# MMD-based Variable Importance for Distributional Random Forest

**Clément Bénard** [1*] **Jeffrey Näf** [2*] **Julie Josse** [2]

[1]Safran Tech, Digital Sciences & Technologies, 78114 Magny-Les-Hameaux, France
[2]Inria, PreMeDICaL Team, University of Montpellier
* Equal contribution

## Abstract

Distributional Random Forest (DRF) is a flexible forest-based method to estimate the full conditional distribution of a multivariate output of interest given input variables. In this article, we introduce a variable importance algorithm for DRFs, based on the well-established drop and relearn principle and MMD distance. While traditional importance measures only detect variables with an influence on the output mean, our algorithm detects variables impacting the output distribution more generally. We show that the introduced importance measure is consistent, exhibits high empirical performance on both real and simulated data, and outperforms competitors. In particular, our algorithm is highly efficient to select variables through recursive feature elimination, and can therefore provide small sets of variables to build accurate estimates of conditional output distributions.

## 1 INTRODUCTION

**Context and Objectives.** Distributional Random Forest (DRF) (Ćevid et al., 2022) is an efficient algorithm designed for estimating the conditional distribution of target outputs given a set of input variables. It is inspired from the highly popular Random Forest algorithm proposed by Breiman (2001), which has found widespread use in both classification and regression problems. Unlike Breiman's forests, which provide only the conditional output mean, DRF goes a step further by offering the complete conditional output distribution. This capacity enables users to compute a wide range of quantities of interest with a high accuracy and low computational cost in a subsequent step. These computations encompass the calculation of conditional quantiles, assessments of conditional independence, evaluation of conditional copulas, and the estimation of heterogeneous treatment effects. The main features of DRF are the Maximum Mean Discrepancy (MMD) (Gretton et al., 2007) used as splitting criterion, and the adaptive nearest neighbor formulation of Random Forest (Lin and Jeon, 2006; Scornet, 2016). Unfortunately, DRF also inherits the black-box nature of forest-based methods. Indeed, the large number of operations involved in their prediction mechanisms makes it impossible to grasp how inputs are combined to generate predictions. This lack of interpretability is a strong limitation, in particular for applications with critical decisions at stake, such as healthcare. Therefore, the pursuit of interpretability for black-box algorithms has gained considerable momentum in the machine learning community in recent years, with variable importance measures emerging as one of the main post-hoc method for elucidating these complex models.

The principle of variable importance measures is to quantify the influence of each input variable in a given prediction task. Two importance measures were originally proposed along with Breiman's forests: the Mean Decrease Accuracy (MDA) Breiman (2001), and the Mean Decrease Impurity (MDI) Breiman (2003). The principle of the MDA is to permute the values of a given input variable to break its relation with the output, and compute the associated decrease of accuracy of the forest, defining the importance value for the permuted variable. Although this approach is widely used because of its intuitive definition and small computational cost, many empirical and theoretical studies have shown a strong bias when inputs are correlated (Strobl et al., 2008; Auret and Aldrich, 2011; Gregorutti et al., 2017; Hooker et al., 2021; Bénard et al., 2022). On the other hand, the MDI is defined as the sum of weighted decreases of impurity over all tree nodes that split on a

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

given variable. However, Strobl et al. (2007) highlight several practical flaws of the MDI, and Scornet (2023) shows that the MDI is ill-defined, except in restricted settings. Overall, there is a growing consensus in the machine learning community that other alternatives should be preferred to quantify variable importance for Random Forest. Regarding non-parametric multi-dimensional output regression, there appears to be little literature on variable importance. Recently, Sikdar et al. (2023) develop an importance measure for the multivariate Random Forest (MRF) of Segal and Xiao (2011), based on split improvement criteria. While this is an important first step, this measure suffers from similar limitations as the original MDI, mentioned above. Additionally, the *drf* package (Michel and Ćevid, 2021) also provides an importance measure, defined as the frequency of splits involving a given variable, following the proposal of Athey et al. (2019) for generalized forests, and denoted by vimp-drf throughout the article. This approach is also purely empirical and does not provide a precise quantification of the impact of inputs on the output distribution.

Instead of empirical definitions, variable importance should be first defined through theoretical quantities, and then estimated in a second step with appropriate algorithms, as argued by Williamson et al. (2022) and Bénard et al. (2022) for example. In particular, the drop and relearn principle is often advocated as an efficient approach for variable importance, targeting well-defined theoretical quantities (Mentch and Hooker, 2016; Candes et al., 2018; Lei et al., 2018; Williamson et al., 2022; Hooker et al., 2021). More precisely, the forest is retrained without a given input variable, and the decrease of accuracy with respect to the initial forest with all variables, provides the importance value. In the case of regression problems, this measure gives the proportion of explained output variance lost when a variable is removed, and has a well-grounded theoretical definition as the total Sobol index (Sobol, 1993). Formally, Sobol sensitivity indices quantify the variance of output means conditional on input variables, and were recently extended to output distributions using the MMD (Da Veiga, 2016, 2021). Thus, we build on the drop and relearn principle and the generalized MMD-sensitivity indices, to propose a variable importance measure for DRF, defined as the MMD distance between the conditional output distribution given all inputs, and given all but one input. Such importance measure therefore quantifies how the output conditional distribution changes when a variable is removed, and can be estimated by refitting DRF removing variables one by one.

The definition of a relevant importance measure hinges on the ultimate practical objective, which is typically categorized into two groups: (1) finding a small number of variables with a maximized accuracy, or (2) detecting and ranking all influential variables to focus on for further exploration (Genuer et al., 2010). These two goals differ when variables are dependent. For example, if two variables are highly correlated together and with the output, one of the two inputs can be removed without hurting accuracy for objective (1), since both variables convey the same information. However, both should be included for objective (2), since these two variables may have different meanings in practice for domain experts. In this article, we focus on objective (1), since we use the drop and relearn principle, which is only adapted in this case. For objective (2) other strategies can be used, such as Shapley effects (Owen, 2014; Lundberg and Lee, 2017; Bénard et al., 2022). The dependence between variables not only play a role in the definition of importance measures, but also present significant challenges when it comes to designing efficient algorithms for estimating it. In the case of MMD-sensitivity indices, Da Veiga (2021) essentially introduces estimates adapted for the field of computer experiments, where inputs are independent, or assuming the input distribution is known, or using $k$-nearest neighbors, which struggle in non-trivial input dimensions. Using DRF, we can tackle realistic settings with dependent inputs, higher dimensions, and when only a data sample is available.

**Motivating example.** We consider the following scenario, combining two examples given in Athey et al. (2019), where a Gaussian output $Y$ depends on two uniform variables $X^{(1)}$ and $X^{(2)}$, respectively through a shift in mean and a shift in variance, defined by

$$Y \sim \mathcal{N}(0.8 \cdot \mathbb{1}(X^{(1)} > 0), (1 + \mathbb{1}(X^{(2)} > 0))^2). \quad (1)$$

In addition, $\mathbf{X}$ contains $X^{(3)}$ that is correlated with $X^{(1)}$, but does not influence $Y$, and also seven independent uniform variables. Previous variable importance measures for regression problems are designed to quantify the effect of $\mathbf{X}$ on the conditional mean of $Y$. As such they cannot detect the influence of $X^{(2)}$ on the output distribution, as opposed to the measure introduced in this article. In particular, it correctly quantifies that the effect of $\mathbf{X}$ on $Y$ is divided between $X^{(1)}$ and $X^{(2)}$, as seen in Table 1. This is of critical importance, if the goal is to predict the distribution of $Y$ itself, or if one is generally interested in more targets than conditional expectation, such as quantile estimates. Moreover, the importance values of $X^{(3)}, \ldots, X^{(10)}$ are negligible, showing that the importance measure correctly identifies the irrelevant variables, despite the correlation of $X^{(3)}$ with $X^{(1)}$.

Table 1: Variable importance for data distribution defined in Equation (1).

| $X^{(1)}$ | $X^{(2)}$ | $X^{(3)}$ | $X^{(4)}, \dots, X^{(10)}$ |
|---|---|---|---|
| 0.21 | 0.76 | 0.007 | $< 0.006$ |

**Contributions.** We formally define the new variable importance measure for DRF inspired by Da Veiga (2021), in Section 2, and show how it can be easily estimated by refitting DRF with variables removed one by one. We show that this estimator is consistent for the MMD-based sensitivity index in Section 3. To reduce the computational complexity, we also discuss an estimator based on the Projected Distributional Random Forest, extending the Sobol-MDA algorithm from Bénard et al. (2022), and show that this also leads to a consistent estimator. Our approach can thus be seen as a natural extension of the Sobol-MDA developed for standard Random Forest to DRF. In particular, this allows for a principled variable importance measure for a multivariate dependent variable $\mathbf{Y}$. While the theoretical definition of the above importance measure is close to existing proposals (Da Veiga, 2021), the core of our contribution is the introduction of an estimate of this importance measure using DRF, which is efficient when input variables are dependent, the output is multi-dimensional, and only a data sample is available, with unknown data distributions. Finally in Section 4, we analyze a broad range of simulated and real examples, showing the versatility of the new method. The examples raise from one or low-dimensional dependent variables $\mathbf{Y}$ to functional dependent data. In particular, we show the efficiency of our importance measure for recursive feature elimination on real datasets.

## 2 DRF VARIABLE IMPORTANCE

We first need to introduce several notations and concepts to formalize our variable importance measure for DRF. Throughout, we assume an underlying probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and denote by $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ the reproducing kernel Hilbert space (RKHS) induced by the positive definite, bounded, and continuous kernel $k \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (Hsing and Eubank, 2015, Chapter 2.7), with dimension $d \in \mathbb{N}^\star$. The kernel embedding function $\Phi$ maps any probability measure $\mathbb{Q}$ on $\mathbb{R}^d$ to an element $\Phi(\mathbb{Q}) \in \mathcal{H}$, defined by $\Phi(\mathbb{Q}) = \mathbb{E}[k(\mathbf{Z}, \cdot)]$, with $\mathbf{Z} \sim \mathbb{Q}$, which is well-defined by continuity and boundedness of $k$. For two probability measures $\mathbb{Q}_1$ and $\mathbb{Q}_2$ on $\mathbb{R}^d$, the well-known Maximum Mean Discrepancy (MMD) distance (Gretton et al., 2012) is given by

$$\mathrm{MMD}(\mathbb{Q}_1, \mathbb{Q}_2) = \|\Phi(\mathbb{Q}_1) - \Phi(\mathbb{Q}_2)\|_{\mathcal{H}}.$$

If the kernel $k$ is characteristic, then $\Phi$ is injective, and the MMD is a distance between probability measures. Next, we consider an output vector of interest $\mathbf{Y} = (Y^{(1)}, Y^{(2)}, \dots, Y^{(d)})^T \in \mathbb{R}^d$, and an input vector $\mathbf{X} = (X^{(1)}, X^{(2)}, \dots, X^{(p)})^T \in \mathbb{R}^p$ of dimension $p \in \mathbb{N}^\star$. Finally, we focus on $\mu(\mathbf{x})$, the Hilbert space embedding of the multivariate conditional distribution $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}$ for a given input point $\mathbf{x} \in \mathbb{R}^p$, i.e.,

$$\mu(\mathbf{x}) \overset{\text{def}}{=} \Phi(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H}.$$

**Theoretical importance measure.** Now, in the same spirit as Da Veiga (2021), we show how to obtain a variable importance measure based on the MMD embedding of the conditional distribution and the drop and relearn principle. We first consider the estimation of the conditional mean $\tau(\mathbf{x}) \overset{\text{def}}{=} \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$, for $d = 1$. In this case, one may define the total Sobol index (Sobol, 1993) as

$$\mathrm{ST}^{(j)} \overset{\text{def}}{=} \frac{\mathbb{E}[\mathbb{V}[\tau(\mathbf{X}) \mid \mathbf{X}^{(-j)}]]}{\mathbb{V}[Y]}, \tag{2}$$

where $\mathbf{X}^{(-j)} = \big(X^{(\ell)}\big)_{\ell \neq j}$, and $\mathbb{V}[\cdot]$ is the variance. As mentioned above, this is the expected reduction in output explained variance, once the $j$-th input variable is removed. Another way to see this measure is to write the numerator of $\mathrm{ST}^{(j)}$ as

$$\mathbb{E}[\mathbb{V}[\tau(\mathbf{X}) \mid \mathbf{X}^{(-j)}]] = \mathbb{E}[d_E(\tau(\mathbf{X}), \mathbb{E}[\tau(\mathbf{X}) \mid \mathbf{X}^{(-j)}])^2],$$

where $d_E$ is the Euclidean distance. That is, we consider the distance $d_E$ between the estimates conditional on respectively $\mathbf{X}$ and $\mathbf{X}^{(-j)}$. For other target quantity than $\tau$, we need another relevant distance. For $\mu(\mathbf{x}) \in \mathcal{H}$, a natural choice is the distance induced by the norm $\| \cdot \|_{\mathcal{H}}$, i.e. $d(\xi, \xi') = \|\xi - \xi'\|_{\mathcal{H}}$ for $\xi, \xi' \in \mathcal{H}$, which leads to the variance operator $\mathbb{V}_{\mathcal{H}}$ in $\mathcal{H}$, defined by $\mathbb{V}_{\mathcal{H}}[\xi \mid \mathbf{X}] = \mathbb{E}[\|\xi - \mathbb{E}[\xi \mid \mathbf{X}]\|^2_{\mathcal{H}} \mid \mathbf{X}]$. We can now formalize our theoretical importance measure as the generalized total Sobol index, defined by

$$\mathrm{I}^{(j)} \overset{\text{def}}{=} \frac{\mathbb{E}[\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X}) \mid \mathbf{X}^{(-j)}]]}{\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X})]}, \tag{3}$$

which can also be written using the MMD, as stated in the following proposition: $\mathrm{I}^{(j)}$ quantifies the distance between the conditional distribution $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}}$ with all input variables involved and $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}}$ when one variable is dropped, with respect to the distance between $\mathbb{P}_{\mathbf{Y}}$ and $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}}$. Therefore, this measure has different goals than the traditional Sobol indices. While the latter is designed to detect changes in the conditional expectation of the response variable, $\mathrm{I}^{(j)}$ is designed to detect any change in the distribution of $\mathbf{Y}$, as we will formally prove in the following section. All proofs of propositions and theorems are gathered in Appendix B.

**Proposition 1.** *If $\mathrm{I}^{(j)}$ is the generalized total Sobol index defined by Equation (3), then we have*

$$\mathrm{I}^{(j)} = \frac{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}}, \mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}^{(-j)}})]}{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}})]},$$

$$\mathrm{I}^{(j)} = 1 - \frac{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}^{(-j)}})]}{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}})]}.$$

Importantly, our importance measure is defined with a different normalization constant than in Da Veiga (2021), since we use $\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X})]$ instead of $\mathbb{V}_{\mathcal{H}}[k(\mathbf{Y},\cdot)] = \mathbb{E}[k(\mathbf{Y},\mathbf{Y})] - \mathbb{E}[k(\mathbf{Y},\mathbf{Y}')]$, where $\mathbf{Y}'$ is and independent copy of $\mathbf{Y}$. Indeed, Da Veiga (2021) introduces MMD-based sensitivity indices in the specific settings of computer experiments, where outputs are deterministic functions of inputs. In this case, the two normalization constants coincide, since $\mu(\mathbf{X}) = k(\mathbf{Y},\cdot)$. On the other hand, we consider distributions where the variability of $\mathbf{Y}$ is only partially explained by $\mathbf{X}$, with potentially an explained variability $\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X})]$ much smaller than the total output variations $\mathbb{V}_{\mathcal{H}}[k(\mathbf{Y},\cdot)]$. Therefore, using this last quantity as normalization constant would often lead to small importance values on real data, and we rather define $\mathrm{I}^{(j)}$ with respect to the variability $\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X})]$ explained by all inputs. Notice that, if $\mathbf{Y}$ and $\mathbf{X}$ are independent, $\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X})] = 0$, and no input is influential by construction.

**Variable importance estimate.** We assume that an independent and identically distributed data sample $\mathcal{Z}_n = \{\mathbf{Z}_i\}_{i=1}^n$ of size $n$ is available, where the $i$-th observation is defined by $\mathbf{Z}_i = (\mathbf{X}_i, k(\mathbf{Y}_i,\cdot)) \in \mathbb{R}^p \times \mathcal{H}$. DRF provides nonparametric estimates of the distribution of the multivariate response $\mathbf{Y}$, conditional on the potentially high-dimensional input vector $\mathbf{X}$. That is, for a given query point $\mathbf{x} \in \mathbb{R}^p$, DRF estimates the Hilbert space embedding $\mu(\mathbf{x})$ of $\mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x}}$, denoted by $\mu_{N,n}(\mathbf{x})$, and defined as the average of the $N$ tree estimates. Formally, $\mu_{N,n}(\mathbf{x})$ writes

$$\mu_{N,n}(\mathbf{x}) = \frac{1}{N}\sum_{\ell=1}^{N} T_n(\mathbf{x}; \varepsilon_\ell, \mathcal{Z}_\ell),$$

where $\mathcal{Z}_\ell = \{\mathbf{Z}_{\ell_1}, \ldots, \mathbf{Z}_{\ell_{s_n}}\}$ is a random subset of $\mathcal{Z}_n$ of size $s_n$ chosen for constructing the $\ell$-th tree, $\varepsilon_\ell$ is a random variable capturing the randomness in growing the $\ell$-th tree such as the choice of the splitting candidates, and $T_n(\mathbf{x}; \varepsilon_\ell, \mathcal{Z}_\ell)$ denotes the output of a single tree. More precisely, the tree estimate at query point $\mathbf{x}$, constructed from $\varepsilon_\ell$ and $\mathcal{Z}_\ell$, is given by the average of the terms $k(\mathbf{Y}_i,\cdot)$ over all data points $\mathbf{X}_i$ contained in the leaf $\mathcal{L}_\ell(\mathbf{x})$ where $\mathbf{x}$ falls, i.e.,

$$T_n(\mathbf{x}; \varepsilon_\ell, \mathcal{Z}_\ell) = \sum_{i=1}^{s_n} \frac{\mathbb{1}(\mathbf{X}_{\ell_i} \in \mathcal{L}_\ell(\mathbf{x}))}{|\mathcal{L}_\ell(\mathbf{x})|} k(\mathbf{Y}_{\ell_i},\cdot).$$

Then, we can express the DRF output $\mu_{N,n}(\mathbf{x})$ as an adaptive nearest neighbor estimate, defined by

$$\mu_{N,n}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} w_i(\mathbf{x}) k(\mathbf{Y}_i,\cdot),$$

where the weight for each training observation $\mathbf{X}_i$ writes $w_i(\mathbf{x}) = 1/N \sum_{\ell=1}^N \mathbb{1}(\mathbf{X}_i \in \mathcal{L}_\ell(\mathbf{x}))/|\mathcal{L}_\ell(\mathbf{x})|$.

Finally, we build an estimate $\mathrm{I}_n^{(j)}$ of $\mathrm{I}^{(j)}$ using an initial DRF estimate $\mu_{N,n}(\mathbf{x})$ fit with all inputs involved, combined with the DRF estimate retrained with the $j$-th variable removed, denoted by $\mu_{N,n}(\mathbf{x}^{(-j)})$. Thus, using an independent sample $\mathbf{X}_1', \ldots \mathbf{X}_n'$, we define

$$\mathrm{I}_n^{(j)} = \frac{\sum_{i=1}^n \|\mu_{N,n}(\mathbf{X}_i') - \mu_{N,n}(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2}{\sum_{i=1}^n \|\mu_{N,n}(\mathbf{X}_i') - \overline{\mu_{N,n}}\|_{\mathcal{H}}^2} - \mathrm{I}_n^{(0)}, \quad (4)$$

where $\overline{\mu_{N,n}} = \sum_{i=1}^n \mu_{N,n}(\mathbf{X}_i')/n$, and $\mathrm{I}_n^{(0)}$ is defined as the first term of $\mathrm{I}_n^{(j)}$, but with the DRF $\mu_{N,n}'(\mathbf{X}_i')$, retrained with still all inputs involved but new independent randomizations of the trees $\varepsilon_1', \ldots, \varepsilon_N'$, i.e.

$$\mathrm{I}_n^{(0)} = \frac{\sum_{i=1}^n \|\mu_{N,n}(\mathbf{X}_i') - \mu_{N,n}'(\mathbf{X}_i')\|_{\mathcal{H}}^2}{\sum_{i=1}^n \|\mu_{N,n}(\mathbf{X}_i') - \overline{\mu_{N,n}}\|_{\mathcal{H}}^2}.$$

This is used in Equation (4) to mitigate the finite sample bias. In practice, $\mathrm{I}_n^{(j)}$ is simply computed through vector and matrix multiplications. To state the formula, we introduce the kernel matrix $\mathbf{K} = (k(\mathbf{Y}_i, \mathbf{Y}_j))_{i,j\in\{1,\ldots n\}}$, the DRF weight vectors $\mathbf{w}(\mathbf{x}) = (w_1(\mathbf{x}), \ldots, w_n(\mathbf{x}))$, and $\mathbf{w}(\mathbf{x}^{(-j)}) = (w_1(\mathbf{x}^{(-j)}), \ldots, w_n(\mathbf{x}^{(-j)}))$ for the retrained DRF without the $j$-th variable. Moreover, we consider the vector $\mathbf{k} = (k(\mathbf{Y}_1,\cdot), \ldots, k(\mathbf{Y}_n,\cdot))^\top$, and the mean weight over the independent sample $\bar{\mathbf{w}} = \sum_{i=1}^n \mathbf{w}(\mathbf{X}_i')/n$. Then, the forest estimates writes $\mu_{N,n}(\mathbf{x}) = \mathbf{w}(\mathbf{x})^\top\mathbf{k}$ and $\mu_{N,n}(\mathbf{x}^{(-j)}) = \mathbf{w}(\mathbf{x}^{(-j)})^\top\mathbf{k}$, and $\mathrm{I}_n^{(j)}$ is calculated with the following formula,

$$\begin{aligned}\mathrm{I}_n^{(j)} = \Big\{ \sum_{i=1}^n \big[\mathbf{w}(\mathbf{X}_i') - \mathbf{w}(\mathbf{X}_i'^{(-j)})\big]^\top \mathbf{K} \qquad\qquad (5)\\ \times \big[\mathbf{w}(\mathbf{X}_i') - \mathbf{w}(\mathbf{X}_i'^{(-j)})\big]\Big\}\\ \times \Big\{ \sum_{i=1}^n \big[\mathbf{w}(\mathbf{X}_i') - \bar{\mathbf{w}}\big]^\top \mathbf{K} \big[\mathbf{w}(\mathbf{X}_i') - \bar{\mathbf{w}}\big]\Big\}^{-1} - \mathrm{I}_n^{(0)},\end{aligned}$$

where $\mathrm{I}_n^{(0)}$ takes the same form as the first term. Thus, we in fact consider the difference in weights, with each element weighted by the kernel matrix $\mathbf{K}$. In turn, this is standardized by the estimated variance of the embedding of $\mathbf{Y} \mid \mathbf{X}$. For the sake of clarity, $\mathrm{I}_n^{(j)}$ is formalized with an independent dataset, but out-of-bag predictions can also be used instead. We will see in the next section that this variable importance algorithm for DRF is consistent with respect to the theoretical importance measure defined in Equation (3), and thus provides an efficient assessment of the impact of each variable on the output conditional distribution.

# 3 THEORETICAL PROPERTIES

The construction of our variable importance algorithm is based on well-defined quantities from Equations (2) and (3) and Proposition 1, and therefore enjoys good theoretical properties as we show throughout this section. To state our results, we need to formalize several assumptions, and we first characterize the required kernel properties.

(**K1**) The kernel $k$ is bounded, and the function $(\mathbf{x}, \mathbf{y}) \mapsto k(\mathbf{x}, \mathbf{y})$ is (jointly) continuous.

(**K2**) The kernel $k$ is characteristic.

In particular, Assumption (**K2**) implies that the kernel embedding function $\Phi$ is injective, and then, for two probability measures $\mathbb{Q}_1$ and $\mathbb{Q}_2$, $\|\Phi(\mathbb{Q}_1) - \Phi(\mathbb{Q}_2)\|_{\mathcal{H}} = 0$ implies $\mathbb{Q}_1 = \mathbb{Q}_2$, as explained in Sriperumbudur (2016); Simon-Gabriel et al. (2020). For example, all these assumptions are met for the Gaussian kernel, which is the standard kernel in DRF, see e.g., Ćevid et al. (2022, Appendix A). When the chosen kernel $k$ satisfies these assumptions, our importance measure defined in Equation (3) detects any change in the output conditional distribution when a variable is removed, as stated in the following proposition.

**Proposition 2.** *Assume that Assumptions (**K1**)-(**K2**) holds and that for each* $\mathbf{x}$ *in a set with nonzero probability,* $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} \neq \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}=\mathbf{x}^{(-j)}}$*, then* $0 < \mathrm{I}^{(j)} \leq 1$*, and otherwise* $\mathrm{I}^{(j)} = 0$.

To deepen our discussion, we also need assumptions about the data distribution, given below.

(**D1**) The observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and identically distributed on $[0, 1]^p$, with a density bounded from below and above by strictly positive constants.

(**D2**) The mapping $\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H}$ is Lipschitz.

Assumption **(D1)** is standard when analyzing Random Forest, see e.g., Wager and Athey (2017); Athey et al. (2019); Ćevid et al. (2022); Näf et al. (2023). Assumption **(D2)** can be restated as

$$\mathrm{MMD}(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}_1}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}_2}) \leq L\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathbb{R}^p},$$

for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$, and some Lipschitz constant $L > 0$. Thus, whenever $\mathbf{x}_1$ and $\mathbf{x}_2$ are close, the corresponding distributions $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}_1}$ and $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}_2}$ need to be close in terms of MMD distance. This corresponds to a natural generalization of the Lipschitz condition usually assumed for standard Random Forest (Wager and Athey, 2018; Athey et al., 2019).

**Consistency of DRF importance.** To develop our theory, we consider infinite forests, a standard simplification also employed by Scornet et al. (2015); Wager and Athey (2018); Athey et al. (2019), where the considered forest is defined as the limit when the number of trees $N$ grows to infinity. Such forest estimator $\mu_n(\mathbf{x})$ is obtained by averaging all $\binom{n}{s_n}$ possible subsets of $\{\mathbf{Z}_i\}_{i=1}^n$ of size $s_n$, and taking the expectation over the tree randomization $\varepsilon$. This idealized version of our DRF predictor, which we will denote by $\mu_n(\mathbf{x})$ from now onwards, is given by

$$\mu_n(\mathbf{x}) = \binom{n}{s_n}^{-1} \sum_{i_1 < \cdots < i_{s_n}} \mathbb{E}\left[T_n(\mathbf{x}; \varepsilon, \{\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}\}) \mid \mathcal{Z}_n\right]. \tag{6}$$

Following Wager and Athey (2018); Athey et al. (2019); Ćevid et al. (2022); Näf et al. (2023), the forest construction enforces that trees are honest, symmetric, $\alpha$-regular, each node may split on all variables with a positive probability, and the subsample size $s_n$ is defined by $s_n = n^\beta$, with $0 < \beta < 1$. Theses characteristics are formalized in Assumptions **(F1)**–**(F5)** in Appendix B. We now prove the consistency of the proposed DRF variable importance algorithm, stated in Equation (4). First, we slightly strengthen the result of consistency in Ćevid et al. (2022, Theorem 1).

**Proposition 3.** *Assume that the forest construction satisfies the properties **(F1)**-**(F5)**. Additionally, assume that k meets Assumption (**K1**), and that (**D1**) and (**D2**) hold. Then, we have consistency of* $\mu_n(\mathbf{x})$ *in* (6) *with respect to the RKHS norm in mean, that is*

$$\mathbb{E}[\|\mu_n(\mathbf{X}) - \mu(\mathbf{X})\|_{\mathcal{H}}] \longrightarrow 0.$$

The consistency of $\mu_n(\mathbf{X}^{(-j)})$ directly follows from Proposition 3, since $\mu_n(\mathbf{X}^{(-j)})$ is trained on the data with the $j$-th input variable removed. We only need the following additional Lipschitz assumption to satisfy Assumption (**D2**) with a reduced set of inputs.

(**D3**) The mapping $\mathbf{x}^{(-j)} \mapsto \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}] \in \mathcal{H}$ is Lipschitz.

**Proposition 4.** *Under the assumptions of Proposition 3, and provided that Assumption (**D3**) is satisfied, we have* $\mathbb{E}[\|\mu_n(\mathbf{X}^{(-j)}) - \mu(\mathbf{X}^{(-j)})\|_{\mathcal{H}}] \longrightarrow 0$.

We deduce the consistency of our variable importance algorithm from the last two propositions.

**Theorem 1.** *Assume that the forest construction satisfies the properties **(F1)**-**(F5)**. Additionally, assume that k meets Assumption (**K1**), and that (**D1**)-(**D3**) hold. Then, we have consistency of* $\mathrm{I}_n^{(j)}$ *in* (4)*, that is*

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

**Projected DRF.** The computation of our importance measure for all input variables involves a DRF retrain for each variable, i.e. $p+1$ training. While this approach is efficient for moderate dimensions, it is not tractable in high-dimensional settings. Indeed, Ćevid et al. (2022) state that the computational complexity of a single DRF fit is $\mathcal{O}(B \times N \times p \times n \log n)$, where $B$ is the number of random features to approximate the MMD statistics for the splitting criterion. Therefore, our importance algorithm has a quadratic complexity with respect to the input dimension $p$. In fact, this limitation also occurs for Breiman's forest in the case of regression problems. A solution was proposed by Bénard et al. (2022) with the Projected Random Forest, whose complexity is independent of $p$, once the initial forest with all inputs is trained, and is therefore strongly efficient in high dimensional settings. The principle of projected forests is to eliminate a given input variable from the prediction mechanism of the forest by the projection of the tree partitions on the subspace generated by all other variables. In practice, the associated predictions can be easily computed using the original trees, and simply ignoring splits involving the discarded variable by sending data points on both sides of such splits. Following this procedure, both training points and the new query point fall in multiple terminal leaves of the original partition. Then, cells are intersected to get the training points falling in the exact same collections of leaves as the considered query point, and finally used to compute the projected tree prediction. The projected forest was later extended to Shapley effects in Bénard et al. (2022), and to local importance measures by I Amoukou and Brunel (2022).

As DRF is essentially a Random Forest with the dependent variable taking values in the RKHS $\mathcal{H}$, the same approach can be adapted as well, to obtain a Projected Distributional Random Forest. Applying the arguments of Bénard et al. (2022), the projection approach reduces the computational complexity from the $\mathcal{O}(B \times N \times p^2 \times n \log(n)^2)$ of a DRF fit for each variable to $\mathcal{O}(B \times N \times n \log(n)^3)$. Here, we show that this approach still leads to a consistent estimator of $\mathrm{I}^{(j)}$. In the sequel, we denote by $\mu_n^{(-j)}(\mathbf{X})$ the projected DRF estimate, indicating that the projection was done after fitting DRF on the full data. This is in contrast to $\mu_n(\mathbf{X}^{(-j)})$ which was already fitted with $X^{(j)}$ removed.

**Proposition 5.** *Assume that the forest construction satisfies the properties **(F1)**-**(F5)**. Additionally, assume that $k$ meets Assumption **(K1)**, and that **(D1)**-**(D3)** hold. Then, we have consistency of $\mu_n^{(-j)}(\mathbf{x})$ in probability,*

$$\|\mu_n^{(-j)}(\mathbf{x}) - \mu(\mathbf{x}^{(-j)})\|_{\mathcal{H}} = \mathcal{O}_p\left(n^{-\gamma}\right),$$

*for any $\gamma \leq \frac{1}{2}\min\left(1-\beta, \frac{\pi \log(1-\alpha)}{p\log(\alpha)} \cdot \beta\right)$, where $\alpha$ and $\beta$ are chosen in **(F4)** and **(F5)** respectively. Moreover,*

$$\mathbb{E}[\|\mu_n^{(-j)}(\mathbf{X}) - \mu(\mathbf{X}^{(-j)})\|_{\mathcal{H}}] \longrightarrow 0.$$

We then redefine the variable importance by simply exchanging $\mu_n(\mathbf{X}_i'^{(-j)})$ by $\mu_n^{(-j)}(\mathbf{X}_i')$, and obtain

$$\mathrm{I}_{n,\mathrm{proj}}^{(j)} = \frac{\sum_{i=1}^{n}\|\mu_n(\mathbf{X}_i') - \mu_n^{(-j)}(\mathbf{X}_i')\|_{\mathcal{H}}^2}{\sum_{l=1}^{n}\|\mu_n(\mathbf{X}_i') - \frac{1}{n}\sum_{i=1}^{n}\mu_n(\mathbf{X}_i')\|_{\mathcal{H}}^2}, \quad (7)$$

which is also consistent, as stated in this last result.

**Theorem 2.** *Assume that the forest construction satisfies the properties **(F1)**-**(F5)**. Additionally, assume that $k$ meets Assumption **(K1)**, and that **(D1)**-**(D3)** hold. Then, $\mathrm{I}_{n,proj}^{(j)}$ in (7) is consistent, that is*

$$\mathrm{I}_{n,proj}^{(j)} \xrightarrow{p} \mathrm{I}^{(j)}.$$

In addition, we note that the last step of our algorithm is to compute $\mathrm{I}_n^{(j)}$ through Equation (5), which involves a matrix multiplication of complexity $\mathcal{O}(n^2)$. Following Ćevid et al. (2022), we can use the Random Fourier approximation to compute the MMD of Equation (5), with a linear complexity with respect to the sample size $n$. Alternatively, once all predictions are computed, it is also possible to subsample the data to compute Equation (5), with a size of typically 1000 points for large samples. This leads to a $O(1)$ complexity, and a minor impact on the accuracy of $\mathrm{I}_n^{(j)}$, since Equation (5) is simply an average over a large number of points.

# 4 EXPERIMENTS

We run several batches of experiments to show the high performance of our importance measure on both simulated and real data, especially with respect to the main competitors. In particular, we run comparisons with the native DRF importance (Michel and Ćevid, 2021), based on split frequencies and denoted by vimp-drf. For univariate output cases, we also add the MDA (Breiman, 2001) and Sobol-MDA (Bénard et al., 2022) algorithms. While Sobol-MDA also focuses on objective (1), MDA and vimp-drf are not specifically tailored for objective (1) or (2). Nonetheless, vimp-drf appears to be the only existing competing variable importance measure for DRF, to our best knowledge. Therefore, we only show that our method performs better than vimp-drf for our objective of interest. Additionally, the MDA is the most widely used importance measure for Breiman's forest, and is thus a useful baseline. Finally, we use the Gaussian kernel with the median heuristic, and a number of random features $B = 10$, which is standard in the DRF implementation (Michel and Ćevid, 2021; Ćevid et al., 2022). Code to reproduce the experiments is available in the Supplementary Material.

### 4.1 Simulated data

**Univariate output.** We run a first experiment with a univariate output to compare our proposed variable importance algorithm to the existing vimp-drf based on split frequencies, and standard methods for regression forests. Hence, we consider a Gaussian input vector of dimension $p = 10$, where all variables have unit variance, and each pair of distinct variables have a correlation of 0.5, except that $\text{Cov}(X^{(1)}, X^{(10)}) = 0.9$. Then, the output is defined as

$$Y \sim \mathcal{N}(2X^{(1)} + X^{(2)}, (2|X^{(3)}| + 2|X^{(4)}| + 2|X^{(5)}|)^2).$$

Based on this data distribution, we run the following experiment: a data sample of size $n = 3000$ is drawn, a DRF is fit with $N = 500$ trees, and both $\mathrm{I}_n^{(j)}$ and vimp-drf are computed. We also fit a regression forest and compute the MDA (Breiman, 2001) and Sobol-MDA (Bénard et al., 2022). This procedure is repeated 10 times for uncertainties, and the average importance values are reported in Table 2. Most standard deviations are small, and displayed in Table 1 in Appendix A. Clearly, $\mathrm{I}_n^{(j)}$ is the only algorithm to identify the five relevant variables as the most important ones. On the other hand, both vimp-drf and MDA rank $X^{(10)}$ in second position, because of its strong correlation with $X^{(1)}$, although $X^{(10)}$ is not involved in the distribution of $Y$. Since variables $X^{(3)}$, $X^{(4)}$, and $X^{(5)}$ are not involved in the mean of $Y$ but only in its variance, they are only identified as important by $\mathrm{I}_n^{(j)}$ and vimp-drf, but not by the MDA and Sobol-MDA, as expected. While the Sobol-MDA gives a negligible importance to all variables not involved in the mean of $Y$, the MDA gives high negative values to the relevant variables $X^{(3)}$, $X^{(4)}$, and $X^{(5)}$. This phenomenon is not really surprising given the MDA's flaws, mentioned in the introduction, and extensively discussed in the literature. Finally, the regression forest has an explained variance of 13%, because of the strong noise involved in the definition of $Y$. This explains the quite small values of $X^{(1)}$, $X^{(2)}$ given by the Sobol-MDA, which estimates the proportion of output variance lost when a given input variable is removed.

**Bivariate output.** A main feature of DRF is to handle multivariate outputs, and we therefore focus our second simulated experiment on such a case. We consider $p = 10$ input variables, following a uniform distribution on the unit cube, and two uniform outputs defined by $Y^{(1)} \sim \mathcal{U}(X^{(1)}, 1 + X^{(1)})$ and $Y^{(2)} \sim \mathcal{U}(0, X^{(2)})$. Next, we draw a sample of size $n = 500$, fit a DRF of $N = 500$ trees, and finally compute our importance measure, as well as vimp-drf. Table 3 displays the mean importance over 10 repetitions for each variable, where standard deviations are small, and thus omitted.

| $X^{(j)}$ | $\mathrm{I}_n^{(j)}$ | $X^{(j)}$ | v-drf | $X^{(j)}$ | MDA | $X^{(j)}$ | S-MDA |
|---|---|---|---|---|---|---|---|
| $\boldsymbol{X^{(1)}}$ | 0.181 | $\boldsymbol{X^{(1)}}$ | 0.649 | $\boldsymbol{X^{(1)}}$ | 6.47 | $\boldsymbol{X^{(1)}}$ | 0.014 |
| $\boldsymbol{X^{(4)}}$ | 0.073 | $X^{(10)}$ | 0.096 | $X^{(10)}$ | 2.56 | $\boldsymbol{X^{(2)}}$ | 0.012 |
| $\boldsymbol{X^{(5)}}$ | 0.073 | $\boldsymbol{X^{(2)}}$ | 0.062 | $\boldsymbol{X^{(2)}}$ | 1.33 | $X^{(8)}$ | -0.001 |
| $\boldsymbol{X^{(2)}}$ | 0.072 | $\boldsymbol{X^{(5)}}$ | 0.059 | $X^{(6)}$ | 0.25 | $X^{(7)}$ | -0.003 |
| $\boldsymbol{X^{(3)}}$ | 0.065 | $\boldsymbol{X^{(4)}}$ | 0.056 | $X^{(8)}$ | 0.24 | $X^{(9)}$ | -0.003 |
| $X^{(10)}$ | 0.010 | $\boldsymbol{X^{(3)}}$ | 0.050 | $X^{(7)}$ | 0.18 | $X^{(6)}$ | -0.003 |
| $X^{(7)}$ | 0.005 | $X^{(6)}$ | 0.007 | $X^{(9)}$ | 0.18 | $\boldsymbol{X^{(5)}}$ | -0.003 |
| $X^{(6)}$ | 0.005 | $X^{(9)}$ | 0.007 | $\boldsymbol{X^{(3)}}$ | -0.52 | $\boldsymbol{X^{(3)}}$ | -0.004 |
| $X^{(9)}$ | 0.005 | $X^{(7)}$ | 0.007 | $\boldsymbol{X^{(5)}}$ | -0.62 | $\boldsymbol{X^{(4)}}$ | -0.004 |
| $X^{(8)}$ | 0.005 | $X^{(8)}$ | 0.007 | $\boldsymbol{X^{(4)}}$ | -0.90 | $X^{(10)}$ | -0.006 |

Table 2: Variable importance for the univariate output experiment for $\mathrm{I}_n^{(j)}$, vimp-drf (v-drf), MDA, and Sobol-MDA (S-MDA).

| | $\boldsymbol{X^{(1)}}$ | $\boldsymbol{X^{(2)}}$ | $X^{(6)}$ | $X^{(8)}$ | $X^{(3)}$ | $X^{(10)}$ |
|---|---|---|---|---|---|---|
| $\mathrm{I}_n^{(j)}$ | 0.68 | 0.41 | $9.10^{-4}$ | $8.10^{-4}$ | $7.10^{-4}$ | $6.10^{-4}$ |
| vimp-drf | 0.19 | 0.70 | 0.01 | 0.02 | 0.01 | 0.02 |

Table 3: Top six variables for the bivariate output experiment for $\mathrm{I}_n^{(j)}$ and vimp-drf.

Clearly, both methods identify the relevant variables $X^{(1)}$ and $X^{(2)}$ as the most relevant ones. However, $\mathrm{I}_n^{(j)}$ identifies $X^{(1)}$ as more important than $X^{(2)}$, as opposed to vimp-drf. By definition of $\mathrm{I}_n^{(j)}$ and the MMD distance, if $X^{(1)}$ is removed from the training data, the conditional distribution estimated by DRF is closer to the true target than when $X^{(2)}$ is removed. This shows that vimp-drf based on split frequencies can be misleading. Also notice that the importance given by $\mathrm{I}_n^{(j)}$ is negligible for all irrelevant variables, while vimp-drf gives higher values. Finally, we also take advantage of this second experiment to show the scalability of our algorithm with respect to the sample size $n$. We run the same experiment with increasing sample sizes, up to $n = 10000$. For the last step of the procedure, where we aggregate predictions using Equation (5), we use a subsample of fixed size 1000, to preserve a linear complexity—see the end of Section 3. Figure 1 displays the results and shows that the importance values (again averaged over 10 repetitions) are roughly constant as the sample size increases, with a slight increase for $X^{(1)}$ and $X^{(2)}$.

**High-dimensional case.** Variable selection is frequently performed in high-dimensional settings. The goal of this third experiment is to show the good behavior of our algorithm in such cases. We consider again the experiment of the previous paragraph with a bivariate output. We simply set $p = 1000$ instead of $p = 10$, by adding uniform input variables, and keep all the other settings untouched. We obtain the results

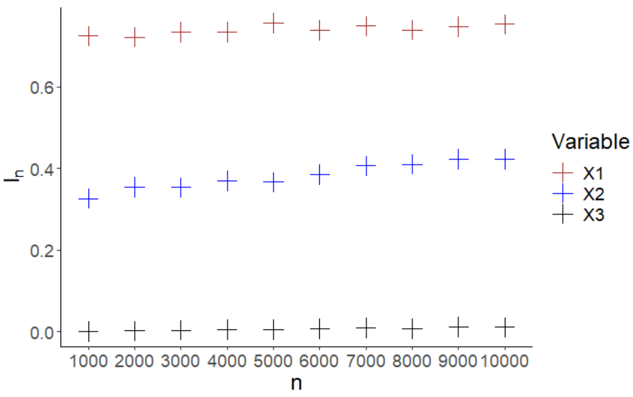

Figure 1: Values of $\mathrm{I}_n^{(1)}$, $\mathrm{I}_n^{(2)}$, and $\mathrm{I}_n^{(3)}$, with an increasing sample size for the bivariate output experiment.

| | $\boldsymbol{X^{(1)}}$ | $\boldsymbol{X^{(2)}}$ | $X^{(371)}$ | $X^{(698)}$ | $X^{(866)}$ | $X^{(705)}$ |
|---|---|---|---|---|---|---|
| $\mathrm{I}_n^{(j)}$ | 0.57 | 0.29 | 0.02 | 0.02 | 0.02 | 0.02 |
| vimp-drf | 0.03 | 0.05 | 0.001 | 0.001 | 0.001 | 0.001 |

Table 4: Top six variables for the bivariate output experiment with $p = 1000$.

displayed in Table 4. The good performance of our algorithm is preserved, and it still outperforms the existing competitor, which wrongly identifies variable $X^{(2)}$ as more important than $X^{(1)}$, and dilutes the importance of the two relevant variables compared to the original case with $p = 10$—see Table 3. Variables are ordered in decreasing order of $\mathrm{I}_n^{(j)}$ in Table 4: although there are 1000 variables involved, the highest value among the irrelevant variables is still small.

**Functional output.** For this last simulated experiment, we address the more challenging case of a functional output. That is, we assume that conditional on $\mathbf{X}$, $Y$ is a Gaussian process (Rasmussen and Williams, 2006): $Y(t) = X^{(1)} + f(t)$, where for all $t \in \mathbb{R}$, $f(t) \sim \mathcal{GP}\big(0, k_{\mathbf{X}}(t,t)\big)$, with $k_{\mathbf{X}}$ being a Gaussian kernel with bandwidth parameter $1/X^{(2)}$. Then, the vector $\mathbf{Y}$ given $\mathbf{X}$ is obtained by sampling from this Gaussian process on a fixed regular grid of $[-5, 5]$ of size $d = 30$. As before, $\mathbf{X}$ is uniformly distributed on the unit cube, with $p = 10$, and we also set $n = 2000$ and $N = 500$ trees. Table 5 shows that the DRF importance measures obtain the right ordering, and especially detect the dependence of $\mathbf{Y}$ on $X^{(2)}$, which is a notoriously difficult problem.

| | $\boldsymbol{X^{(1)}}$ | $\boldsymbol{X^{(2)}}$ | $X^{(6)}$ | $X^{(3)}$ | $X^{(4)}$ | $X^{(9)}$ |
|---|---|---|---|---|---|---|
| $\mathrm{I}_n^{(j)}$ | 0.70 | 0.34 | 0.03 | 0.03 | 0.03 | 0.03 |
| vimp-drf | 0.61 | 0.26 | 0.02 | 0.02 | 0.02 | 0.02 |

Table 5: Top six variables for the functional output experiment for $\mathrm{I}_n^{(j)}$ and vimp-drf.

### 4.2 Recursive feature elimination for real data

In this subsection, we show the efficiency of our importance measure to perform backward variable selection on real data with multivariate outputs. Hence, we use the recursive feature elimination algorithm (RFE), originally introduced by Guyon et al. (2002) for variable selection with support vector machines, and first adapted to random forests for regression problems by Gregorutti et al. (2017). The main principle of the RFE is to iteratively remove the less important variable. At each step, the learning algorithm is rerun on the data with the reduced set of inputs, and the importance measure is computed for all variables involved. Then, the less important variable is removed from the data, and the algorithm moves on to the next iteration. In our case, a DRF is fit at each step of the RFE, and we compare our method introduced in Section 2, and vimp-drf. As explained in the introduction, our algorithm focuses on objective (1) of finding a small number of variables with a maximized accuracy. The recursive feature elimination procedure built from $\mathrm{I}_n^{(j)}$, removes highly correlated redundant features. Indeed, all variables of a highly correlated and important group have low importance, until only one variable of this group remains in the data over the backward selection, by definition of $\mathrm{I}_n^{(j)}$, which estimates the information loss when a given input is removed. To assess the quality of the variable selection, we need an appropriate loss to quantify the accuracy of the empirical conditional measure $\hat{\mathbb{P}}_{\mathbf{Y}\,|\,\mathbf{X}^{(\mathbf{J})}}$ given by DRF fit with a subset of inputs $\mathbf{J} \subset \{1, \ldots, p\}$, with respect to the theoretical target measure $\mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}}$. The MMD provides a natural distance between these two quantities of interest, defined by $\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}}, \hat{\mathbb{P}}_{\mathbf{Y}\,|\,\mathbf{X}^{(\mathbf{J})}})]$. We simplify this metric by subtracting $\mathbb{E}[\|\Phi(\mathbb{P}_{\mathbf{Y}\,|\,\mathbf{X}})\|^2_{\mathcal{H}}]$, since this last term does not rely on the estimated distribution, and we are only interested in relative comparisons of our approach with competitors. Therefore, our evaluation metric writes $\mathbb{E}[\|\mu_{N,n}(\mathbf{X}^{(\mathbf{J})})\|^2_{\mathcal{H}}] - 2\mathbb{E}[\langle k(\mathbf{Y}, \cdot), \mu_{N,n}(\mathbf{X}^{(\mathbf{J})})\rangle_{\mathcal{H}}]$. Then our final empirical loss $\mathcal{L}_n^{(\mathbf{J})}$ is defined as a Monte-Carlo estimate of the above quantity using an independent sample $(\mathbf{X}'_1, \mathbf{Y}'_1), \ldots, (\mathbf{X}'_n, , \mathbf{Y}'_n)$, i.e.

$$\mathcal{L}_n^{(\mathbf{J})} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{w}(\mathbf{X}'_i)^\top \mathbf{K}\mathbf{w}(\mathbf{X}'_i) - 2\mathbf{w}(\mathbf{X}'_i)^\top \mathbf{k}(\mathbf{Y}'_i),$$

where the index $\mathbf{J}$ of $\mathbf{X}_i'^{(\mathbf{J})}$ is dropped to lighten notations, and $\mathbf{k}(\mathbf{Y}'_i) = (k(\mathbf{Y}_1, \mathbf{Y}'_i), \ldots, k(\mathbf{Y}_n, \mathbf{Y}'_i))^\top$. Notice that this loss can be negative since we subtract a constant positive term, but is obviously still a valid metric for comparisons. Finally, this loss $\mathcal{L}_n^{(\mathbf{J})}$ provides a value at each step of the RFE, and to get a

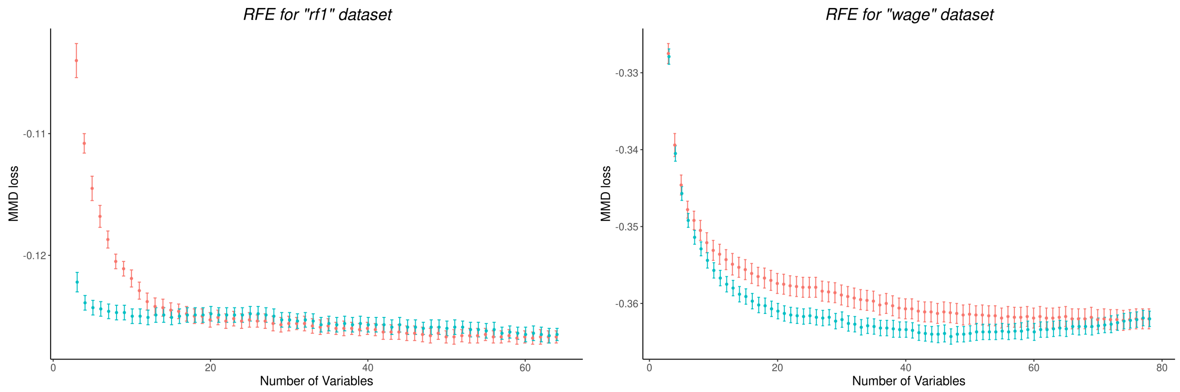

Figure 2: RFE for 'rf1' (left panel) and 'wage' (right panel) datasets, using $\mathrm{I}_n^{(j)}$ (blue) or vimp-drf (red). The RFE procedure is repeated 40 times to compute the standard error of the loss at each step, displayed as error bars.

| | $n$ | $p$ | $d$ | $\mathrm{I}_n^{(j)}$ (std) | vimp-drf (std) |
|---|---|---|---|---|---|
| **enb** | 768 | 8 | 2 | **-3.478** (0.02) | -3.105 (0.02) |
| jura | 359 | 15 | 3 | -0.921 (0.01) | -0.906 (0.01) |
| **wq** | 1060 | 16 | 14 | **-0.164** (0.003) | -0.155 (0.003) |
| air | 2000 | 15 | 6 | -0.146 (0.003) | -0.143 (0.002) |
| births | 2000 | 24 | 4 | -0.328 (0.003) | -0.329 (0.003) |
| **rf1** | 2000 | 64 | 8 | **-7.970** (0.04) | -7.886 (0.04) |
| **scm20d** | 2000 | 61 | 16 | **0.289** (0.001) | 0.292 (0.001) |
| **wage** | 2000 | 78 | 2 | **-28.032** (0.08) | -27.850 (0.1) |
| **oes97** | 334 | 263 | 16 | **1.159** (0.004) | 1.183 (0.004) |

Table 6: Cumulated MMD-loss over RFE steps, using our importance measure $\mathrm{I}_n^{(j)}$ or vimp-drf. Datasets are in bold when the gap between the two losses is higher than the sum of the two standard deviations.

unique score, we sum $\mathcal{L}_n^{(\mathbf{J})}$ over all steps of the RFE. Next, we consider a wide variety of real datasets, inspired from Ćevid et al. (2022), which mainly come from Tsoumakas et al. (2011)—all dataset details are provided in Appendix A. DRF is fit using $N = 500$ trees, the RFE is stopped when only 3 variables remain (forests struggle in very small dimensions), the procedure is repeated 40 times for uncertainties, and each dataset is split in two halves: one to fit the forest and compute the importance measure with out-of-bag predictions, and the other half to estimate the MMD loss.

Results are displayed in Table 6 for the considered datasets. In each case, we provide the cumulative MMD-loss, averaged over the 40 repetitions, along with the standard deviation of this mean value. Table 6 clearly shows a significant improvement of our importance measure over vimp-drf for most datasets. The variable ranking provided by $\mathrm{I}_n^{(j)}$ and vimp-drf can be quite different. Considering the dataset 'enb' with 8 inputs for example, the top three variables given by the two methods do not overlap. Additionally, Figure 2 shows the full path of the MMD-loss at each step of the RFE for several cases. For example, we observe a major improvement in variable selection for 'rf1' and 'wage' datasets. Figure 1 in Appendix A displays results for the 'jura' dataset, an example where the cumulative loss gap is not really significant, according to Table 6. Nevertheless, this figure shows that the loss is flat over all iterations of the RFE, except the last one, which tells us that removing most variables does not hurt DRF estimates, and our method still outperforms vimp-drf at the final step of the RFE. Overall, our proposed importance measure improves variable selection for all datasets, except "Births" and "Air". In these two cases, the MMD-loss is constant over most iterations of the RFE, meaning that variables are removed without hurting the quality of the conditional output distribution estimates. This means that few variables are detected as influential by the DRF, making different competitors of equal efficiency to select variables.

## 5 CONCLUSION

We have introduced a variable importance algorithm for Distributional Random Forest, which generalizes total Sobol indices using the MMD, to quantify the influence of each input variable on a multivariate output distribution. The method enjoys good theoretical properties with provable consistency, and shows high performance on experiments with both simulated and real data, especially for recursive feature elimination. Besides, the extension of this approach to MMD-based Shapley effects seems an interesting research direction for future work, since Shapley effects are strongly valuable to interpret various learning algorithms, and are so far limited to the detection of inputs impacting output means.

## References

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. The Annals of Statistics, 47(2):1148–1178.

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

Bénard, C., Biau, G., Da Veiga, S., and Scornet, E. (2022). Shaff: Fast and consistent shapley effect estimates via random forests. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, Proceedings of The 25th International Conference on Artificial Intelligence and Statistics, volume 151 of Proceedings of Machine Learning Research, pages 5563–5582. PMLR.

Breiman, L. (2001). Random forests. Machine learning, 45(1):5–32.

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

Bénard, C., Da Veiga, S., and Scornet, E. (2022). Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. Biometrika, 109(4):881–900.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society Series B: Statistical Methodology, 80(3):551–577.

Ćevid, D., Michel, L., Näf, J., Meinshausen, N., and Bühlmann, P. (2022). Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. Journal of Machine Learning Research, 23(333):1–79.

Da Veiga, S. (2016). New perspectives for sensitivity analysis. In Proceedings of Mascot-Num 2016 conference, Toulouse, France.

Da Veiga, S. (2021). Kernel-based anova decomposition and shapley effects – application to global sensitivity analysis.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. Pattern Recognition Letters, 31(14):2225–2236.

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

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. (2007). A kernel method for the two-sample-problem. In Advances in neural information processing systems, pages 513–520.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. The Journal of Machine Learning Research, 13(1):723–773.

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

Hooker, G., Mentch, L., and Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Statistics and Computing, 31:1–16.

Hsing, T. and Eubank, R. (2015). Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators. Wiley Series in Probability and Statistics. Wiley.

I Amoukou, S. and Brunel, N. (2022). Consistent sufficient explanations and minimal local rules for explaining the decision of any classifier or regressor. Advances in Neural Information Processing Systems, 35:8027–8040.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523):1094–1111.

Lin, Y. and Jeon, Y. (2006). Random forests and adaptive nearest neighbors. Journal of the American Statistical Association, 101(474):578–590.

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

Mentch, L. and Hooker, G. (2016). Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests. Journal of Machine Learning Research, 17(1):841–881.

Michel, L. and Ćevid, D. (2021). drf: Distributional Random Forests. R package version 1.1.0.

Näf, J., Emmenegger, C., Bühlmann, P., and Meinshausen, N. (2023). Confidence and uncertainty assessment for distributional random forests. Preprint arXiv:2302.05761.

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

Rasmussen, C. E. and Williams, C. K. I. (2006). Gaussian processes for machine learning. Adaptive computation and machine learning. MIT Press.

Scornet, E. (2016). Random forests and kernel methods. IEEE Transactions on Information Theory, 62(3):1485–1500.

Scornet, E. (2023). Trees, forests, and impurity-based variable importance in regression. Annales de l'Institut Henri Poincare (B) Probabilites et statistiques, 59:21–52.

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

Segal, M. and Xiao, Y. (2011). Multivariate random forests. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1):80–87.

Sikdar, S., Hooker, G., Kadiyali, V., et al. (2023). Variable importance measures for variable selection and statistical inference in multivariate random forests. PREPRINT (Version 1) available at Research Square.

Simon-Gabriel, C.-J., Barp, A., and Mackey, L. (2020). Metrizing weak convergence with maximum mean discrepancies. Preprint arXiv:2006.09268.

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

Sriperumbudur, B. (2016). On the optimal estimation of probability measures in weak and strong topologies. Bernoulli, 22(3):1839–1893.

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

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

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. (2011). Mulan: A java library for multi-label learning. The Journal of Machine Learning Research, 12:2411–2414.

Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. Preprint arXiv:1510.04342.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242.

Williamson, B. D., Gilbert, P. B., Simon, N. R., and Carone, M. (2022). A general framework for inference on algorithm-agnostic variable importance. Journal of the American Statistical Association, 0(0):1–14.

## Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Supplementary Material for “MMD-based Variable Importance for Distributional Random Forest”

Appendix A provides additional experiments and details for Section 4 of the main article. Then, Appendix B states the proofs of all theorems and propositions.

## A Experiments

The first subsection is dedicated to additional details for experiments with simulated data, while the second subsection focuses on RFE for real datasets. In particular, we provide figures of the RFE path for the nine datasets, along with data sources.

### A.1 Simulated data

For simulated data, all experiment settings are given in the main article. For the univariate case, we provide an extended version of Table 2 with standard deviations of the mean importance values in small font. Notice that these uncertainties are small, and are thus omitted in the paper.

### A.2 RFE for real datasets

We first recall Table 5 of the main article in Table 8 below. The choice of these datasets was driven by the original DRF paper (Ćevid et al., 2022), originally inspired by Tsoumakas et al. (2011), which provides data with multivariate outputs. The data was found in the following repository : `https://github.com/tsoumakas/mulan`, accessed in August 2023. For the details and sources of “Births”, “Air quality”, and “wage” datasets, see Appendix C in Ćevid et al. (2022). Notice that our “Births” data corresponds to “birth2” in Ćevid et al. (2022). When the sample size exceeds $n = 2000$, a random subsampling of 2000 observations is done at each repetition of the RFE, to keep a reasonable computational cost. The RFE experiments were conducted on a cluster with the following characteristics: 27 computational nodes Ice Lake with 48 cores (CPU: Intel Xeon Gold 6342, 2x24 cores, 2.80 Ghz. Memory : 263 GB (5.4 GB/cores)).

Then, Figure 3 provides the RFE paths for “enb” and “jura” datasets, Figure 4 for “wq” and “scm20d” datasets, Figure 5 for “Births” and “oes97” datasets, and Figure 6 for “Air quality” dataset. Overall, our proposed importance measure improves variable selection for all datasets, except “Births” and “Air”. In these two cases, the MMD-loss is constant over most iterations of the RFE, meaning that variables are removed without hurting the

| $X^{(j)}$ | $\mathrm{I}_n^{(j)}$ | $X^{(j)}$ | v-drf | $X^{(j)}$ | MDA | $X^{(j)}$ | S-MDA |
|---|---|---|---|---|---|---|---|
| $\boldsymbol{X}^{(\mathbf{1})}$ | $0.181\ {\scriptstyle 0.009}$ | $\boldsymbol{X}^{(\mathbf{1})}$ | $0.649\ {\scriptstyle 0.01}$ | $\boldsymbol{X}^{(\mathbf{1})}$ | $6.47\ {\scriptstyle 0.6}$ | $\boldsymbol{X}^{(\mathbf{1})}$ | $0.014\ {\scriptstyle 0.003}$ |
| $\boldsymbol{X}^{(\mathbf{4})}$ | $0.073\ {\scriptstyle 0.007}$ | $X^{(10)}$ | $0.096\ {\scriptstyle 0.01}$ | $X^{(10)}$ | $2.56\ {\scriptstyle 0.5}$ | $\boldsymbol{X}^{(\mathbf{2})}$ | $0.012\ {\scriptstyle 0.003}$ |
| $\boldsymbol{X}^{(\mathbf{5})}$ | $0.073\ {\scriptstyle 0.008}$ | $\boldsymbol{X}^{(\mathbf{2})}$ | $0.062\ {\scriptstyle 0.004}$ | $\boldsymbol{X}^{(\mathbf{2})}$ | $1.33\ {\scriptstyle 0.4}$ | $X^{(8)}$ | $-0.001\ {\scriptstyle 0.001}$ |
| $\boldsymbol{X}^{(\mathbf{2})}$ | $0.072\ {\scriptstyle 0.004}$ | $\boldsymbol{X}^{(\mathbf{5})}$ | $0.059\ {\scriptstyle 0.007}$ | $X^{(6)}$ | $0.25\ {\scriptstyle 0.1}$ | $X^{(7)}$ | $-0.003\ {\scriptstyle 0.001}$ |
| $\boldsymbol{X}^{(\mathbf{3})}$ | $0.065\ {\scriptstyle 0.004}$ | $\boldsymbol{X}^{(\mathbf{4})}$ | $0.056\ {\scriptstyle 0.005}$ | $X^{(8)}$ | $0.24\ {\scriptstyle 0.2}$ | $X^{(9)}$ | $-0.003\ {\scriptstyle 0.0006}$ |
| $X^{(10)}$ | $0.010\ {\scriptstyle 0.001}$ | $\boldsymbol{X}^{(\mathbf{3})}$ | $0.050\ {\scriptstyle 0.003}$ | $X^{(7)}$ | $0.18\ {\scriptstyle 0.2}$ | $X^{(6)}$ | $-0.003\ {\scriptstyle 0.0007}$ |
| $X^{(7)}$ | $0.005\ {\scriptstyle 0.0004}$ | $X^{(6)}$ | $0.007\ {\scriptstyle 0.0007}$ | $X^{(9)}$ | $0.18\ {\scriptstyle 0.1}$ | $\boldsymbol{X}^{(\mathbf{5})}$ | $-0.003\ {\scriptstyle 0.002}$ |
| $X^{(6)}$ | $0.005\ {\scriptstyle 0.0005}$ | $X^{(9)}$ | $0.007\ {\scriptstyle 0.0005}$ | $\boldsymbol{X}^{(\mathbf{3})}$ | $-0.52\ {\scriptstyle 0.2}$ | $\boldsymbol{X}^{(\mathbf{3})}$ | $-0.004\ {\scriptstyle 0.001}$ |
| $X^{(9)}$ | $0.005\ {\scriptstyle 0.0003}$ | $X^{(7)}$ | $0.007\ {\scriptstyle 0.0007}$ | $\boldsymbol{X}^{(\mathbf{5})}$ | $-0.62\ {\scriptstyle 0.1}$ | $\boldsymbol{X}^{(\mathbf{4})}$ | $-0.004\ {\scriptstyle 0.002}$ |
| $X^{(8)}$ | $0.005\ {\scriptstyle 0.0002}$ | $X^{(8)}$ | $0.007\ {\scriptstyle 0.0005}$ | $\boldsymbol{X}^{(\mathbf{4})}$ | $-0.90\ {\scriptstyle 0.1}$ | $X^{(10)}$ | $-0.006\ {\scriptstyle 0.001}$ |

Table 7: Variable importance for the univariate output experiment for $\mathrm{I}_n^{(j)}$, vimp-drf (v-drf), MDA, and Sobol-MDA (S-MDA), with standard deviations of mean importance values in small font.

|  | $n$ | $p$ | $d$ | $\mathrm{I}_n^{(j)}$ | std | vimp-drf | std |
|---|---|---|---|---|---|---|---|
| **enb** | 768 | 8 | 2 | **-3.478** | 0.020 | -3.105 | 0.017 |
| jura | 359 | 15 | 3 | -0.921 | 0.014 | -0.906 | 0.014 |
| **wq** | 1060 | 16 | 14 | **-0.164** | 0.003 | -0.155 | 0.003 |
| air | 2000 | 15 | 6 | -0.146 | 0.003 | -0.143 | 0.002 |
| births | 2000 | 24 | 4 | -0.328 | 0.003 | -0.329 | 0.003 |
| **rf1** | 2000 | 64 | 8 | **-7.970** | 0.037 | -7.886 | 0.035 |
| **scm20d** | 2000 | 61 | 16 | **0.289** | 0.001 | 0.292 | 0.001 |
| **wage** | 2000 | 78 | 2 | **-28.032** | 0.077 | -27.850 | 0.098 |
| **oes97** | 334 | 263 | 16 | **1.159** | 0.004 | 1.183 | 0.004 |

Table 8: Cumulated MMD-loss over RFE steps, using our importance measure $\mathrm{I}_n^{(j)}$ or vimp-drf. Datasets are in bold when the gap between the two losses is higher than the sum of the two standard deviations.

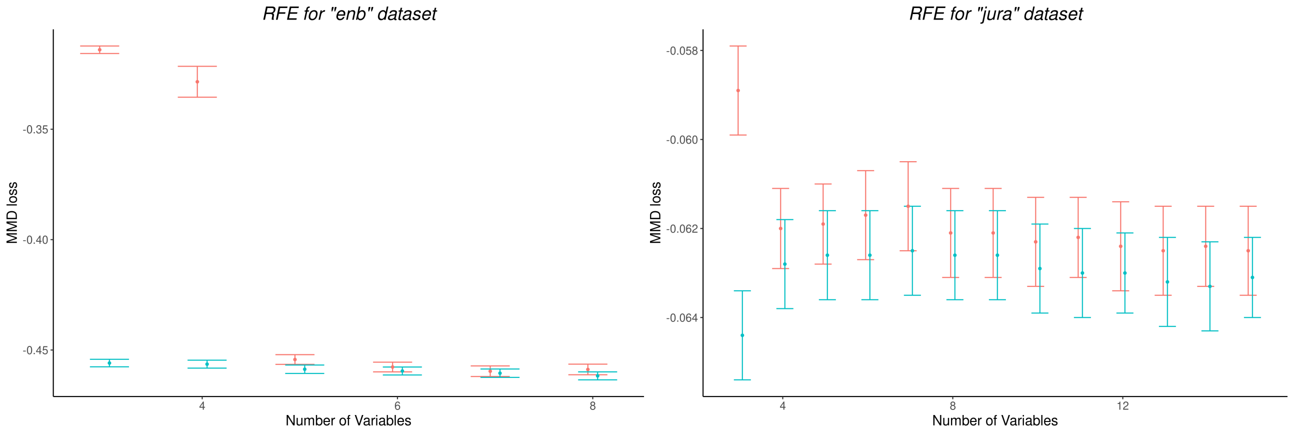

Figure 3: RFE for 'enb' (left panel) and 'jura' (right panel) datasets, using our importance measure $\mathrm{I}_n^{(j)}$ (blue) or vimp-drf (red).

quality of the conditional output distribution estimates. This means that few variables are detected as influential by the DRF, making different competitors of equal efficiency to select variables.

## B Proofs

We first recall the main equations of the article.

$$\mathrm{I}^{(j)} \overset{\text{def}}{=} \frac{\mathbb{E}[\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X}) \mid \mathbf{X}^{(-j)}]]}{\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X})]}, \tag{8}$$

$$\mathrm{I}_n^{(j)} = \frac{\sum_{i=1}^n \|\mu_{N,n}(\mathbf{X}_i') - \mu_{N,n}(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2}{\sum_{i=1}^n \|\mu_{N,n}(\mathbf{X}_i') - \overline{\mu_{N,n}}\|_{\mathcal{H}}^2} - \mathrm{I}_n^{(0)}, \tag{9}$$

$$\mu_n(\mathbf{x}) = \binom{n}{s_n}^{-1} \sum_{i_1 < \cdots < i_{s_n}} \mathbb{E}\left[T_n(\mathbf{x}; \varepsilon, \{\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}\}) \mid \mathcal{Z}_n\right]. \tag{10}$$

We put the following assumptions on the forest construction.

(**F1**) (*Honesty*) The data used for constructing each tree is split into two halves; the first is used for determining the splits and the second for populating the leaves and thus for estimating the response. The covariates in the second sample may be used for the splits, to enforce the subsequent assumptions, but not the response.

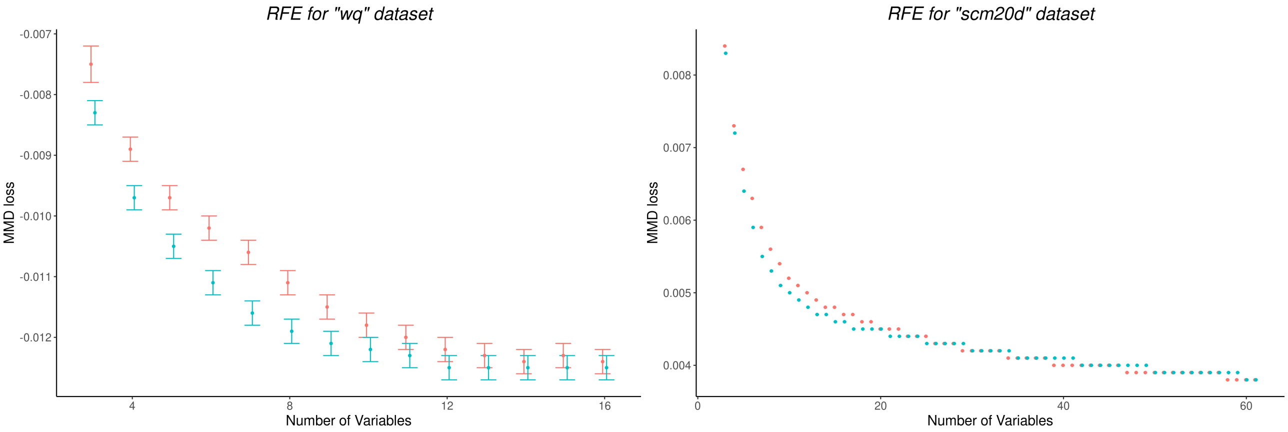

Figure 4: RFE for 'wq' (left panel) and 'scm20d' (right panel) datasets, using our importance measure $\mathrm{I}_n^{(j)}$ (blue) or vimp-drf (red).

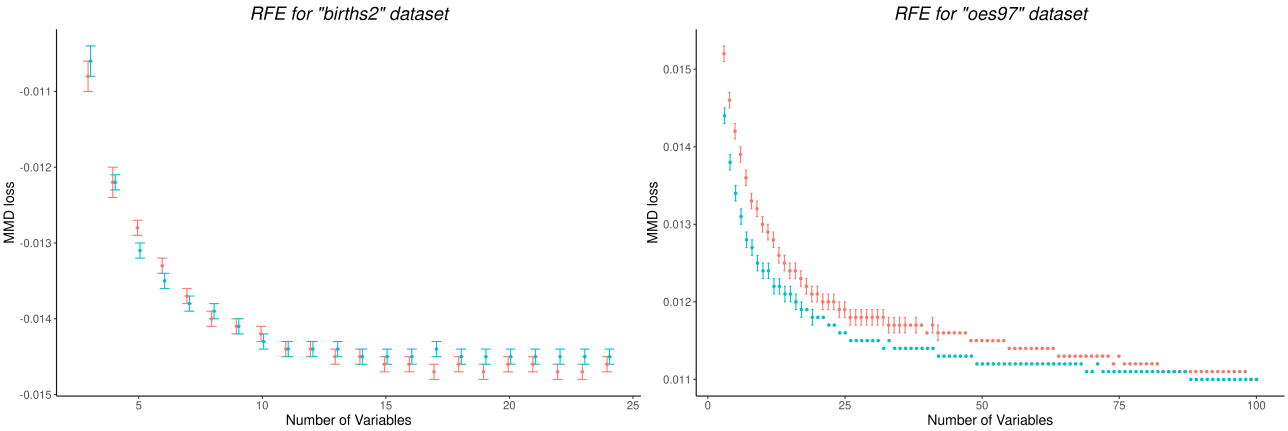

Figure 5: RFE for 'Births' (left panel) and 'oes97' (right panel) datasets, using our importance measure $\mathrm{I}_n^{(j)}$ (blue) or vimp-drf (red).

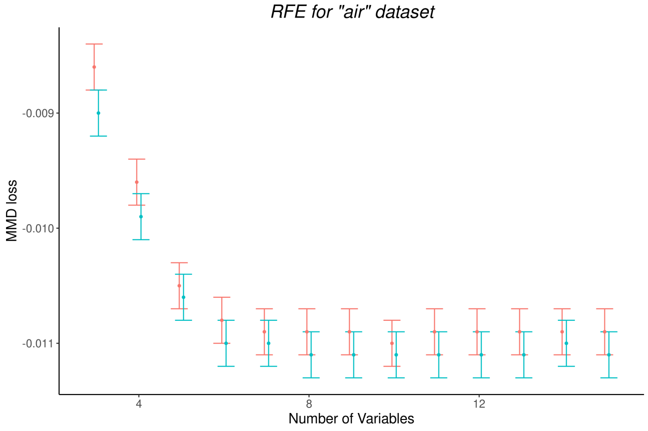

Figure 6: RFE for 'Air' dataset, using our importance measure $\mathrm{I}_n^{(j)}$ (blue) or vimp-drf (red).

**(F2)** (*Random-split*) At every split point and for all feature dimensions $j = 1, \dots, p$, the probability that the split occurs along the feature $X_j$ is bounded from below by $\pi/p$ for some $\pi > 0$.

**(F3)** (*Symmetry*) The (randomized) output of a tree does not depend on the ordering of the training samples.

**(F4)** (*$\alpha$-regularity*) After splitting a parent node, each child node contains at least a fraction $\alpha \leq 0.2$ of the parent's training samples. Moreover, the trees are grown until every leaf contains between $\kappa$ and $2\kappa - 1$ many observations for some fixed tuning parameter $\kappa \in \mathbb{N}$.

**(F5)** (*Data sampling*) To grow a tree, a subsample of size $s_n$ out of the $n$ training data points is sampled. We consider $s_n = n^{\beta}$ with $0 < \beta < 1$.

We also recall the assumptions of the article.

**(K1)** The kernel $k$ is bounded, and the function $(\mathbf{x}, \mathbf{y}) \mapsto k(\mathbf{x}, \mathbf{y})$ is (jointly) continuous.

**(K2)** The kernel $k$ is characteristic.

**(D1)** The observations $\mathbf{X}_1, \dots, \mathbf{X}_n$ are independent and identically distributed on $[0,1]^p$, with a density bounded from below and above by strictly positive constants.

**(D2)** The mapping $\mathbf{x} \mapsto \mu(\mathbf{x}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X} = \mathbf{x}] \in \mathcal{H}$ is Lipschitz.

**(D3)** The mapping $\mathbf{x}^{(-j)} \mapsto \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}] \in \mathcal{H}$ is Lipschitz.

For a sequence of random variables $X_n \colon \Omega \to \mathbb{R}$ and $a_n \in (0, +\infty)$, $n \in \mathbb{N}$, we write $X_n = \mathcal{O}_p(a_n)$ if

$$\lim_{M \to \infty} \sup_n \mathbb{P}(a_n^{-1} |X_n| > M) = 0.$$

**Proposition 1.** *If* $\mathrm{I}^{(j)}$ *is the generalized total Sobol index defined by Equation (3), then we have*

$$\mathrm{I}^{(j)} = \frac{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}})]}{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}})]},$$
$$\mathrm{I}^{(j)} = 1 - \frac{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}})]}{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}})]}.$$

*Proof.* By definition of our importance measure given in Equation (8), we have

$$\mathrm{I}^{(j)} \overset{\text{def}}{=} \frac{\mathbb{E}[\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X}) \mid \mathbf{X}^{(-j)}]]}{\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X})]}.$$

We first consider the numerator, and write

$$\begin{aligned}\mathbb{E}[\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X}) \mid \mathbf{X}^{(-j)}]] &= \mathbb{E}[\|\mu(\mathbf{X}) - \mathbb{E}[\mu(\mathbf{X}) \mid \mathbf{X}^{(-j)}]\|_{\mathcal{H}}^2] = \mathbb{E}[\|\mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X}] - \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X}^{(-j)}]\|_{\mathcal{H}}^2] \\ &= \mathbb{E}[\|\Phi(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}}) - \Phi(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}})\|_{\mathcal{H}}^2] = \mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}})].\end{aligned}$$

For the denominator, we have

$$\begin{aligned}\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X})] &= \mathbb{E}[\|\mu(\mathbf{X}) - \mathbb{E}[\mu(\mathbf{X})]\|_{\mathcal{H}}^2] = \mathbb{E}[\|\mu(\mathbf{X})\|_{\mathcal{H}}^2] - \|\mathbb{E}[\mu(\mathbf{X})]\|_{\mathcal{H}}^2 \\ &= \mathbb{E}[\|\mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X}]\|_{\mathcal{H}}^2] - \|\mathbb{E}[k(\mathbf{Y}, \cdot)]\|_{\mathcal{H}}^2.\end{aligned}$$

Using the RKHS properties of $\mathcal{H}$, and with $\mathbf{Y}'$ an independent copy of $\mathbf{Y}$, we can write

$$\|\mathbb{E}[k(\mathbf{Y}, \cdot)]\|_{\mathcal{H}}^2 = \langle \mathbb{E}[k(\mathbf{Y}, \cdot)], \mathbb{E}[k(\mathbf{Y}', \cdot)] \rangle_{\mathcal{H}} = \mathbb{E}[\langle k(\mathbf{Y}, \cdot), k(\mathbf{Y}', \cdot) \rangle_{\mathcal{H}}] = \mathbb{E}[k(\mathbf{Y}, \mathbf{Y}')].$$

For the other term, we introduce $\tilde{\mathbf{Y}}$ distributed as $\mathbf{Y}$, and independent and identically distributed as $\mathbf{Y}$ conditional on $\mathbf{X}$, then

$$\|\mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X}]\|_{\mathcal{H}}^2 = \langle \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X}], \mathbb{E}[k(\tilde{\mathbf{Y}}, \cdot) \mid \mathbf{X}] \rangle_{\mathcal{H}} = \mathbb{E}[\langle k(\mathbf{Y}, \cdot), k(\tilde{\mathbf{Y}}, \cdot) \rangle_{\mathcal{H}} \mid \mathbf{X}] = \mathbb{E}[k(\mathbf{Y}, \tilde{\mathbf{Y}}) \mid \mathbf{X}].$$

Overall, $\mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X})] = \mathbb{E}[k(\mathbf{Y}, \tilde{\mathbf{Y}})] - \mathbb{E}[k(\mathbf{Y}, \mathbf{Y}')]$. Next, by definition of the MMD distance, we have

$$\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}})] = \mathbb{E}[k(\mathbf{Y}, \mathbf{Y}')] + \mathbb{E}[k(\mathbf{Y}, \tilde{\mathbf{Y}})] - 2\mathbb{E}[k(\mathbf{Y}', \tilde{\mathbf{Y}})],$$

and since $\mathbf{Y}'$ and $\mathbf{X}$ are independent, $\mathbb{E}[k(\mathbf{Y}', \tilde{\mathbf{Y}})] = \mathbb{E}[k(\mathbf{Y}', \mathbf{Y})]$, and we get

$$\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}})] = \mathbb{E}[k(\mathbf{Y}, \tilde{\mathbf{Y}})] - \mathbb{E}[k(\mathbf{Y}, \mathbf{Y}')] = \mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X})].$$

Finally, we obtain

$$\mathrm{I}^{(j)} = \frac{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}})]}{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}})]}.$$

Similarly, it holds that

$$\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}})] = \mathbb{E}[k(\mathbf{Y}, \tilde{\mathbf{Y}})] - \mathbb{E}[k(\mathbf{Y}, \tilde{\mathbf{Y}}^{(-j)})],$$

where $\tilde{\mathbf{Y}}^{(-j)}$ is distributed as $\mathbf{Y}$, and independent and identically distributed as $\mathbf{Y}$ conditional on $\mathbf{X}^{(-j)}$. Consequently, we get

$$\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}})] - \mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}})] = \mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}})],$$

and we also obtain

$$\mathrm{I}^{(j)} = 1 - \frac{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}})]}{\mathbb{E}[\mathrm{MMD}^2(\mathbb{P}_{\mathbf{Y}}, \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}})]}.$$

□

**Proposition 2.** *Assume that Assumptions (**K1**)-(**K2**) holds and that for each* $\mathbf{x}$ *in a set with nonzero probability,* $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} \neq \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}=\mathbf{x}^{(-j)}}$*, then* $0 < \mathrm{I}^{(j)} \leq 1$*, and otherwise* $\mathrm{I}^{(j)} = 0$*.*

*Proof.* The condition that $\mathbb{P}_{\mathbf{Y} \mid \mathbf{X}=\mathbf{x}} \neq \mathbb{P}_{\mathbf{Y} \mid \mathbf{X}^{(-j)}=\mathbf{x}^{(-j)}}$ over a set of non-null probability, combined with Assumption (**K2**), implies that $\|\mu(\mathbf{x}) - \mu(\mathbf{x}^{(-j)})\|_{\mathcal{H}} > 0$, for $\mathbf{x}$ in a set with nonzero probability. This in turn implies $\mathbb{E}[\|\mu(\mathbf{X}) - \mu(\mathbf{X}^{(-j)})\|_{\mathcal{H}}^2] > 0$, and thus $\mathrm{I}^{(j)} > 0$. Additionally, $\mathrm{I}^{(j)} \leq 1$ is a direct consequence of Proposition 1, since MMD takes non-negative values. □

**Proposition 3.** *Assume that the forest construction satisfies the properties **(F1)-(F5)**. Additionally, assume that* $k$ *meets Assumption (**K1**), and that (**D1**) and (**D2**) hold. Then, we have consistency of* $\mu_n(\mathbf{x})$ *in* (6) *with respect to the RKHS norm in mean, that is*

$$\mathbb{E}[\|\mu_n(\mathbf{X}) - \mu(\mathbf{X})\|_{\mathcal{H}}] \longrightarrow 0.$$

*Proof.* Ćevid et al. (2022, Theorem 1) implies that

$$\|\mu_n(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} \xrightarrow{p} 0. \tag{11}$$

As a first consequence, we have for a random test point $\mathbf{X}$

$$\|\mu_n(\mathbf{X}) - \mu(\mathbf{X})\|_{\mathcal{H}} \xrightarrow{p} 0, \tag{12}$$

as

$$\mathbb{P}(\|\mu_n(\mathbf{X}) - \mu(\mathbf{X})\|_{\mathcal{H}} > \varepsilon) \leq \mathbb{E}[\mathbb{P}(\|\mu_n(\mathbf{X}) - \mu(\mathbf{X})\|_{\mathcal{H}} > \varepsilon \mid \mathbf{X})] \to 0,$$

by boundedness of $\mathbb{P}(\|\mu_n(\mathbf{x}) - \mu(\mathbf{x})\|_{\mathcal{H}} > \varepsilon \mid \mathbf{X} = \mathbf{x})$. Crucially, since $\sup_{\mathbf{y}_1, \mathbf{y}_2} k(\mathbf{y}_1, \mathbf{y}_2) \leq C$ by assumption **(K1)**, $\|\mu_n(\mathbf{X}) - \mu(\mathbf{X})\|_{\mathcal{H}}$ is bounded as well. Indeed, $\mu_n(\mathbf{X})$ in (10) can also be written as a convex combination of $k(\mathbf{Y}_i, \cdot)$,

$$\mu_n(\mathbf{X}) = \sum_{i=1}^{n} w_i(\mathbf{X}) k(\mathbf{Y}_i, \cdot).$$

Thus, for all $x \in [0,1]$, $n \in \mathbb{N}^\star$,

$$\begin{aligned}\|\mu_n(\mathbf{X}) - \mu(\mathbf{X})\|_{\mathcal{H}} &= \|\sum_{i=1}^{n} w_i(\mathbf{X})(k(\mathbf{Y}_i,\cdot) - \mathbb{E}[k(\mathbf{Y},\cdot) \mid \mathbf{X}])\|_{\mathcal{H}} \\ &\leq \max_i \|k(\mathbf{Y}_i,\cdot) - \mathbb{E}[k(\mathbf{Y},\cdot) \mid \mathbf{X}]\|_{\mathcal{H}} \\ &\leq \max_i \|k(\mathbf{Y}_i,\cdot)\|_{\mathcal{H}} + \mathbb{E}[\|k(\mathbf{Y},\cdot)\|_{\mathcal{H}} \mid \mathbf{X}] \\ &\leq 2\sqrt{C},\end{aligned}$$

since $\|k(\mathbf{Y}_i,\cdot)\|_{\mathcal{H}}^2 = k(\mathbf{Y}_i,\mathbf{Y}_i) \leq C$.

As for a bounded random variable convergence in probability implies convergence in expectations, the result follows. □

**Proposition 4.** *Under the assumptions of Proposition 3, and provided that Assumption **(D3)** is satisfied, we have* $\mathbb{E}[\|\mu_n(\mathbf{X}^{(-j)}) - \mu(\mathbf{X}^{(-j)})\|_{\mathcal{H}}] \longrightarrow 0$.

*Proof.* Direct application of Proposition 3. □

**Theorem 1.** *Assume that the forest construction satisfies the properties **(F1)-(F5)**. Additionally, assume that k meets Assumption **(K1)**, and that **(D1)-(D3)** hold. Then, we have consistency of* $\mathrm{I}_n^{(j)}$ *in (4), that is*

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

*Proof.* We recall that for the number of trees $N \to \infty$, $\mu_n$ is defined in (10), and

$$I_n^{(j)} = \frac{\sum_{i=1}^{n} \|\mu_n(\mathbf{X}_i') - \mu_n(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2}{\sum_{i=1}^{n} \|\mu_n(\mathbf{X}_i') - \frac{1}{n}\sum_{i'=1}^{n} \mu_n(\mathbf{X}_{i'}')\|_{\mathcal{H}}^2}$$

We first study the upper part and show

Claim:

$$\mathbb{E}[|\|\mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}^2 - \|\mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}^2|] \to 0. \tag{13}$$

Proof: First,

$$\begin{aligned}&\mathbb{E}[|\|\mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}^2 - \|\mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}^2|] = \\ &\mathbb{E}[|\langle \mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)}), \mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)})\rangle_{\mathcal{H}} - \|\mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}^2|] = \\ &\mathbb{E}[|\langle \mu_n(\mathbf{X}_1') - \mu(\mathbf{X}_1') + \mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)}) + \mu(\mathbf{X}_1'^{(-j)}) - \mu_n(\mathbf{X}_1'^{(-j)}), \mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)})\rangle_{\mathcal{H}} - \\ &\|\mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}^2|] \leq \mathbb{E}[|\langle \mu_n(\mathbf{X}_1') - \mu(\mathbf{X}_1'), \mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)})\rangle_{\mathcal{H}}|] + \\ &\mathbb{E}[|\langle \mu(\mathbf{X}_1'^{(-j)}) - \mu_n(\mathbf{X}_1'^{(-j)}), \mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)})\rangle_{\mathcal{H}}|] + \\ &\mathbb{E}[|\langle \mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)}), \mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)})\rangle_{\mathcal{H}} - \|\mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}^2|]\end{aligned}$$

Now using that all norms of estimates are uniformly bounded by $K < \infty$ and the Cauchy-Schwarz inequality,

$$\begin{aligned}\mathbb{E}[|\langle \mu_n(\mathbf{X}_1') - \mu(\mathbf{X}_1'), \mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)})\rangle_{\mathcal{H}}|] &\leq \mathbb{E}[\|\mu_n(\mathbf{X}_1') - \mu(\mathbf{X}_1')\|_{\mathcal{H}}]K \to 0 \\ \mathbb{E}[|\langle \mu(\mathbf{X}_1'^{(-j)}) - \mu_n(\mathbf{X}_1'^{(-j)}), \mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)})\rangle_{\mathcal{H}}|] &\leq \mathbb{E}[\|\mu(\mathbf{X}_1'^{(-j)}) - \mu_n(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}]K \to 0,\end{aligned}$$

by Propositions 3 and 4. Moreover,

$$\begin{aligned}&\mathbb{E}[|\langle \mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)}), \mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)})\rangle_{\mathcal{H}} - \|\mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}^2|] = \\ &\mathbb{E}[|\langle \mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)}), \mu_n(\mathbf{X}_1') - \mu_n(\mathbf{X}_1'^{(-j)}) - (\mu(\mathbf{X}_1') - \mu(\mathbf{X}_1'^{(-j)}))\rangle_{\mathcal{H}}|],\end{aligned}$$

since again from Propositions 3 and 4,

$$\begin{aligned}&\mathbb{E}[|\langle \mu(\mathbf{X}'_1) - \mu(\mathbf{X}_1'^{(-j)}), \mu_n(\mathbf{X}'_1) - \mu_n(\mathbf{X}_1'^{(-j)}) - (\mu(\mathbf{X}'_1) - \mu(\mathbf{X}_1'^{(-j)}))\rangle_{\mathcal{H}}|] \leq \\ &K\mathbb{E}[\|\mu_n(\mathbf{X}'_1) - \mu_n(\mathbf{X}_1'^{(-j)}) - \mu(\mathbf{X}'_1) + \mu(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}] \leq \\ &K\left(\mathbb{E}[\|\mu_n(\mathbf{X}'_1) - \mu(\mathbf{X}'_1)\|_{\mathcal{H}}] + \mathbb{E}[\|\mu(\mathbf{X}_1'^{(-j)}) - \mu_n(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}]\right) \to 0.\end{aligned}$$

Thus, we have that (13) holds.

□

Given (13), we can now show that

Claim:

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\|\mu_n(\mathbf{X}'_i) - \mu_n(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2 - \mathbb{E}[\|\mu(\mathbf{X}_1) - \mu(\mathbf{X}_1^{(-j)})\|_{\mathcal{H}}^2]\right|\right] \to 0. \tag{14}$$

Proof: First,

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\|\mu_n(\mathbf{X}'_i) - \mu_n(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2 - \mathbb{E}[\|\mu(\mathbf{X}_1) - \mu(\mathbf{X}_1^{(-j)})\|_{\mathcal{H}}^2]\right|\right] \leq \\ &\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\|\mu_n(\mathbf{X}'_i) - \mu_n(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2 - \frac{1}{n}\sum_{i=1}^{n}\|\mu(\mathbf{X}'_i) - \mu(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2\right|\right] + \\ &\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\|\mu(\mathbf{X}'_i) - \mu(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2 - \mathbb{E}[\|\mu(\mathbf{X}_1) - \mu(\mathbf{X}_1^{(-j)})\|_{\mathcal{H}}^2]\right|\right]\end{aligned}$$

For the first term, by the triangle inequality,

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\|\mu_n(\mathbf{X}'_i) - \mu_n(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2 - \frac{1}{n}\sum_{i=1}^{n}\|\mu(\mathbf{X}'_i) - \mu(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2\right|\right] \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\left|\|\mu_n(\mathbf{X}'_i) - \mu_n(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2 - \|\mu(\mathbf{X}'_i) - \mu(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2\right|\right] \\ &= \mathbb{E}\left[\left|\|\mu_n(\mathbf{X}'_1) - \mu_n(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}^2 - \|\mu(\mathbf{X}'_1) - \mu(\mathbf{X}_1'^{(-j)})\|_{\mathcal{H}}^2\right|\right],\end{aligned}$$

the latter goes to zero due to (13).

For the second term,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\|\mu(\mathbf{X}'_i) - \mu(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2 - \mathbb{E}[\|\mu(\mathbf{X}_1) - \mu(\mathbf{X}_1^{(-j)})\|_{\mathcal{H}}^2]\right| \xrightarrow{p} 0,$$

by the law of large numbers. Since the sequence is again uniformly bounded, the same arguments as in Proposition (3), also imply

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\|\mu(\mathbf{X}'_i) - \mu(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2 - \mathbb{E}[\|\mu(\mathbf{X}_1) - \mu(\mathbf{X}_1^{(-j)})\|_{\mathcal{H}}^2]\right|\right] \xrightarrow{p} 0,$$

proving the claim. □

In turn, (14) implies weak convergence of the upper part of $I_n^{(-j)}$, that is,

$$\frac{1}{n}\sum_{i=1}^{n}\|\mu_n(\mathbf{X}'_i) - \mu_n(\mathbf{X}_i'^{(-j)})\|_{\mathcal{H}}^2 \xrightarrow{p} \mathbb{E}[\|\mu(\mathbf{X}_1) - \mu(\mathbf{X}_1^{(-j)})\|_{\mathcal{H}}^2].$$

For the lower part, we first proof

Claim:

$$\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\mu_n(\mathbf{X}_i') - \mathbb{E}[\mu(\mathbf{X}_1')]\|_{\mathcal{H}}] \to 0. \tag{15}$$

Proof: To show (15), note that

$$\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\mu_n(\mathbf{X}_i') - \mathbb{E}[\mu(\mathbf{X}_1')]\|_{\mathcal{H}}] \leq$$
$$\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\mu_n(\mathbf{X}_i') - \frac{1}{n}\sum_{i=1}^{n}\mu(\mathbf{X}_i')\|_{\mathcal{H}}] + \mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\mu(\mathbf{X}_i') - \mathbb{E}[\mu(\mathbf{X}_1')]\|_{\mathcal{H}}]$$

For the first term,

$$\begin{aligned}\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\mu_n(\mathbf{X}_i') - \frac{1}{n}\sum_{i=1}^{n}\mu(\mathbf{X}_i')\|_{\mathcal{H}}] &\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\mu_n(\mathbf{X}_i') - \mu(\mathbf{X}_i')\|_{\mathcal{H}}] \\ &= \mathbb{E}[\|\mu_n(\mathbf{X}_1) - \mu(\mathbf{X}_1)\|_{\mathcal{H}}] \to 0.\end{aligned}$$

Moreover, it follows by the Law of Large numbers on $\mathcal{H}$ (see e.g., Hsing and Eubank (2015, Chapter 7)), that

$$\|\frac{1}{n}\sum_{i=1}^{n}\mu(\mathbf{X}_i') - \mathbb{E}[\mu(\mathbf{X}_1')]\|_{\mathcal{H}} \xrightarrow{p} 0. \tag{16}$$

As $\|\frac{1}{n}\sum_{i=1}^{n}\mu(\mathbf{X}_i') - \mathbb{E}[\mu(\mathbf{X}_1')]\|_{\mathcal{H}}$ is uniformly bounded, we have

$$\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\mu(\mathbf{X}_i') - \mathbb{E}[\mu(\mathbf{X}_1')]\|_{\mathcal{H}}] \to 0.$$

□

Thus, $\frac{1}{n}\sum_{i=1}^{n}\mu_n(\mathbf{X}_i')$ is a mean-consistent estimate of $\mathbb{E}[\mu(\mathbf{X}_1')]$ and the same argument to prove (14) applied again shows that

$$\frac{1}{n}\sum_{i=1}^{n}\|\mu_n(\mathbf{X}_i') - \frac{1}{n}\sum_{i'=1}^{n}\mu_n(\mathbf{X}_{i'}')\|_{\mathcal{H}}^2 \xrightarrow{p} \mathbb{E}[\|\mu(\mathbf{X}_1) - \mathbb{E}[\mu(\mathbf{X}_1)]\|_{\mathcal{H}}^2] = \mathbb{V}_{\mathcal{H}}[\mu(\mathbf{X}_1)].$$

□

We now consider the consistency of the projected DRF,

$$\mu_n^{(-j)}(\mathbf{x}) = \binom{n}{s_n}^{-1}\sum_{i_1<i_2<\ldots<i_{s_n}}\mathbb{E}_{\varepsilon}\left[T^{(-j)}(\mathbf{x}^{(-j)}, \varepsilon; \mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}})\right], \tag{17}$$

where the sum is taken over all $\binom{n}{s_n}$ possible subsamples $\mathbf{Z}_{i_1}, \ldots, \mathbf{Z}_{i_{s_n}}$ of $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ and $s_n \to \infty$ with $n$ and where

$$T^{(-j)}(\mathbf{x}^{(-j)}, \varepsilon; \mathbf{Z}_1, \ldots, \mathbf{Z}_{s_n}) = \sum_{i=1}^{s_n}\frac{\mathbb{1}(\mathbf{X}_i^{(-j)} \in \mathcal{L}^{(-j)}(\mathbf{x}^{(-j)}))}{|\mathcal{L}^{(-j)}(\mathbf{x}^{(-j)})|}k(\mathbf{Y}_i, \cdot).$$

For simplicity we write here the sum from $j = 1, \ldots, s_n$, though it should be understood that $\mathbb{1}(\mathbf{X}_i^{(-j)} \in \mathcal{L}^{(-j)}(\mathbf{x}^{(-j)})) = 0$ for $i$ that are used for tree building and not to populate the leaves, according to **(F1)**.

It should be noted that (17) is not the same as fitting a forest on the data $((\mathbf{Y}_1, \mathbf{X}_1^{(-j)}), \ldots, (\mathbf{Y}_n, \mathbf{X}_n^{(-j)}))$, as growing a tree includes $X_j$ implicitly, while populating the leaves or predicting does not. Nonetheless, key assumptions about the estimator translate from $\mu_n(\mathbf{x})$ to $\mu_n^{(-j)}(\mathbf{x})$:

(**F1**') (*Honesty*) The data used for constructing $T^{(-j)}$ is split into two halves; the first is used for determining the splits and the second for populating the leaves and thus for estimating the response. The covariates in the second sample may be used for the splits, to enforce the subsequent assumptions, but not the response.

(**F2**') (*Random-split*) At every split point and for all feature dimensions $l \in \{1, \dots, p\} \setminus j$, the probability that the split occurs along the feature $X_l$ is bounded from below by $\pi/p$ for some $\pi > 0$.

(**F3**') (*Symmetry*) The (randomized) output of $T^{(-j)}$ does not depend on the ordering of the training samples.

(**F4**') (*Data sampling*) To grow $T^{(-j)}$, a subsample of size $s_n$ out of the $n$ training data points is sampled. We consider $s_n = n^{\beta}$ with $0 < \beta < 1$.

**Lemma 1.** ***(F1)**, **(F2)**, **(F3)**, **(F5)** for $\mu_n(\mathbf{x})$ imply respectively **(F1')**, **(F2')**, **(F3')** and **(F4')** for $\mu_n^{(-j)}(\mathbf{x})$.*

*Proof.* (**F1**')–(**F4**') are simply restatements of **(F1)**–**(F3)**, **(F5)**, with tree replaced by $T^{(-j)}$. As each is not impacted by the projection, they continue holding for the projected DRF. □

In particular, the conditional independence statements derived from honesty **(F1)**, crucial for the proofs in Ćevid et al. (2022), remain the same. Moreover, $T^{(-j)}$ is still a weighted mean involving $k(\mathbf{y}_i, \cdot)$. As such most results follow in exactly the same way as in Ćevid et al. (2022) and are thus mostly stated for completeness. Throughout we assume that expectations on $\mathcal{H}$ are well-defined. In particular, without always explicitly stating it we assume **(K1)** holds, such that $\mathcal{H}$ is separable and measurability issues do not arise, as in Ćevid et al. (2022); Näf et al. (2023).

Applying the decomposition in Ćevid et al. (2022, Lemma 9) to $\mu_n^{(-j)}(\mathbf{x})$, we obtain

**Lemma 2.** *Assume $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$ satisfies **(F3)**. Then,*

$$\mathbb{V}_{\mathcal{H}}(\mu_n^{(-j)}(\mathbf{x})) \leq \left(\frac{s_n}{n} + \frac{s_n^2}{n^2}\right) \mathbb{V}_{\mathcal{H}}(T^{(-j)}), \tag{18}$$

*Proof.* Since **(F3)** implies (**F3**'), the proof of this result is analogous to the one of Lemma 10 in Ćevid et al. (2022), using the ANOVA decomposition in Lemma Ćevid et al. (2022, Lemma 9). □

We then need a previous result, which we restate for convenience:

**Lemma 3.** *Let $T$ be a tree satisfying **(F2)** and **(F4)** that is trained on data $\mathcal{Z}_{s_n}$. Suppose that assumption **(D1)** holds for $\mathbf{X}_1, \dots, \mathbf{X}_{s_n}$. Then,*

$$\mathbb{P}\left(\operatorname{diam}(\mathcal{L}^{(-j)}(\mathbf{x}^{(-j)})) \geq \sqrt{p}\left(\frac{s_n}{2k-1}\right)^{-0.51\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}\right) \leq p\left(\frac{s_n}{2k-1}\right)^{-1/2\frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}\frac{\pi}{p}}. \tag{19}$$

*Proof.* Using Lemma 2 of Wager and Athey (2017) for $\operatorname{diam}(\mathcal{L}(\mathbf{x}))$ and using the fact that,

$$\operatorname{diam}(\mathcal{L}^{(-j)}(\mathbf{x}^{(-j)})) \leq \operatorname{diam}(\mathcal{L}(\mathbf{x})),$$

see e.g., Bénard et al. (2022, Proof of Lemma 6), gives the result. □

**Lemma 4.** *Let $T$ be a tree satisfying **(F1)** and **(F5)**. Then,*

$$\mathbb{E}[T^{(-j)}(\mathcal{Z}_{s_n})] = \mathbb{E}[\mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X}^{(-j)} \in \mathcal{L}^{(-j)}(\mathbf{x}^{(-j)})]]. \tag{20}$$

*Proof.* As **(F1)** and **(F5)** imply (**F1**') and (**F4**') the proof is completely analogous to the proof of Ćevid et al. (2022, Lemma 12).[1] □

[1] Though $d$ needs to be exchanged by $p$, a typo in the original paper.

**Corollary 1.** *In addition to the conditions of Lemma 3, assume* ***(D2)*** *and that the trees* $T(\mathbf{x}, \varepsilon; \mathcal{Z}_{s_n})$ *in the forest satisfy* ***(F1)****. Then, we have*

$$\|\mathbb{E}[\mu_n^{(-j)}(\mathbf{x})] - \mu(\mathbf{x}^{(-j)})\|_{\mathcal{H}} = \mathcal{O}\left(s_n^{-1/2 \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})} \frac{\pi}{p}}\right). \tag{21}$$

*Proof.* Again the proof follows the exact same steps as in Ćevid et al. (2022, Corollary 13), using the fact that $\mathcal{L}^{(-j)}(\mathbf{x}^{(-j)})$ gets smaller in all dimensions from Lemma 3 and (20) (recalling that $\mu(\mathbf{x}^{(-j)}) = \mathbb{E}[k(\mathbf{Y}, \cdot) \mid \mathbf{X}^{(-j)} = \mathbf{x}^{(-j)}]$). □

**Proposition 5.** *Assume that the forest construction satisfies the properties* ***(F1)-(F5)****. Additionally, assume that k meets Assumption* ***(K1)****, and that* ***(D1)-(D3)*** *hold. Then, we have consistency of* $\mu_n^{(-j)}(\mathbf{x})$ *in probability,*

$$\|\mu_n^{(-j)}(\mathbf{x}) - \mu(\mathbf{x}^{(-j)})\|_{\mathcal{H}} = \mathcal{O}_p\left(n^{-\gamma}\right),$$

*for any* $\gamma \leq \frac{1}{2}\min\left(1-\beta, \frac{\pi \log(1-\alpha)}{p\log(\alpha)} \cdot \beta\right)$*, where* $\alpha$ *and* $\beta$ *are chosen in* ***(F4)*** *and* ***(F5)*** *respectively. Moreover,*

$$\mathbb{E}[\|\mu_n^{(-j)}(\mathbf{X}) - \mu(\mathbf{X}^{(-j)})\|_{\mathcal{H}}] \longrightarrow 0.$$

*Proof.* Again the proof works in the same way as the proof of Ćevid et al. (2022, Theorem 2), but since the result is more important than the previous ones, we state it here for completeness:

We first note that **(K1)** implies $\mathbb{V}_{\mathcal{H}}(T^{(-j)}) < \infty$. Thus, from Markov's inequality and Lemma 2,

$$\mathbb{P}\left(n^{\gamma}\|\mu_n^{(-j)}(\mathbf{x}) - \mathbb{E}[\mu_n^{(-j)}(\mathbf{x})]\|_{\mathcal{H}} > \varepsilon\right) \leq \frac{n^{2\gamma}}{\varepsilon^2}(s/n + s^2/n^2)\mathbb{V}_{\mathcal{H}}(T^{(-j)}) = \frac{1}{\varepsilon^2}\mathcal{O}(n^{2\gamma+\beta-1}),$$

where the last step followed from **(F5)**. Thus

$$n^{\gamma}||\mu_n^{(-j)}(\mathbf{x}) - \mathbb{E}[\mu_n^{(-j)}(\mathbf{x})]||_{\mathcal{H}} = \mathcal{O}_p(1),$$

for $\gamma \leq (1-\beta)/2$. In particular, it goes to zero for any $\varepsilon > 0$, if $\gamma < (1-\beta)/2$. Since,

$$n^{\gamma}\left\|\mu_n^{(-j)}(\mathbf{x}) - \mu(\mathbf{x})\right\|_{\mathcal{H}} \leq n^{\gamma}\left\|\mu_n^{(-j)}(\mathbf{x}) - \mathbb{E}[\mu_n^{(-j)}(\mathbf{x})]\right\| + n^{\gamma}\left\|\mathbb{E}[\mu_n^{(-j)}(\mathbf{x})] - \mu(\mathbf{x}^{(-j)})\right\|_{\mathcal{H}},$$

the result follows as soon as the second expression goes to zero. Now from Theorem 1, with $C_\alpha = \frac{\log((1-\alpha)^{-1})}{\log(\alpha^{-1})}$,

$$n^{\gamma}\|\mathbb{E}[\mu_n^{(-j)}(\mathbf{x})] - \mu(\mathbf{x}^{(-j)})\|_{\mathcal{H}} = \mathcal{O}\left(n^{\gamma}s_n^{-1/2C_\alpha \frac{\pi}{p}}\right) = \mathcal{O}\left(n^{\gamma - 1/2\beta C_\alpha \frac{\pi}{p}}\right).$$

This is bounded provided that,

$$1/2\beta C_\alpha \frac{\pi}{p} \geq \gamma.$$

This proves convergence in probability. Using **(K1)** convergence in expectation follows in the same way as argued in Proposition 3. □

Finally, given Proposition 5, Theorem 2 can be proven with the same arguments as in Theorem 1.