Targeted Learning for Variable Importance

Xiaohan Wang[1] **Yunzhe Zhou**[2] **Giles Hooker**[3]

[1]Department of Statistics and Data Science, Cornell University, Ithaca, New York, USA
[2]Department of Biostatistics, University of California, Berkeley, Berkeley, California, USA
[3]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Abstract

Variable importance is one of the most widely used measures for interpreting machine learning with significant interest from both statistics and machine learning communities. However, attention has only recently been directed toward uncertainty quantification in these metrics. Current approaches largely rely on one-step procedures, which, while asymptotically efficient, can present higher sensitivity and instability in finite sample settings. To address these limitations, we propose a novel method by employing the *targeted learning* (TL) framework, designed to enhance robustness in inference for variable importance metrics. Our approach is particularly suited for conditional permutation variable importance. We show that it (i) retains the asymptotic efficiency of traditional methods, (ii) maintains comparable computational complexity, and (iii) delivers improved accuracy, especially in finite sample contexts. We further support these findings with numerical experiments that illustrate the practical advantages of our method and validate the theoretical results.

1 INTRODUCTION

Machine Learning (ML) models offer high-quality predictions for complex data structures and have become indispensable across various fields, including civil engineering [Lu et al., 2023], sociology [Molina and Garip, 2019], and archaeology [Bickler, 2021], due to their versatility and predictive power. However, due to their complexity, humans find the internal structures of ML models challenging to turn into real-world interpretation [Hooker and Hooker, 2017, Hooker et al., 2021, Freiesleben et al., 2024]. To address this, a considerable suite of post hoc interpretable machine learning (IML) tools have been developed. Among these tools, variable importance, which measures the contribution of individual covariates to the response variable, is a widely adopted measure in IML [Molnar, 2020]. Traditionally, this has been applied to assess the behavior of fixed models, such as random forests [Breiman, 2001] and linear models [Grömping, 2007]. Additionally, efforts have been made to create model-specific uncertainty quantification methods, as seen in Gan et al. [2022]. Building on these advances, there is a growing interest in exploring model-agnostic variable importance using nonparametric techniques [van der Laan, 2006, Lei et al., 2018, Williamson et al., 2021, Donnelly et al., 2023, Verdinelli and Wasserman, 2024a].

Despite substantial efforts devoted to developing new methodologies, little attention has been given to fully understanding these tools. Specifically, there are few methodological developments around uncertainty quantification for variable importance metrics. Some recent work has started to develop such methods, [Williamson et al., 2021, 2023, Wolock et al., 2023, Freiesleben et al., 2024, Fauvel et al., 2025]. However, these have focused on utilizing one-step de-biasing procedures.

In this paper, we introduce a novel method to quantify the uncertainty of variable importance metrics. Employing the targeted learning framework of van der Laan [2006], our method provides a robust algorithm for conducting inference on variable importance. Our approach is statistically efficient within the class of regular estimators as well as computationally cheap. Particularly, we focus on conditional permutation importance, as this method avoids potential issues with extrapolation [Hooker et al., 2021].

This paper is organized as follows: In Section 2, we formally state the problem setup and introduce some key concepts related to variable importance. In Section 3, we give an overview of the existing methodology and its justification, present our methodology, and illustrate it using conditional permutation importance. In Section 4, we introduce the efficiency theory and the theoretical guarantees of

our methodology. Lastly, we illustrate the effectiveness of our method through simulation studies and two real-world data applications. We make our code publicly available at `https://github.com/xw547/TL4VI`.

2 VARIABLE IMPORTANCE

2.1 PROBLEM SETUP

Suppose that we observe n independent and identically distributed (i.i.d.) observations $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ drawn from the unknown joint distribution $P^*_{Y,X,Z} \in \mathcal{M}$, where $\mathcal{M}$ is the class of nonparametric distributions. That is,

$$(Y_i, X_i, Z_i) \overset{\text{i.i.d.}}{\sim} P_{Y,X,Z}, \quad i = 1, \ldots, n.$$

We aim to investigate the relationship between the response $Y \in \mathbb{R}$ and the covariate of interest $X \in \mathcal{X}$ in the presence of other covariates $Z \in \mathcal{Z}$ through some predefined variable importance, make sure it is "efficient", and then conduct inference. We define variable importance with respect to performance on some loss $L(\cdot)$, but our estimator $\hat{f} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ is assumed to approximate $E(Y|X, Z)$. For simplicity, we'll focus on the case where $\mathcal{X} \subseteq \mathbb{R}, \mathcal{Z} \subseteq \mathbb{R}^{d-1}$, yet we note that our method can be generalized to the cases where X_i is a vector.

2.2 NOTATION

We use $\mathbb{P}_n$ to denote the empirical measure, that is, suppose $f : \mathcal{X} \to \mathbb{R}$, $\mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^n f(X_i)$. In contrast, we use $\mathbb{P}$ to denote the probability measure, that is, $\mathbb{P}(f) = \int f(X)d\mathbb{P}$. And $L_2^0(P)$ denotes the collection of functions such that $Pf = 0$ and $Pf^2 < \infty$. O_P and o_P are used as follows: $X_n = O_P(r_n)$ denotes X_n/r_n is bounded in probability and $X_n = o_P(r_n)$ indicates $X_n/r_n \xrightarrow{P} 0$, respectively. Lastly, we denote the $L_2(P)$ norm as $\|\cdot\|$.

2.3 VARIABLE IMPORTANCE

A number of variable importance metrics have been suggested. Here we include a brief description of some of the most commonly studied.

2.3.1 Permutation Importance

Variable importance is obtained by considering the out-of-bag (OOB) loss of a certain feature [Breiman, 2001]. First, we permute the feature(s) that we are interested in quantifying the importance. That is, we randomly permute the index of the column of X, denoted by X^π, and then put it together with the remaining features, which results in the final data (Y, X^π, Z). Then, for model $\hat{f} : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$:

$$VI_X^\pi = \frac{1}{N}\sum_{i=1}^N L(Y_i, \hat{f}(X_i^\pi, Z_i) - L(Y_i, \hat{f}(X_i, Z_i))).$$

While providing a starting point, this metric has been critiqued in Strobl et al. [2008], Hooker et al. [2021] as resulting in extrapolation when (X_i^π, Z_i) are far from observed data.

2.3.2 Conditional Permutation Importance

This metric is obtained by conditional permutation copy of X such that: $X_i^C \sim X_i|Z_i$, $X_i^C \perp Y_i|Z_i$. With a similar notation defined above, we may thus have the plug-in estimator, defined as:

$$VI_X^C = \frac{1}{N}\sum_{i=1}^N L(Y_i, \hat{f}(X_i^C, Z_i)) - L(Y_i, \hat{f}(X_i, Z_i)).$$

This approach was first proposed by Strobl et al. [2008] for the random forest, where they obtain the conditional permuted version by conducting the permutation within each leaf. A similar idea is also present in Fisher et al. [2019], Chamma et al. [2024]. Hooker et al. [2021] observes that both conditional permutation, as well as Leave-One-Covariate-Out (LOCO) and other retraining methods, have the same population estimand that serves as our target.

2.3.3 Leave-One-Covariate-Out

LOCO can be considered a nonparametric extension of the classical R^2 statistic [Williamson et al., 2023]. In addition to $\hat{f}$, we training another model $\hat{f}_{-X} : \mathcal{Z} \to \mathbb{R}$, which has no access to X. The plug-in estimator is defined as:

$$VI_X^d = \frac{1}{N}\sum_{i=1}^N L(Y_i, \hat{f}_{-X}(Z_i) - L(Y_i, \hat{f}(X_i, Z_i))).$$

This approach was first proposed by Lei et al. [2018], and Williamson et al. [2023] quantified the uncertainty of the method through the efficient influence function. See Mentch and Zhou [2022] for cautionary results.

3 METHODOLOGY

Existing literature on quantifying the uncertainty of variable importance is scarce. There are two main trends in the uncertainty quantification of IML.

The first is a de-biasing approach utilizing the influence function, in which a bias correction and confidence intervals are constructed from a one-step method [Williamson et al.,

2023, Wolock et al., 2023]. In Williamson et al. [2021] and Williamson et al. [2023], the efficient influence function is leveraged to construct confidence intervals, since the plug-in estimator is shown to be efficient under mild assumptions. While Wolock et al. [2023] applies this concept to de-bias the variable importance for survival analysis, and then construct the confidence interval, an approach also seen in Ning and Liu [2017], Chernozhukov et al. [2018].

Alternatively, Molnar et al. [2023] and Freiesleben et al. [2024] propose a bootstrap-like approach for uncertainty quantification. In these methods, models are refitted on different subsets of the data, and the variance is estimated from the ensemble of models and their associated metrics. This approach is conceptually similar to the bootstrap variance estimation described by DiCiccio and Efron [1996]. However, these bootstrap-based methods require significant computational effort.

In the following subsections, we first review efficient influence function. Next, we briefly introduce the theory behind the de-biasing approach. We then formally present the proposed methodology within the targeted learning framework. Lastly, we present an implementation of the proposed methodology.

3.1 EFFICIENT INFLUENCE FUNCTION

Influence functions characterize the first-order behavior of pathwise differentiable functionals. By naively appending the empirical estimator of the influence function, we can "de-bias" the estimator, which will be discussed in detail in the next section.

Definition 1. *Let $\mathcal{M}$ be a class of probability distributions and $\Psi : \mathcal{M} \to \mathbb{R}$ be a functional. We say that Ψ is pathwise differentiable at $P^* \in \mathcal{M}$ with tangent space $\dot{P}_0$ if there exists a bounded linear function ψ_{P^*}, called the influence function, such that for $P_{\varepsilon,g} = (1+\varepsilon)P^* + \varepsilon g \in \dot{P}^*$, the following holds:*

$$\frac{d}{d\epsilon}\Psi(P_{\epsilon,g})\bigg|_{\epsilon=0} = \mathbb{E}_{P^*}\left[\psi_{P^*}(X)\cdot\frac{d}{d\epsilon}\log\frac{dP_{\epsilon,g}}{dP^*}(X)\bigg|_{\epsilon=0}\right].$$

Following Hines et al. [2022], for many targets Ψ the influence function can be calculated from a Gâteaux derivative in the direction of a point-mass contamination at each x:

$$\psi_P(x) = \frac{d}{d\epsilon}\Psi\left((1-\epsilon)P + \epsilon\delta_{X=x}\right).$$

For distributions $\hat{P}, P^* \in \mathcal{M}$, if Ψ is a pathwise differentiable functional, we can consider the von Mises expansion:

$$\Psi(\hat{P}) - \Psi(P^*) = \int \psi_{\hat{P}}(x)d(\hat{P}-P^*)(x) + R_2(\hat{P}, P^*), \tag{1}$$

where $R_2(\hat{P}, P^*)$ is the second-order remainder term. Intuitively, the von Mises expansion can act as a distribution version of a Taylor expansion. In particular, if $\mathcal{M}$ is the class of nonparametric distributions, the influence function ψ is also the efficient influence function [Hines et al., 2022, Kennedy, 2022]. The efficient influence function characterizes the optimal attainable asymptotic variance (See lemma 25.19 of van der Vaart [2000] and theorem 5.2.1 of Bickel et al. [1993].

3.2 DE-BIASING APPROACH

Based on the above explanation, a natural idea would be to seek a bias correction. A naive bias correction estimator is:

$$\hat{\Psi}_{naive} = \Psi(\hat{P}) + \mathbb{P}_n(\psi_{\hat{P}}).$$

With the same von Mises expansion, we may then have:

$$\begin{aligned}\hat{\Psi}_{naive} - \Psi(P^*) = &(\mathbb{P}_n - \mathbb{P})(\psi_{P^*}) \\ &+ (\mathbb{P}_n - \mathbb{P})(\psi_{\hat{P}} - \psi_{P^*}) \\ &+ R_2(\hat{P}, P^*)\end{aligned}$$

The first term is a simple average of fixed functions, where we can then apply the central limit theorem. The second term is usually referred to as the *empirical process term*. If $\psi_{\hat{P}} \xrightarrow{P} \psi_{P^*}$, it can be shown to be of order $o_P(1/\sqrt{n})$ under either Donsker class assumptions on $\hat{P}$, or in the sample-splitting regime which we adopt. The last term, also called *second order term*, is generally assumed to be of order $o_P(1/\sqrt{n})$, which is typically determined in a case-by-case manner [Cheng, 1984, Luo et al., 2016, Benkeser and van der Laan, 2016, Farrell et al., 2018, Wei et al., 2023].

Wolock et al. [2023] employs this first-order correction to provide uncertainty quantification for variable importance of survival analysis. This gives rise to the construction of a confidence interval from

$$\hat{\Psi}_{naive} \pm z_{\alpha/2}s_{\mathbb{P}_n}(\psi_{\hat{P}})$$

in which $z_{\alpha/2}$ are the quantile of a normal distribution and $s_{\mathbb{P}_n}(\psi_{\hat{P}})$ indicates the standard deviation over the values of the influence function. However, we note from the non-asymptotic perspective, that the instability of empirical distributions can hinder the effectiveness of both methods, as highlighted by Booth and Sarkar [1998] and van der Laan et al. [2011b].

3.3 PROPOSED METHODOLOGY

In contrast to current de-biasing methods, our method provides an iterative update to remove the bias and produces a more refined estimator than the one-step versions. The

targeted learning framework, first proposed by van der Laan and Rubin [2006], originates from semiparametric statistics and causal inference. To obtain an asymptotically linear estimator, they proposed to perturb the empirical distribution in the direction of the influence function to obtain an efficient estimator. Specifically, we create a one-dimensional family of densities starting from $\hat{P}$ and moving in the direction of the influence function: $P_\varepsilon = (1 + \varepsilon\hat{\psi})\hat{P}$ and find the maximum likelihood estimate of ε:

$$\hat{\varepsilon} = \underset{\varepsilon}{\operatorname{argmin}}\, \mathbb{P}_n \log P_\varepsilon.$$

This defines a new estimated density $\hat{P}_{\hat{\varepsilon}}$ for which we can calculate an efficient influence function, a new one-dimensional family and a corresponding update. Repeating this yields a sequence of estimates:

$$\hat{\varepsilon}^{j+1} = \underset{\varepsilon_j}{\operatorname{argmin}}\, \mathbb{P}_n \log P_\varepsilon^j$$
$$P_\varepsilon^{j+1} = (1 + \varepsilon^{j+1}\hat{\psi}_{P^j})P^j$$

We continue updating the distribution until we obtain and update $\hat{\varepsilon}^k = 0$ at iterate k and obtain a final debiased estimator $\hat{\Psi}_n = \Psi(P^k)$. Many extensions of this framework have been proposed for causal inference: cross-validation TL, dynamic treatment regimes, and time-to-event outcome van der Laan et al. [2011a], Luedtke and van der Laan [2016], Cai and van der Laan [2020]. Instead, we employ TL to conduct uncertainty quantification for variable importance.

The intention of this iterative definition is to ensure that the likelihood is maximized and that the plug-in bias term ($\mathbb{P}_n\psi(\hat{P})$) is zero. From this, the first order error of $\hat{\Psi}_n - \Psi(P^*)$ is described by $\mathbb{P}_n\psi_{P^k}$ which admits a central limit theorem and in common with the naive method we base confidence intervals on the standard deviation $s_{\mathbb{P}_n}(\psi_{P^k})$. The improved accuracy of this framework over the naive implementation is due both to a more exact control of bias and because the naive method does not account for uncertainty due to the plug-in bias. The iterative scheme that we propose requires a single representation of the distribution P and thus cannot directly be employed within LOCO-type models that rely on re-training estimators.

We note that in order to obtain theoretical results with weaker conditions, we adopted the sample splitting strategy implemented in van der Laan et al. [2011a], Chernozhukov et al. [2018] and Newey and Robins [2018]. That is, the plug-in estimate is obtained from the first set of data I_1 and the iterative update is conducted using an independent data set I_2. In theoretical results, we also assume a third set I_3 used to quantify uncertainty, although we do not believe this is strictly necessary. In the next section, we present an implementation of the proposed methodology through CPI.

3.4 ILLUSTRATION

In this section we apply our methodology to conditional variable importance. Detailed steps of our implementation can be found at Algorithm 1. We begin by observing that the estimand of conditional permutation metric is defined as:

$$\Psi^C(X, Y, Z) = \mathbb{E}\left[L(y, \hat{y}(X^C, Z)) - L(y, \hat{y}(X, Z))\right], \tag{2}$$

where $\hat{y}(x, z) = \mathbb{E}[Y|X = x, Z = z]$, $X^C \sim X|Z$, and $X^C \perp X|Z$.

Here the first term measures the "conditional permuted" performance. The second term in the estimand is the reference loss, which serves as the "benchmark" of our importance. We therefore decompose the conditional permutation importance into:

$$\Psi^C(X, Y, Z) = \Psi_0^C(X, Y, Z) - \Psi_0(X, Y, Z),$$

where $\Psi_0^C(X, Y, Z) = \mathbb{E}\left[L(y, \hat{y}(X^C, Z))\right]$ and $\Psi_0(X, Y, Z) = \mathbb{E}[L(y, \hat{y}(X, Z))]$.

These each have a corresponding influence function:

Lemma 1. *The efficient influence function for*

$$\Psi_0(X, Y, Z) = \mathbb{E}[L(y, \hat{y}(X, Z))]$$

is:

$$\begin{aligned}&\psi_0(X, Y, Z)\\ &= (Y - \hat{y}(X, Z)) \int L'(y, \hat{y}(X, Z))P(y|X, Z)dy\\ &+ L(Y, \hat{y}(X, Z)) - \Psi_0(P).\end{aligned}$$

Lemma 2. *The efficient influence function for*

$$\Psi_0^C(X, Y, Z) = \mathbb{E}\left[L(y, \hat{y}(X^C, Z))\right]$$

is:

$$\begin{aligned}\psi_0^C(X, Y, Z) = &\int L'(y, \hat{y}(X, Z))(Y - \hat{y}(X, Z))p(y|Z)dy\\ &+ \int L\left(y, \hat{y}(X, Z)\right)p(y|Z)dy\\ &- \int L\left(y, \hat{y}(x, Z)\right)p(y|Z)p(x|Z)dxdy\\ &+ \int L\left(Y, \hat{y}(x, Z)\right)p(x|Z)dx - \Psi_0^C(P).\end{aligned}$$

The algorithm for calculating conditional permutation importance is shown in Algorithm 1, following the methodology outlined in Section 3.3.

Implementing a TL update for CPI requires, in addition to and estimate of $\hat{y}(X, Z)$, which we obtain from $\hat{f}$, auxiliary

estimates of $p(y|Z)$ and $p(x|Z)$. In practice, we implement these via a weighted empirical distribution on I_1 and calculate the integrals above via Monte Carlo simulation. In particular, we fit a random forest (RF) to predict X or Y from Z and use OOB data to derive the conditional distributions for $p(y|Z)$ and $p(x|Z)$ from the tree kernel defined by in-leaf proximities, following Lu and Hardin [2021]. We express this as $P(y = Y_i|Z) = w_i(Z)$ where the $w_i(Z)$ are initially obtained the frequency with which Z_i appears in the same leaf as Z across trees in the RF for which (Y_i, Z_i) is out of bag. To construct an updated TL distribution we simply need to multiply the weights $w_i(Z)$ by $(1 + \hat{\epsilon})\hat{\psi}$.

We can easily generalize this algorithm into $K-$folds rather than a single split, which would result in the same asymptotic result; choosing $K = 10$ produces to a more numerically stable result. We refer the readers to Smith et al. [2023] for a more comprehensive review on the selection of folds for targeted learning.

Algorithm 1 Conditional permutation calculation on I_1 with mean squared error loss

Require: $\{Y_i, X_i, Z_i\}$ for $i = 1, \ldots, n$, I_1, I_2, I_3 such that $I_1 \cup I_2 \cup I_3 = \{1, \ldots, n\}$ and $I_1 \cap I_2 \cap I_3 = \emptyset$.

Train an initial estimate $\hat{f}_{I_1}$.

Estimate $\hat{P}(x|z), \hat{P}(y|z)$.

for each iteration t **do**

  Sample from $\hat{P}(x|z), \hat{P}(y|z)$, denoted as $\{X_j^*\}_{j=1,\ldots,m}$ and $\{Y^*\}_{k=1,\ldots,m}$ respectively.

  Calculate

$$\hat{\Psi}^C_{I_2,0} = \frac{1}{|I_2|} \sum_{i \in I_2} (Y_i - \hat{f}(X_i^C, Z_i))^2.$$

  and

$$\begin{aligned}\hat{\psi}^C_{I_2,0}(X_i, Y_i, Z_i; \hat{P}) &= \frac{1}{m} \sum_{j=1}^{n} L(Y_i, \hat{f}(X_j^*, Z_i)) \\ &\quad - \frac{1}{m^2} \sum_{j=1}^{n} \sum_{k=1}^{n} L(Y_k^*, \hat{f}(X^*, Z_i)) \\ &\quad + \frac{1}{m} \sum_{k=1}^{n} L(Y_k^*, \hat{f}(X_i, Z_i))\end{aligned}$$

  Find $\hat{\epsilon}$ to maximize the likelihood of $\sum_{i \in I_2} c(\hat{\epsilon}) \hat{P}(X_i, Y_i, Z_i)(1 + \hat{\epsilon}\hat{\psi}^C_{I_2,0}(X_i, Y_i, Z_i; \hat{P}))$.

  Update $\hat{P} = c(\hat{\epsilon})(1 + \hat{\epsilon}\hat{\psi}^C_{I_2,0})\hat{P}$

Repeat the above iteration until convergence.

Return: $\hat{\Psi}(\hat{f}_{I_1}, P_{\varepsilon^{k_n}})$ and variance $\sqrt{\frac{1}{n} \sum_{i \in I_3} \hat{\psi}^C_{I_2,0}}$ based on I_3.

3.5 WHY CONDITIONAL PERMUTATION IMPORTANCE?

We explore conditional permutation importance for two reasons: CPI provides an orthogonal factorization and CPI avoids density ratio estimation, as mentioned in Verdinelli and Wasserman [2024a].

Example: CPI factorization

Traditional targeted learning relies on factorizing the influence function into orthogonal components in order to conduct updates, as seen in average treatment effect estimation [van der Laan et al., 2011a].

For conditional permutation importance we consider components corresponding to $P_{X,Y|Z}$ and P_Z. Then, we have:

- $P_{X,Y|Z}$:

$$\begin{aligned}\psi_0^C(X, Y|Z) = &\int L(y, \hat{y}(X, Z))\, p(y|Z) dy \\ &- \int L(y, \hat{y}(x, Z))\, p(y|Z) p(x|Z) dx dy \\ &+ \int L(Y, \hat{y}(x, Z))\, p(x|Z) dx\end{aligned}$$

- P_Z:

$$\begin{aligned}\psi_0^C(Z) = &\int L'(y, \hat{y}(X, Z))(Y - \hat{y}(X, Z)) p(y|Z) dy \\ &- \Psi_0^C(P).\end{aligned}$$

Notice that for P_Z the empirical log-likelihood of the data points Z is already maximized at the empirical distribution and thus no update is needed.

For $P_{X,Y|Z}$, we employ the iterative update methodology, which follows a similar structure as algorithm 1, but with a much simpler form.

As a contrast, we consider the efficient influence function of the LOCO importance. Following the similar construction as conditional permutation importance, for LOCO importance, we have:

$$\Psi^d(X, Y, Z) = \Psi_0^d(Y, Z) - \Psi_0(X, Y, Z),$$

where $\Psi_0^d(Y, Z) = \mathbb{E}[L(y, \hat{y}(Z))]$. The corresponding influence function is:

Lemma 3 (Williamson and Feng [2020])**.** *Let* $\Psi_0^d(X, Y, Z) = \mathbb{E}[L(y, \hat{y}(Z))]$*, the efficient influence function is:*

$$\begin{aligned}\psi_0^d(X, Y, Z) = (Y - \hat{y}(Z)) &\int L'(y, \hat{y}(X, Z)) P(y|Z) dy \\ &+ L(Y, \hat{y}(Z)) - \Psi_0^d(P).\end{aligned}$$

Compared to CPI, the estimation of the LOCO involves the estimation of two models $\hat{y}(Z), \hat{y}(X, Z)$, where $\hat{y}(Z)$ is based on the retraining of a new model based on perturbed data. This is described as having variational dependent models and thus need a more subtle treatment.

Finally, we also examine the influence function of the traditional permutation importance metric.

Lemma 4. *Let*

$$\Psi_0^{\pi L}(X, Y, Z) = \mathbb{E}\left[L(y, \hat{y}(X^\pi, Z))\right].$$

The efficient influence function is:

$$\begin{aligned}
&\psi_0^{\pi L}(X, Y, Z) \\
&= (Y - \hat{y}(X, Z)) \int L'(y, \hat{y}(X, Z)) \frac{P(X)P(y, Z)}{P(X, Z)} dy \\
&+ \int L(Y, \hat{y}(x', Z)) P(x') dx' \\
&+ \int L(y, \hat{y}(X, z)) P(y, z) dy dz - 2\Psi_0^{\pi L}(P),
\end{aligned}$$

where $X^\pi \sim X$, and $X^\pi \perp X$.

We note that the performance of the density ratio estimation in the first term can be unstable, due to extrapolation and inherent low density at certain regions – a problem that also applies to the decorrelated LOCO as mentioned in Verdinelli and Wasserman [2024a].

4 THEORETICAL RESULTS

In this section, we present the theoretical results of our methodology. To formally introduce the theoretical results, we start with a brief introduction to a few concepts that would be helpful in developing our method. We start with the efficiency theory and methodology of targeted learning in section 4.1, then we present the theoretical result for the estimator obtained in algorithm 1.

4.1 EFFICIENCY THEORY AND TARGETED LEARNING

By considering the variable importance as *general parameter* $\Psi : P \to \mathbb{R}, P \in \mathcal{M}$, where $\mathcal{M}$ is the class of nonparametric distributions, our aim is to find a "good" estimator of the true value $\Psi(P^*)$, and then construct the corresponding confidence interval. We define "good" using three criteria:

- Consistency: we would like to construct an estimator that is statistically consistent, which can be guaranteed by *asymptotically linearity*.
- Robustness: we would like to construct an estimator that is robust to small perturbations of the data distribution, which can be guaranteed with *regularity*.
- Efficiency: we hope to have an estimator that has minimum-possible variance given the available data, which will be ensured by the TL methodology, based on the two other requirements.

Building upon these three objectives, our goal is to construct an efficient, regular, and asymptotically linear estimator. In the following sections, we will rigorously define our concepts and then introduce the targeted learning methodology to construct such an estimator.

4.1.1 Regular Asymptotically Linear (RAL) Estimators

To begin with, we at least hope we can estimate with guaranteed consistency. One such class of estimators is the *asymptotically linear* estimators, where classical *asymptotically linear* estimators for parametric models include maximum likelihood estimation and generalized method of moments under mild conditions. In addition to the fact that the influence function characterizes the first-order term of a pathwise differentiable estimand, it also determines the asymptotic distribution of asymptotic linear estimator. Formally, asymptotically linear estimator is defined as:

Definition 2. *An estimator sequence $\{\hat{\Psi}_n(P^*)\}$ is said to be asymptotically linear with influence function $\psi \in L_2^0(P^*)$ at distribution P^* if*

$$\sqrt{n}(\hat{\Psi}_n(P^*) - \Psi(P^*)) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i) = o_P(1).$$

We note that when the influence function is the same as *efficient influence function*, the asymptotically linear estimator is efficient.

4.1.2 Tangent Space

The *tangent space* characterizes the collection of possible functions to locally construct a path between distributions, defined by score function $h = \frac{d}{d\varepsilon} \log dP^*\big|_{\varepsilon=0}$ and their linear combinations at distribution $P^* \in \mathcal{M}$ [Bickel et al., 1993, van der Vaart, 2000]. Formally, *tangent space* is defined as:

Definition 3. *Let $\{V_1, \ldots, V_n\}$ denote the collection of score functions of $P^* \in \mathcal{M}$, then the tangent space $\dot{P}^*$ of P^* is defined as the linear span of $V_1, \ldots, V_n$.*

For the class of nonparametric distributions $\mathcal{M}$, the tangent space is $\dot{\mathcal{P}}^* := L_2^0(P^*)$ [Bickel et al., 1993].

With the tangent space defined, we can say that a sequence of estimators $\hat{\Psi}_n$ at P^* is *regular* if there exists a probability measure L such that:

$$\sqrt{n}\left(\hat{\Psi}_n - \Psi\left(P_{1/\sqrt{n},g}\right)\right) \overset{P_{1/\sqrt{n},g}}{\rightsquigarrow} L, \quad \text{for each } g \in \dot{\mathcal{P}}^*,$$

where $P_{1/\sqrt{n},g} = (1 + \frac{1}{\sqrt{n}}g)P^*$.

4.2 ASYMPTOTIC RESULTS

In this section, we outline the assumptions necessary to establish the efficiency of our final estimator and then present our main theorem.

Assumption 1 (Convergence)**.** *Let k_n denote the number of iterations until the algorithm converges. Assume that there exists $k_n = k(\hat{P}) > 0$ such that $P(k(\hat{P}) < k_0) \to 1$ for some $k_0 \equiv k(P^*)$ and*

$$\frac{1}{|I_2|}\mathbb{P}_{n,I_2}\psi_{P_\varepsilon^{k_n}} = o_P(1/\sqrt{n}),$$

The same equation holds if we consider the empirical distribution of I_3. In addition, we assume that the k_0-th step of estimate $P_{\varepsilon^{k_0}}$ converges to P^* almost surely, where $P^* \in \mathcal{M}$ is the least favorable model.

This assumption is standard for cross-validation TL [van der Laan et al., 2011a]. Here we assume that the algorithm will converge in at most the k_0 steps and that the efficient influence function will be small. In addition, the assumption ensures the limiting distribution is within the nonparametric model class and the first-order optimality. Lastly, Assumption 1 implicitly places an assumption on the initial estimator, as a poorly chosen initial estimator could result in divergence. In practice, initial estimators based on either the plug-in or Z-estimation approach have been shown to perform well.

Assumption 2 (Differentiability and Optimality)**.** *Given a variable importance metric Ψ, we assume that it is pathwise differentiable for the class of nonparametric distributions $\mathcal{M}$. In addition, the von Mises expansion satisfies:*

$$\Psi(\hat{f}_{I_1}, \hat{P}) - \Psi(f^*, P^*) = \int \psi(\hat{f}_{I_1}, \hat{P})d(\hat{P} - P^*) + O_P(\|\Psi(\hat{f}_{I_1}, \hat{P}) - \Psi(f^*, P^*)\|^2),$$

where $\hat{P} \in \mathcal{M}$, $f^* \equiv \hat{y}$, and $\hat{y}$ is defined as in equation 2.

This assumption restricts the differentiability of the variable importance measure and imposes an assumption on the asymptotic performance of the second-order remainder term, originating from van der Laan et al. [2011a]. Together with Assumption 1, the above two results guarantee the asymptotic efficiency of the TL estimator.

Remark 1. *We note that assumption 2 functions in a similar manner as the (A1) and (B1) given in Williamson et al. [2023] or Assumption 5 of Wei et al. [2023]. In both cases, the aim is to control the second-order term. By considering the second-order term as the order of the bias directly, the proof is greatly simplified. For conditional permutation variable importance, we can alternatively have the order of $\mathbb{E}_{I_2}\left[\|\hat{p}(y|z) - p(y|z)\right]\|$, $\mathbb{E}_{I_2}\left[\|\hat{p}(x|z) - p(x|z)\|\right], \mathbb{E}_{I_2}\left[\|\hat{f}(x,z) - f^*(x,z)\|\right]$ be $o_P(n^{-1/4})$, where f is the estimator and $\hat{p}(x|z)$ is the estimator of density $p(x|z)$. A similar assumption can be defined for I_3 as well.*

Assumption 3 (Consistency)**.**

$$\int \left(\Psi(\hat{f}_{I_1}, P^*) - \Psi(f^*, P^*)\right)^2 dP^* = o_P(1)$$

This assumption ensures the consistency of the plug-in estimator, which is also given in van der Laan et al. [2011a]. Without such, the result wouldn't be efficient.

Assumption 4 (Sample-Splitting)**.** *Let ε_*^j be the limit of ε_n^j, that is $\varepsilon_n^j \xrightarrow{P} \varepsilon_*^j$ for $j \in 1, \ldots, k_0$. We assume that the final efficient influence function $\psi_{P_{\varepsilon_n^{k_n}}}$ is estimated from I_1, I_2, independent from empirical measure of I_3, denoted as $\mathbb{P}_{n,I_3}$. And our final estimator is obtained through I_3. To ensure the consistency, we assume that $\sup_{j \le k_0} \|\psi_{P_{\varepsilon_n^j}} - \psi_{P^*}\| = o_P(1)$.*

Remark 2. *Assumption 4 is specifically designed to address the empirical process term. In particular, we introduce an additional subset of the data, I_3, to ensure independence between the efficient influence function and the final estimator. Although this approach differs from the classical method described by Chernozhukov et al. [2018], it is necessitated by the iterative nature of our procedure, in contrast to their one-step framework.*

As an alternative to Assumption 4, Donsker assumptions similar to A2 of van der Laan et al. [2011a] can also be made to establish the asymptotic results, which we refer to as Assumption 5; details can be found in the supplementary materials.

Theorem 1. *Assume that Assumptions 1-3 hold, and Assumption 4 or 5 hold. Our final estimator $\hat{\Psi}(\hat{f}_{I_1}, P_{\varepsilon_n^{k_n}})$ is asymptotically linear and satisfies:*

$$\hat{\Psi}(\hat{f}_{I_1}, P_{\varepsilon_n^{k_n}}) - \Psi(P^*) = \mathbb{P}_n\psi_{P^*} + o_P(1/\sqrt{n}),$$

where $\psi(P^*)$ is the efficient influence function.

Note that the above three assumptions are defined for $\hat{f}_{I_1}$, where similar assumptions can easily be defined for $\hat{f}_{I_2}, \hat{f}_{I_3}$. As a result, we can also average over swapping the rolls of subsets and $\bar{\Psi} \equiv 1/3(\hat{\Psi}(\hat{f}_{I_1}, P_{\varepsilon_n^{k_n}}) + \hat{\Psi}(\hat{f}_{I_2}, P_{\varepsilon_n^{k_n}}) + \hat{\Psi}(\hat{f}_{I_3}, P_{\varepsilon_n^{k_n}}))$ obtains the same asymptotic results while achieving better sample efficiency.

Our result implies that our estimator is asymptotically normal and efficient. We note that under similar assumptions,

the plug-in estimator with bias correction may obtain the same asymptotic performance, yet preserve worse finite data performance. These asymptotic results can also be easily extended to $K-$fold case and the asymptotic performance would remain the same.

Remark 3. *Compared to van der Laan et al. [2011a], the only difference in assumptions is that we do not impose Donsker conditions on $\hat{f}$ by considering the sample splitting methodology if we adopt Assumptions 1-3 and 4. Relative to Williamson et al. [2023], we additionally imposed Assumption, Assumption 1 to ensure the stochastic convergence of the algorithm.*

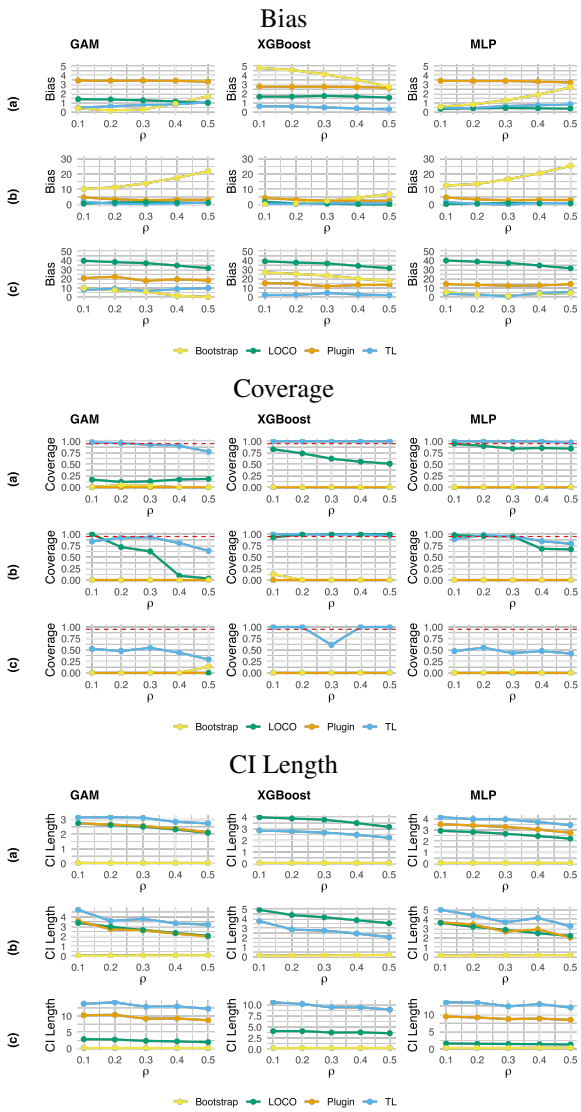

Figure 1: Bias, Coverage and Length of Confidence Intervals of targeted learning and plug-in estimators using three different initial estimators: General Additive Model (left), XGBoost (middle), and Multi-Layer Perceptron (right) based on 240 simulated data, each with 1000 observations

5 SIMULATION STUDY

In this section, we present the simulation results, starting by comparing the bias of our estimator with that of the plug-in estimator Strobl et al. [2008], as well as examining a bootstrap correction to the bias.

To construct the initial estimator $\hat{f}$, we consider three models: Generalized Additive Model (GAM) via `pyGAM` package, Multi-Layer Perceptron (MLP) implemented in `scikit-learn`, and eXtreme Gradient Boosting (XGBoost) from `xgboost` package. To mitigate the impact of hyperparameter tuning, we employ the default parameters for both GAM and XGBoost; additional technical details are provided in the supplementary materials.

For each simulation setting, we generate 1,000 observations and repeat the entire procedure 240 times. The outcome y_i is generated following one of the following three designs:

(a) $y_i = 3x_{i1} + \epsilon_i$, following Verdinelli and Wasserman [2024b].

(b) $y_i = 3x_{i1} + x_{i2} + x_{i3} + x_{i4} + x_{i5} + 0x_{i6} + 0.5x_{i7} + 0.8x_{i8} + 1.2x_{i9} + 1.5x_{i10} + \epsilon_i$, following Hooker et al. [2021].

(c) $y_i = 10\sin(x_{i1}) + 10\cos(x_{i2}) + 3\,x_{i3}x_{i6} + 3x_{i10} + \epsilon_i$.

In all cases$\epsilon_i \sim \mathcal{N}(0,1)$. Following Hooker et al. [2021], the 10-dimensional covariate vector X is sampled from a multivariate normal distribution with mean vector 0 and covariance matrix Σ. For setting (a), the covariance matrix is defined as $\Sigma_{ii} = 1, \Sigma_{12} = \Sigma_{21} = \rho$ and 0 otherwise. In settings (b) and (c), the covariance matrix is $\Sigma_{ij} = \rho^{|i-j|}$.

Consistent with Verdinelli and Wasserman [2024b] and Hooker et al. [2021], we vary $\rho \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ to examine the performance of the proposed algorithm under different correlation structures. For settings (a) and (b), the true value is derived theoretically, while for setting (c) it is approximated via Monte Carlo integration.

As illustrated in Figure 1, our estimator consistently exhibits lower bias compared to the plug-in estimator and offers a consistently good estimator compared to LOCO and bootstrap estimators. The proposed estimator generally achieves superior coverage compared to the plug-in estimator. This improvement is primarily attributable to a reduction in bias, as evidenced in Figure 1. In particular, the plug-in estimator exhibits considerably higher bias than the TL estimator, which in turn leads to substantially compromised coverage. Moreover, when the correlation is strong, the performance of both the initial estimator $\hat{f}$ and the conditional density estimator deteriorates, potentially violating Assumption 1. Although the confidence interval (CI) length for our proposed estimator is slightly longer, it achieves significantly better coverage than the plug-in estimator, indicating superior overall performance. For the XGboost model, the CI lengths for both estimators are nearly identical, making the

corresponding lines in the plots indistinguishable. A table of computational costs is provided in the supplementary material.

5.1 REAL WORLD DATA APPLICATION

5.1.1 Bike sharing

In our real-data application, we examine the variable importance scores for the hourly bike share dataset obtained from the UCI repository [Fanaee-T, 2013]. We employ XGBoost to generate the initial estimates, and the results are presented in Figure 2.

From the plot, we see that `workingday` and `yr` sit well above the rest in terms of importance, suggesting they explain a larger share of the variability in the response than other predictors. Meanwhile, features like holiday and `weathersit` occupy the next tier of influence, although their bars are noticeably shorter. At the lower end, variables such as month and `wdspd` barely rise above zero, implying they may add little explanatory power. A noteworthy takeaway is how `Temp` and `atemp` rank surprisingly low, despite one might expect temperature-related variables to matter more. Hence, even though many features cluster in a middle range of importance, the disagreements at the extremes illustrate why a nuanced approach to screening (beyond raw importance scores alone) is often necessary for sound statistical analysis.

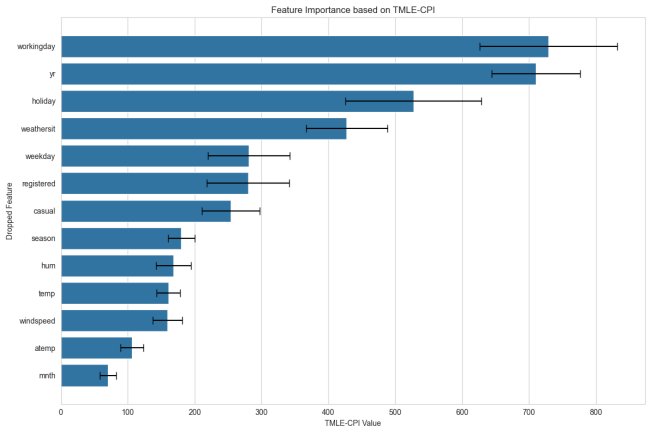

Figure 2: Conditional variable importance scores for the hourly bike share dataset, obtained using TL with an XGBoost-based initial estimate.

5.1.2 Wine quality

In addition, we included the wine quality dataset to illustrate the application of our method in classification settings through the wine quality dataset. We employ random forest to generate the initial estimates.

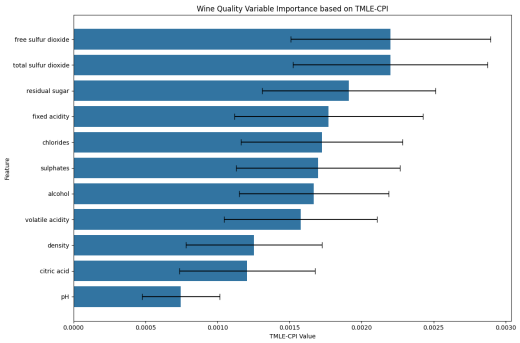

Figure 3: Conditional variable importance scores for the wine quality, obtained using TL with an Random Forest-based initial estimate.

From the conditional permutation importance plot, we see that `free sulfur dioxide` and `total sulfur dioxide` sit well above the rest in terms of importance, suggesting they explain a larger share of the variability in wine quality than other chemical measures. Meanwhile, `residual sugar` and `fixed acidity` occupy the next tier of influence, although their bars are noticeably shorter. In the middle range, variables such as `chlorides`, `sulphates`, `alcohol`, and `volatile acidity` cluster with moderately high importance, indicating their meaningful but not dominant contribution. Toward the lower end, features like `density` and `citric acid` display only modest importance, while `pH` barely rises above zero, implying it adds little explanatory power in this context. A noteworthy takeaway is how preservative-related variables dominate the ranking even though one might expect acidity or alcohol content to matter more strongly. Here we note that width of our confidence intervals suggest that the data only provide a highly uncertain ranking of variable importance.

6 CONCLUSION

In this paper, we study uncertainty quantification in IML using the targeted learning framework, illustrated through conditional permutation importance. Under mild assumptions, our methodology achieves asymptotic efficiency, maintains comparable computational complexity, and delivers improved finite-sample accuracy.

Future work includes developing methodology for estimating the overlap model as mentioned in Section 3.5, since we cannot factorize the subspace into orthogonal ones. It is also interesting to consider problems involving density ratios, which might be more approachable using methods that bypass the calculation of influence function, such as Cho et al. [2023], van der Laan et al. [2024].

References

David Benkeser and Mark van der Laan. The highly adaptive lasso estimator. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 689–696. IEEE, 2016.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer, 1993.

Simon H Bickler. Machine learning arrives in archaeology. *Advances in Archaeological Practice*, 9(2):186–191, 2021.

James G Booth and Somnath Sarkar. Monte carlo approximation of bootstrap variances. *The American Statistician*, 52(4):354–357, 1998.

Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

Weixin Cai and Mark J van der Laan. One-step targeted maximum likelihood estimation for time-to-event outcomes. *Biometrics*, 76(3):722–733, 2020.

Ahmad Chamma, Denis A Engemann, and Bertrand Thirion. Statistically valid variable importance assessment through conditional permutations. *Advances in Neural Information Processing Systems*, 36, 2024.

Philip E Cheng. Strong consistency of nearest neighbor regression function estimators. *Journal of Multivariate Analysis*, 15(1):63–72, 1984.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

Brian M Cho, Yaroslav Mukhin, Kyra Gan, and Ivana Malenica. Kernel debiased plug-in estimation: Simultaneous, automated debiasing without influence functions for many target parameters. In *Forty-first International Conference on Machine Learning*, 2023.

Thomas J DiCiccio and Bradley Efron. Bootstrap confidence intervals. *Statistical science*, 11(3):189–228, 1996.

Jon Donnelly, Srikar Katta, Cynthia Rudin, and Edward Browne. The rashomon importance distribution: Getting rid of unstable, single model-based variable importance. *Advances in Neural Information Processing Systems*, 36: 6267–6279, 2023.

Hadi Fanaee-T. Bike Sharing. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5W894.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference: application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*, 20, 2018.

Kevin Fauvel, Marine Morvan, Baptiste Cadre, and François Subtil. Sobol-cpi: a doubly robust conditional permutation importance statistic. *arXiv preprint arXiv:2501.17520*, 2025.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Timo Freiesleben, Gunnar König, Christoph Molnar, and Álvaro Tejero-Cantero. Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. *Minds and Machines*, 34 (3):32, 2024.

Luqin Gan, Lili Zheng, and Genevera I Allen. Model-agnostic confidence intervals for feature importance: A fast and powerful approach using minipatch ensembles. *arXiv preprint arXiv:2206.02088*, 2022.

Ulrike Grömping. Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147, 2007.

Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3):292–304, 2022.

Giles Hooker and Cliff Hooker. Machine learning and the future of realism. *arXiv preprint arXiv:1704.04688*, 2017.

Giles Hooker, Lucas Mentch, and Siyu Zhou. 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, 2021.

Edward H Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Benjamin Lu and Johanna Hardin. A unified framework for random forest prediction error estimation. *Journal of Machine Learning Research*, 22(8):1–41, 2021.

Yanle Lu, Xu-Hui Zhou, Heng Xiao, and Qi Li. Using machine learning to predict urban canopy flows for land surface modeling. *Geophysical Research Letters*, 50(1): e2022GL102313, 2023.

Alexander R Luedtke and Mark J van der Laan. Optimal targeted learning: confidence intervals for a median parameter. *Statistical Methods in Medical Research*, 25(3): 897–917, 2016.

Ye Luo, Martin Spindler, and Jannis Kück. High-dimensional l_2 boosting: Rate of convergence. *arXiv preprint arXiv:1602.08927*, 2016.

Lucas Mentch and Siyu Zhou. Getting better from worse: Augmented bagging and a cautionary tale of variable importance. *Journal of Machine Learning Research*, 23 (224):1–32, 2022.

Mario Molina and Filiz Garip. Machine learning for sociology. *Annual Review of Sociology*, 45(1):27–45, 2019.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Christoph Molnar, Timo Freiesleben, Gunnar König, Julia Herbinger, Tim Reisinger, Giuseppe Casalicchio, Marvin N Wright, and Bernd Bischl. Relating the partial dependence plot and permutation feature importance to the data generating process. In *World Conference on Explainable Artificial Intelligence*, pages 456–479. Springer, 2023.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *The Annals of Statistics*, 2017.

Matthew J Smith, Rachael V Phillips, Miguel Angel Luque-Fernandez, and Camille Maringe. Application of targeted maximum likelihood estimation in public health and epidemiological studies: a systematic review. *Annals of epidemiology*, 86:34–48, 2023.

Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC bioinformatics*, 9:1–11, 2008.

Lars van der Laan, Alex Luedtke, and Marco Carone. Automatic doubly robust inference for linear functionals via calibrated debiased machine learning. *arXiv preprint arXiv:2411.02771*, 2024.

Mark J van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2 (1), 2006.

Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

Mark J van der Laan, Sherri Rose, and Wenjing Zheng. Cross-validated targeted minimum-loss-based estimation. *Targeted learning: causal inference for observational and experimental data*, pages 459–474, 2011a.

Mark J van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011b.

Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Isabella Verdinelli and Larry Wasserman. Decorrelated variable importance. *Journal of Machine Learning Research*, 25(7):1–27, 2024a.

Isabella Verdinelli and Larry Wasserman. Feature importance: A closer look at shapley values and loco. *Statistical Science*, 39(4):623–636, 2024b.

Waverly Wei, Maya Petersen, Mark J van der Laan, Zeyu Zheng, Chong Wu, and Jingshen Wang. Efficient targeted learning of heterogeneous treatment effects for multiple subgroups. *Biometrics*, 79(3):1934–1946, 2023.

Brian Williamson and Jean Feng. Efficient nonparametric statistical inference on population feature importance using shapley values. In *International conference on machine learning*, pages 10282–10291. PMLR, 2020.

Brian D Williamson, Peter B Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 77 (1):9–22, 2021.

Brian D Williamson, Peter B Gilbert, Noah R Simon, and Marco Carone. A general framework for inference on algorithm-agnostic variable importance. *Journal of the American Statistical Association*, 118(543):1645–1658, 2023.

Charles J Wolock, Peter B Gilbert, Noah Simon, and Marco Carone. Nonparametric variable importance for time-to-event outcomes with application to prediction of hiv infection. *arXiv preprint arXiv:2311.12726*, 2023.

Targeted Learning for Variable Importance
Supplementary Material

Xiaohan Wang[1] **Yunzhe Zhou**[2] **Giles Hooker**[3]

[1]Department of Statistics and Data Science, Cornell University, Ithaca, New York, USA
[2]Department of Biostatistics, University of California, Berkeley, Berkeley, California, USA
[3]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA

Here, we present the proof of our main theorem, along with additional simulation results that could not be included in the main text due to space constraints. We begin by providing more details of simulation results.

A SIMULATIONS AND REAL WORLD DATA

A.1 TECHNICAL DETAILS: PARAMETER SELECTION IMPLEMENTATION DETAILS

In this section, we mostly adopt default parameter settings to maintain consistency with `R` and to avoid any performance improvements arising from parameter tuning. Specifically, for the Generalized Additive Model (GAM), we use the default parameters. For the Random Forest model employed in conditional density estimation, we utilize the default settings provided by `randomForest` package in `R`. In the case of XGBoost, we adjust the number of estimators to match that of the Random Forest. Lastly, for the MLP regressor, we design a two-layer network with 64 neurons in the first layer and 32 neurons in the second layer, applying the ReLU activation function for non-linearity and the Adam optimizer for weight updates. Training is configured to run for a maximum of 3000 iterations to ensure convergence.

To estimate densities $p(x|Z)$ and $p(y|Z)$ we started by fitting a random forest of B trees to predict each from Z using I_1. We then use this to provide an initial estimate based on a weighted empirical distribution from I_1. $P(y = Y_i|Z) = w_i(Z)$ calculated as the fraction of trees for which (Z_i, Y_i) was out-of-bag, in which Z and Z_i fall into the same leaf. In the targeted learning update, the w_i are multiplied by $(1 + \hat{\epsilon})\psi_{\hat{P}}(X_i, Y_i, Z_i)$ allow us to keep track of the updated distribution, and later apply it to I_3. The same procedure was employed to update $p(x|Z)$.

For settings (a) and (b), theoretical calculations based on Theorem 2 in Hooker et al. [2021] show that the true value is $9(1 - \rho^2)$. In contrast, due to the complexity of the nonlinear model in setting (c), the true value is estimated via Monte Carlo integration.

A.2 COMPUTATIONAL COST COMPARISON

Table 1: A Comparison of Computational Runtimes

Method	**Runtime** (s)
Plug-in	1.52
TL	3.47
Bootstrap	57.20

Note: Runtimes were measured on a machine with eighty Intel® Xeon® Gold 6230 CPUs @ 2.10GHz.

B PROOF

To facilitate these proofs, we first introduce the necessary notation.

B.1 NOTATIONS

We use $\mathbb{P}_n$ to denote the empirical measure, that is, suppose $f : \mathcal{X} \to \mathbb{R}$, $\mathbb{P}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$. In contrast, we use $\mathbb{P}$ to denote the probability measure, that is, $\mathbb{P}(f) = \int f(X)d\mathbb{P}$. $L_2^0(P)$ denotes the collection of functions such that $Pf = 0$ and $Pf^2 < \infty$. O_P and o_P are used as follows: $X_n = O_P(r_n)$ denotes X_n/r_n is bounded in probability and $X_n = o_P(r_n)$ indicates $X_n/r_n \xrightarrow{P} 0$, respectively. Additionally, we denote the $L_2(P)$ norm as $\|\cdot\|$. In addition, we denote the conditional mean of y given x as $\hat{y}(x) \equiv \mathbb{E}[y|X = x]$. We assume that the dataset I is divided into three mutually exclusive sets I_1, I_2, I_3. To facilitate the calculation of the efficient influence function, we denote $\delta_O(o)$ as the Dirac delta function with respect to O, i.e., the density of an idealized point mass at O, which equals zero everywhere except at O and integrates to 1.

B.2 EFFICIENT INFLUENCE FUNCTION

We note that the derivation of the efficient influence function largely follows the work of Hines et al. [2022], adopting the "point mass contamination" methodology.

Lemma 5. *Let*

$$\Psi_0(X, Y, Z) = \mathbb{E}\left[L(Y, \hat{y}(X, Z))\right].$$

The efficient influence function is:

$$\begin{aligned}\psi_0(X, Y, Z) = (Y - \hat{y}(X, Z)) \int L'(y, \hat{y}(X, Z))P(y|X, Z)dy \\ + L(Y, \hat{y}(X, Z)) - \Psi_0(P).\end{aligned}$$

Proof. We start by considering the integration form of the estimand, which can be expressed as:

$$\begin{aligned}\Psi_0(P) &= \mathbb{E}\left[L(y, \hat{y}(x, z))\right] \\ &= \int L(y, \hat{y}(x, z))P(x, y, z)dxdydz.\end{aligned}$$

Then, by considering the product rule, we may have:

$$\begin{aligned}\psi_0(X, Y, Z) = \int L'\left(y, \hat{y}(x, z)\right)\psi^{\hat{y}(x,z)}(X, Y, Z)P(x, z, y)dxdydz \\ + \int\int L\left(y, \hat{y}(x, z)\right)\left[\delta_{XYZ}(x, y, z) - P(x, y, z)\right]dxdyz,\end{aligned}$$

where $L'\left(y, \hat{y}(x, z)\right)$ is the derivative of $L\left(y, \hat{y}(x, z)\right)$, and $\psi^{\hat{y}(x,z)}(X, Y, Z)$ is the efficient influence function of $\hat{y}(x, z)$.

From example 6 of Hines et al. [2022], we have $\psi^{\hat{y}(x,z)}(X, Y, Z) = (Y - \hat{y}(x, z))\frac{\delta_{X,Z}(x,z)}{P(x,z)}$.

Then, we have:

$$\psi_0(X, Y, Z) = (Y - \hat{y}(X, Z)) \int L'(y, \hat{y}(X, Z))P(y|X, Z)dy + L(Y, \hat{y}(X, Z)) - \Psi_0(P)$$

□

Lemma 6. *Let*

$$\Psi_0^C(X, Y, Z) = \mathbb{E}\left[L(Y, \hat{y}(X^C, Z))\right].$$

The efficient influence function is:

$$\begin{aligned}\psi_0^C(X,Y,Z) = &\int L'(y,\hat{y}(X,Z))(Y-\hat{y}(X,Z))p(y|Z)dy \\ &+ \int L\left(y,\hat{y}(X,Z)\right)p(y|Z)dy \\ &- \int L\left(y,\hat{y}(x,Z)\right)p(y|Z)p(x|Z)dxdy \\ &+ \int L\left(Y,\hat{y}(x,Z)\right)p(x|Z)dx - \Psi_0^C(P).\end{aligned}$$

Proof. We can start off with the integral form as well, where we shall then get:

$$\Psi_0^C = \int L(y,\hat{y}(x',z))P(x'|z)P(x,y,z)dx'dxdydz$$

Then, to consider the derivative, we may have:

$$\begin{aligned}\phi_0^C &= \int \left[L(y,\hat{y}(x',z)\right]' \frac{p(x',z)p(z,y)}{p(z)}\,dx'\,dz\,dy + \int L(y,\hat{y}(x',z')\left[\frac{p(x',z)p(z,y)}{p(z)}\right]'\,dx'\,dy\,dz \\ &= R_1 + R_2\end{aligned}$$

From here, we start with the first term and adopt a similar treatment as Lemma 5, which yields

$$\begin{aligned}R_1 &= \int L'(y,\hat{y}(x',z))(Y-\hat{y}(x',z))\frac{\delta_{X,Z}(x',z)}{p(x',z)}\frac{p(x',z)p(z,y)}{p(z)}\,dx\,dx'\,dz\,dy \\ &= (Y-\hat{y}(X,Z))\int L'(y,\hat{y}(X,Z)\frac{p(y,Z)}{p(Z)}dy\end{aligned}$$

Then, for the second term, we may have to consider decomposing it into three terms.

$$\int L\left(y,\hat{y}(x',z)\right)\frac{\delta_{XZ}(x',z)-p(x',z)}{p(z)}p(y,z)\,dy\,dz = \int L\left(y,\hat{y}(x',Z)p(y|Z)\right)dy - \Psi_0^C(P)$$

Also, we may have:

$$\begin{aligned}&\int L\left(y,\hat{y}(x',z')\right)\frac{\delta_Z(z)-p(z))}{p(z)^2}p(x',z)p(y,z)\,dx'\,dy\,dz\,dt \\ &= \int L\left(y,\hat{y}(x',Z)\right)p(x'|Z)p(y|Z)dx\,dy - \Psi(P) \\ &= \int L\left(y,\hat{y}(x,Z)\right)p(x|Z)p(y|Z)dx\,dy - \Psi_0^C(P)\end{aligned}$$

Lastly, we then have:

$$\begin{aligned}&\int L\left(y,\hat{y}(x,z)\right)\frac{p(x',z)}{p(z)}\left(\delta_{Y,Z}(y,z)-p(y,z)\right)dx\,dy\,dz \\ &= \int L\left(Y,\hat{y}(x',Z)\right)p(x'|Z)dx' - \Psi(P) \\ &= \int L\left(Y,\hat{y}(x,Z)\right)p(x|Z)dx - \Psi_0^C(P)\end{aligned}$$

Putting the three terms together, we shall have:

$$
\begin{aligned}
R_2 = & \int L\left(y, \hat{y}(x', Z) p(y|Z)\right) dy \\
& - \int L\left(y, \hat{y}(x, Z)\right) p(x|Z) p(y|Z) dx\, dy \\
& + \int L\left(Y, \hat{y}(x, Z)\right) p(x|Z) dx - 2\Psi_0^C(P)
\end{aligned}
$$

Putting everything together, we may then obtain the desired result. □

Lemma 7. *Let*

$$\Psi_0^{\pi L}(X, Y, Z) = \mathbb{E}\left[L(Y, \hat{y}(X^\pi, Z))\right].$$

The efficient influence function is:

$$
\begin{aligned}
\psi_0^{\pi L}(X, Y, Z) = & (Y - \hat{y}(X, Z)) \int L'(y, \hat{y}(X, Z)) \frac{P(X)P(y, Z)}{P(X, Z)} dy \\
& + \int L(Y, \hat{y}(x', Z)) P(x') dx' \\
& + \int L(y, \hat{y}(X, Z)) P(y, Z) dy dz - 2\Psi_0^{\pi L}(P),
\end{aligned}
$$

where $X^\pi \sim X$ and $X^\pi \perp X$.

Proof. Using a similar approach as Lemma 5, we have

$$
\begin{aligned}
\psi^{\pi L}(X, Y, Z) = & \int \psi^{\hat{y}(x', z)(X, Z)} L'(y, \hat{y}(x', z)) P(x') P(y, z) dx' dy dz \\
& + \int L(Y, \hat{y}(x', Z)) P(x') dx' + \int L(y, \hat{y}(X, z)) P(y, z) dy dz - 2\Psi^{\pi L}(P) \\
= & (Y - \hat{y}(X, Z)) \int L'(y, \hat{y}(X, Z)) \frac{P(X)P(y, Z)}{P(X, Z)} dy \\
& + \int L(Y, \hat{y}(x', Z)) P(x') dx' + \int L(y, \hat{y}(X, z)) P(y, z) dy dz - 2\Psi^{\pi L}(P)
\end{aligned}
$$

□

B.3 ADDITIONAL ASSUMPTIONS

We note that Assumptions 5 below and Assumption 4 essentially play the same role in eliminating the empirical process term. In Assumption 4, we used an additional share of data I_3 to ensure the independence of the efficient influence function and the final estimator. Though this is different from the classical approach described by Chernozhukov et al. [2018], we note that our method is iterative, whereas theirs is a one-step method. And Assumption 5 is a replicate Assumption A4 of van der Laan et al. [2011a]. Here we define $\overrightarrow{\epsilon}_n^{k_0}$ to be the sequence of ε_n^j, padded with zeros if needed to create a k_0 vector.

Assumption 5 (Donsker Condition; A2 of Theorem 5 in van der Laan et al. [2011a])**.** *Let $\varepsilon_{k_0}^*$ be the limit of $\overrightarrow{\epsilon}_n^{k_0}$, that is, $\overrightarrow{\epsilon}_n^{k_0} \xrightarrow{P} \varepsilon_{k_0}^*$. Condition on P_{n,I_2} and consider a class of measurable functions f estimated on I_1:*

$$\mathcal{F}(P_{n,I_2}) \equiv \left\{\psi(\hat{f}_{I_1}, P_\varepsilon) - \psi(f^*, P_{\varepsilon_{k_0^*}}) : \varepsilon\right\},$$

where the set over which ϵ varies is chosen so that it is a subset of $\mathbb{R}^{k_0}$ and contains $\overrightarrow{\epsilon}_n^{k_0}$ with probability tending to 1. Define the subclasses

$$\mathcal{F}_{\delta_n}(P_{n,I_2}) \equiv \left\{f_\epsilon \in \mathcal{F}(P_{n,I_2}) : \|\epsilon - \varepsilon_{k_0}^*\| < \delta_n\right\}.$$

If for deterministic sequence $\delta_n \to 0$, we have

$$E\left\{\textit{Entro}(\mathcal{F}_{\delta_n}(P_{n,I_2}))\sqrt{P^*F(\delta_n, P_{n,I_2})^2}\right\} \to 0 \quad \textit{as} \quad n \to \infty,$$

where $F(\delta_n, P_{n,I_2})$ is the envelope of $\mathcal{F}_{\delta_n}(P_{n,I_2})$ and Entro$(\mathcal{F}_{\delta_n}(P_{n,I_2}))$ is the entropy of $\mathcal{F}_{\delta_n}(P_{n,I_2})$.

This condition is the same as A2 given in van der Laan et al. [2011a], to which we refer the reader for further details.

B.4 PROOF OF THEOREM 1

Proof. If Assumptions 1,2,3 and 4 are satisfied, this is exactly the same result as Theorem 5 of van der Laan et al. [2011a], and so will be the proof.

If Assumptions 1,2,3 and 5 are satisfied, the only thing we need to do is create a similar lemma as Lemma 2 of van der Laan et al. [2011a]. We can start by considering the empirical process term, that is

$$(\mathbb{P}_{n,I_3} - \mathbb{P})\left(\psi(\hat{f}_{I_1}, P_{\overrightarrow{\varepsilon}_n^{k_n}}) - \psi(f^*, P^*)\right)$$

For the conditional variance of the term on I_3, we have:

$$\begin{aligned} var\left((\mathbb{P}_{n,I_3} - \mathbb{P})\left(\psi(\hat{f}_{I_1}, P_{\overrightarrow{\varepsilon}_n^{k_n}}) - \psi(f^*, P^*)\right)\right) &= var\left(\mathbb{P}_{n,I_3}\left(\psi(\hat{f}_{I_1}, P_{\overrightarrow{\varepsilon}_n^{k_n}}) - \psi(f^*, P^*)\right)\right) \\ &= \frac{1}{n} var(\psi(\hat{f}_{I_1}, P_{\overrightarrow{\varepsilon}_n^{k_n}}) - \psi(f^*, P^*)) \\ &\leq \frac{1}{n}\|\psi(\hat{f}_{I_1}, P_{\overrightarrow{\varepsilon}_n^{k_n}}) - \psi(f^*, P^*)\| \\ &= o_P(1/n) \end{aligned}$$

Then, by Chebyshev's inequality, we have:

$$(\mathbb{P}_{n,I_3} - \mathbb{P})\left(\psi(\hat{f}_{I_1}, P_{\overrightarrow{\varepsilon}_n^{k_n}}) - \psi(f^*, P^*)\right) = O_p\left(\sqrt{\frac{1}{n}\|\psi(\hat{f}_{I_1}, P_{\overrightarrow{\varepsilon}_n^{k_n}}) - \psi(f^*, P^*)\|}\right)$$

We can then obtain the desired result by Assumption 4, that is:

$$(\mathbb{P}_{n,I_3} - \mathbb{P})\left(\psi(\hat{f}_{I_1}, P_{\overrightarrow{\varepsilon}_n^{k_n}}) - \psi(f^*, P^*)\right) = o_P(1/\sqrt{n})$$

The rest of the proof follows in the same manner as Theorem 5 of van der Laan et al. [2011a]. □