}[m_j(X'^{(j)})]\}^2] \\ &= \mathbb{V}[m_j(X^{(j)})] \\ &= \mathbb{V}[Y] \times ST^{(j)}_{mg},\end{aligned}$$

which gives the final result.

□

3.7 Proof of Property 1

Property 1 (Marginal Total Sobol Index) *If Assumption* 1 *is satisfied, the marginal total Sobol index* $ST^{(j)}_{mg}$ *satisfies the following properties.*

(a) $ST^{(j)}_{mg} = 0 \iff ST^{(j)} = 0$.

(b) If the components of X *are independent, then we have* $ST^{(j)}_{mg} = ST^{(j)}$.

(c) If m *is additive, i.e.* $m(X) = \sum_k m_k(X^{(k)})$*, then we have* $ST^{(j)}_{mg} = \mathbb{V}[m_j(X^{(j)})]/\mathbb{V}[Y]$*, and* $ST^{(j)}_{mg} \geq ST^{(j)}$.

of Property 1. We assume that Assumption 1 is satisfied.

(a) First, we assume that $ST^{(j)} = 0$. Using the definition of the total Sobol index, we get that

$$\mathbb{E}[(m(\mathbf{X}) - \mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}])^2] = 0.$$

By Assumption 1, the density of $\mathbf{X}$ is strictly positive on its support $[0,1]^p$, and since the square function is positive, the previous equation gives that, almost surely,

$$(m(\mathbf{X}) - \mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}])^2 = 0,$$

which gives

$$m(\mathbf{X}) = \mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}] \quad \text{a.s.}$$

Therefore, m($\mathbf{X}$) does not depend on the j-th component almost surely, and we have

$$m(X_{\pi_j}) = m(\mathbf{X}) \quad \text{a.s.},$$

and consequently $ST_{mg}^{(j)} = ST^{(j)} = 0$. The reverse case follows the same proof.

(b) By construction, X_{π_j} and $\mathbf{X}$ have the same joint distribution when $\mathbf{X}$ has independent components, and the result follows.

(c) We assume that m is additive and writes

$$m(\mathbf{X}) = \sum_{k=1}^{p} m_k(X^{(k)}).$$

We expand the definition of the marginal total Sobol index using the above expression of m and obtain

$$\begin{aligned}
\mathbb{V}[Y] \times ST_{mg}^{(j)} &= \mathbb{E}[\mathbb{V}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]] \\
&= \mathbb{E}[\{m(X_{\pi_j}) - \mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]\}^2] \\
&= \mathbb{E}[\{m_j(X'^{(j)}) + \sum_{k\neq j}^{p} m_k(X^{(k)}) - \mathbb{E}[m_j(X'^{(j)}) + \sum_{k\neq j}^{p} m_k(X^{(k)})|\mathbf{X}^{(-j)}]\}^2] \\
&= \mathbb{E}[\{m_j(X'^{(j)}) - \mathbb{E}[m_j(X'^{(j)})]\}^2] \\
&= \mathbb{V}[m_j(X^{(j)})].
\end{aligned}$$

For the second part of the statement, we similarly derive

$$\begin{aligned}
\mathbb{V}[Y] \times ST^{(j)} &= \mathbb{E}[\{m_j(X^{(j)}) - \mathbb{E}[m_j(X^{(j)})|\mathbf{X}^{(-j)}]\}^2] \\
&= \mathbb{E}[\mathbb{V}[m_j(X^{(j)})|\mathbf{X}^{(-j)}]],
\end{aligned}$$

and the law of total variance gives that $ST_{mg}^{(j)} \geq ST^{(j)}$.

□

4 Proof of the Sobol-MDA Consistency

For the sake of clarity, we recall Assumptions 5, 6, and Theorem 2.

Assumption 5 *A node split is constrained to generate child nodes with at least a small fraction $\gamma > 0$ of the parent node observations. Secondly, the split selection is slightly modified: at each tree node, the number* `mtry` *of candidate variables drawn to optimize the split is set to* `mtry` $= 1$ *with a small probability $\delta > 0$. Otherwise, with probability $1 - \delta$, the default value of* `mtry` *is used.*

Assumption 6 *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 $\kappa > 0$, $\lim_{n\to\infty} a_n = \infty$, $\lim_{n\to\infty} t_n = \infty$, and $\lim_{n\to\infty} 2^{t_n} \frac{(\log(a_n))^9}{a_n} = 0$.*

Theorem 2 *If Assumptions 1, 5, and 6 are satisfied, for all $M \in \mathbb{N}^\star$ and $j \in \{1, \dots, p\}$*

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

The consistency of the Sobol-MDA relies on the consistency of the projected random forest, stated in Lemma 6, and Lemma 7 for the corresponding OOB estimate. Lemma 5 is an intermediate result on the asymptotic behavior of the original forest. Under the small modifications of the random forest algorithm defined by Assumption 5, Lemma 5 states that the cells of a random tree in the empirical forest become infinitely small as the sample size increases. For a cell $A \in [0,1]$, we define $\text{diam}(A)$ the diameter of a cell as

$$\text{diam}(A) = \sup_{\mathbf{x},\mathbf{x}' \in A} ||\mathbf{x} - \mathbf{x}'||_2.$$

Recall that $A_n(\mathbf{X}, \Theta)$ is the cell of the original Θ-random CART where $\mathbf{X}$ falls.

Lemma 5 *If Assumptions 1, 5, and 6 are satisfied, we have in probability*

$$\lim_{n\to\infty} diam(A_n(\boldsymbol{X}, \Theta)) = 0.$$

The following lemma states that the Projected-CART estimate is consistent. Recall that $A_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)$ is the cell of the projected partition where $\mathbf{X}^{(-j)}$ falls, $m_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)$ is the associated projected tree, and $m_n^{(-j)}(\mathbf{X}^{(-j)}) = \mathbb{E}[m_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)|\mathscr{D}_n, \mathbf{X}^{(-j)}]$ is the projected infinite forest estimate. We also define $m^{(-j)}(\mathbf{z}) = \mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)} = \mathbf{z}]$ for $\mathbf{z} \in [0,1]^{p-1}$.

Lemma 6 *If Assumptions 1, 5, and 6 are satisfied, we have for $j \in \{1, \dots, p\}$*

$$\lim_{n\to\infty} \mathbb{E}[(m_n^{(-j)}(\boldsymbol{X}^{(-j)}) - m^{(-j)}(\boldsymbol{X}^{(-j)}))^2] = 0.$$

Lemma 7 *If Assumptions 1, 5, and 6 are satisfied, for all $i \in \{1, \dots, n\}$, $j \in \{1, \dots, p\}$, and $M \in \mathbb{N}^\star$ we have*

$$\lim_{n\to\infty} \mathbb{E}[(m_{M,n}^{(-j,OOB)}(\boldsymbol{X}_i^{(-j)}, \Theta_M) - m(\boldsymbol{X}_i^{(-j)}))^2 \mathbb{1}_{|\Lambda_{n,i}|>0}] = 0.$$

of Theorem 2. We assume that Assumptions 1, 5, and 6 are satisfied and consider $j \in \{1, \dots, p\}$. We can exactly follow the proof of Theorem 1-(iii) by only replacing $\mathbb{E}[m(X_{\pi_j})|\mathbf{X}^{(-j)}]$ by $\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}]$ in the main decomposition, and get the $\mathbb{L}^1$-consistency of the unnormalized Sobol-MDA using Lemmas 3 and 7. Finally, the Sobol-MDA is normalized by the standard variance estimate $\hat{\sigma}_Y$ of the output Y, which is consistent by the Law of Large Numbers. Next, according to the continuous mapping theorem $1/\hat{\sigma}_Y \xrightarrow{p} 1/\mathbb{V}[Y]$. Overall, the Sobol-MDA is the product of two random quantities which convergence in probability, and we have

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

□

of Lemma 5. The proof is inspired by Lemma 2 from Meinshausen (2006). We define $s_n(\mathbf{X}, \Theta)$ as the number of splits to reach the terminal cell $A_n(\mathbf{X}, \Theta)$ where $\mathbf{X}$ falls. The asymptotic regime of the tree growing is controlled by Assumption 6 by setting the number of terminal leaves to t_n. Since $A_n(\mathbf{X}, \Theta)$ is a terminal leave, there are two possible cases: further splitting $A_n(\mathbf{X}, \Theta)$ will necessarily lead to cells with a number of observations smaller than the algorithm parameter `minimum node size`, that we call N_{min}, and is typically equal to 5 in practice. Formally, it means that

$$N_n(\mathbf{X}, \Theta) < 2N_{min}, \tag{4.1}$$

where $N_n(\mathbf{X}, \Theta)$ is the number of observations in $A_n(\mathbf{X}, \Theta)$. The other possibility is that the total number of leaves t_n is reached, which implies that

$$2^{s_n(\mathbf{X},\Theta)} \geq t_n,$$

the equality case happening if the tree is balanced. Next, according to Assumption 5, all children nodes have at least a fraction $0.5 > \gamma > 0$ of the parent node observations. Then we have $a_n \gamma^{s_n(\mathbf{X},\Theta)} \leq N_n(\mathbf{X}, \Theta)$. Combining this last inequality with (4.1), we obtain $a_n \gamma^{s_n(\mathbf{X},\Theta)} < 2N_{min}$. Overall, at least one of the two following inequalities is satisfied

$$\begin{aligned} s_n(\mathbf{X}, \Theta) &\geq \log_2(t_n) \\ s_n(\mathbf{X}, \Theta) &> \frac{\log_2(a_n/2N_{min})}{\log_2(1/\gamma)}. \end{aligned}$$

From Assumption 6, $a_n \to \infty$ and $t_n \to \infty$. Therefore, we can conclude that

$$s_n(\mathbf{X}, \Theta) \xrightarrow{p} \infty. \tag{4.2}$$

Now, we fix $j \in \{1, \ldots, p\}$, and define $s_n^{(j)}(\mathbf{X}, \Theta)$ as the number of splits involving the j-th variable in the path to $A_n(\mathbf{X}, \Theta)$. According to Assumption 5, variable j can be selected at each node with probability at least δ/p. Combined with result (4.2), we consequently have

$$s_n^{(j)}(\mathbf{X}, \Theta) \xrightarrow{p} \infty. \tag{4.3}$$

Next, we break down the cell $A_n(\mathbf{X}, \Theta)$ with a collection of intervals for each of the p directions:

$$A_n(\mathbf{X}, \Theta) = \bigotimes_{j=1}^{p} A_n^{(j)}(\mathbf{X}, \Theta),$$

where each $A_n^{(j)}(\mathbf{X}, \Theta)$ is an interval and can be written as $A_n^{(j)}(\mathbf{X}, \Theta) = [l_n^{(j)}(\mathbf{X}, \Theta), u_n^{(j)}(\mathbf{X}, \Theta)]$. Then, we can bound from above the number $N_n^{(j)}(\mathbf{X}, \Theta)$ of observations whose j-th coordinate belongs to $A_n^{(j)}(\mathbf{X}, \Theta)$ using Assumption 2,

$$N_n^{(j)}(\mathbf{X}, \Theta) \leq a_n (1 - \gamma)^{s_n^{(j)}(\mathbf{X},\Theta)},$$

and using (4.3), we get that

$$N_n^{(j)}(\mathbf{X}, \Theta)/a_n \xrightarrow{p} 0.$$

Next, we introduce $F_{a_n}^{(j)}$ the empirical cdf of $X^{(j)}$, estimated with the $\Theta^{(S)}$-subsample of $\mathscr{D}_n$. Similarly, $F^{(j)}$ denotes the cdf of $X^{(j)}$. By definition, we have

$$N_n^{(j)}(\mathbf{X}, \Theta)/a_n = F_{a_n}^{(j)}(u_n^{(j)}(\mathbf{X}, \Theta)) - F_{a_n}^{(j)}(l_n^{(j)}(\mathbf{X}, \Theta)) \xrightarrow{p} 0. \tag{4.4}$$

On the other hand, we can write

$$\begin{aligned} F^{(j)}(u_n^{(j)}(\mathbf{X}, \Theta)) - F^{(j)}(l_n^{(j)}(\mathbf{X}, \Theta)) =& F_{a_n}^{(j)}(u_n^{(j)}(\mathbf{X}, \Theta)) - F_{a_n}^{(j)}(l_n^{(j)}(\mathbf{X}, \Theta)) \\ &- [F_{a_n}^{(j)}(u_n^{(j)}(\mathbf{X}, \Theta)) - F^{(j)}(u_n^{(j)}(\mathbf{X}, \Theta))] \\ &+ [F_{a_n}^{(j)}(l_n^{(j)}(\mathbf{X}, \Theta)) - F^{(j)}(l_n^{(j)}(\mathbf{X}, \Theta))], \end{aligned}$$

and we get the following bound

$$\begin{aligned} F^{(j)}(u_n^{(j)}(\mathbf{X}, \Theta)) - F^{(j)}(l_n^{(j)}(\mathbf{X}, \Theta)) \leq & F_{a_n}^{(j)}(u_n^{(j)}(\mathbf{X}, \Theta)) - F_{a_n}^{(j)}(l_n^{(j)}(\mathbf{X}, \Theta)) \\ &+ 2 \sup_{z \in [0,1]} |F_{a_n}^{(j)}(z) - F^{(j)}(z)|. \end{aligned}$$

The Glivenko-Cantelli Theorem gives that

$$\sup_{z \in [0,1]} |F_{a_n}^{(j)}(z) - F^{(j)}(z)| \xrightarrow{p} 0,$$

and combined with (4.4), we obtain

$$F^{(j)}(u_n^{(j)}(\mathbf{X}, \Theta)) - F^{(j)}(l_n^{(j)}(\mathbf{X}, \Theta)) \xrightarrow{p} 0. \tag{4.5}$$

Finally, using the integral form of the difference above, we have

$$F^{(j)}(u_n^{(j)}(\mathbf{X}, \Theta)) - F^{(j)}(l_n^{(j)}(\mathbf{X}, \Theta)) = \int_{A_n^{(j)}(\mathbf{X}, \Theta)} f^{(j)}(x) dx,$$

and since $f^{(j)}$ is lower bounded by c_1 according to Assumption 1,

$$F^{(j)}(u_n^{(j)}(\mathbf{X}, \Theta)) - F^{(j)}(l_n^{(j)}(\mathbf{X}, \Theta)) \geq c_1 \text{diam}(A_n^{(j)}(\mathbf{X}, \Theta)).$$

This last inequality combined with limit (4.5) gives

$$\text{diam}(A_n^{(j)}(\mathbf{X}, \Theta)) \xrightarrow{p} 0,$$

and since this is true for each direction $j = 1, \ldots, p$, the final result follows. Then, we have in probability

$$\lim_{n \to \infty} \text{diam}(A_n(\mathbf{X}, \Theta)) = 0.$$

□

The proof of Lemma 6 is based on Theorem 10.2 from Györfi et al. (2006) and Theorem 1 from Scornet et al. (2015). First, we introduce several notations following Scornet et al. (2015). The partition of $[0,1]^{p-1}$ obtained with the Θ-random tree projected along the j-th direction is denoted by $\mathcal{P}_n^{(-j)}(\mathscr{D}_n, \Theta)$. We define the family of all achievable partitions with Θ as

$$\Pi_n^{(-j)}(\Theta) = \{\mathcal{P}^{(-j)}((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n), \Theta) : (\mathbf{x}_i, y_i) \in [0,1]^{p-1} \times \mathbb{R}\},$$

and the associated maximal number $M(\Pi_n^{(-j)}(\Theta))$ of terminal nodes among all partitions in $\Pi_n^{(-j)}(\Theta)$ is

$$M(\Pi_n^{(-j)}(\Theta)) = \max\{|\mathcal{P}| : \mathcal{P} \in \Pi_n^{(-j)}(\Theta)\}.$$

Next, we consider $\mathbf{z}_1, \dots, \mathbf{z}_n \in [0,1]^{p-1}$ and denotes $\Gamma(\mathbf{z}_1, \dots, \mathbf{z}_n, \Pi_n^{(-j)}(\Theta))$ the number of distinct partitions of $\mathbf{z}_1, \dots, \mathbf{z}_n$ induced by the elements of $\Pi_n^{(-j)}(\Theta)$. Then, the partitioning number $\Gamma(\Pi_n^{(-j)}(\Theta))$ is defined as

$$\Gamma(\Pi_n^{(-j)}(\Theta)) = \max\{\Gamma(\mathbf{z}_1, \dots, \mathbf{z}_n, \Pi_n^{(-j)}(\Theta)) : \mathbf{z}_1, \dots, \mathbf{z}_n \in [0,1]^{p-1}\}.$$

We define the truncated operator T_L for $L > 0$. Thus, the truncated tree estimate $T_L m_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)$ returns the constant L whenever $|m_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)| > L$. Finally, we define $\mathcal{F}_n^{(-j)}(\Theta)$ the set of piecewise constant functions over the partition $\mathcal{P}_n^{(-j)}(\mathscr{D}_n, \Theta)$. Then, the projected tree estimate $m_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)$ is defined as the element of $\mathcal{F}_n^{(-j)}(\Theta)$ which minimizes the quadratic risk.

For the sake of clarity, we recall Theorem 10.2 from Györfi et al. (2006), as presented in Scornet et al. (2015) in the case of random forests.

Theorem 3 (Theorem 10.2 in Györfi et al. (2006)) *Assume that*

(*i*) $\lim_{n\to\infty} \beta_n = \infty$,

(*ii*) $\lim_{n\to\infty} \mathbb{E}\Big[\inf_{f \in \mathcal{F}_n^{(-j)}(\Theta), \|f\|_\infty \le \beta_n} \mathbb{E}[(f(\boldsymbol{X}^{(-j)}) - m^{(-j)}(\boldsymbol{X}^{(-j)}))^2]\Big] = 0$,

(*iii*) *for all* $L > 0$,

$$\lim_{n\to\infty} \mathbb{E}\Big[\sup_{\substack{f \in \mathcal{F}_n^{(-j)}(\Theta), \\ \|f\|_\infty \le \beta_n}} \Big|\frac{1}{a_n} \sum_{i \in \Theta^{(S)}} [f(\boldsymbol{X}_i^{(-j)}) - Y_{i,L}]^2 - \mathbb{E}[(f(\boldsymbol{X}^{(-j)}) - Y_L)^2]\Big|\Big] = 0.$$

Then, we have

$$\lim_{n\to\infty} \mathbb{E}[(T_{\beta_n} m_n^{(-j)}(\boldsymbol{X}^{(-j)}) - m^{(-j)}(\boldsymbol{X}^{(-j)}))^2] = 0.$$

of Lemma 6. We assume that Assumptions 1, 5, and 6 are satisfied, and we fix $j \in \{1, \dots, p\}$. We closely follow the proof of Theorem 1 from Scornet et al. (2015) to adapt it to the case of projected forest.

(i) We set $\beta_n = ||m||_\infty + \mathbb{V}[\varepsilon]\sqrt{2}\log^2(a_n)$. By definition, $\beta_n \to \infty$ and (i) is satisfied.

(ii) Approximation Error. Fix $\xi > 0$. We can show that (see Scornet et al. (2015, page 17) for the details), for n large enough such that $\beta_n > ||m||_\infty$,

$$\mathbb{E}\Big[\inf_{\substack{f \in \mathcal{F}_n^{(-j)}(\Theta), \\ ||f||_\infty \leq \beta_n}} \mathbb{E}[(f(\mathbf{X}^{(-j)}) - m^{(-j)}(\mathbf{X}^{(-j)}))^2] \Big] < \xi^2 + 4||m||_\infty^2 \mathbb{P}(\Delta(m, A_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)) > \xi).$$

On the other hand, observe that $A_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)$ is included in the projection of $A_n(\mathbf{X}, \Theta)$ along the j-th direction by construction—see Figure 5 for an illustration. Furthermore, when a cell is projected, its diameter is smaller than the original one. Thus, we have

$$\operatorname{diam}(A_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)) \leq \operatorname{diam}(A_n(\mathbf{X}, \Theta)).$$

and consequently Lemma 5 implies that in probability

$$\lim_{n\to\infty} \operatorname{diam}(A_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)) = 0.$$

Since m is continuous, the control on the cell diameter implies that

$$\Delta(m, A_n^{(-j)}(\mathbf{X}^{(-j)}, \Theta)) \xrightarrow{p} 0.$$

This enables to control the approximation error, i.e., for n large enough

$$\mathbb{E}\Big[\inf_{\substack{f \in \mathcal{F}_n^{(-j)}(\Theta), \\ ||f||_\infty \leq \beta_n}} \mathbb{E}[(f(\mathbf{X}^{(-j)}) - m^{(-j)}(\mathbf{X}^{(-j)}))^2] \Big] < 2\xi^2,$$

and therefore (ii) is satisfied.

(iii) Estimation Error. The number of terminal leaves in the original tree is t_n. Consequently, the number of leaves in the projected tree is upper bounded by 2^{t_n}. Thus, by definition $M(\Pi_n^{(-j)}(\Theta)) \leq 2^{t_n}$, and simple calculations give $\Gamma(\Pi_n^{(-j)}(\Theta)) \leq [(p-1)a_n]^{2^{t_n}}$. Since Assumption 6 ensures that $\lim_{n\to\infty} 2^{t_n} \frac{(\log(a_n))^9}{a_n} = 0$, we can show (iii) exactly as in Scornet et al. (2015, page 17-18).

Since (i), (ii), and (iii) are satisfied, Theorem 3 gives the consistency of the truncated projected tree estimate,

$$\lim_{n\to\infty} \mathbb{E}[(T_{\beta_n} m_n^{(-j)}(\mathbf{X}^{(-j)}) - m^{(-j)}(\mathbf{X}^{(-j)}))^2] = 0.$$

Finally, the extension to the untruncated projected tree estimate strictly follows Scornet et al. (2015, pages 18-19) when the noise is Gaussian, and is still valid for our case of a sub-Gaussian noise (Assumption 1). Overall, we have

$$\lim_{n\to\infty} \mathbb{E}[(m_n^{(-j)}(\mathbf{X}^{(-j)}) - m^{(-j)}(\mathbf{X}^{(-j)}))^2] = 0.$$

□

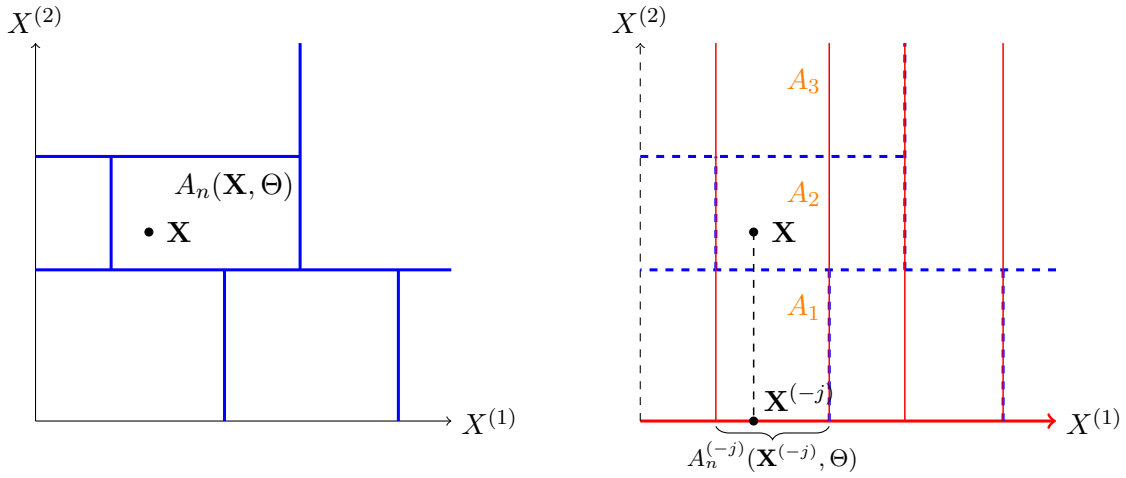

Figure 5: Example of the partition of $[0,1]^2$ by a random CART tree (left side) projected on the subspace span by $\mathbf{X}^{(-2)} = X^{(1)}$ (right side). Here, $p = 2$ and $j = 2$.

of Lemma 7. We assume that Assumptions 1, 5, and 6 are satisfied, and we fix $j \in \{1, \dots, p\}$. First, we expand the considered risk

$$\begin{aligned}\mathbb{E}[(m_{M,n}^{(-j,OOB)}(\mathbf{X}_i, \Theta_M) - m(\mathbf{X}_i^{(-j)}))^2 \mathbb{1}_{|\Lambda_{n,i}|>0}] \\ = \mathbb{E}\Big[\Big(\frac{1}{|\Lambda_{n,i}|} \sum_{\ell \in \Lambda_{n,i}} [m_n^{(-j)}(\mathbf{X}_i^{(-j)}, \Theta_\ell) - m(\mathbf{X}_i^{(-j)})] \mathbb{1}_{|\Lambda_{n,i}|>0}\Big)^2\Big].\end{aligned}$$

Then, identically to the proof of Lemma 2, we can handle the randomness of the selected batch of trees $\Lambda_{n,i}$, and bound the OOB risk with the risk of the standard projected forest, i.e.,

$$\begin{aligned}\mathbb{E}\big[\big(m_{M,n}^{(-j,OOB)}(\mathbf{X}_i^{(-j)}, \Theta_M) - m(\mathbf{X}_i^{(-j)})\big)^2 \mathbb{1}_{|\Lambda_{n,i}|>0}\big] \\ \leq \frac{2}{1 - a_n/n} \mathbb{E}\big[\big(m_{M,a_n,n-1}^{(-j)}(\mathbf{X}^{(-j)}, \Theta_M) - m(\mathbf{X}^{(-j)})\big)^2\big].\end{aligned}$$

Lemma 6 gives the consistency of the infinite projected forest, which also implies the consistency of the finite projected forest, that is

$$\mathbb{E}\big[\big(m_{M,a_n,n-1}^{(-j)}(\mathbf{X}^{(-j)}, \Theta_M) - m(\mathbf{X}^{(-j)})\big)^2\big] \longrightarrow 0.$$

Additionally, from Assumption 6, $a_n/n < 1 - \kappa$ with $\kappa > 0$, and thus

$$\lim_{n \to \infty} \mathbb{E}[(m_{M,n}^{(-j,OOB)}(\mathbf{X}_i^{(-j)}, \Theta_M) - m(\mathbf{X}_i^{(-j)}))^2 \mathbb{1}_{|\Lambda_{n,i}|>0}] = 0.$$

□

5 MDA Software Implementations

We provide detailed references of the MDA implementations of the main random forest packages:

1. `scikit-learn 0.24`
   (*https://scikit-learn.org/stable/*)

2. `randomForest 4.6-14`
   (*https://cran.r-project.org/web/packages/randomForest/index.html*)

3. `ranger 0.12.1`
   (*https://cran.r-project.org/web/packages/ranger/index.html*)

4. `randomForestSRC 2.9.3`
   (*https://cran.r-project.org/web/packages/randomForestSRC/index.html*)

5.1 `scikit-learn 0.24`

In `scikit-learn`, the MDA is not specific for random forests, but is a generic procedure taking a trained model and an independent testing sample as inputs. The MDA implementation is located in the file: "scikit-learn/sklearn/inspection/_permutation_importance.py".

The method *_calculate_permutation_scores(estimator, X, y, sample_weight, col_idx, random_state, n_repeats, scorer)* computes the error of the model *estimator* when the column of index *col_idx* of the testing sample X is permuted, over multiple repetitions defined by the parameter *n_repeats*. The model error is defined by *scorer*, and *random_state* defines the random seed. Finally, the permuted and the original errors are subtracted and the multiple repetitions are aggregated in the method *permutation_importance(estimator, X, y, *, scoring=None, n_repeats=5, n_jobs=None, random_state=None)* which thus implements the Train/Test MDA.

5.2 `randomForest 4.6-14`

The `R` script "randomForest/R/importance.R" implements the function *importance.randomForest <- function(x, type=NULL, class=NULL, scale=TRUE, ...)* between lines 6 and 44, where x is a fitted forest, which as the attribute *x$importance* storing the Breiman-Cutler MDA and the standard deviation of the risk differences across trees, computed with the script "randomForest/src/regrf.c" for regression forests. The function *importance.randomForest* handles exceptions and normalizes the MDA with the standard deviations, and thus implements the normalized Breiman-Cutler MDA.

For regression forests, the `C` script "randomForest/src/regrf.c" computes the difference between the permuted and original errors for each tree between lines 262 and 295. The associated means and standard deviations across all trees are computed between lines 327 and 338. These computations are done right after the forest construction at the end of the method *void regRF*.

5.3 `ranger 0.12.1`

In `ranger`, the MDA is computed during the forest growing by specifying the paramater *importance = 'permutation'* in the call to the main function *ranger*. For each tree of the forest,

the accuracy decrease is computed in the `C++` file "ranger/src/Tree.cpp" with the method *void Tree::computePermutationImportance()*, located between lines 206 and 255. Next, the importance measures are averaged over all trees with the method *void Forest:: computePermutationImportance()* between lines 646 and 763 of the `C++` file "ranger/src/Forest.cpp", and thus the BC-MDA is computed. If the paramater *scale.permutation.importance* is set to *True*, then the normalized BC-MDA is computed (default value is *False*).

5.4 `randomForestSRC 2.9.3`

The package `randomForestSRC` can compute the three types of MDA. The function *vimp.rfsrc* (lines 1 to 82 of file "randomForestSRC/R/vimp.rfsrc.R") computes the MDA, and takes a fitted forest *object* as an input. If an independent testing sample is provided as the input *newdata*, TT-MDA is computed. Otherwise if *importance = 'permute'*, the IK-MDA by blocks is estimated: the trees of the forest are divided in multiple blocks and the IK-MDA is computed for each block and averaged. The parameter *block.size* set the number of trees in each block, 10 by default. If *block.size* = 1, this procedure is the BC-MDA.

The function *vimp.rfsrc* computes the MDA calling a chain of `C` subroutines, located in the file "randomForestSRC/src/randomForestSRC.c" between lines 2026 and 2564: *permute*, *getPermuteMembership*, *getVimpMembership*, *updateVimpEnsemble*, *summarizePerturbedPerformance*, and *finalizeVimpPerformance*.

6 Analytical Example Computations

We first recall the analytical example definition, and all computations are provided next. The input $\mathbf{X}$ is a Gaussian vector of dimension $p = 5$. Its covariance matrix is defined by $\mathbb{V}[X^{(j)}] = \sigma_j^2$ for $j \in \{1, \dots, 5\}$, and all covariance terms are null except

$$\mathrm{Cov}[X^{(1)}, X^{(2)}] = \rho_{1,2}\sigma_1\sigma_2,$$

and

$$\mathrm{Cov}[X^{(4)}, X^{(5)}] = \rho_{4,5}\sigma_4\sigma_5.$$

The regression function m is given by

$$m(\mathbf{X}) = \alpha X^{(1)}X^{(2)}\mathbb{1}_{X^{(3)}>0} + \beta X^{(4)}X^{(5)}\mathbb{1}_{X^{(3)}<0}.$$

6.1 Total Sobol Index $ST^{(1)}$.

By definition, $\mathbb{V}[Y] \times ST^{(1)} = \mathbb{E}[\mathbb{V}[m(\mathbf{X})|\mathbf{X}^{(-1)}]]$. Since $X^{(1)}$ and $X^{(2)}$ are independent of $X^{(3)}$, $X^{(4)}$, and $X^{(5)}$, we have

$$\begin{aligned}
\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-1)}] &= \mathbb{E}[\alpha X^{(1)}X^{(2)}\mathbb{1}_{X^{(3)}>0} + \beta X^{(4)}X^{(5)}\mathbb{1}_{X^{(3)}<0}|\mathbf{X}^{(-1)}] \\
&= \mathbb{E}[\alpha X^{(1)}X^{(2)}\mathbb{1}_{X^{(3)}>0}|X^{(2)}] + \beta X^{(4)}X^{(5)}\mathbb{1}_{X^{(3)}<0} \\
&= \alpha X^{(2)}\mathbb{E}[X^{(1)}|X^{(2)}]\mathbb{1}_{X^{(3)}>0} + \beta X^{(4)}X^{(5)}\mathbb{1}_{X^{(3)}<0}.
\end{aligned}$$

Since $(X^{(1)}, X^{(2)})$ is a bivariate centered Gaussian vector,

$$\mathbb{E}[X^{(1)}|X^{(2)}] = \rho_{1,2}\frac{\sigma_1}{\sigma_2}X^{(2)},$$

and then

$$\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-1)}] = \alpha\rho_{1,2}\frac{\sigma_1}{\sigma_2}X^{(2)2}\mathbb{1}_{X^{(3)}>0} + \beta X^{(4)}X^{(5)}\mathbb{1}_{X^{(3)}<0}.$$

Next, we compute

$$\begin{aligned}
\mathbb{E}[\mathbb{V}[m(\mathbf{X})|\mathbf{X}^{(-1)}]] &= \mathbb{E}[(m(\mathbf{X}) - \mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-1)}])^2] \\
&= \mathbb{E}[(\alpha X^{(1)}X^{(2)}\mathbb{1}_{X^{(3)}>0} - \alpha\rho_{1,2}\frac{\sigma_1}{\sigma_2}X^{(2)2}\mathbb{1}_{X^{(3)}>0})^2] \\
&= \frac{\alpha^2}{2}\mathbb{E}[(X^{(1)}X^{(2)} - \rho_{1,2}\frac{\sigma_1}{\sigma_2}X^{(2)2})^2] \\
&= \frac{\alpha^2}{2}\big(\mathbb{E}[(X^{(1)}X^{(2)})^2] + (\rho_{1,2}\frac{\sigma_1}{\sigma_2})^2\mathbb{E}[X^{(2)4}] - 2\rho_{1,2}\frac{\sigma_1}{\sigma_2}\mathbb{E}[X^{(1)}X^{(2)3}]\big).
\end{aligned}$$

Standard formulas give

$$\mathbb{E}[(X^{(1)}X^{(2)})^2] = (1 + 2\rho_{1,2}^2)\sigma_1^2\sigma_2^2,$$

$$\mathbb{E}[X^{(2)4}] = 3\sigma_2^4,$$

and

$$\mathbb{E}[X^{(1)}X^{(2)3}] = \mathbb{E}[X^{(2)3}\mathbb{E}[X^{(1)}|X^{(2)}]] = \rho_{1,2}\frac{\sigma_1}{\sigma_2}\mathbb{E}[X^{(2)4}].$$

Using these last three formulas in the previous result, we get

$$\begin{aligned}
\mathbb{E}[\mathbb{V}[m(\mathbf{X})|\mathbf{X}^{(-1)}]] &= \frac{\alpha^2}{2}\big[(1 + 2\rho_{1,2}^2)\sigma_1^2\sigma_2^2 + (\rho_{1,2}\frac{\sigma_1}{\sigma_2})^2 3\sigma_2^4 - 2(\rho_{1,2}\frac{\sigma_1}{\sigma_2})^2 3\sigma_2^4\big] \\
&= \frac{\alpha^2}{2}\big[(1 + 2\rho_{1,2}^2)\sigma_1^2\sigma_2^2 + 3(\rho_{1,2}\sigma_1\sigma_2)^2 - 6(\rho_{1,2}\sigma_1\sigma_2)^2\big] \\
&= \boldsymbol{\frac{1}{2}(\alpha\sigma_1\sigma_2)^2(1 - \rho_{1,2}^2)}.
\end{aligned}$$

6.2 Marginal Total Sobol Index $ST_{mg}^{(1)}$.

By definition, $\mathbb{V}[Y] \times ST_{mg}^{(1)} = \mathbb{E}[\mathbb{V}[m(\mathbf{X}_{\pi_1})|\mathbf{X}^{(-1)}]]$.

$$\begin{aligned}
\mathbb{E}[\mathbb{V}[m(\mathbf{X}_{\pi_1})|\mathbf{X}^{(-1)}]] &= \mathbb{E}[(m(\mathbf{X}_{\pi_1}) - \mathbb{E}[m(\mathbf{X}_{\pi_1})|\mathbf{X}^{(-1)}])^2] \\
&= \mathbb{E}[(\alpha X'^{(1)}X^{(2)}\mathbb{1}_{X^{(3)}>0} - \alpha\mathbb{E}[X'^{(1)}|\mathbf{X}^{(-1)}]X^{(2)}\mathbb{1}_{X^{(3)}>0})^2],
\end{aligned}$$

where $X'^{(1)}$ is an iid copy of $X^{(1)}$. Therefore $X'^{(1)}$ is independent of $\mathbf{X}$ and $\mathbb{E}[X'^{(1)}|\mathbf{X}^{(-1)}] = 0$, and we get

$$\begin{aligned}
\mathbb{E}[\mathbb{V}[m(\mathbf{X}_{\pi_1})|\mathbf{X}^{(-1)}]] &= \frac{\alpha^2}{2}\mathbb{E}[(X'^{(1)}X^{(2)})^2] = \frac{\alpha^2}{2}\mathbb{E}[(X'^{(1)}]\mathbb{E}[X^{(2)})^2] \\
&= \boldsymbol{\frac{1}{2}(\alpha\sigma_1\sigma_2)^2}.
\end{aligned}$$

6.3 Third MDA Component $MDA_3^{(1)}$.

By definition,

$$MDA_3^{(1)} = \mathbb{E}[(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-1)}] - \mathbb{E}[m(\mathbf{X}_{\pi_1})|\mathbf{X}^{(-1)}])^2]$$

As computed above for the marginal total Sobol index, $\mathbb{E}[m(\mathbf{X}_{\pi_1})|\mathbf{X}^{(-1)}] = \beta X^{(4)}X^{(5)}\mathbb{1}_{X^{(3)}>0}$, thus

$$\begin{aligned}
MDA_3^{(1)} &= \mathbb{E}[(\alpha X^{(1)}\mathbb{E}[X^{(2)}|\mathbf{X}^{(-1)}]\mathbb{1}_{X^{(3)}>0})^2] \\
&= \frac{1}{2}\alpha^2\mathbb{E}[(X^{(1)}\mathbb{E}[X^{(2)}|X^{(1)}])^2] \\
&= \frac{1}{2}\alpha^2(\rho_{1,2}\frac{\sigma_1}{\sigma_2})^2\mathbb{E}[X^{(2)4}] \\
&= \boldsymbol{\frac{3}{2}\rho_{1,2}^2(\alpha\sigma_1\sigma_2)^2}.
\end{aligned}$$

6.4 Final MDA Limits

Overall, using Proposition 2, we obtain

$$\begin{aligned}
\mathrm{MDA}^{\star(1)} =& \underbrace{\frac{1}{2}(\alpha\sigma_1\sigma_2)^2(1-\rho_{1,2}^2)}_{\mathrm{MDA}_1^{\star(1)}} + \underbrace{\frac{1}{2}(\alpha\sigma_1\sigma_2)^2}_{\mathrm{MDA}_2^{\star(1)}} + \underbrace{\frac{3}{2}\rho_{1,2}^2(\alpha\sigma_1\sigma_2)^2}_{\mathrm{MDA}_3^{\star(1)}} \\
\mathrm{MDA}^{\star(1)} =& \boldsymbol{(\alpha\sigma_1\sigma_2)^2(1+\rho_{1,2}^2)}.
\end{aligned}$$

By symmetry, $\mathrm{MDA}^{\star(2)} = \mathrm{MDA}^{\star(1)} = \boldsymbol{(\alpha\sigma_1\sigma_2)^2(1+\rho_{1,2}^2)}$, and

$$\mathrm{MDA}^{\star(4)} = \mathrm{MDA}^{\star(5)} = \boldsymbol{(\beta\sigma_4\sigma_5)^2(1+\rho_{4,5}^2)}.$$

Finally, since $X^{(3)}$ is independent of the other variables, Corollary 1 gives

$$\mathrm{MDA}^{\star(3)} = 2\mathrm{MDA}_1^{\star(3)} = 2\mathbb{E}[\mathbb{V}[m(\mathbf{X})|\mathbf{X}^{(-3}]].$$

Next,

$$\begin{aligned}
\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-3)}] &= \mathbb{E}[\alpha X^{(1)}X^{(2)}\mathbb{1}_{X^{(3)}>0} + \beta X^{(4)}X^{(5)}\mathbb{1}_{X^{(3)}<0}|\mathbf{X}^{(-3)}] \\
&= \frac{1}{2}\alpha X^{(1)}X^{(2)} + \frac{1}{2}\beta X^{(4)}X^{(5)},
\end{aligned}$$

and

$$\mathbb{V}[\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-3)}]] = \frac{1}{4}\alpha^2\mathbb{V}[X^{(1)}X^{(2)}] + \frac{1}{4}\beta^2\mathbb{V}[X^{(4)}X^{(5)}].$$

Since

$$\begin{aligned}
\mathbb{V}[X^{(1)}X^{(2)}] =& \mathbb{E}[(X^{(1)}X^{(2)})^2] - \mathbb{E}[X^{(1)}X^{(2)}]^2 \\
=& (1+2\rho_{1,2}^2)\sigma_1^2\sigma_2^2 - (\rho_{1,2}\sigma_1\sigma_2)^2 \\
=& (1+\rho_{1,2}^2)\sigma_1^2\sigma_2^2,
\end{aligned}$$

we obtain

$$\mathbb{V}[\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-3)}]] = \frac{1}{4}\alpha^2(1+\rho_{1,2}^2)\sigma_1^2\sigma_2^2 + \frac{1}{4}\beta^2(1+\rho_{4,5}^2)\sigma_4^2\sigma_5^2.$$

On the other hand,

$$\begin{aligned}
\mathbb{V}[m(\mathbf{X})] =& \alpha^2\mathbb{V}[X^{(1)}X^{(2)}\mathbb{1}_{X^{(3)}>0}] + \beta^2\mathbb{V}[X^{(4)}X^{(5)}\mathbb{1}_{X^{(3)}<0}] \\
&+ 2\text{Cov}[\alpha X^{(1)}X^{(2)}\mathbb{1}_{X^{(3)}>0}, \beta X^{(4)}X^{(5)}\mathbb{1}_{X^{(3)}<0}] \\
=& \frac{\alpha^2}{2}(1+2\rho_{1,2}^2)\sigma_1^2\sigma_2^2 - \frac{\alpha^2}{4}(\rho_{1,2}\sigma_1\sigma_2)^2 + \frac{\beta^2}{2}(1+2\rho_{4,5}^2)\sigma_4^2\sigma_5^2 - \frac{\beta^2}{4}(\rho_{4,5}\sigma_4\sigma_5)^2 \\
&- 2\alpha\beta\frac{1}{4}\mathbb{E}[X^{(1)}X^{(2)}]\mathbb{E}[X^{(4)}X^{(5)}] \\
=& \frac{\alpha^2}{2}(1+\frac{3}{2}\rho_{1,2}^2)\sigma_1^2\sigma_2^2 + \frac{\beta^2}{2}(1+\frac{3}{2}\rho_{4,5}^2)\sigma_4^2\sigma_5^2 - 2\alpha\beta\frac{1}{4}\rho_{1,2}\sigma_1\sigma_2\rho_{4,5}\sigma_4\sigma_5.
\end{aligned}$$

Finally,

$$\begin{aligned}
\text{MDA}^{\star(3)} &= 2\mathbb{E}[\mathbb{V}[m(\mathbf{X})|\mathbf{X}^{(-3}]] = 2(\mathbb{V}[m(\mathbf{X})] - \mathbb{V}[\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-3)}]]) \\
&= 2(\frac{\alpha^2}{4}(1+2\rho_{1,2}^2)\sigma_1^2\sigma_2^2 + \frac{\beta^2}{4}(1+2\rho_{4,5}^2)\sigma_4^2\sigma_5^2 - 2\alpha\beta\frac{1}{4}\rho_{1,2}\sigma_1\sigma_2\rho_{4,5}\sigma_4\sigma_5) \\
&= \boldsymbol{\frac{1}{2}(\alpha\sigma_1\sigma_2)^2(1+\rho_{1,2}^2) + \frac{1}{2}(\beta\sigma_4\sigma_5)^2(1+\rho_{4,5}^2) + \frac{1}{2}(\alpha\rho_{1,2}\sigma_1\sigma_2 - \beta\rho_{4,5}\sigma_4\sigma_5)^2}.
\end{aligned}$$

6.5 High Correlation Setting.

In a high correlation setting, the third term becomes the main MDA contribution for variables $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, $\mathbf{X}^{(4)}$, and $\mathbf{X}^{(5)}$. Since computations are similar, we only consider $\mathbf{X}^{(1)}$:

$$\begin{aligned}
\text{MDA}_3^{\star(1)} &> \text{MDA}_1^{\star(1)} + \text{MDA}_2^{\star(1)} \\
\frac{3}{2}\rho_{1,2}^2(\alpha\sigma_1\sigma_2)^2 &> \frac{1}{2}(\alpha\sigma_1\sigma_2)^2(1-\rho_{1,2}^2) + \frac{1}{2}(\alpha\sigma_1\sigma_2)^2 \\
3\rho_{1,2}^2(\alpha\sigma_1\sigma_2)^2 &> 2(\alpha\sigma_1\sigma_2)^2 - (\alpha\sigma_1\sigma_2)^2\rho_{1,2}^2 \\
4\rho_{1,2}^2(\alpha\sigma_1\sigma_2)^2 &> 2(\alpha\sigma_1\sigma_2)^2 \\
\rho_{1,2}^2 &> \frac{1}{2} \\
\boldsymbol{\rho_{1,2}} &> \boldsymbol{\frac{\sqrt{2}}{2}}.
\end{aligned}$$