Research Article

Clément Bénard and Julie Josse*

Variable importance for causal forests: breaking down the heterogeneity of treatment effects

https://doi.org/10.1515/jci-2023-0062
Received September 18, 2023; accepted September 20, 2025; published online December 22, 2025

Abstract: Causal random forests provide efficient estimates of heterogeneous treatment effects. However, forest algorithms are also well-known for their black-box nature, and therefore, do not characterize how input variables are involved in treatment effect heterogeneity, which is a strong practical limitation. In this article, we develop a new importance variable algorithm for causal forests, to quantify the impact of each input on the heterogeneity of treatment effects. The proposed approach is inspired from the drop and relearn principle, widely used for regression problems. Importantly, we show how to handle the case where the forest is retrained without a confounding variable. If the confounder is not involved in the treatment effect heterogeneity, the local centering step enforces consistency of the importance measure. Otherwise, when a confounder also impacts heterogeneity, we introduce a corrective term in the retrained causal forest to recover consistency. Additionally, experiments on simulated, semi-synthetic, and real data show the good performance of our importance measure, which outperforms competitors on several test cases. Experiments also show that our approach can be efficiently extended to groups of variables, providing key insights in practice.

Keywords: causal inference; heterogeneous treatment effects; causal random forests; variable importance; interpretability

MSC 2020: 62D20; 62G05; 62G20

1 Introduction

1.1 Context and objectives

Estimating heterogeneous treatment effects has recently attracted a great deal of interest in the machine learning community, particularly for medical applications [1] and in the social sciences. Over the past few years, numerous efficient algorithms have been developed to estimate such effects, including double robust methods [2], R-learners [3], X-learners [4], causal forests [5, 6], the lasso [7], BART [8], or neural networks [9]. Besides, let us also mention policy learning, which aims at selecting relevant individuals to treat [10–13]. However, most of these methods remain black boxes, and it is therefore difficult to grasp how input variables impact treatment effects. This understanding is crucial for optimizing treatment policies, for instance, so that this shortcoming clearly limits their practical use. While the accuracy of treatment effect estimates has significantly improved recently, little effort has been dedicated to improve their interpretability, and quantifying the impact of variables involved in treatment effect heterogeneity. In this regard, we can mention the importance measure of the causal forest package `grf` [14], the double robust approach of [15], and the algorithm from [16] for high dimensional linear

***Corresponding author: Julie Josse**, PreMeDICaL Project Team, INRIA-Inserm, Idesp, University of Montpellier, Montpellier, France, E-mail: julie.josse@inria.fr
Clément Bénard, Thales CortAIx-Labs, Palaiseau, France, E-mail: clement-l.benard@thalesgroup.com

cases. The main purpose of this article is to introduce a variable importance measure for heterogeneous treatment effects, improving over the existing algorithms, to better identify the sources of heterogeneity. We focus on causal random forests, defined as a specific case of generalized forests [6], and well-known to be one of most powerful algorithm to estimate heterogeneous treatment effects.

1.1.1 Contributions

Our main contribution is thus the introduction of a variable importance algorithm for causal random forests, following the drop and retrain principle, which is well-established for regression problems [17–20]. The main idea is to retrain the learning algorithm without a given input variable, and measure the drop of accuracy to get its importance. In particular, such approach ensures that irrelevant variables get a null importance asymptotically. In the context of causal inference, the main obstacle is to retrain the causal forest without a confounding variable, since the unconfoundedness assumption can be violated, leading to inconsistent forest estimates and biased importance values, as explained in Section 2. However, we will see that the local centering of the outcome and treatment assignment leads to consistent estimates, provided that the removed variable is not involved in the treatment effect heterogeneity. Otherwise, to handle a confounder involved in heterogeneity, we introduce a corrective term in the retrained causal forest. Overall, we will show in Section 3, that our proposed variable importance algorithm is consistent, under standard assumptions in the literature about the theoretical analysis of random forests. Next, in Section 4, we run several batches of experiments on simulated, semi-synthetic, and real data to show the good performance of the introduced method compared to the existing competitors. Additionally, we take advantage of the experimental section to illustrate that the extension of our approach to group of variables is straightforward and provides powerful insights in practice. The remaining of this first section is dedicated to the mathematical formalization of the problem.

1.2 Definitions

To define heterogeneous treatment effects, we first introduce a standard causal setting with an input vector $\mathbf{X} = (X^{(1)}, \ldots, X^{(p)}) \in \mathbb{R}^p$ with $p \in \mathbb{N}^\star$, the binary treatment assignment $W \in \{0, 1\}$, the potential outcome $Y(1) \in \mathbb{R}$ for the subject receiving the treatment, and the potential outcome without treatment $Y(0) \in \mathbb{R}$. We denote by $\mathbf{X}^{(\mathcal{H})}$ the subvector with only the components in $\mathcal{H} \subset \{1, \ldots, p\}$, and $\mathbf{X}^{(-j)}$ the vector $\mathbf{X}$ with the j-th component removed. The observed outcome is given by $Y = WY(1) + (1 - W)Y(0)$, which is known as the SUTVA assumption in the literature. More precisely, the potential outcomes are defined by

$$\begin{aligned} Y(0) &= \mu(\mathbf{X}) + \varepsilon(0), \\ Y(1) &= \mu(\mathbf{X}) + \tau(\mathbf{X}^{(\mathcal{H})}) + \varepsilon(1), \end{aligned}$$

where $\mu(\mathbf{X})$ is a baseline function, $\tau(\mathbf{X}^{(\mathcal{H})})$ is the conditional average treatment effect (CATE) only depending on variables in $\mathcal{H} \subset \{1, \ldots, p\}$, and $\varepsilon(0)$, $\varepsilon(1)$ are some noise variables satisfying $\mathbb{E}[\varepsilon(0) \mid \mathbf{X}] = \mathbb{E}[\varepsilon(1) \mid \mathbf{X}] = 0$. Notice that the CATE is also defined as the mean difference between potential outcomes, conditional on $\mathbf{X}$, i.e., $\mathbb{E}[Y(1) - Y(0) \mid \mathbf{X}] = \tau(\mathbf{X}^{(\mathcal{H})})$, by construction. Overall, the observed outcome Y also writes

$$Y = \mu(\mathbf{X}) + \tau(\mathbf{X}^{(\mathcal{H})}) \times W + \varepsilon(W). \tag{1}$$

The cornerstone of causal treatment effect identifiability is the assumption of unconfoundedness given in Assumption 1, which states that all confounding variables are observed in the data. By definition, the responses $Y(0)$, $Y(1)$, and the treatment assignment W simultaneously depend on the confounding variables. If all confounding variables are observed, then the responses and the treatment assignment are independent conditional on the inputs. Consequently, the treatment effect is identifiable, as stated in Proposition 1 below—all proofs of propositions and theorems stated throughout the article are gathered in Appendix A. Notice that Assumption 1 below enforces that the input vector $\mathbf{X}$ contains all confounding variables, but $\mathbf{X}$ may also contain

non-confounding variables. Consequently, $\mathbf{X}^{(\mathcal{H})}$ can also be a mix of confounding and non-confounding variables, or contain only variables of one type. Ideally, all variables impacting the treatment effect heterogeneity should be involved in the analysis, even if they are not confounding variables, to better estimate and interpret the treatment effect.

Assumption 1. Potential outcomes are independent of the treatment assignment conditional on the observed input variables, i.e., $Y(0), Y(1) \perp\!\!\!\perp W|\mathbf{X}$.

Proposition 1. If the unconfoundedness Assumption 1 is satisfied, then we have

$$\tau(\mathbf{X}^{(\mathcal{H})}) = \mathbb{E}[Y \mid \mathbf{X}, W = 1] - \mathbb{E}[Y \mid \mathbf{X}, W = 0].$$

Note that we define above the treatment effect as the expected difference between potential outcomes, conditioned on input variables. However, the heterogeneity properties strongly depend on how we define the treatment effect [21–23]. The ratio between the means of potential outcomes may also define a treatment effect, leading to potential heterogeneity while our original outcome difference remains constant. A thorough discussion of this topic is out of scope of this article, and we take the difference of potential outcomes as treatment effect, the widely used metric for many applications [21]. We refer to Colnet et al. [23] for a comparison of treatment effect measures.

VanderWeele and Robins [21] defined treatment effect heterogeneity as follows.

Definition 1: (VanderWeele and Robins [21]). The treatment effect τ is said to be heterogeneous with respect to $\mathbf{X}$ if it exists $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^p$ such that $\tau(\mathbf{x}^{(\mathcal{H})}) \neq \tau(\mathbf{x}'^{(\mathcal{H})})$.

We strengthen this definition in two directions, formalized in Definition 2 below. First, we require τ to be heterogeneous with respect to each variable in $\mathcal{H}$, to enforce $\mathcal{H}$ to be the subset of variables impacting treatment effect heterogeneity. Secondly, notice that Definition 1 can be satisfied while having an homogeneous treatment effect in probability, i.e., $\mathbb{P}(\tau(\mathbf{X}^{(\mathcal{H})}) = \tau(\mathbf{X}'^{(\mathcal{H})})) = 1$,, with $\mathbf{X}'^{(\mathcal{H})}$ an independent copy of $\mathbf{X}^{(\mathcal{H})}$. In such cases, heterogeneity is not detectable from a data sample, and has a negligible impact in practice. Therefore, we enforce τ to take distinct values with respect to all variables in $\mathcal{H}$ on sets of non-null Lebesgue measure.

Definition 2. The treatment effect τ is said to be heterogeneous with respect to all variables in $\mathcal{H}$, if for all $j \in \mathcal{H}$, it exists $\mathcal{X}_{p-1} \subset \mathbb{R}^{p-1}$ and $\mathcal{X}_1, \mathcal{X}_1' \subset \mathbb{R}$, such that for all $\mathbf{x}^{(-j)} \in \mathcal{X}_{p-1}$, $x^{(j)} \in \mathcal{X}_1$, $x'^{(j)} \in \mathcal{X}_1'$, we have

$$\tau(\mathbf{x}^{(\mathcal{H})}) \neq \tau(\mathbf{x}'^{(\mathcal{H})}),$$

with $\mathbf{x}'^{(-j)} = \mathbf{x}^{(-j)}$, and $\mathcal{X}_{p-1}$, $\mathcal{X}_1$, and $\mathcal{X}_1'$ have a non-null Lebesgue measure.

In the sequel, we assume that the treatment effect τ is heterogeneous in the sense of Definition 2, and that $\mathbf{X}$ admits a strictly positive density, to enforce heterogeneity with a positive probability, as stated in the proposition below. Our objective is to quantify the influence of the input variables $\mathbf{X}$ on the treatment heterogeneity using an available sample $\mathscr{D}_n = \{(\mathbf{X}_i, Y_i, W_i)\}_{i=1}^n$, made of $n \in \mathbb{N}^*$ independent and identically distributed (iid) observations.

Assumption 2. The treatment effect τ is heterogeneous according to Definition 2, and $\mathbf{X}$ admits a strictly positive density.

Proposition 2. If Assumption 2 is satisfied, and $\mathbf{X}'$ is an independent copy of $\mathbf{X}$, then

$$\mathbb{P}(\tau(\mathbf{X}^{(\mathcal{H})}) \neq \tau(\mathbf{X}'^{(\mathcal{H})})) > 0.$$

2 Variable importance for heterogeneous treatment effects

2.1 Theoretical definition

To propose a variable importance measure, we build on Sobol [24] and Williamson et al. [18], which define variable importance in the case of regression as the proportion of output explained variance lost when a given input variable is removed. Hines et al. [15] extend this idea to treatment effects, and introduce the theoretical importance measure $\mathrm{I}^{(j)}$ of $X^{(j)}$, defined by

$$\mathrm{I}^{(j)} = \frac{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})] - \mathbb{V}[\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})|\mathbf{X}^{(-j)}]]}{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})]} = \frac{\mathbb{E}[(\tau(\mathbf{X}^{(\mathcal{H})}) - \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})|\mathbf{X}^{(-j)}])^2]}{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})]}, \tag{2}$$

which is well-defined under Assumption 2, since $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})] > 0$. Otherwise, when $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})] = 0$, the treatment is homogeneous, i.e. constant with respect to all input variables, and does not satisfy Definition 2. This importance measure gives the proportion of treatment effect variance lost when a given input variable is removed, and is called the total Sobol index of τ in sensitivity analysis. Additionally, the following proposition shows that $\mathrm{I}^{(j)}$ properly identifies variables in $\mathcal{H}$, which have an impact on treatment heterogeneity, where the proof in Appendix A is a consequence of Assumption 2.

Proposition 3. Let Assumption 2 be satisfied. If $j \notin \mathcal{H}$, then we have $\mathrm{I}^{(j)} = 0$. Otherwise, if $j \in \mathcal{H}$, we have $0 < \mathrm{I}^{(j)} \leq 1$.

Note that by definition of $\mathrm{I}^{(j)}$, a variable strongly correlated to another variable involved in the heterogeneity, has a low importance value. This is due to the fact that, owing to this strong dependence, there is minimal loss of information regarding the treatment effect heterogeneity when such a variable is removed. As suggested by both Williamson et al. [18] and Hines et al. [15], one possible approach involves extending the importance measure to a group of variables, where strongly dependent variables are grouped together. For the sake of clarity, we focus on the case of a single variable in the following sections. However, extending this approach to groups of variables is straightforward, and we will present such examples in the experimental section.

More importantly, Hines et al. [15] highlight that a key problem to estimate the above quantity $\mathrm{I}^{(j)}$, is that the unconfoundedness Assumption 1 does not imply unconfoundedness for the reduce set of input variables $\mathbf{X}^{(-j)}$, i.e., we may have $Y(0), Y(1) \not\perp\!\!\!\perp W|\mathbf{X}^{(-j)}$. Hines et al. [15] overcome this issue using double robust approaches [2, 3] to estimate τ with all input variables in a first step, and then regress the obtained treatment effect on $\mathbf{X}^{(-j)}$ to estimate $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})|\mathbf{X}^{(-j)}]$. Actually, the generalized random forest framework from Athey et al. [6] enables to get closer to the original proposal of Williamson et al. [18] by retraining the causal forest without variable $X^{(j)}$ and still get consistent estimates of $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})|\mathbf{X}^{(-j)}]$, as we will see. Therefore, we focus on causal forests [5, 6], one of the state-of-the-art algorithm to estimate heterogeneous treatment effects, to propose efficient estimates of $\mathrm{I}^{(j)}$. Hence, the proposed approach differs from Hines et al. [15], since we estimate $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})|\mathbf{X}^{(-j)}]$ from scratch, whereas Hines et al. [15] reuse the initial estimate of $\tau(\mathbf{X}^{(\mathcal{H})})$ and solve a regression problem to obtain the treatment effect with a variable removed.

2.2 Causal random forests

Generalized random forests [6] are a generic framework to build efficient estimates of quantities defined as solutions of local moment equations. As opposed to original Breiman's forests, generalized forests are not the average of tree outputs. Instead, trees are aggregated to generate weights for each observation of the training data, used in a second step to build a weighted estimate of the target quantity. Causal forests are a specific case of generalized forest, where the following local moment equation identifies the treatment effect under the unconfoundedness Assumption (1),

$$\tau(\mathbf{X}^{(\mathcal{H})}) \times \mathbb{V}[W \mid \mathbf{X}] - \mathrm{Cov}[W, Y \mid \mathbf{X}] = 0. \tag{3}$$

The local moment Equation (3) is thus used to define the causal forest estimate $\tau_{M,n}(\mathbf{x})$ at a new query point $\mathbf{x}$, built from the data $\mathscr{D}_n$ with $M \in \mathbb{N}^\star$ trees, and formally defined in Athey et al. [6, Section 6.1] by

$$\tau_{M,n}(\mathbf{x}) = \frac{\sum_{i=1}^n \alpha_i(\mathbf{x}) W_i Y_i - \overline{W}_\alpha \overline{Y}_\alpha}{\sum_{i=1}^n \alpha_i(\mathbf{x}) (W_i - \overline{W}_\alpha)^2}, \tag{4}$$

where $\overline{Y}_\alpha = \sum_{i=1}^n \alpha_i(\mathbf{x}) Y_i$, $\overline{W}_\alpha = \sum_{i=1}^n \alpha_i(\mathbf{x}) W_i$, and the weights $\alpha_i(\mathbf{x})$ are generated by the forest to quantify the frequency of $\mathbf{x}$ and the training observation $\mathbf{X}_i$ both falling in the same terminal leaves of trees. Notice that the ℓ-th tree of the forest is randomized by Θ_ℓ, which defines the resampling of the data prior to the tree growing, as well as the random variable selection at each node for the split optimization. We write the causal forest estimate $\tau_{M,n}(\mathbf{x}, \mathbf{\Theta}_M)$ when it improves clarity, where $\mathbf{\Theta}_M = (\Theta_1, \ldots, \Theta_M)$. Besides, notice that the local moment Equation (3) is also used to define an efficient splitting criterion of the tree nodes.

Finally, the causal forest algorithm first performs a local centering step in practice, by regressing Y and W on $\mathbf{X}$ using regression forests, fit with $\mathscr{D}_n$. The obtained out-of-bag forest estimates of $m(\mathbf{X}_i) = \mathbb{E}[Y_i \mid \mathbf{X}_i]$ and $\pi(\mathbf{X}_i) = \mathbb{E}[W_i \mid \mathbf{X}_i]$ are denoted by $\widehat{m}_n(\mathbf{X}_i)$ and $\widehat{\pi}_n(\mathbf{X}_i)$. Then, these quantities are subtracted to get the centered outcome $\widetilde{Y}_i = Y_i - \widehat{m}_n(\mathbf{X}_i)$, and centered treatment $\widetilde{W}_i = W_i - \widehat{\pi}_n(\mathbf{X}_i)$, used to fit the causal forest $\tau_{M,n}(\mathbf{x})$.

2.3 Variable importance algorithm

We take advantage of causal forests to build an estimate of our variable importance measure $\mathrm{I}^{(j)}$, defined in Equation (2). The forest estimate $\tau_{M,n}(\mathbf{x})$, described in the previous subsection, provides a plug-in estimate for the first term $\tau(\mathbf{X}^{(\mathcal{H})})$ of $\mathrm{I}^{(j)}$. Next, we need to estimate the second term $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})|\mathbf{X}^{(-j)}]$ involved in $\mathrm{I}^{(j)}$, and then, a Monte-Carlo method will provide an efficient algorithm for our importance measure. Hence, a natural approach is to drop the j-th variable and retrain the forest to estimate $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})|\mathbf{X}^{(-j)}]$. As we deepen below and summarize in Algorithm 1, a critical feature of this procedure is that all input variables are used in the local centering of Y_i and W_i, before the j-th variable is dropped to build $\tau^{(-j)}_{M,n}(\mathbf{x})$. Therefore, the causal forest is retrained using the observations $\left\{(\mathbf{X}_i^{(-j)}, \widetilde{Y}_i, \widetilde{W}_i)\right\}_{i=1}^n$ to generate new weights $\alpha'(\mathbf{x}^{(-j)})$ and build $\tau^{(-j)}_{M,n}(\mathbf{x})$ through Equation (4).

2.3.1 Identifiability of treatment effect

When a variable $X^{(j)}$ is removed from the input variables, the moment Equation (3) does not necessarely hold anymore, since unconfoundedness Assumption (1) may be violated with a reduced set of inputs, even for $j \notin \mathcal{H}$. However, an important feature of causal forests is the preliminary step of local centering of the observed outcome and treatment assignment, explained above. The following proposition shows that the treatment effect is well identified by the local moment equation of causal forests including only variables in $\mathcal{H}$, provided that the data is centered with all inputs. We recall that $m(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$ and $\pi(\mathbf{X}) = \mathbb{E}[W \mid \mathbf{X}]$.

Proposition 4. If Assumption 1 is satisfied, we have

$$\tau(\mathbf{X}^{(\mathcal{H})}) \times \mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{H})}] - \mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{H})}] = 0,$$

which is the local moment equation defining causal forests, with input variables $\mathbf{X}^{(\mathcal{H})}$, centered outcome $Y - m(\mathbf{X})$, and centered treatment assignment $W - \pi(\mathbf{X})$.

On the other hand, removing an influential and confounding variable $j \in \mathcal{H}$ to learn a causal forest is more delicate. Indeed, a local moment equation to identify the mean CATE over $X^{(j)}$ exists if the treatment effect is uncorrelated to the squared centered treatment assignment.

Proposition 5. If Assumption 1 is satisfied, then we have for $j \in \mathcal{H}$

$$\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)}] \quad \times \mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)}] - \mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(-j)}] + \mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), \pi(\mathbf{X})(1 - \pi(\mathbf{X})) \mid \mathbf{X}^{(-j)}] = 0.$$

Then, for a query point $\mathbf{x}^{(-j)} \in [0,1]^{p-1}$, if $\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), \pi(\mathbf{X})(1 - \pi(\mathbf{X})) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}] = 0$, $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]$ is identified by the original local moment equation of causal forests, with $\mathbf{X}^{(-j)}$ as input variables, centered outcome $Y - m(\mathbf{X})$, and centered treatment assignment $W - \pi(\mathbf{X})$.

Athey and Wager [25, Footnote 5, page 42] conduct an empirical analysis using causal forests, and state in a footnote, that local centering "eliminates confounding effects. Thus, we do not need to give the causal forest all features $X^{(j)}$ that may be confounders. Rather, we can focus on features that we believe may be treatment modifiers". However, Propositions 4 and 5 show that this statement must be completed. Indeed, Proposition 4 states that confounders not involved in the heterogeneity of the treatment effect, i.e. confounders that do no belong to $\mathcal{H}$, may be dropped without hurting the identifiability of τ, thanks the local centering step. On the other hand, Proposition 5 shows that this is clearly not the case for confounders involved in heterogeneity, as the treatment effect is not properly identified by the local moment equation of causal forests, even with local centering. To overcome this problem, we introduce a corrective term in the retrained forest.

2.3.2 Corrected causal forests

The additional covariance term in Proposition 5 can be estimated using the original causal forest fit with all inputs. Therefore, we propose the corrected causal forest estimate when removing a confounding variable $X^{(j)}$ with $j \in \mathcal{H}$. Recall that the weights $\alpha'(\mathbf{x}^{(-j)})$ are generated by the causal forest using centered data and dropping variable $X^{(j)}$, to define $\tau^{(-j)}_{M,n}(\mathbf{x})$. We define the corrected causal forest estimate $\theta^{(-j)}_{M,n}(\mathbf{x})$ as

$$\theta^{(-j)}_{M,n}(\mathbf{x}) = \tau^{(-j)}_{M,n}(\mathbf{x}) - \frac{\sum_{i=1}^n \alpha'_i(\mathbf{x}^{(-j)})\widetilde{W}_i^2 \tau_{M,n}(\mathbf{X}_i) - \overline{W^2_{\alpha'}}\,\overline{\tau}_{\alpha'}}{\overline{W^2_{\alpha'}} - \left(\overline{W}_{\alpha'}\right)^2}, \tag{5}$$

where $\overline{W^2_{\alpha'}} = \sum_{i=1}^n \alpha'_i(\mathbf{x}^{(-j)})\widetilde{W}_i^2$, $\overline{W}_{\alpha'} = \sum_{i=1}^n \alpha'_i(\mathbf{x}^{(-j)})\widetilde{W}_i$, and the mean treatment effect is $\overline{\tau}_{\alpha'} = \sum_{i=1}^n \alpha'_i(\mathbf{x}^{(-j)})\tau_{M,n}(\mathbf{X}_i)$. More precisely, the corrective term of Equation (5) is the forest estimate of the third term of the equation of Proposition 5 divided by $\mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)}]$. Consequently, the corrected causal forest $\theta^{(-j)}_{M,n}(\mathbf{x})$ retrained without a confounding variable is an estimate of $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]$, which is the targeted quantity, and is consistent, as we will show in Section 3. However, note that the correction term can be small in practice, as demonstrated in the experimental Section 4.

2.3.3 Variable importance estimate

Using $\mathscr{D}'_n = \left\{\left(\mathbf{X}'_i, Y'_i, W'_i\right)\right\}_{i=1}^n$ an independent copy of $\mathscr{D}_n$, we define

$$\mathrm{I}^{(j)}_n = \frac{\sum_{i=1}^n \left[\tau_{M,n}(\mathbf{X}'_i) - \theta^{(-j)}_{M,n}(\mathbf{X}'_i)\right]^2}{\sum_{i=1}^n \left[\tau_{M,n}(\mathbf{X}'_i) - \overline{\tau_{M,n}}\right]^2} - \mathrm{I}^{(0)}_n, \tag{6}$$

where $\overline{\tau_{M,n}} = \sum_{i=1}^n \tau_{M,n}(\mathbf{X}'_i)/n$, and $\mathrm{I}^{(0)}_n$ is the mean squared difference between the initial forest predictions and the predictions of the corrected forest $\theta^{(0)}_{M,n}\left(\mathbf{X}'_i, \Theta'_M\right)$, retrained with still all the inputs variables involved but a new randomization Θ'_M, i.e.,

$$\mathrm{I}^{(0)}_n = \frac{\sum_{i=1}^n \left[\tau_{M,n}(\mathbf{X}'_i, \Theta_M) - \theta^{(0)}_{M,n}\left(\mathbf{X}'_i, \Theta'_M\right)\right]^2}{\sum_{i=1}^n \left[\tau_{M,n}(\mathbf{X}'_i) - \overline{\tau_{M,n}}\right]^2}. \tag{7}$$

In fact, $\mathrm{I}^{(0)}_n$ partially removes the bias of the first term of $\mathrm{I}^{(j)}_n$, due to the randomization of the forest training, and vanishes as the sample size increases if the causal forest converges. Notice that the above definition is formalized

with $\mathscr{D}_n^{'}$ for the sake of clarity, but that such additional data is usually not available in practice. Instead, out-of-bag causal forest estimates are rather used to define $\mathrm{I}_n^{(j)}$, as summarized in Algorithm 1 below.

Algorithm 1 Variable importance algorithm for causal forests

Require: A dataset $\mathscr{D}_n = \{(\mathbf{X}_i, Y_i, W_i)\}_{i=1}^n$ containing all confounding variables.

1: Perform local centering of outputs Y_i and treatment assignments W_i to get the centered dataset $\{(\mathbf{X}_i, \tilde{Y}_i, \tilde{W}_i)\}_{i=1}^n$, using regression forests and out-of-bag estimates.
2: Train a causal forest with the centered data $\{(\mathbf{X}, \tilde{Y}_i, \tilde{W}_i)\}_{i=1}^n$ containing all variables.
3: Retrain a corrected causal forest with the same data as the previous step.
4: Compute $\mathrm{I}_n^{(0)}$ according to Equation (7) and using the forests trained at Steps 2 and 3.
5: **for** $j \in \{1, \ldots, p\}$ **do**
6: Train a corrected causal forest with the centered data $\{(\mathbf{X}^{(-j)}, \tilde{Y}_i, \tilde{W}_i)\}_{i=1}^n$, where the j-th variable is removed.
7: Compute $\mathrm{I}_n^{(j)}$ according to Equation (6) using the initial forest of Step 2 and the retrained forest of the previous Step 6, and with $\mathrm{I}_n^{(0)}$ computed at Step 4.
8: **end for**
9: **return** $\left\{\mathrm{I}_n^{(j)}\right\}_{j=1}^p$

3 Theoretical properties

Propositions 4 and 5 are the cornerstones of the consistency of our variable importance algorithm. This result relies on the asymptotic analysis of Athey et al. [6], which states the consistency of causal forests in Theorem 1. Several mild assumptions are required, mainly about the input distribution, the regularity of the involved functions, and the forest growing. Then, the core of our mathematical analysis is the extension to the case of a causal forest fit without a given input variable. When the removed input is a confounding variable, consistency is obtained thanks to the corrective term introduced in Equation (5) of the previous section. Then, the convergence of our variable importance algorithm follows using a standard asymptotic analysis. We first formalize the required assumptions and specifications on the tree growing from Athey et al. [6], that are frequently used in the theoretical analysis of random forests [5, 26, 27].

Assumption 3. The input $\mathbf{X}$ takes value in $[0,1]^p$, and admits a density bounded from above and below by strictly positive constants.

Assumption 4. The functions π, m, and τ are Lipschitz, $0 < \pi(\mathbf{x}) < 1$ for $\mathbf{x} \in [0,1]^p$, and μ and τ are bounded.

Specification 1. Tree splits are constrained to put at least a fraction $\gamma > 0$ of the parent node observations in each child node. The probability to split on each input variable at every tree node is greater than $\delta > 0$. The forest is honest, and built via subsampling with subsample size a_n, satisfying $a_n/n \to 0$ and $a_n \to \infty$.

The first part of Specification 1 is originally introduced by Meinshausen [26]. The idea is to enforce the diameter of each cell of the trees to vanish as the sample size increases, by adding a constraint on the minimum size of children nodes, and slightly increasing the randomization of the variable selection for the split at each node. Then, vanishing cell diameters combined to Lipschitz functions lead to the forest convergence. Additionally, honesty is a key property of the tree growing, extensively discussed in Wager and Athey [5], where half of the data is used to optimize the splits, and the other half to estimate the cell outputs. With these assumptions satisfied, we state below the causal forest consistency proved in Athey et al. [6]. Notice that the original proof is conducted for generalized forests, for any local moment equation satisfying regularity assumptions, automatically fulfilled for the moment Equation (3)

involved in our analysis. In Appendix A, we give a specific proof of Theorem 1 in the case of causal forests. We built on this proof to further extend the consistency result when a confounding variable is removed.

Theorem 1: (Theorem 3 from Athey et al. [6]). If Assumptions 1–4 and Specification 1 are satisfied, and the causal forest $\tau_{M,n}(\mathbf{x})$ is built with $\mathscr{D}_n$ without local centering, then we have for $\mathbf{x} \in [0,1]^p$,

$$\tau_{M,n}(\mathbf{x}) \xrightarrow{p} \tau(\mathbf{x}^{(\mathcal{H})}).$$

Next, we need a slight simplification of our variable importance algorithm to alleviate the mathematical analysis. We assume that a centered dataset $\mathscr{D}_n^\star = \{(\mathbf{X}_i, W_i^\star, Y_i^\star)\}$ is directly available, where $W_i^\star = W_i - \pi(\mathbf{X}_i)$ and $Y_i^\star = Y_i - m(\mathbf{X}_i)$. A causal forest grown with this dataset where a given input variable $j \in \{1, \ldots, p\} \setminus \mathcal{H}$ is dropped, consistently estimates the treatment effect as stated below. Consistency also holds for variables $j \in \mathcal{H}$ in specific cases, whereas in the general case, the corrected term introduced in Equation (5) is required. Theorem 2 states the consistency of causal forests when an input variable is removed.

Theorem 2. If Assumptions 1–4 and Specification 1 are satisfied, and the causal forest $\tau_{M,n}^{(-j)}(\mathbf{x})$ is fit with the centered data $\mathscr{D}_n^{\star(-j)}$ without the j-th variable,

(i) for $j \in \{1, \ldots, p\} \setminus \mathcal{H}$ and $\mathbf{x} \in [0,1]^p$, we have

$$\tau_{M,n}^{(-j)}(\mathbf{x}) \xrightarrow{p} \tau(\mathbf{x}^{(\mathcal{H})}),$$

(ii) for $j \in \mathcal{H}$ and $\mathbf{x} \in [0,1]^p$, if $\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), \pi(\mathbf{X})(1 - \pi(\mathbf{X})) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}] = 0$, we have

$$\tau_{M,n}^{(-j)}(\mathbf{x}) \xrightarrow{p} \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}].$$

Theorem 2 is a direct consequence of Propositions 4 and 5 combined with Theorem 1. Indeed, provided that the outcome and treatment assignment are centered, if the removed variable j is not involved in the treatment heterogeneity, i.e. $j \notin \mathcal{H}$, consistency holds. On the other hand, if $j \in \mathcal{H}$, we need an additional assumption that $\tau(\mathbf{X}^{(\mathcal{H})})$ and $\pi(\mathbf{X})(1 - \pi(\mathbf{X}))$ are not correlated conditional on $\mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}$, where $\mathbf{x}^{(-j)}$ is the new query point. Otherwise, consistency is obtained with the corrective term defined in Equation (5), as we will see. However, we need an additional small modification of causal forests to enforce the generated estimates to be bounded, and to limit the number of observations in each terminal leave of trees, as stated in the specification below. Notice that such modifications are quite mild. Indeed, the true treatment effect is bounded by assumption. For the second part, the number of observations in each terminal leave may not be bounded in specific cases, because of honest tree growing. Nevertheless, it is still possible to comply with this specification, by randomly splitting cells that exceed the number of observation threshold.

Specification 2. The causal forest estimates are truncated from below and above by $-K$ and K, where $K \in \mathbb{R}$ is an arbitrarily large constant. The number of observations in each terminal leave of trees is smaller than a threshold $t_0 \in \mathbb{N}^\star$.

Theorem 3. Let the initial causal forest $\tau_{M,n}(\mathbf{x})$ fit with the centered data $\mathscr{D}_n^\star$, and the corrected causal forest $\theta_{M,n}^{(-j)}(\mathbf{x})$ fit using $\tau_{M,n}(\mathbf{x})$ and $\mathscr{D}_n^{\star(-j)}$, an independent copy of the centered data with the j-th variable dropped. If Assumptions 1–4, and Specifications 1 and 2 are satisfied, then for $j \in \{1, \ldots, p\}$ and $\mathbf{x} \in [0,1]^p$, we have

$$\theta_{M,n}^{(-j)}(\mathbf{x}) \xrightarrow{p} \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}].$$

Since Theorems 1 and 3 give the consistency of causal forests respectively fit with all input variables, and when a given variable is removed, we can deduce the consistency of our variable importance algorithm from standard asymptotic arguments.

Theorem 4. Under the same assumptions than Theorem 3, we have for all $j \in \{1, \ldots, p\}$

$$\mathrm{I}_n^{(j)} \xrightarrow{p} \mathrm{I}^{(j)}.$$

Theorem 4 states that the introduced variable importance algorithm gets arbitrarily close to the true theoretical value, provided that the sample size is large enough. Combining this result with Proposition 3, we get that, for $j \notin \mathcal{H}$, $\mathrm{I}_n^{(j)} \to^p 0$, which means that the variables not involved in the treatment heterogeneity by construction get a null importance. Finally, we conclude our theoretical analysis with a focus on the corrective term of the retrained causal forests. In particular, we quantify the positive asymptotic bias introduced in the importance measure without this correction. We thus denote by $\mathcal{I}_n^{(j)}$ the estimated importance measure following the same procedure as for $\mathrm{I}_n^{(j)}$, except that the corrected forest $\theta_{M,n}^{(-j)}(\mathbf{x})$ is replaced by the raw retrained forest $\tau_{M,n}^{(-j)}(\mathbf{x})$.

Theorem 5. Under the same assumptions than Theorem 3, with $\mathcal{I}_n^{(j)}$ the importance measure estimated without the corrective term in the causal forests, we have for all $j \in \mathcal{H}$,

$$\mathcal{I}_n^{(j)} \xrightarrow{p} \mathrm{I}^{(j)} + \frac{1}{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})]}\mathbb{E}\left[\frac{\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), \pi(\mathbf{X})(1-\pi(\mathbf{X})) \mid \mathbf{X}^{(-j)}]^2}{\mathbb{E}[\pi(\mathbf{X})(1-\pi(\mathbf{X})) \mid \mathbf{X}^{(-j)}]^2}\right].$$

4 Experiments

We assess the performance of the introduced algorithm through three batches of experiments. First, we use simulated data, where the theoretical importance values are known by construction, to compare our algorithm to the existing competitors. Secondly, we test our procedure with the semi-synthetic cases of the ACIC data challenge 2019, where the variables involved in the heterogeneity are known, but not the importance value. Finally, we present cases with real data to show examples of an analysis conducted with our procedure. Our approach is compared to the importance of the `grf` package and TE-VIM, the double robust approach of Hines et al. [15]. For TE-VIM, any learning method can be used, and we report the performance of GAM models, which outperform regression forests in the presented experiments. Otherwise, we use the default settings of TE-VIM. When reading the results, recall that TE-VIM targets the same theoretical quantities $\mathrm{I}^{(j)}$ as our algorithm, whereas the `grf` importance `grf-vimp` is the frequency of variable occurrence in tree splits. The `grf-vimp` algorithm is fast to compute, since it only requires to count the variables involved in node splits, but does not provide a precise quantification of the impact of each variable on the output variability. Besides, the algorithm of Boileau et al. [16] is designed for high dimensional cases and linear treatment effects, and is thus not appropriate to our goal of precisely quantifying variable importance in non-linear settings. The implementation of our variable importance algorithm is available online at https://gitlab.com/random-forests/vimp-causal-forests, along with the code to reproduce experiments with simulated data.

4.1 Simulated Data

4.1.1 Experiment 1

We consider a first example of simulated data to highlight the good performance of the proposed importance measure. The input is a Gaussian vector of dimension $p = 8$, defined by $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where Σ is the identity matrix. The treatment assignment W is simply a Bernoulli random variable of parameter 1/2, and the response Y follows

$$Y = \left(X^{(1)}1_{X^{(1)}>0} + 0.6X^{(2)}1_{X^{(2)}>0}\right) \times W + 0.25\,(X^{(3)} \times X^{(4)})^2 + \varepsilon, \tag{8}$$

where $\varepsilon \sim \mathcal{N}(0, 0.1)$. This first experiment gives a baseline in a quite simple case, since input variables are independent, there are no confounding effects, and the ratio $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})]/\mathbb{V}[Y]$ is large, with a value of about 50 %. The targeted theoretical values can be easily computed from the defined distributions and Equation (8), and are reported in the left column of Table 1. Next, we draw a sample of size $n = 3000$, and the causal forest is fit with all default settings, including the number of trees $M = 2000$. Table 1 shows the estimated mean importance values over 10 repetitions for our proposed algorithm $\mathrm{I}_n^{(j)}$, TE-VIM, and the metric from the `grf` package, along with the standard deviation of the mean importance in brackets. Notice that for each repetition, a new data sample is drawn from the data generating process described above, and then is used to run each variable importance algorithm. As expected, both $\mathrm{I}_n^{(j)}$ and TE-VIM provide accurate estimates of the theoretical importance values $\mathrm{I}^{(j)}$ in this quite simple setting. The fast metric provided by grf-vimp also provides a good approximation of the relative variable importance, even if irrelevant variables get higher values than with $\mathrm{I}_n^{(j)}$ and TE-VIM. Next, Figure 1 displays the importance $\mathrm{I}_n^{(j)}$ with respect to the sample size n. The estimated values get closer to the theoretical quantities, as expected from the convergence results of Section 3. Notice that variables $X^{(3)}, \ldots, X^{(6)}$ all have a small importance, and therefore overlap on Figure 1. We omit $X^{(7)}$ and $X^{(8)}$ on the figure, since they are symmetric with $X^{(6)}$.Finally, we also take advantage of this first experiment to analyze a case of higher dimension. We thus consider the same settings, but we add 32 independent Gaussian variables of unit variance to get a final input dimension of $p = 40$. Table 2 shows that the impact of this large number of noisy variables is small for $\mathrm{I}_n^{(j)}$ and TE-VIM, but strong for grf-vimp, with a high decrease of the importance value of $X^{(1)}$. Such phenomenon is expected, since the forest quite often splits on noisy variables because of the split randomization at each tree node, leading

Table 1: Variable importance of Experiment 1 for $\mathrm{I}_n^{(j)}$, the importance measure of `grf` package, and TE-VIM. Standard deviations are displayed in brackets when greater than 0.002.

I		**I_n**		**TE-VIM**		**grf-vimp**	
$X^{(1)}$	0.74	$X^{(1)}$	0.77 (0.01)	$X^{(1)}$	0.74 (0.02)	$X^{(1)}$	0.70 (0.003)
$X^{(2)}$	0.26	$X^{(2)}$	0.25 (0.01)	$X^{(2)}$	0.28 (0.01)	$X^{(2)}$	0.19 (0.003)
$X^{(3)}$	0	$X^{(3)}$	0.001	$X^{(3)}$	0.007	$X^{(3)}$	0.03
$X^{(4)}$	0	$X^{(4)}$	0.001	$X^{(4)}$	−0.005	$X^{(4)}$	0.03
$X^{(5)}$	0	$X^{(5)}$	0.0005	$X^{(5)}$	0.003	$X^{(5)}$	0.01
$X^{(6)}$	0	$X^{(6)}$	0.0003	$X^{(6)}$	0.003	$X^{(6)}$	0.01
$X^{(7)}$	0	$X^{(7)}$	0.0003	$X^{(7)}$	0.003	$X^{(7)}$	0.01
$X^{(8)}$	0	$X^{(8)}$	0.0003	$X^{(8)}$	0.003	$X^{(8)}$	0.01

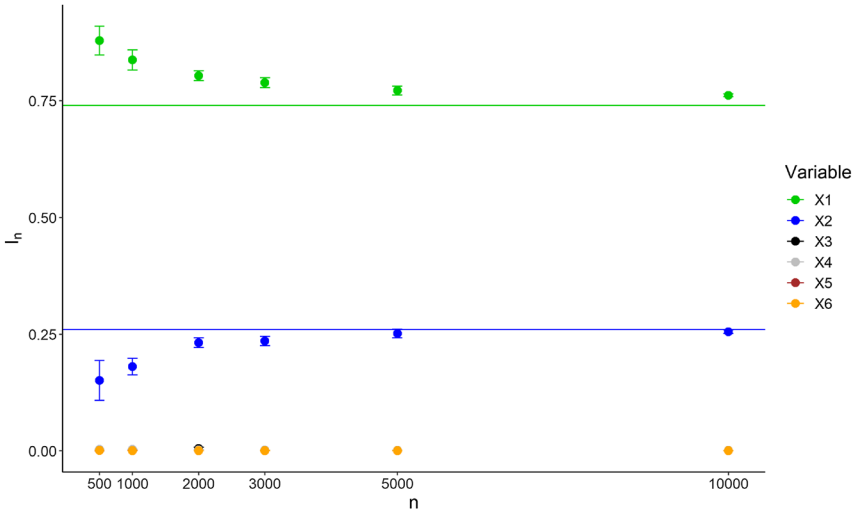

Figure 1: Importance values $\mathrm{I}_n^{(j)}$ with respect to the sample size n for Experiment 1. Non-null theoretical importance are displayed as solid lines.

Table 2: Variable importance of Experiment 1 with the dimension p set to 40 by adding noisy variables, for $\mathrm{I}_n^{(j)}$, the importance measure of `grf` package, and TE-VIM. Standard deviations are displayed in brackets when greater than 0.002.

I		I_n		TE-VIM		grf-vimp	
$X^{(1)}$	0.74	$X^{(1)}$	0.79 (0.01)	$X^{(1)}$	0.80 (0.01)	$X^{(1)}$	0.56 (0.004)
$X^{(2)}$	0.26	$X^{(2)}$	0.22 (0.01)	$X^{(2)}$	0.24 (0.01)	$X^{(2)}$	0.23 (0.003)
$X^{(3)}$	0	$X^{(3)}$	0.001	$X^{(3)}$	−0.04 (0.01)	$X^{(3)}$	0.03 (0.002)
$X^{(4)}$	0	$X^{(4)}$	0.001	$X^{(4)}$	−0.02 (0.01)	$X^{(4)}$	0.03 (0.003)
$X^{(5)}$	0	$X^{(5)}$	0.00004	$X^{(5)}$	−0.02 (0.004)	$X^{(5)}$	0.004
$X^{(6)}$	0	$X^{(6)}$	0.00003	$X^{(6)}$	−0.01 (0.003)	$X^{(6)}$	0.005
$X^{(7)}$	0	$X^{(7)}$	0.00003	$X^{(7)}$	−0.01 (0.005)	$X^{(7)}$	0.004
$X^{(8)}$	0	$X^{(8)}$	0.00002	$X^{(8)}$	−0.01 (0.004)	$X^{(8)}$	0.005

to a lower split frequency of the relevant inputs. Besides, we can also notice that TE-VIM assigns a negative bias to irrelevant variables in this noisy setting, whereas $\mathrm{I}_n^{(j)}$ still provides values very close to 0.

4.1.2 Experiment 2

We consider the same settings as in Experiment 1 with the original dimension $p = 8$, but we introduce three modifications to make the problem more difficult. In particular, we set $W \sim$Bernouilli$(0.4 + 0.21_{X^{(1)}>0})$ to introduce a confounding factor, the covariance Σ is now the identity matrix except that $\mathrm{Cov}(X^{(1)}, X^{(5)}) = 0.9$ to get a strong correlation between two variables, and we finally increase the weight of $(X^{(3)} \times X^{(4)})^2$ to 1 to reduce the ratio $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})]/\mathbb{V}[Y]$ to about 5 %. Such a quite small ratio is realistic, and makes the treatment effect quite difficult to estimate in practice. The response Y is now defined by

$$Y = \left(X^{(1)}1_{X^{(1)}>0} + 0.6X^{(2)}1_{X^{(2)}>0}\right) \times W + (X^{(3)} \times X^{(4)})^2 + \varepsilon, \tag{9}$$

where $\varepsilon \sim \mathcal{N}(0, 0.1)$. We then take a sample size $n = 3000$, and the causal forest is again fit with default settings and a number of trees $M = 2000$. Here, both $X^{(1)}$ and $X^{(2)}$ are involved in heterogeneity, i.e. $\mathcal{H} = \{1, 2\}$, but only $X^{(1)}$ is also a confounder. Results are averaged over 50 repetitions, and are reported in Table 3. Additionally, the standard deviation of the mean importance for each variable is displayed in brackets, except for negligible values (<0.001). The first column of Table 3 is the oracle importance value, precisely estimated using Equation (2), the closed-form of τ given by Equation (9), and a Monte-Carlo method with a large sample drawn from the joint distribution of $(Y, W, \mathbf{X})$, known by construction.

The results displayed in Table 3 show that both our algorithm and TE-VIM provide the accurate variable ranking, where $X^{(2)}$ is the most important variable, and $X^{(1)}$ the second most important one. However, the standard

Table 3: Variable importance ranking of Experiment 2 for $\mathrm{I}_n^{(j)}$, the importance measure of `grf` package, and TE-VIM. Standard deviations are displayed in brackets when greater than 0.001.

I		I_n		TE-VIM		grf-vimp	
$X^{(2)}$	0.26	$X^{(2)}$	0.21 (0.008)	$X^{(2)}$	0.25 (0.06)	$X^{(1)}$	0.49 (0.01)
$X^{(1)}$	0.18	$X^{(1)}$	0.19 (0.004)	$X^{(1)}$	0.13 (0.07)	$X^{(4)}$	0.12 (0.006)
$X^{(3)}$	0	$X^{(4)}$	0.03 (0.005)	$X^{(6)}$	−0.22 (0.14)	$X^{(5)}$	0.12 (0.007)
$X^{(4)}$	0	$X^{(3)}$	0.03 (0.003)	$X^{(8)}$	−0.23 (0.15)	$X^{(2)}$	0.11 (0.005)
$X^{(5)}$	0	$X^{(5)}$	0.005	$X^{(5)}$	−0.24 (0.15)	$X^{(3)}$	0.11 (0.005)
$X^{(6)}$	0	$X^{(6)}$	0.001	$X^{(7)}$	−0.28 (0.16)	$X^{(7)}$	0.02 (0.001)
$X^{(7)}$	0	$X^{(7)}$	0.001	$X^{(3)}$	−0.32 (0.28)	$X^{(8)}$	0.02 (0.001)
$X^{(8)}$	0	$X^{(8)}$	0.001	$X^{(4)}$	−0.55 (0.32)	$X^{(6)}$	0.02

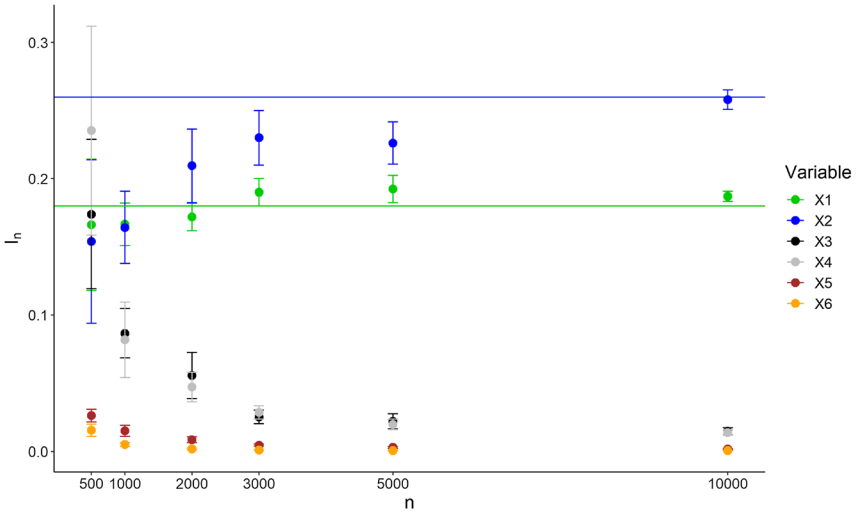

Figure 2: Importance values $I_n^{(j)}$ with respect to the sample size n for Experiment 2. Non-null theoretical importance are displayed as solid lines.

deviations of the mean importance values over 50 repetitions are higher for TE-VIM, which is induced by a high instability across repetitions. This can be a limitation in practice with real data, as importance values are computed only once. We also observe that variables with a theoretical null importance get a quite strong negative bias. On the other hand, the importance measure from the `grf` package underestimates the importance of variable $X^{(2)}$, and identifies $X^{(3)}$, $X^{(4)}$, and $X^{(5)}$ as slightly more important than $X^{(2)}$, although these three variables are not involved in the treatment heterogeneity by construction. In particular, $X^{(5)}$ is not involved at all in the response Y, but is strongly correlated to the influential input $X^{(1)}$. Because of this dependence, $X^{(5)}$ is frequently used in the causal forests splits, leading to this quite high importance given by the `grf` package. On the other hand, $I_n^{(j)}$ gives an importance close to 0 for $X^{(5)}$. This result is expected, since the removal of $X^{(5)}$ does not lead to any loss of information regarding the treatment heterogeneity, by definition. An additional interesting phenomenon is the non-negligible importance for variables $X^{(3)}$ and $X^{(4)}$ given by all procedures. In fact, the interaction term in the baseline function μ, which takes the form of a squared product, is rather difficult to estimate by regression forests. Then, the local centering of Y is only partial, and $X^{(3)}$ and $X^{(4)}$ still have impact on the variance of treatment estimates. Besides, notice that the corrective term of Equation (5) is negligible in this experiment, and that using the original causal forest retrained with one variable removed, gives the same result as in Table 3 for $I_n^{(j)}$, up to the displayed digits. Finally, Figure 2 also displays the importance values $I_n^{(j)}$ with respect to the sample size n. Results are consistent with the convergence results of the previous section. In particular, for a large sample of n = 10000, the relative errors of $I_n^{(j)}$ happen to be really small, whereas irrelevant variables get high importance values for n = 500.

4.1.3 Experiment 3

This third experiment has the same setting than Experiment 2, except that variable $X^{(1)}$ is only a confounder and is not involved in the treatment effect heterogeneity anymore. Now, the response writes

$$Y = \left(0.6X^{(2)}1_{X^{(2)}>0}\right) \times W + X^{(1)}1_{X^{(1)}>0} + (X^{(3)} \times X^{(4)})^2 + \varepsilon.$$

The results are provided in Table 4. Clearly, $I_n^{(j)}$ outperforms the competitors. Indeed, $X^{(2)}$ is well-identified by $I_n^{(j)}$ as responsible for most of the heterogeneity of the treatment effect, whereas TE-VIM is strongly biased, with some values exceeding 1 because of the variance of estimates involved in the ratio of TE-VIM. The importance procedure of the `grf` package outputs quite close values for $X^{(2)}$, $X^{(4)}$, and $X^{(3)}$. As expected, the importance of these last two variables is relatively larger than in Experiment 1, since the ratio $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})]/\mathbb{V}[Y]$ drops to 1 % in this

Table 4: Variable importance ranking of Experiment 3 for $\mathrm{I}_n^{(j)}$, the importance measure of `grf` package, and TE-VIM. Standard deviations are displayed in brackets when greater than 0.001.

I		I_n		TE-VIM		grf-vimp	
$X^{(2)}$	1	$X^{(2)}$	0.84 (0.02)	$X^{(2)}$	1.12 (0.6)	$X^{(2)}$	0.36 (0.01)
$X^{(1)}$	0	$X^{(4)}$	0.16 (0.02)	$X^{(3)}$	0.63 (0.4)	$X^{(4)}$	0.24 (0.008)
$X^{(3)}$	0	$X^{(3)}$	0.15 (0.02)	$X^{(4)}$	0.37 (0.2)	$X^{(3)}$	0.23 (0.008)
$X^{(4)}$	0	$X^{(7)}$	0.007 (0.001)	$X^{(1)}$	0.29 (0.2)	$X^{(6)}$	0.04 (0.003)
$X^{(5)}$	0	$X^{(6)}$	0.005	$X^{(8)}$	0.29 (0.2)	$X^{(8)}$	0.04 (0.004)
$X^{(6)}$	0	$X^{(8)}$	0.005	$X^{(7)}$	0.29 (0.2)	$X^{(7)}$	0.04 (0.002)
$X^{(7)}$	0	$X^{(5)}$	0.003	$X^{(5)}$	0.25 (0.2)	$X^{(1)}$	0.03 (0.002)
$X^{(8)}$	0	$X^{(1)}$	0.002	$X^{(6)}$	0.17 (0.1)	$X^{(5)}$	0.03 (0.002)

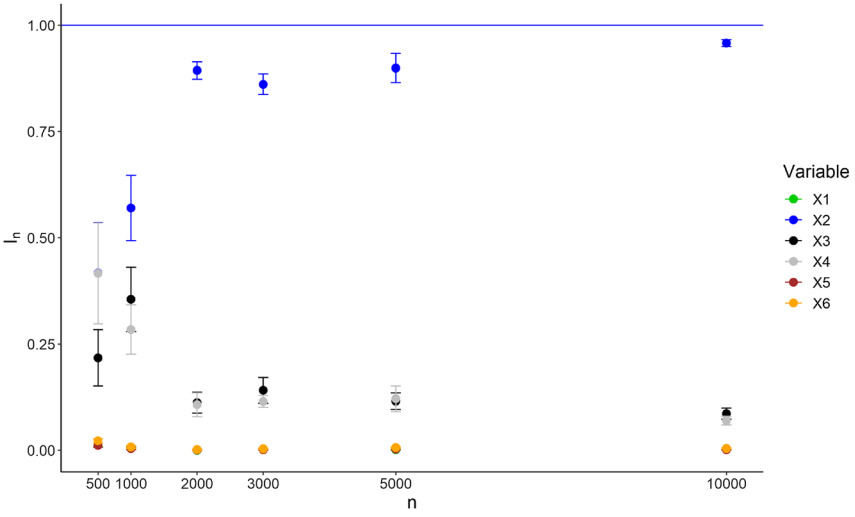

Figure 3: Importance values $\mathrm{I}_n^{(j)}$ with respect to the sample size n for Experiment 3. Non-null theoretical importance are displayed as solid lines.

case. As in the previous experiments, Figure 3 displays the importance values $\mathrm{I}_n^{(j)}$ with respect to the sample size n, and shows that the errors decrease as n increases, following the theory.

4.1.4 Experiment 4

The goal of this fourth simulated experiment is to highlight a case where the corrective term in the retrained causal forest has a strong influence, as opposed to Experiments 1, 2, and 3. We consider $p = 5$ inputs uniformly distributed over [0,1], except $X^{(1)}$ defined as $X^{(1)} = U^3$, where $U \sim \mathcal{U}(0,1)$. The treatment assignment W is a Bernoulli variable defined from $\pi(\mathbf{X}) = X^{(1)}$, and the response is given by

$$Y = 10X^{(1)}(1 - X^{(1)}) \times W + X^{(2)} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 0.1)$. We still use $n = 3000$ and $M = 2000$ trees in the causal forests. Next, we compute our importance measure $\mathrm{I}_n^{(j)}$ for all inputs, as well as its counterpart $\mathcal{I}_n^{(j)}$, where the corrective term is removed, and with 10 repetitions for uncertainties. Results are reported in Table 5, and clearly show the high bias of the importance of $X^{(1)}$ when the corrective term in the retrained forest is removed. Indeed, we get $\mathcal{I}_n^{(1)} = 1.57$, whereas the target quantity is $\mathrm{I}^{(1)} = 1$, since $X^{(1)}$ is the only variable involved in the treatment effect heterogeneity and $X^{(1)}$ is independent of the other inputs. With the correction, we recover an importance value of 0.98 for $X^{(1)}$ as expected. Notice that the asymptotic bias exhibited in Theorem 5 takes values 0.72 for this case, which explains the empirical results. Importantly, this bias takes small values in practice in most cases. Here, we take the treatment effect as $\tau(\mathbf{X}^{(\mathcal{H})}) = 10\pi(\mathbf{X})(1 - \pi(\mathbf{X}))$ to maximize the covariance term involved in the bias of Theorem 5.

Table 5: Variable importance ranking of Experiment 4 for $\mathrm{I}_n^{(j)}$ (with corrected causal forests) and $\mathcal{I}_n^{(j)}$ (without correction). Standard deviations are displayed in brackets when greater than 0.001.

$\mathbf{I}_n$		$\mathcal{I}_n$	
$X^{(1)}$	0.98 (0.002)	$X^{(1)}$	1.57 (0.01)
$X^{(2)}$	0.0003	$X^{(2)}$	0.001
$X^{(3)}$	0.001	$X^{(3)}$	0.001
$X^{(4)}$	0.0002	$X^{(4)}$	0.002
$X^{(5)}$	0.0002	$X^{(5)}$	0.001

4.2 ACIC Data Challenge 2019

We run a second batch of experiments using the data from the ACIC data challenge 2019 (https://sites.google.com/view/acic2019datachallenge/data-challenge), where the goal was to estimate ATEs in various settings. The input data is taken from real datasets available online on the UCI repository. Next, outcomes are simulated with different scenarios, and the associated code scripts were released after the challenge. Since the data generating mechanism is available, we have access to the variables involved in the heterogeneous treatment effect. In each scenario, a hundred datasets were randomly sampled.

We first use the "student performance 2" data with 29 input variables, considering Scenario 4 defined in the ACIC challenge, involving heterogeneity of the treatment effect with respect to $X^{(3)}$. Each dataset is of size $n = 649$, and we run 50 repetitions with independent datasets for uncertainties. Table 6 gives the top 5 variables ranked by $\mathrm{I}_n^{(j)}$, which accurately identifies $X^{(3)}$ as the only variable involved in the treatment heterogeneity, since other variables all have a negligible importance value. Both TE-VIM and the `grf` importance measure also identify $X^{(3)}$ as the most important variable. However, the bias of TE-VIM is strong, and the `grf` importance of many irrelevant variables is not negligible, as opposed to $\mathrm{I}_n^{(j)}$.

Secondly, we use the "spam email" data, made of 22 input variables. We also consider Scenario 4, where variables $X^{(8)}$ and $X^{(19)}$ are involved in the heterogeneous treatment effect. In this case, we merge 20 datasets to get a quite large sample of size $n = 10000$, and run 5 repetitions to compute standard deviations. The two relevant variables are properly identified as the most important ones by $\mathrm{I}_n^{(j)}$ and grf-vimp, as shown in Table 7. On the other

Table 6: Top 5 variables for "Student performance 2 (Scenario 4)" dataset using $\mathrm{I}_n^{(j)}$, TE-VIM, and the importance measure of `grf` package. Standard deviations are displayed in brackets.

$\mathrm{I}_n^{(j)}$		**TE-VIM**		**grf-vimp**	
$X^{(3)}$	0.85 (0.02)	$X^{(3)}$	0.82 (1.1)	$X^{(3)}$	0.44 (0.01)
$X^{(29)}$	0.013 (0.006)	$X^{(26)}$	0.22 (0.6)	$X^{(29)}$	0.06 (0.004)
$X^{(28)}$	0.007 (0.002)	$X^{(20)}$	0.06 (0.5)	$X^{(28)}$	0.04 (0.002)
$X^{(27)}$	0.006 (0.003)	$X^{(25)}$	−0.09 (0.8)	$X^{(25)}$	0.04 (0.003)
$X^{(25)}$	0.005 (0.002)	$X^{(19)}$	−0.25 (0.3)	$X^{(27)}$	0.03 (0.003)

Table 7: Top 6 variables for "Spam email (Scenario 4)" dataset using $\mathrm{I}_n^{(j)}$, TE-VIM, and the importance measure of `grf` package. Standard deviations are displayed in brackets.

$\mathbf{I}_n$		**TE-VIM**		**grf-vimp**	
$X^{(8)}$	0.83 (0.001)	$X^{(8)}$	0.88 (0.01)	$X^{(8)}$	0.85 (4.10^{-3})
$X^{(19)}$	0.011 (0.002)	$X^{(1)}$	0.04 (0.007)	$X^{(19)}$	0.064 (6.10^{-3})
$X^{(22)}$	0.003 (4.10^{-4})	$X^{(19)}$	0.04 (0.007)	$X^{(1)}$	0.013 (3.10^{-3})
$X^{(12)}$	0.002 (4.10^{-4})	$X^{(11)}$	0.006 (0.007)	$X^{(22)}$	0.013 (1.10^{-3})
$X^{(15)}$	0.001 (3.10^{-4})	$X^{(14)}$	0.006 (0.006)	$X^{(15)}$	0.010 (8.10^{-4})
$X^{(17)}$	0.0004 ($<10^{-4}$)	$X^{(3)}$	0.005 (0.006)	$X^{(17)}$	0.009 (2.10^{-3})

hand, TE-VIM ranks the noisy variable $X^{(1)}$ as slightly more important than the relevant input $X^{(19)}$. Again, the `grf` importance gives rather higher values to irrelevant variables than $\mathrm{I}_n^{(j)}$. Notice that the impact of $X^{(19)}$ on heterogeneity is really small, and if we use only few datasets of size $n = 500$ in the forest training, $X^{(19)}$ is not identified as more important than noisy variables by any method. Thus, a large sample size is required to detect its influence, and therefore we use $n = 10000$.

4.3 Real data

4.3.1 Welfare data

For a first experiment with real data, we use the "Welfare" dataset from a GSS survey, introduced in Green and Kern [28] and available at https://github.com/gsbDBI/ExperimentData. The goal of this survey is to analyze the impact of question wording about the support of Americans to the government welfare spending. Respondents are randomly assigned one of two possible questions, with the same introduction and response options, but using the phrasing "welfare" or "assistance to the poor". In fact, this slight wording difference has a quite strong impact on the survey answers, and defines the treatment. The output of interest indicates if respondents have answered that "too much" is spent. Our objective is to identify the main characteristics of individuals that have an impact on the heterogeneity of the treatment effect. The considered dataset is of size $n = 13198$ with $p = 31$ input variables, and basic data preparation steps were used to drop rows with missing values. Notice that imputing missing values may improve estimates. We leave this topic for future work, as handling missing values for variable importance is of high practical interest.

Table 8 displays the top 10 most important variables for Welfare data using our algorithm I_n and also the importance from the `grf` package. The ranking provided by the two algorithms are close, but I_n has a clear meaning as the variance proportion of the treatment effect lost when a given variable is removed, whereas grf-vimp can only be used as a relative importance between covariates, without an intrinsic meaning.

Notice that the sum of the importance of all input variables, i.e. $\sum_j \mathrm{I}_n^{(j)}$, adds to 0.45, which is far from 1. Indeed, when inputs are independent, we have $\sum_j \mathrm{I}^{(j)} \geq 1$. Such a low value is explained by the correlation within input variables. We run a simple hierarchical clustering of the input variables in 10 groups based on correlation, to enforce a small correlation between these groups. More precisely, the hierarchical clustering uses the dissimilarity matrix defined as $1 - \mathbf{C}$, with $\mathbf{C}$ the correlation matrix of the inputs, and the cutting threshold is set to get 10 groups of variables. Then, we run the group variable importance $\mathrm{I}_n^{(J)}$ for each group of variables $J \subset \{1, \ldots, p\}$. The results are displayed in the following Table 9, and are quite straightforward to read. Indeed, half of the treatment heterogeneity is explained by political orientations of individuals, almost a quarter of the heterogeneity is given

Table 8: Top 10 most important variables with respect to I_n and grf-vimp for Welfare data.

$\mathbf{I}_n$		**grf-vimp**	
polviews	0.18	polviews	0.31
partyid	0.09	partyid	0.17
hrs1	0.04	educ	0.09
indus80	0.03	indus80	0.07
maeduc	0.02	hrs1	0.07
educ	0.02	marital	0.04
marital	0.01	degree	0.04
age	0.01	maeduc	0.04
occ80	0.01	occ80	0.02
reg16	0.01	age	0.02

Table 9: Group variable importance for Welfare data.

Variable group	$I_n^{(j)}$
partyid, polviews	0.51
educ, sibs, occ80, prestg80, maeduc, degree	0.23
hrs1, income, rincome, wrkstat	0.07
age, marital, childs, babies	0.04
wrkslf, indus80, sex	0.03
reg16, mobile16	0.01
race, res16, parborn, born	0.00
family16	0.00
earnrs, hompop, adults	0.00
preteen, teens	0.00

by variables mostly related to education and degrees. Then, several groups have a small impact, especially a group about income and working status, and a second one about family information.

4.3.2 NHEFS health data

For the second case study, we use the NHEFS real data about body weight gain following a smoking cessation, extensively described in the causal inference book of Hernan and Robins [29]. As highlighted in the introduction of Chapter 12, these data help to answer the question "what is the average causal effect of smoking cessation on body weight gain?". According to the authors, the unconfoundedness assumption holds. Here, we go a step further to analyze the heterogeneity of this causal effect with respect to health and personal data of individuals who have stopped smoking, using causal forests and our variable importance algorithm. The data record the weight of individuals, first measured in 1971, and then in 1982. The treatment assignment W indicates whether people have stopped smoking during this period, and the observed output Y is the weight difference between 1971 and 1982. We take the dataset of size $n = 1566$ used in Hernan and Robins [29, Chapter 12]. Notice that 63 rows with the output missing were removed, introducing a small bias, as discussed by the authors. They include 9 variables in their analysis, sufficient for unconfoundedness. To better estimate heterogeneity, we also include all variables of the original dataset, that do not contain missing values and are not related to the response, and obtain $p = 41$ input variables. As already mentioned, handling missing values is out of scope of this article, and is left for future work. We run our variable importance algorithm and the grf importance, using $M = 4000$ trees.

The results are displayed in Table 10. Clearly, the original weight of individuals in 1971 has a strong causal effect on weight gain following smoking cessation, with half of the treatment effect variance lost when this

Table 10: Top 10 most important variables with respect to I_n and grf-vimp for NHEFS data.

$\mathbf{I_n}$		**grf-vimp**	
wt71	0.52	wt71	0.26
smokeyrs	0.09	smokeyrs	0.13
smokeintensity	0.07	age	0.10
ht	0.06	ht	0.10
age	0.05	smokeintensity	0.07
alcoholfreq	0.01	school	0.07
active	0.01	active	0.03
tumor	0.01	alcoholfreq	0.03
asthma	0.01	chroniccough	0.02
alcoholtype	0.01	marital	0.02

Table 11: Group variable importance for NHEFS data.

Variable group	$I_n^{(j)}$
sex, ht, wt71, birthcontrol	0.67
age, smokeyrs	0.26
school, education	0.03
alcoholpy, alcoholfreq, alcoholtype	0.02
hbp, diabetes, pica, hbpmed, boweltrouble	0.02

variable is removed. The intensity and duration of smoking, as well as personal characteristics, such as height and age are also involved in treatment heterogeneity, according to both algorithms. Notice that grf-vimp underestimates the importance of wt71 with respect to other variables. Next, we group together variables that are highly correlated, to compute group variable importance. We use the same clustering procedure as the Welfare case, except that we increase the number of groups to 30, since most variables have a very weak dependence with the others. Sex, height, and birth control are highly correlated with the weight in 1971, and this group explains two third of the treatment effect heterogeneity. In fact, age and smoke years also have a quite strong impact with a quarter of heterogeneity explained. Also notice that sex and birth control belong to the most influential group, but have a low importance $\mathrm{I}_n^{(j)}$, which means that they do not contain unique information regarding the treatment effect. This shows how single and group variable importance provide complementary information (Table 11).

5 Conclusions

We introduced a new variable importance algorithm for causal forests, based on the drop and relearn principle, widely used for regression problems. The proposed method has both theoretical and empirical solid groundings. Indeed, we show that our algorithm is consistent, under standard assumptions in the mathematical analysis of random forests. Additionally, we run extensive experiments on simulated, semi-synthetic, and real data, to show the practical efficiency of the method. Notice that the implementation of our variable importance algorithm is available online at https://gitlab.com/random-forests/vimp-causal-forests.

Let us summarize the main guidelines for practitioners using our variable importance algorithm. First, all confounders must be included in the initial data, as it is always necessary to fulfill the unconfoundedness assumption to obtain consistent estimates. Secondly, it is also recommended to include all variables impacting heterogeneity in the data as well. However, leaving aside a non-confounding variable impacting heterogeneity, does not bias the analysis, as opposed to a missing confounder. Thirdly, practitioners must also keep in mind that adding a large number of irrelevant variables, i.e. non-confounding and not impacting heterogeneity, may hurt the accuracy of causal forests. Finally, it is recommended to group correlated variables together, and then compute group variable importance to get additional relevant insights.

We only consider binary treatment assignments throughout the article for the sake of clarity, but notice that the extension to conditional average partial effects with continuous treatment assignments is straightforward, since causal forests natively handle continuous assignments [6]. In this case, the outcome and treatment effect are directly defined through Equation (1) as $Y = \mu(\mathbf{X}) + \tau(\mathbf{X}^{(\mathcal{H})}) \times W + \varepsilon$, where the noise ε is a random variable eventually dependent on W, and the unconfoundedness Assumption 1 is extended to: $\varepsilon \perp\!\!\!\perp W|\mathbf{X}$, as explained in Section 6 of [6]. Then, all the propositions and theorems stated in the article also apply to this continuous settings (except Proposition 1), since they do not rely on the binary assignment assumption. However, further investigations are required to precisely analyze the introduced variable importance algorithm in this continuous setting, which is left for future work.

To conclude, we highlight two further research directions of interest. First, handling missing values in variable importance algorithms is barely discussed in the literature, but is strongly useful in practice, since

observational databases often have missing values, which should be handled carefully to avoid misleading results. Secondly, developing a testing procedure to detect significantly non-null importance values, would enable to identify the set $\mathcal{H}$ of variables involved in heterogeneity, an insight of high practical value. The asymptotic normality of causal forests is probably a promising starting point to develop such testing algorithms.

Research funding: Julie Josse is supported in part by the French National Research Agency ANR16-IDEX-0006.
Author contributions: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results and approved the final version of the manuscript.
Conflict of interest: Authors state no conflict of interest.
Data availability: The datasets from the ACIC data challenge 2019 are available at https://sites.google.com/view/acic2019datachallenge/data-challenge. The "Welfare" dataset is available at https://github.com/gsbDBI/ExperimentData. The NHEFS dataset is available at https://miguelhernan.org/whatifbook.

A Proofs of Propositions 1–5 and Theorems 1–5

Proof of Proposition 1.

Using the observed outcome definition with SUTVA (line 1), and the unconfoundedness Assumption 1 (line 2 to 3), we have

$$\begin{aligned}
\mathbb{E}[Y \mid \mathbf{X}, W] &= \mathbb{E}[WY(1) + (1-W)Y(0) \mid \mathbf{X}, W] \\
&= W\mathbb{E}[Y(1) \mid \mathbf{X}, W] + (1-W)\mathbb{E}[Y(0) \mid \mathbf{X}, W] \\
&= W\mathbb{E}[Y(1) \mid \mathbf{X}] + (1-W)\mathbb{E}[Y(0) \mid \mathbf{X}] \\
&= \mathbb{E}[Y(0) \mid \mathbf{X}] + W(\mathbb{E}[Y(1) \mid \mathbf{X}] - \mathbb{E}[Y(0) \mid \mathbf{X}]) \\
&= \mathbb{E}[Y(0) \mid \mathbf{X}] + W\mathbb{E}[Y(1) - Y(0) \mid \mathbf{X}]) \\
&= \mathbb{E}[\mu(\mathbf{X}) + \varepsilon(0) \mid \mathbf{X}] + W\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) + \varepsilon(1) - \varepsilon(0) \mid \mathbf{X}]\Big) \\
&= \mu(\mathbf{X}) + W\tau(\mathbf{X}^{(\mathcal{H})}),
\end{aligned}$$

and the final result follows. □

Proof of Proposition 2.

From Assumption 2, $\mathbf{X}$ admits a strictly positive density, denoted by f. Then, from Definition 2,

$$\mathbb{P}(\tau(\mathbf{X}^{(\mathcal{H})}) \neq \tau(\mathbf{X}^{'(\mathcal{H})})) > \int_{\mathcal{X}_1 \times \mathcal{X}_1' \times \mathcal{X}_{p-1}} f(x^{(j)}, \mathbf{x}^{(-j)}) f(x^{'(j)}, \mathbf{x}^{(-j)}) \mathrm{d}x^{(j)} \times \mathrm{d}x^{'(j)} \mathrm{d}\mathbf{x}^{(-j)},$$

which is strictly positive, since f is strictly positive and $\mathcal{X}_1$, $\mathcal{X}_1'$, and $\mathcal{X}_{p-1}$ have a non-null Lebesgue measure. □

Proof of Proposition 3.

Assumption 2 implies that $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})] > 0$. By definition,

$$\mathrm{I}^{(j)} = \frac{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})] - \mathbb{V}[\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) | \mathbf{X}^{(-j)}]]}{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})]}, \tag{10}$$

which also writes using the law of total variance

$$\mathrm{I}^{(j)} = \frac{\mathbb{E}[\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})}) | \mathbf{X}^{(-j)}]]}{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})]} = \frac{\mathbb{E}[(\tau(\mathbf{X}^{(\mathcal{H})}) - E[\tau(\mathbf{X}^{(\mathcal{H})}) | \mathbf{X}^{(-j)}])^2]}{\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})]}. \tag{11}$$

If $j \notin \mathcal{H}$, we clearly have $E[\tau(\mathbf{X}^{(\mathcal{H})}) | \mathbf{X}^{(-j)}] = \tau(\mathbf{X}^{(\mathcal{H})})$, and then Equation (11) gives that $\mathrm{I}^{(j)} = 0$.

We now consider the case where $j \in \mathcal{H}$. First, since $\mathbb{V}[\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) | \mathbf{X}^{(-j)}]] \geq 0$, we directly get that $\mathrm{I}^{(j)} \leq 1$ from Equation (10). Secondly, from Definition 2, for $\mathbf{x}^{(-j)} \in \mathcal{X}_{p-1}$, the function $x^{(j)} \to \tau(x^{(j)}, \mathbf{x}^{(-j)})$ takes different values over $\mathcal{X}_1$ and $\mathcal{X}_1'$, and therefore $(\tau(\mathbf{X}^{(\mathcal{H})}) - E[\tau(\mathbf{X}^{(\mathcal{H})}) | \mathbf{X}^{(-j)}])^2 > 0$ with a positive probability, since $\mathcal{X}_1$, $\mathcal{X}_1'$, and $\mathcal{X}_{p-1}$ have a non-null Lebesgue measure. It implies that $\mathrm{I}^{(j)} > 0$. □

Proof of Proposition 4.

We first expand the covariance term

$$\begin{aligned}&\mathrm{Cov}[W-\pi(\mathbf{X}), Y-m(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{H})}]\\ &\quad=\mathbb{E}[(W-\pi(\mathbf{X}))(Y-m(\mathbf{X})) \mid \mathbf{X}^{(\mathcal{H})}]-\mathbb{E}[W-\pi(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{H})}]\mathbb{E}[Y-m(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{H})}].\end{aligned}$$

Notice that the second term is null since $\mathbb{E}[Y-m(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{H})}]=\mathbb{E}[\mathbb{E}[Y-m(\mathbf{X}) \mid \mathbf{X}] \mid \mathbf{X}^{(\mathcal{H})}]=0$. Additionally, by definition,

$$m(\mathbf{X})=\mathbb{E}[Y \mid \mathbf{X}]=\mathbb{E}[\mu(\mathbf{X})+\tau(\mathbf{X}^{(\mathcal{H})})\times W+\varepsilon(W) \mid \mathbf{X}]=\mu(\mathbf{X})+\tau(\mathbf{X}^{(\mathcal{H})})\pi(\mathbf{X}),$$

then $Y-m(\mathbf{X})=(W-\pi(\mathbf{X}))\tau(\mathbf{X}^{(\mathcal{H})})+\varepsilon(W)$, and we get

$$\begin{aligned}&\mathrm{Cov}[W-\pi(\mathbf{X}), Y-m(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{H})}]\\ &\qquad=\mathbb{E}\Big[(W-\pi(\mathbf{X}))\Big((W-\pi(\mathbf{X}))\tau(\mathbf{X}^{(\mathcal{H})})+\varepsilon(W)\Big) \mid \mathbf{X}^{(\mathcal{H})}\Big]\\ &\qquad=\tau(\mathbf{X}^{(\mathcal{H})})\times\mathbb{E}[(W-\pi(\mathbf{X}))^2 \mid \mathbf{X}^{(\mathcal{H})}]+\mathbb{E}[\varepsilon(W)(W-\pi(\mathbf{X})) \mid \mathbf{X}^{(\mathcal{H})}]\\ &\qquad=\tau(\mathbf{X}^{(\mathcal{H})})\times\mathbb{E}[(W-\pi(\mathbf{X}))^2 \mid \mathbf{X}^{(\mathcal{H})}]+\mathbb{E}\Big[(W-\pi(\mathbf{X}))\mathbb{E}[\varepsilon(W) \mid \mathbf{X}, W] \mid \mathbf{X}^{(\mathcal{H})}\Big]\Big]\\ &\qquad=\tau(\mathbf{X}^{(\mathcal{H})})\times\mathbb{V}[W-\pi(\mathbf{X}) \mid \mathbf{X}^{(\mathcal{H})}],\end{aligned}$$

which gives the final local moment equation in $\mathbf{X}^{(\mathcal{H})}$. □

Proof of Proposition 5.

As in the proof of Proposition 4, we obtain

$$\mathrm{Cov}[W-\pi(\mathbf{X}), Y-m(\mathbf{X}) \mid \mathbf{X}^{(-j)}]=\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})(W-\pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}].$$

Notice that

$$\begin{aligned}\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), (W-\pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}]&=\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})(W-\pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}]\\ &\quad-\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)}]\mathbb{E}[(W-\pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}].\end{aligned}$$

Combining the above two equations, we have

$$\begin{aligned}\mathrm{Cov}[W-\pi(\mathbf{X}), Y-m(\mathbf{X}) \mid \mathbf{X}^{(-j)}]&=\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), (W-\pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}]\\ &\quad+\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)}]\times\mathbb{V}[W-\pi(\mathbf{X}) \mid \mathbf{X}^{(-j)}],\end{aligned}$$

which gives the final result since

$$\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), (W-\pi(\mathbf{X}))^2 \mid \mathbf{X}^{(-j)}]=\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), \pi(\mathbf{X})(1-\pi(\mathbf{X})) \mid \mathbf{X}^{(-j)}].$$

□

Proof of Theorem 1.

The result is obtained by applying Theorem 3 from Athey et al. [6]. The first paragraph of Section 3 of Athey et al. [6] provides conditions to apply Theorem 3, that are satisfied by our Assumptions 3 and 4: $\mathbf{X}\in[0,1]^p$, $\mathbf{X}$ admits a density bounded from below and above by strictly positive constants, and μ and τ are bounded.

Next, Assumptions 1–6 from Athey et al. [6] must be verified. As stated at the end of Section 6.1, Assumptions 3–6 always hold for causal forests, the first assumption holds because the functions m, μ, and τ are Lispschitz from our Assumption 4 (the product of Lipschitz functions is Lipschitz), and Assumption 2 is satisfied because $0<\mathbb{V}[W \mid \mathbf{X}]=\pi(\mathbf{X})(1-\pi(\mathbf{X}))<1$ from our Assumption 4.

Finally, the forest is grown from Specification 1, and the treatment effect is identified by Equation (3) since Assumption 1 enforces unconfoundedness. Overall, we apply Theorem 3 from Athey et al. [6] to get the consistency of the causal forest estimate, i.e., for $\mathbf{x}\in[0,1]^p$

$$\tau_{M,n}(\mathbf{x})\xrightarrow{p}\tau(\mathbf{x}^{(\mathcal{H})}).$$

Notice that Theorem 3 from Athey et al. [6] states the consistency of generalized forests. As it will be useful for further results, we give below a proof of the weak consistency in the specific case of causal forests, using arguments of Athey et al. [6]. In particular, we take advantage of Specification 1, which enforces the honesty property, and that the diameters of tree cells vanish as the sample size n increases. First, in our case of binary treatment W, the causal forest estimate writes

$$\tau_{M,n}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i Y_i - \left(\sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i\right)\left(\sum_{i=1}^{n} \alpha_i(\mathbf{x}) Y_i\right)}{\sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i^2 - \left(\sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i\right)^2},$$

where the weight $\alpha_i(\mathbf{x})$ is defined by Equation (3) of Athey et al. [6], as the weight associated to training observation $\mathbf{X}_i$ to form an estimate at the new query point $\mathbf{x}$. The weights $\alpha_i(\mathbf{x})$ sum to 1 over all observations, i.e., $\sum_{i=1}^{n} \alpha_i(\mathbf{x}) = 1$. Also notice that we alleviate notations of $\alpha_i(\mathbf{x})$ throughout the article, but the full expression with all dependencies is $\alpha_i(\mathbf{x}, \mathbf{X}_i, \mathbf{\Theta}_M, \mathscr{D}_n)$, where the causal forest is built with data $\mathscr{D}_n$, and trees are randomized with $\mathbf{\Theta}_M$. Now, we denote by $\Delta_{1,n}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i Y_i$ the first term of the numerator of $\tau_{M,n}(\mathbf{x})$, and derive its convergence. Since the weights sum to 1,

$$\Delta_{1,n}(\mathbf{x}) - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}] = \sum_{i=1}^{n} \alpha_i(\mathbf{x}) (W_i Y_i - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]),$$

and then,

$$\mathbb{E}\left[\Delta_{1,n}(\mathbf{x}) - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]\right] = \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}[\alpha_i(\mathbf{x}) (W_i Y_i - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]) \mid \mathbf{X}_i]].$$

Here, we use a key property of the forest growing given by Specification 1: honesty. Indeed, it enforces that $\mathscr{D}_n$ is randomly split in two halves for each tree, where one part is used to build the splits, and the other half to compute the weights. Therefore, $\alpha_i(\mathbf{x}, \mathbf{X}_i, \mathbf{\Theta}_M, \mathscr{D}_n)$ and $W_i Y_i$ are independent conditional on $\mathbf{X}_i$, for all $\{i, \ldots, n\}$. Then, we have

$$\begin{aligned}
\mathbb{E}\left[\Delta_{1,n}(\mathbf{x}) - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]\right] &= \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}[\alpha_i(\mathbf{x}) \mid \mathbf{X}_i] \mathbb{E}[W_i Y_i - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}] \mid \mathbf{X}_i]] \\
&= \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}[\alpha_i(\mathbf{x}) \mid \mathbf{X}_i] (\mathbb{E}[W_i Y_i \mid \mathbf{X}_i] - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}])].
\end{aligned}$$

Since W and Y are independent conditional on $\mathbf{X}$ from the unconfoundedness Assumption 1, $\mathbb{E}[W_i Y_i \mid \mathbf{X}_i] = \mathbb{E}[W_i \mid \mathbf{X}_i] \mathbb{E}[Y_i \mid \mathbf{X}_i]$. Additionally, Assumption 4 states that the functions π and m are Lipschitz, and since the product of two Lipschitz functions is Lipschitz, $\mathbb{E}[W_i Y_i \mid \mathbf{X}_i]$ is Lipschitz, with a constant $C > 0$. Therefore, we obtain

$$\begin{aligned}
\mathbb{E}\left[\Delta_{1,n}(\mathbf{x}) - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]\right] &\leq \sum_{i=1}^{n} \mathbb{E}[\mathbb{E}[\alpha_i(\mathbf{x}) \mid \mathbf{X}_i] C \parallel \mathbf{X}_i - \mathbf{x} \parallel_2] \\
&\leq C \mathbb{E}\left[\sum_{i=1}^{n} \alpha_i(\mathbf{x}) \parallel \mathbf{X}_i - \mathbf{x} \parallel_2\right] \\
&\leq C \mathbb{E}\left[\sup_i \parallel \mathbf{X}_i - \mathbf{x} \parallel_2 1_{\alpha_i(\mathbf{x}) > 0} \sum_{i=1}^{n} \alpha_i(\mathbf{x})\right] \\
&\leq C \mathbb{E}\left[\sup_i \parallel \mathbf{X}_i - \mathbf{x} \parallel_2 1_{\alpha_i(\mathbf{x}) > 0}\right].
\end{aligned}$$

Since Assumptions 3 and 4 and Specification 1 are satisfied, Equation (26) in the Supplementary Material of Athey et al. [6] states that

$$\mathbb{E}\left[\sup_i \parallel \mathbf{X}_i - \mathbf{x} \parallel_2 1_{\alpha_i(\mathbf{x}) > 0}\right] \to 0,$$

which gives that

$$\mathbb{E}[\Delta_{1,n}(\mathbf{x})] \to \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]. \tag{12}$$

Next, we use Equation (24) in Lemma 7 of the Supplementary Material of Athey et al. [6], to get that $\mathbb{V}[\Delta_{1,n}(\mathbf{x})] = O(a_n/n)$. Since $a_n/n \to 0$ by Specification 1, we finally have $\mathbb{V}[\Delta_{1,n}(\mathbf{x})] \to 0$. Finally, this last limit combined with Equation (12), states that $\Delta_{1,n}(\mathbf{x}) - \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]$ is asymptotically unbiased and of null variance. Using the bias-variance decomposition, we obtain the $\mathbb{L}^2$-consistency of $\Delta_{1,n}(\mathbf{x})$ towards $\mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}]$, which implies the weak consistency

$$\sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i Y_i \xrightarrow{p} \mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}].$$

Identically, we obtain the weak consistency of the other terms involved in $\tau_{M,n}(\mathbf{x})$, i.e., $\sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i \xrightarrow{p} \pi(\mathbf{x})$, $\sum_{i=1}^{n} \alpha_i(\mathbf{x}) Y_i \xrightarrow{p} m(\mathbf{x})$, and $\sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i^2 \xrightarrow{p} \mathbb{E}[W^2 \mid \mathbf{X} = \mathbf{x}]$. The continuous mapping theorem gives for the last term that $\left(\sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i\right)^2 \xrightarrow{p} \mathbb{E}[W \mid \mathbf{X} = \mathbf{x}]^2$. Finally, using Slutsky's Lemma, we obtain

$$\begin{aligned}
\tau_{M,n}(\mathbf{x}) &\xrightarrow{p} \frac{\mathbb{E}[WY \mid \mathbf{X} = \mathbf{x}] - \mathbb{E}[W \mid \mathbf{X} = \mathbf{x}]\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]}{\mathbb{E}[W^2 \mid \mathbf{X} = \mathbf{x}] - \mathbb{E}[W \mid \mathbf{X} = \mathbf{x}]^2} \\
&= \frac{\mathrm{Cov}[W, Y \mid \mathbf{X} = \mathbf{x}]}{\mathbb{V}[W \mid \mathbf{X} = \mathbf{x}]} \\
&= \tau(\mathbf{x}^{(\mathcal{H})}),
\end{aligned}$$

where the last line is given by the local moment Equation (3), which identifies the treatment effect. Finally, notice that this proof applies to any linear local moment equation defining a generalized random forest. □

Proof of Theorem 2.

We consider $j \notin \mathcal{H}$, and follow the same proof as Theorem 1, to show that the causal forest $\tau_{M,n}^{(-j)}(\mathbf{x})$ fit with $\mathscr{D}_n^{\star(-j)}$ converges as

$$\tau_{M,n}^{(-j)}(\mathbf{x}) \xrightarrow{p} \theta(\mathbf{x}^{(-j)}),$$

where $\theta(\mathbf{x}^{(-j)})$ satisfies the following equation by definition of causal forests,

$$\theta(\mathbf{x}^{(-j)}) \times \mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}] - \mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}] = 0.$$

Then, according to Proposition 4, the above moment equation identifies the treatment effect under Assumptions 1 and 2, and we obtain

$$\theta(\mathbf{x}^{(-j)}) = \tau(\mathbf{x}^{(\mathcal{H})}),$$

which gives (i). For (ii), we apply the same proof, except that the obtained local moment equation identifies $\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]$ according to Proposition 5. □

Proof of Theorem 3.

With $j \in \{1, \ldots, p\}$, recall that the causal forest $\tau_{M,n}(\mathbf{x})$ is fit with a centered dataset $\mathscr{D}_n^{\star}$, and the corrected causal forest estimate $\theta_{M,n}^{(-j)}(\mathbf{x})$ is fit with $\mathscr{D}_n^{\star(-j)}$, an independent copy of the centered dataset with the j-th variable dropped, and is formally defined as

$$\theta_{M,n}^{(-j)}(\mathbf{x}) = \tau_{M,n}^{(-j)}(\mathbf{x}) - \frac{\sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 \tau_{M,n}(\mathbf{X}_i) - \overline{W_{\alpha'}^2}\,\overline{\tau}_{\alpha'}}{\sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})\left(W_i - \overline{W}_{\alpha'}\right)^2},$$

where $\overline{W_{\alpha'}^2} = \sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2$, $\overline{\tau}_{\alpha'} = \sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)}) \tau_{M,n}(\mathbf{X}_i)$, and $\overline{W}_{\alpha'} = \sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))$. We first prove the convergence of the first term of the numerator,

$$\begin{aligned}\Delta_n &= \sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 \tau_{M,n}(\mathbf{X}_i) \\ &= \sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 \tau(\mathbf{X}_i) + \sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 (\tau_{M,n}(\mathbf{X}_i) - \tau(\mathbf{X}_i)).\end{aligned}$$

Using the same proof as for Theorem 1, we get that

$$\sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 \tau(\mathbf{X}_i) \xrightarrow{p} \mathbb{E}[(W - \pi(\mathbf{X}))^2 \tau(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}^{(-j)}].$$

For the second term involved in Δ_n, we cannot directly apply the proof of Theorem 1 since the output depends on n through the term $\tau_{M,n}(\mathbf{X}_i)$. We first need to bound $\mathbb{P}(\alpha_1'(\mathbf{x}^{(-j)}) > 0)$. Let us consider a given tree $\ell \in \{1, \ldots, M\}$, and the associated weights $\alpha_{i\ell}'(\mathbf{x}^{(-j)})$ for this tree alone. From Specification 2, we have

$$\sum_{i=1}^{n} 1_{\alpha_{i\ell}'(\mathbf{x}^{(-j)})>0} \le t_0,$$

where t_0 is the maximum number of observations in each terminal leave. Since the weights are identically distributed, we have $n\mathbb{E}\left[1_{\alpha_{1\ell}'(\mathbf{x}^{(-j)})>0}\right] \le t_0$, i.e., $\mathbb{P}(\alpha_{1\ell}'(\mathbf{x}^{(-j)}) > 0) \le t_0/n$. Finally, considering all trees, since $\alpha_1'(\mathbf{x}^{(-j)}) = \sum_{\ell=1}^{M} \alpha_{1\ell}'(\mathbf{x}^{(-j)})/M$, we obtain

$$\mathbb{P}(\alpha_1'(\mathbf{x}^{(-j)}) > 0) \le \frac{M t_0}{n}. \tag{13}$$

Next, for the second term of Δ_n, we write

$$\begin{aligned}\mathbb{E}\left[\left[\left|\sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})(W_i - \pi(\mathbf{X}_i))^2 (\tau_{M,n}(\mathbf{X}_i) - \tau(\mathbf{X}_i))\right|\right]\right] &\le \mathbb{E}\left[\sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)}) |\tau_{M,n}(\mathbf{X}_i) - \tau(\mathbf{X}_i)|\right] \\ &\le n\mathbb{E}\left[\alpha_1'(\mathbf{x}^{(-j)}) |\tau_{M,n}(\mathbf{X}_1) - \tau(\mathbf{X}_1)|\right].\end{aligned}$$

The right hand side of this inequality writes

$$\begin{aligned}&n\mathbb{E}\left[\alpha_1'(\mathbf{x}^{(-j)}) |\tau_{M,n}(\mathbf{X}_1) - \tau(\mathbf{X}_1)|\right] \\ &\quad = n\mathbb{E}\left[\alpha_1'(\mathbf{x}^{(-j)}) |\tau_{M,n}(\mathbf{X}_1) - \tau(\mathbf{X}_1)| \mid \alpha_1'(\mathbf{x}^{(-j)}) > 0\right] \mathbb{P}(\alpha_1'(\mathbf{x}^{(-j)}) > 0) \\ &\quad \le M t_0 \mathbb{E}\left[|\tau_{M,n}(\mathbf{X}_1) - \tau(\mathbf{X}_1)| \mid \alpha_1'(\mathbf{x}^{(-j)}) > 0\right],\end{aligned}$$

where the last inequality is obtained using (13). Finally, since the original causal forest trained with all inputs and the weights $\alpha_1'(\mathbf{x}^{(-j)})$ of the retrained forest are built using independent data, the conditioning event in $\mathbb{E}\left[|\tau_{M,n}(\mathbf{X}_1) - \tau(\mathbf{X}_1)| \mid \alpha_1'(\mathbf{x}^{(-j)}) > 0\right]$ only modifies the distribution of $\mathbf{X}_1$. Therefore, with $\mathbf{Z}_n$ a random variable following this conditional distribution, we have

$$\mathbb{E}\left[|\tau_{M,n}(\mathbf{X}_1) - \tau(\mathbf{X}_1)| \mid \alpha_1'(\mathbf{x}^{(-j)}) > 0\right] = \mathbb{E}\left[|\tau_{M,n}(\mathbf{Z}_n) - \tau(\mathbf{Z}_n)|\right].$$

Since Theorem 1 gives the convergence in probability towards 0 of $\tau_{M,n}(\mathbf{x}) - \tau(\mathbf{x})$ for all $\mathbf{x} \in [0, 1]$ and $\mathbf{Z}_n$ is independent from $\tau_{M,n}(\mathbf{x})$, we get that $\tau_{M,n}(\mathbf{Z}_n) - \tau(\mathbf{Z}_n) \xrightarrow{p} 0$. Since the causal forest is bounded from Specification 2, convergence in probability implies $\mathbb{L}^1$-convergence, and we get that

$$\mathbb{E}\left[|\tau_{M,n}(\mathbf{X}_1) - \tau(\mathbf{X}_1)| \mid \alpha_1'(\mathbf{x}^{(-j)}) > 0\right] = \mathbb{E}\left[|\tau_{M,n}(\mathbf{Z}_n) - \tau(\mathbf{Z}_n)|\right] \to 0.$$

This implies the convergence of the second term of Δ_n, and overall, we obtain that

$$\Delta_n \xrightarrow{p} \mathbb{E}[(W - \pi(\mathbf{X}))^2 \tau(\mathbf{X}) \mid \mathbf{X} = \mathbf{x}^{(-j)}].$$

Next, $\overline{\tau}_{\alpha'}$ is handled similarly as Δ_n, and we follow the same proof as for Theorem 1 to get the weak consistency of the remaining terms involved in $\theta_{M,n}^{(-j)}(\mathbf{x})$, and using Slutsky's lemma, we obtain

$$\frac{\sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})\,(W_i - \pi(\mathbf{X}_i))^2 \tau_{M,n}(\mathbf{X}_i) - \overline{W_\alpha^2}\,\overline{\tau}_{\alpha'}}{\sum_{i=1}^{n} \alpha_i'(\mathbf{x}^{(-j)})\left(W_i - \overline{W}_{\alpha'}\right)^2} \xrightarrow{p} \frac{\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), \pi(\mathbf{X})(1-\pi(\mathbf{X})) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}{\mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}.$$

Then, following the case (ii) of Theorem 2, we get

$$\tau_{M,n}^{(-j)}(\mathbf{x}) \xrightarrow{p} \frac{\mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}{\mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]},$$

which gives the final result

$$\begin{aligned}\theta_{M,n}^{(-j)}(\mathbf{x}) \xrightarrow{p} & \frac{\mathrm{Cov}[W - \pi(\mathbf{X}), Y - m(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}{\mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]} \\ & - \frac{\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), \pi(\mathbf{X})(1-\pi(\mathbf{X})) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}{\mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]} \\ = & \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}],\end{aligned}$$

where the last equality is given by Proposition 5. □

Proof of Theorem 4.

We first consider the case $j \in \{1, \ldots, p\} \setminus \mathcal{H}$ for the sake of clarity. We assume that Assumptions 1–4, and Specifications 1 and 2 are satisfied, and causal forests are trained as specified in Theorem 3. Then, we can apply Theorems 1 and 3 to get that

$$\tau_{M,n}(\mathbf{X}) - \theta_{M,n}^{(-j)}(\mathbf{X}) \xrightarrow{p} 0.$$

According to Specification 2, $\tau_{M,n}(\mathbf{X}) - \theta_{M,n}^{(-j)}(\mathbf{X})$ is bounded, and therefore convergence in probability implies $\mathbb{L}^2$-convergence, i.e.,

$$\mathbb{E}\left[\left(\tau_{M,n}(\mathbf{X}) - \theta_{M,n}^{(-j)}(\mathbf{X})\right)^2\right] \to 0. \tag{14}$$

Next, recall that

$$\mathrm{I}_n^{(j)} = \frac{\sum_{i=1}^{n}\left[\tau_{M,n}(\mathbf{X}_i') - \theta_{M,n}^{(-j)}(\mathbf{X}_i')\right]^2}{\sum_{i=1}^{n}\left[\tau_{M,n}(\mathbf{X}_i') - \overline{\tau_{M,n}}\right]^2} - \mathrm{I}_n^{(0)}.$$

We first consider

$$\Delta_{n,1} = \frac{1}{n}\sum_{i=1}^{n}\left[\tau_{M,n}(\mathbf{X}_i') - \theta_{M,n}^{(-j)}(\mathbf{X}_i')\right]^2,$$

and then

$$\mathbb{E}[\Delta_{n,1}] = \mathbb{E}\left[\left(\tau_{M,n}(\mathbf{X}_1') - \theta_{M,n}^{(-j)}(\mathbf{X}_1')\right)^2\right].$$

Since $|\Delta_{n,1}| = \Delta_{n,1}$, according to Equation (14), we have

$$\mathbb{E}[|\Delta_{n,1}|] \to 0,$$

which also implies the convergence in probability of $\Delta_{n,1}$.

Similarly for the denominator, we write

$$\Delta_{n,2} = \frac{1}{n}\sum_{i=1}^{n}\tau_{M,n}(\mathbf{X}_i')^2 - \overline{\tau_{M,n}}^2.$$

We first show the convergence of $\overline{\tau_{M,n}}$. Hence,

$$\mathbb{E}[\overline{\tau_{M,n}}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\tau_{M,n}(\mathbf{X}_i^{'})\right] = \mathbb{E}[\tau_{M,n}(\mathbf{X})] \to \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})],$$

where the limit is obtained because Theorem 1 gives the weak consistency of $\tau_{M,n}(\mathbf{X})$, which implies the convergence of the first moment since $\tau_{M,n}(\mathbf{X})$ is bounded from Specification 2. Next, we show that the variance of $\overline{\tau_{M,n}}$ vanishes. We use the law of total variance to get

$$\mathbb{V}[\overline{\tau_{M,n}}] = \mathbb{V}[\mathbb{E}[\overline{\tau_{M,n}} \mid \boldsymbol{\Theta}_M, \mathscr{S}_n]] + \mathbb{E}[\mathbb{V}[\overline{\tau_{M,n}} \mid \boldsymbol{\Theta}_M, \mathscr{S}_n]].$$

For $\mathbb{E}[\mathbb{V}[\overline{\tau_{M,n}} \mid \boldsymbol{\Theta}_M, \mathscr{S}_n]]$, notice that $\tau_{M,n}(\mathbf{X}_i^{'})$ are iid conditional on $\boldsymbol{\Theta}_M$ and $\mathscr{S}_n$. Therefore,

$$\mathbb{V}[\overline{\tau_{M,n}} \mid \boldsymbol{\Theta}_M, \mathscr{S}_n] = \frac{\mathbb{V}[\tau_{M,n}(\mathbf{X}) \mid \boldsymbol{\Theta}_M, \mathscr{S}_n]}{n} < \frac{K^2}{n},$$

since $\tau_{M,n}(\mathbf{X})$ is bounded by K from Specification 2. We thus obtain $\mathbb{E}[\mathbb{V}[\overline{\tau_{M,n}} \mid \boldsymbol{\Theta}_M, \mathscr{S}_n]] \to 0$. For the first term, notice that

$$\mathbb{V}[\mathbb{E}[\overline{\tau_{M,n}} \mid \boldsymbol{\Theta}_M, \mathscr{S}_n]] = \mathbb{V}[\mathbb{E}[\tau_{M,n}(\mathbf{X}) \mid \boldsymbol{\Theta}_M, \mathscr{S}_n]] < \mathbb{V}[\tau_{M,n}(\mathbf{X})],$$

where this upper bound converges to 0, since $\tau_{M,n}(\mathbf{X})$ converges towards $\tau(\mathbf{X}^{(\mathcal{H})})$ in $\mathbb{L}^2$. Overall, $\overline{\tau_{M,n}}$ is asymptotically unbiased and its variance vanishes, and therefore converges towards 0 in $\mathbb{L}^2$, and the weak consistency follows, i.e.,

$$\overline{\tau_{M,n}} \xrightarrow{p} \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})].$$

Using the continuous mapping theorem, we conduct the same analysis to get that $\frac{1}{n}\sum_{i=1}^{n}\tau_{M,n}(\mathbf{X}_i^{'})^2 \xrightarrow{p} \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})})^2]$, and then

$$\Delta_{n,2} \xrightarrow{p} \mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})],$$

with $\mathbb{V}[\tau(\mathbf{X}^{(\mathcal{H})})] > 0$ from Assumption 2. Finally, both the numerator $\Delta_{n,1}$ and denominator $\Delta_{n,2}$ of $\mathrm{I}_n^{(j)}$ converge in probability, and we can apply Slutsky's Lemma to obtain

$$\mathrm{I}_n^{(j)} + \mathrm{I}_n^{(0)} \xrightarrow{p} 0,$$

and following the same arguments, we get that $\mathrm{I}_n^{(0)} \xrightarrow{p} 0$, which gives the final result. The proof is similar for the case where $j \notin \mathcal{H}$. □

Proof of Theorem 5.

We can directly deduce from the proof of Theorem 3 that, for $\mathbf{x} \in (0, 1)$,

$$\tau_{M,n}^{(-j)}(\mathbf{x}) \xrightarrow{p} \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}] + \frac{\mathrm{Cov}[\tau(\mathbf{X}^{(\mathcal{H})}), \pi(\mathbf{X})(1-\pi(\mathbf{X})) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}{\mathbb{V}[W - \pi(\mathbf{X}) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]}.$$

We denote by $C_j(\mathbf{x}^{(-j)})$ the second term of the above limit to lighten notations. Next, we follow the proof of Theorem 4 to get the convergence of $\mathcal{I}_n^{(j)}$, given by

$$\mathcal{I}_n^{(j)} \xrightarrow{p} \frac{\mathbb{E}\left[(\tau(\mathbf{X}^{(\mathcal{H})}) - \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)}] - C_j(\mathbf{X}^{(-j)}))^2\right]}{\mathbb{V}[\tau\left(\mathbf{X}^{(\mathcal{H})}\right]}.$$

The numerator writes

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\tau(\mathbf{X}^{(\mathcal{H})}) - \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)}] - C_j(\mathbf{X}^{(-j)})\Big)^2\Big] \\
&\quad = \mathbb{E}\Big[\mathbb{E}\Big[(\tau(\mathbf{X}^{(\mathcal{H})}) - \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)}] - C_j(\mathbf{X}^{(-j)}))^2 \mid \mathbf{X}^{(-j)}\Big]\Big] \\
&\quad = \mathbb{E}[(\tau(\mathbf{X}^{(\mathcal{H})}) - \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)}])^2 + C_j(\mathbf{X}^{(-j)})^2] \\
&\qquad - 2\mathbb{E}\Big[\mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) - \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)}] \mid \mathbf{X}^{(-j)}]\mathbb{E}\Big[C_j(\mathbf{X}^{(-j)}))^2 \mid \mathbf{X}^{(-j)}\Big]\Big] \\
&\quad = \mathbb{E}[(\tau(\mathbf{X}^{(\mathcal{H})}) - \mathbb{E}[\tau(\mathbf{X}^{(\mathcal{H})}) \mid \mathbf{X}^{(-j)}])^2] + \mathbb{E}\Big[C_j(\mathbf{X}^{(-j)})^2\Big].
\end{aligned}$$

Then, we have

$$\mathcal{I}_n^{(j)} \xrightarrow{p} \mathrm{I}^{(j)} + \frac{\mathbb{E}\Big[C_j(\mathbf{X}^{(-j)})^2\Big]}{\mathbb{V}\Big[\tau\Big(\mathbf{X}^{(\mathcal{H})}\Big]},$$

which gives the final result. □

References

1. Obermeyer Z, Emanuel EJ. Predicting the future–big data, machine learning, and clinical medicine. N Engl J Med 2016;375:1216–19.
2. Kennedy EH. Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic J. Statistics 2023;17:3008–49.
3. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika 2021;108:299–319.
4. Künzel S, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Natl Acad Sci 116; 2019. p. 4156–65.
5. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 2018;113:1228–42.
6. Athey S, Tibshirani J, Wager S. Generalized random forests. Ann Stat 2019;47:1148–78.
7. Kosuke I, Marc R. Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat 2013;7:443–70.
8. Hill JL. Bayesian nonparametric modeling for causal inference. J Comput Graph Stat 2011;20:217–40.
9. Shalit U, Johansson FD, Sontag D. Estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning. PMLR; 2017:3076–85 pp.
10. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc 2012; 107:1106–18.
11. Swaminathan A, Joachims T. Batch learning from logged bandit feedback through counterfactual risk minimization. J Mach Learn Res 2015;16:1731–55.
12. Kitagawa T, Tetenov A. Who should be treated? Empirical welfare maximization methods for treatment choice. Econometrica 2018;86: 591–616.
13. Athey S, Wager S. Policy learning with observational data. Econometrica 2021;89:133–61.
14. Tibshirani J, Athey S, Sverdrup E, Wager S. grf: Generalized Random Forests. R package version 2.3.0 2023. https://CRAN.R-project.org/package=grf.
15. Hines O, Diaz-Ordaz K, Vansteelandt S. Variable importance measures for heterogeneous causal effects. arXiv preprint arXiv:2204.06030 2022.
16. Boileau P, Qi NT, Van Der Laan MJ, Dudoit S, Leng N. A flexible approach for predictive biomarker discovery. Biostatistics 2023;24: 1085–105.
17. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. J Am Stat Assoc 2018;113: 1094–111.
18. Williamson BD, Gilbert PB, Simon NR, Carone M. A general framework for inference on algorithm-agnostic variable importance. J Am Stat Assoc 2023;118:1645–58.
19. Hooker G, Mentch L, Zhou S. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Stat Comput 2021;31:1–16.
20. Bénard C, Da Veiga S, Scornet E. Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. Biometrika 2022;109:881–900.

21. VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. Epidemiology 2007;18: 561–8.
22. Rothman KJ. Epidemiology: an introduction. Oxford University Press; 2012.
23. Colnet B, Josse J, Varoquaux G, Scornet E. Risk ratio, odds ratio, risk difference…which causal measure is easier to generalize? arXiv preprint arXiv:2303.16008 2023.
24. Sobol IM. Sensitivity estimates for nonlinear mathematical models. Math Model Comput Experiments 1993;1:407–14.
25. Athey S, Wager S. Estimating treatment effects with causal forests: an application. Observational Studies 2019;5:37–51.
26. Meinshausen N. Quantile regression forests. J Mach Learn Res 2006;7:983–99.
27. Scornet E, Biau G, Vert J-P. Consistency of random forests. Ann Stat 2015;43:1716–41.
28. Green DP, Kern HL. Modeling Heterogeneous treatment effects in survey experiments with bayesian additive regression trees. Public Opin Q 2012;76:491–511.
29. Hernan MA, Robins J. Causal inference: what if. Boca Raton: Chapman & Hall/CRC; 2020.