Variable importance for causal forests

BKT Workshop in Boston University

Tomoshige Nakamura

June 29, 2023

Faculty of Health Data Science, Juntendo University

Table of contents

Introduction

Background

- Recently, the estimation of conditional average treatment effect using machine learning algorithms has been actively researched (Kunzel et al., 2019).

$$\tau(x) = \mathrm{E}[\, Y_{a=1} - \, Y_{a=0} \mid X = x]$$

- For example, the following methods have been proposed:
  - Causal forests (Athey et al., 2018; Wager and Athey, 2019)
  - Bayesian additive regression trees (Hill, 2011; Hahn et al., 2020)
  - Neural networks (Syrgkanis et al., 2019)
- In the problems of estimating causal effect, ML approach have often shown better results compared to cases where researchers fit parametric or nonparametric models (Dorie et al., 2019).

Background : Interpretability of estimation result

- The problem work with machine learning is that they are a black box approach because the estimator is composed by many operators.
- Because of lack of interpretability of estimation result, there may be cases that prevent to apply machine learning approach in practice.
- **Variable importance (also known as feature importance)** is a score that indicates how ”important” a feature is to the model.
- One of the popular use of variable importance is for Random forest that is a criterion for measuring the contribution of variables to predictions.
- Variable importance generally uses two measures: **Mean Decrease Accuracy (MDA)** and **Mean Decrease Impurity (MDI)**.

The aim of this presentation

(1) Propose a **variable importance measure for Generalized random forests (GRF; Athey et al., 2019)** extending exist variable importance measure defined for random forests.

(2) Propose a variable importance for causal forest which is a one example of (1).

(3) Perform the simulation and evaluate empirical performance of proposed method.

Remark
Causal forests that is the method to estimate CATE is the one of the examples of generalized random forests.

What is Variable Importance ?

In the context of regression problems Y on X, variable importance is defined as the difference of explained variance with

$$VI(X^{(j)}) = \int_{x,y} \left\{ y - m^{(-j)}(x^{(-j)}) \right\}^2 - \{y - m(x)\}^2 f_{X,Y}(x,y)\,dxdy$$

where

$$m(x) = \mathrm{E}[\,Y \mid X = x] \quad \text{and} \quad m^{(-j)}(x^{(-j)}) = \mathrm{E}[m(x) \mid X^{(-j)} = x^{(-j)}]$$

Variable importance is a total Sobol index (Bénard et al 2021) which is defined as

$$VI(X^{(j)}) = \frac{\mathrm{Var}(m(X)) - \mathrm{Var}(\mathrm{E}[m(X)|X^{(-j)}])}{\mathrm{Var}(\,Y)}$$

Variable Importance for random forest

In random forest, variable importance for variable $X^{(j)}$ is defined by replacing $m(x)$ with random forest estimates $f^{(RF)}(\cdot)$:

$$\widehat{VI}(X^{(j)}) = \int_{x,y} \left\{ y - \hat{f}^{(RF)}(x^{(-j)}) \right\}^2 - \left\{ y - \hat{f}^{(RF)}(x) \right\}^2 f_{X,Y}(x,y)\, dxdy$$

where

- $f^{(RF)}(x)$ is a random forest estimates of $\mathrm{E}[Y|X]$ using all observed variables.
- $f^{(RF)}(x^{(-j)})$ is a random forest estimates of $\mathrm{E}[Y|X^{(-j)}]$ using all observed variables without $X^{(j)}$.

However, if compute VIs for all p variables, we need to fit random forests $p+1$ times. To prevent this, two types of VI estimates are usually used.

Variable Importance for random forest

- (a) Permutation type Variable importance is defined as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\hat{f}^{(RF,OOB)}(X_{i,\pi_j})\right\}^2-\left\{Y_i-\hat{f}^{(RF,OOB)}(X_i)\right\}^2$$

- where $\hat{f}^{(RF,OOB)}(\cdot)$ is the Out-of-bags random forest predictor and X_{i,π_j} be the variable obtained by permuting the j-th element of vector X_i among the observed data.
- (b) Tree randomization type variable importance is defined as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\hat{f}^{(RF,OOB,-j)}(X_i)\right\}^2-\left\{Y_i-\hat{f}^{(RF,OOB)}(X_i)\right\}^2$$

- where $\hat{f}^{(RF,OOB),(-j)}(X_i)$ is the random forest predictor with noise up trees which defined as later.

(Remark) Variable Importance that uses in Python modules and R packages

- VIs implemented in Each package/module are different.
- In Python's {scikit-learn}, it is implemented as a method called Train-test MDA.
- The R {randomForest} package uses a variable importance based on the Tree estimator for Out-of-bags samples (BC-MDA) proposed by Breiman (2001).
- The R {randomforestSRC} package uses variable importance based on the Random Forest estimator for Out-of-bags samples (IK-MDA) (Ishwaran, 2007).

In terms of accuracy, IK-MDA is generally superior to both Train-test MDA and BC-MDA, so here we extended a method based on IK-MDA to GRF.

Generalized random forest

Problem Setting

- Let $(X_i, O_i) \in \mathcal{X} \times \mathcal{O}, (i = 1, 2, ..., n)$ be i.i.d. observations from $(X, O) \in \mathcal{X} \times \mathcal{O}$
- $X = (X^{(1)}, X^{(2)}, ..., X^{(p)})$ is a features
- Variable O depends on the problem setting:
- In a regression problem, $O = \{Y\}$, where Y is outcomes.
- In a causal inference problem, $O = \{Y, A\}$, where A is the binary treatment and Y is outcomes.

Local estimating equation

Consider the problem of estimating a functional parameter $\theta(x)$, which is defined as the solution to the following local estimation equation.

Definition: Local Estimating Equation

$$\mathrm{E}[\psi_{\theta(x),\nu(x)}(O_i)|X_i = x] = 0 \tag{1}$$

Here, $\psi(\cdot)$ is the score function and $\nu(x)$ is the nuisance parameter.

Local estimating equations - Examples (1)

- In the case of a general random forest: setting $O_i = \{ Y_i \}$ and the score function as

$$\psi_{\mu(x)}(Y_i) = Y_i - \mu(x) \tag{2}$$

- In fact, the estimator of a random forest is the solution of the following equation:

$$\sum_{i=1}^{n} \alpha_i(x)(Y_i - \mu(x)) = 0 \tag{3}$$

Local estimating equations - Examples (2)

- In the case of q **quantile regression model**, setting $O_i = \{ Y_i \}$, we have

$$\psi_{\theta(x)}(Y_i) = q \cdot 1\{ Y_i > \theta(x)\} - (1-q)1\{ Y_i \leq \theta(x)\}$$

- In the case of **regression model with instrumental variables**: $O_i = \{ Y_i, W_i, Z_i\} \in \mathbb{R} \times \{0,1\} \times \{0,1\}$, where Z_i is an instrumental variable, under the assumption that $Z_i \perp\!\!\!\perp \varepsilon_i | X_i$ and $\text{Cov}(Z_i, W_i | X_i) \neq 0$, we have

$$\psi_{\tau(x),\mu(x)}(O_i) = \{ Y_i - W_i\tau(x) - \mu(x)\} \begin{pmatrix} 1 \\ Z_i \end{pmatrix}$$

Generalized random forest

The solution $(\theta(x), \nu(x))$ to the estimating equation (1) is generally estimated as the solution to the kernel-weighted estimating equation.

$$\left(\hat{\theta}(x), \hat{\nu}(x)\right) \in \text{argmin}_{\theta,\nu} \left\{ \left\| \sum_{i=1}^{n} \alpha_i(x) \psi_{\theta,\nu}(O_i) \right\|_2 \right\} \tag{4}$$

The Generalized Random Forest is a framework that estimates the weights $\alpha_i(x)$ using a random forest.

These $\alpha_i(x)$ are called **forest-weights**.

Estimation of forest-weights $\alpha_i(x)$

- Let $T_b, ; b = 1, 2, .., B$ be a Decision/Regression Trees that are base learner of random forest.
- Let $L_b(x)$ be a set of data $\{X_i, i = 1, 2, ..., n\}$ included in the leaf of Tree T_b that contains point x.

Then, we define the forest-weights $\alpha_i(x)$ as follows.

Definition: forest-weights

$$\alpha_i(x) = \frac{1\{X_i \in L_b(x)\}}{|L_b(x)|}, \quad \alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \alpha_{bi}(x) \tag{5}$$

- $\alpha_i(x)$ represents the strength of the relationship between point x and training data X_i.
- To compute $\alpha_i(x)$, we use a Gradient Tree (Athey et al., 2019).

Example: Forest Weights

- The samples that are weighted when predicting the red $\times$ are shown.
- In GRF, the weights $\alpha_i(x)$ are computed for all samples i, and then weighted estimating equation 4 is solved.
- In other words, GRF estimates the kernel weighting function nonparametrically using a forest.

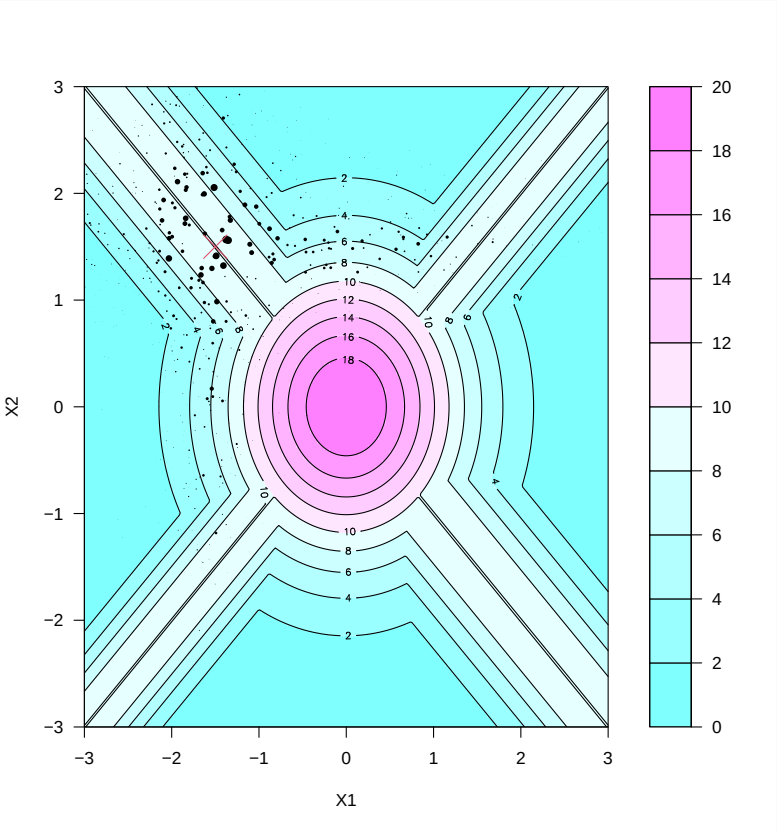


Figure 1: Forest Weights

Variable Importance for Generalized random forests

Projected local estimating equation

To define Variable importance for generalized random forest, we first define a *projected local estimating equation*.

Def : Projected local estimating equation

For original local estimating equation

$$\mathrm{E}[\psi_{\theta(x),\nu(x)}(O_i)|X_i = x] = 0,$$

define the $X^{(-j)}$-projected local estimating equation as:

$$\mathrm{E}\left[\psi_{\theta^{(-j)}(x^{(-j)}),\nu^{(-j)}(x^{(-j)})}(O_i) \mid X_i^{(-j)} = x^{(-j)}\right] = 0. \tag{6}$$

Under some regularity conditions, the following equation holds.

$$\theta^{(-j)}(x^{(-j)}) = \mathrm{E}[\theta(X) \mid X^{(-j)} = x^{(-j)}]$$

Variable Importance for functional parameter $\theta(x)$

We extend the definition of variable importance measure to functional parameter $\theta(x)$ as:

$$VI(X^{(j)}) = \int_{x,y} \left\{\theta(x) - \theta^{(-j)}(x^{(-j)})\right\}^2 - \{\theta(x) - \theta(x)\}^2 f_{X,Y}(x,y)\,dxdy$$

and define a total Sobol index for functional parameter $\theta(x)$ as

$$\begin{aligned} ST^{(j)} &= \frac{\text{Var}(\theta(X)) - \text{Var}(\text{E}[\theta(X)|X^{(-j)}])}{\text{Var}(\theta(X))} \\ &= \frac{\text{Var}(\theta(X)) - \text{Var}\left(\theta^{(-j)}(X^{(-j)})\right)}{\text{Var}(\theta(X))} \end{aligned}$$

- $ST(X^{(j)})$ is the amount of explained output variance lost when $X^{(-j)}$ is removed from the model.
- This quantity is the information that one finds a small group of the most predictive covariates.

VI for generalized random forests estimator

We replace $\theta^{(-j)}(x^{(-j)})$ in the previous slide with GRF estimator $\hat{\theta}^{(-j)}(x^{(-j)})$ and define the variable importance measure for GRF as:

$$\widehat{VI}(X^{(j)}) = \int_{x,y} \left\{\theta(x) - \hat{\theta}^{(-j)}(x^{(-j)})\right\}^2 - \left\{\theta(x) - \hat{\theta}(x)\right\}^2 f_{X,Y}(x,y)\,dxdy$$

However, as a random forest, to compute $VI(X^{(j)})$ for all $j = 1, 2, ..., p$ is computationally (more) expensive compared to RF.

Alternative approach are permutation and noise-up.

Remark
GRF estimator $\hat{\theta}(x)$ has the consistency for $\theta(x)$, then $\widehat{VI}(X^{(j)})$ converge to $VI(X^{(j)})$ in probability.

out-of-bags forest-weights

- Let $\mathcal{D}_n = \{(X_i, O_i), i = 1, 2, ..., n\}$ be observations.
- Let $S^{(b)}$ $(b = 1, 2, ..., B)$ be a random subset of $\mathcal{D}_n$ with size s_n.
- Let T_b be a fitted gradient trees on $S^{(b)}$, $(b = 1, 2, ..., B)$.
- Define $\Lambda_{i'}$ as the index set of trees that do not include the sample $i' \in \{1, 2, ..., n\}$ in $S^{(b)}$:

$$\Lambda_{i'} = \left\{ b \in \{1, 2, ..., B\} \mid (X_{i'}, O_{i'}) \neq S^{(b)} \right\} \tag{7}$$

- For each $b \in \Lambda_{i'}$, define $L_b(X_{i'})$ be the leaf of tree T_b that contains $X_{i'}$.
- This means the leaf $L_b(x)$ is the subset of feature space $\mathcal{X}$ that contains x which generated by tree T_b.

Def: Out-of-bags forest Weights

Def: out-of-bags forest weights (OOB-FW)

Define the **out-of-bags forest weights (OOB-FW)** for the sample i' as:

$$\alpha_i^{(OOB)}(X_{i'}) = \frac{1}{|\Lambda_{i'}|} \sum_{b \in \Lambda_{i'}} \alpha_{bi}(X_{i'}), \quad \alpha_{bi}(X_{i'}) = \frac{1\{X_i \in L_b(X_{i'})\}}{|L_b(X_{i'})|} \tag{8}$$

OOB-FW is defined through only trees T_b which does not use sample $i' \in \{1, 2, ..., n\}$ for learning.

Def: out-of-bags generalized random forest estimator

Def: Out-of-bags GRF Estimator

Define the solution of (4) under OOB-FW as **Out-of-bags GRF estimator (OOB-GRF estimator)** for $(\theta(X_{i'}), \nu(X_{i'}))$:

$$\left(\hat{\theta}^{(OOB)}(X_{i'}), \hat{\nu}^{(OOB)}(X_{i'})\right) \in \underset{\theta,\nu}{\operatorname{argmin}} \left\{ \left\| \sum_{i=1, i\neq i'}^{n} \alpha_i^{(OOB)}(X_{i'}) \psi_{\theta,\nu}(O_i) \right\|_2 \right\} \tag{9}$$

Out-of-bags GRF estimators $\left(\hat{\theta}^{(OOB)}(X_{i'}), \hat{\nu}^{(OOB)}(X_{i'})\right)$ does not depend on sample i' itself.

Permutation type out-of-bags GRF estimators

- Let X_{i,π_j} be the variable obtained by permuting the j-th element of vector X_i among the observed data.
- Let X_{π_j} be the random variable vector obtained by replacing the j-th element of the vector X with a random variable following the distribution of $X^{(j)}$.

Def: Permutation type out-of-bags GRF Estimator

We define permutation type out-of-bags GRF Estimator for $(\theta(X_{i'}), \nu(X_{i'}))$ as follows.

$$\left(\hat{\theta}(X_{i',\pi_j}), \hat{\nu}(X_{i',\pi_j})\right) \in \mathrm{argmin}_{\theta,\nu} \left\{ \left\| \sum_{i=1, i\neq i'}^{n} \alpha_i^{OOB}(X_{i',\pi_j}) \psi_{\theta,\nu}(O_i) \right\|_2 \right\} \tag{10}$$

Permutation type Variable importance for GRF

- Substitute permutation type out-of-bags GRF Estimator to the definition of VI of GRF, Permutation type VI for GRF can be defined.

Permutation type VI for GRF with respect to $X^{(j)}$

$$\widehat{\mathrm{VI}}^{(P)}(X^{(j)}) = \frac{1}{N_{B,n}} \sum_{i=1}^{n} \left\{ \theta(X_i) - \hat{\theta}^{(OOB)}(X_{i,\pi_j}) \right\}^2 - \left\{ \theta(X_i) - \hat{\theta}^{(OOB)}(X_i) \right\}^2 \tag{11}$$

where $N_{B,n} = \sum_{i=1}^{n} 1\{|\Lambda_i| > 0\}$.

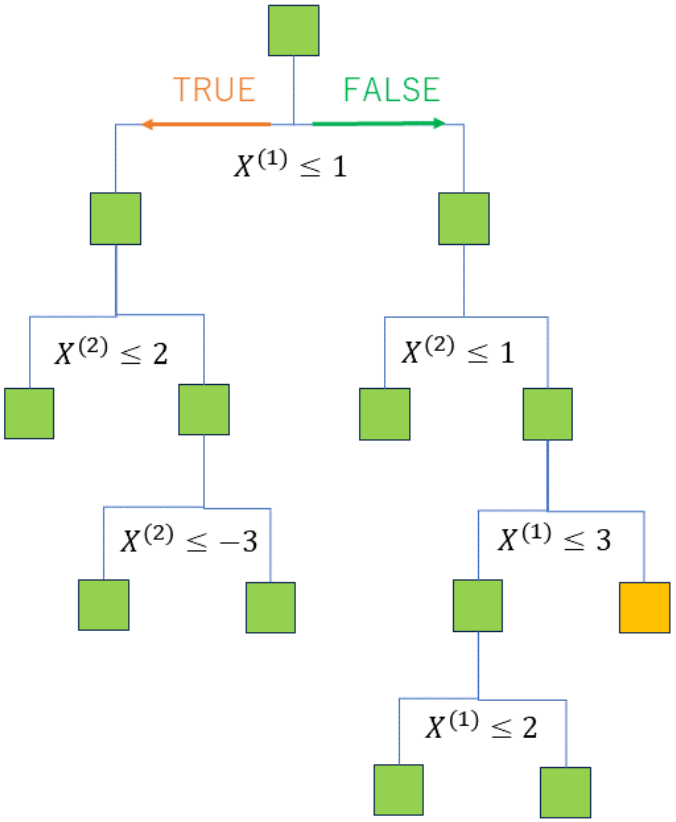
Noise-up Gradient tree ans its out-of-bags estimator

- Let $X^{(-j)}$ be the random variable obtained by removing the j-th element from the random variable vector X.

$$X^{(-j)} = (X^{(1)}, ..., X^{(j-1)}, X^{(j+1)}, ..., X^{(p)})$$

- The Noise-up Tree for the variable $X^{(j)}$ is a method to marginalize the Tree estimator T_b with respect to $X^{(j)}$ by making all subsequent divisions random whenever the split rule contains $X^{(j)}$ in T_b while searching for the leaf containing the point x (Ishwaran, 2007).
- We define a Noise-up gradient tree $T_b^{(-j)}(x)$ that adapt previous procedure on a gradient tree.
- Then we can define Noise-up out-of-bags forest weights using noise-up tree $T_b^{(-j)}(x)$ for $X^{(j)}$, and denoted by $\alpha_i^{(OOB,-j)}(X_i)$ by same procedure for permutation type weights.

(Image) Noise-up Gradient tree ans its out-of-bags estimator



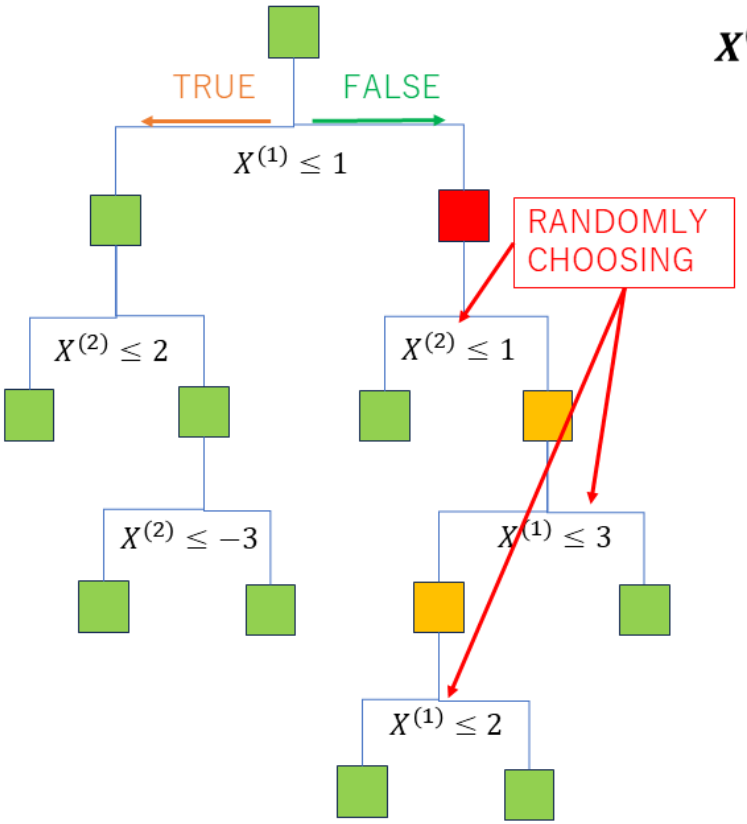
Tree prediction

When we estimate tree estimator for test point $x \in \mathcal{X}$, we drop the sample from the root node and then follows to splitting rules of descending branches.

Left figure is the example of a tree with two dimensional feature variables $X \in (X^{(1)}, X^{(2)})$

Let consider the predictor for test point $x \in (4,0)$. In this case, x contains in the leaf with orange colors.

(Image) Noise-up Gradient tree ans its out-of-bags estimator



$X^{(j)}$ Noise-Up tree prediction

To assign an leaf value to test point x,
x follow tree path until
(a) x reaches a node that has a branch dependent on $X^{(j)}$, or
(b) x reaches the terminal node without encountering a node that has a branch dependent on $X^{(j)}$.

In the case (a), choose the left or right child node with equal probability.

Then, descend the tree to the terminal node, randomly choosing the left or right child node at each subsequent branch, regardless of whether that branch depends on $X^{(j)}$.

So, $X^{(2)}$-Noise-up Tree predictor for $x = (4,0)$, at red branch, random choosing is occurred.

Noise-up Gradient tree ans its out-of-bags estimator

- This noise-up mechanism is designed to deteriorate the terminal prediction value when passing through a node that branches on $X^{(j)}$.
- Let's denote the tree after noise-up on T_b as $T_b^{(-j)}$.
- The predictive performance of $T_b^{(-j)}$ is closely related to the Variable Importance (VI) of $X^{(j)}$, which is tightly connected to the split position of xv in T.
- The more information $X^{(j)}$ holds, the $X^{(j)}$ split appears in near the root node (at a shallow location) of the tree T_b
- $T_b^{(-j)}$ have worse accuracy compared to T_b. As a result, the VI of $X^{(j)}$ increases.

Out-of-bags Noise-Up GRF Estimator

Def: out-of-bags Noise-Up GRF Estimator

We define the out-of-bags Noise-Up GRF estimator for $(\theta(X_i'), \nu(X_i'))$ as follows.

$$\left(\hat{\theta}^{(OOB,-j)}(X_{i'}), \hat{\nu}^{(OOB,-j)}(X_{i'})\right) \in \operatorname{argmin}_{\theta,\nu} \left\{ \left| \sum_{i=1, i\neq i'}^{n} \alpha_i^{(OOB,-j)}(X_{i'}) \psi_{\theta,\nu}(O_i) \right|_2 \right\} \tag{12}$$

Noise-up type variable importance for GRF

Def: Noise-Up type variable importance for GRF with respect to $X^{(j)}$

We define the Noise-up variable importance for the GRF estimator with respect to variable $X^{(j)}$ as follows.

$$\widehat{\mathrm{VI}}^{(NU)}(X^{(j)}) = \frac{1}{N_{B,n}} \sum_{i=1}^{n} \left\{ \theta(X_i) - \hat{\theta}^{(OOB,-j)}(X_i) \right\}^2 - \left\{ \theta(X_i) - \hat{\theta}^{(OOB)}(X_i) \right\}^2 \tag{13}$$

Approximate variable importance measure for GRF

Approximate variable importance measure for GRF

- In the definition of two VIs contains the true $\theta(X_i)$, So this estimator is computational infeasible.
- (Remark) original VIs for random forest are able to compute because it defined only through observed Y_i.
- We approximate two types of VIs for GRF that be able to compute from observed data.
- By following theorem, mean squared error for parameter $\theta(x)$ is approximated by score function.

Proposal : Permutation type approximate variable importance

Here we define the approximate Permutation MDA and Noise-up MDA for the variable $X^{(j)}$. (For clarity, superscript (OOB) are omitted)

Approximated Permutation type VI for $X^{(j)}$

$$\widehat{\mathrm{AVI}}^{(P)}(X^{(j)}) = \frac{1}{N_{B,n}} \sum_{i=1}^{n} \left\{ \rho^2(X_{i,\pi_j}) - \rho^2(X_i) \right\} \tag{14}$$

where

$$\rho(X_i) = \sum_{k=1, k \neq i}^{n} \alpha_k(X_i) \xi^T \widehat{V}^{-1}_{\hat{\theta}(X_i), \hat{\nu}(X_i)}(X_i) \psi_{\hat{\theta}(X_i), \hat{\nu}(X_i)}(O_k)$$

Proposal : Noise-up type approximate variable importance

Approximated Noise-up type VI for $X^{(j)}$

$$\widehat{\mathrm{AVI}}^{(NU)}(X^{(j)}) = \frac{1}{N_{B,n}} \sum_{i=1}^{n} \left\{ \left\{ \rho^{(-j)}(X_i^{(-j)}) \right\}^2 - \rho^2(X_i) \right\} \tag{15}$$

where

$$\rho^{(-j)}(X_i^{(-j)}) = \sum_{k=1, k\neq i}^{n} \alpha_k(X_i) \xi^T \{ \widehat{V}^{(-j)}(X_i) \}^{-1} \psi_i^{(-j)}(O_k)$$

and

$$\widehat{V}^{(-j)}(X_i) = \widehat{V}_{\hat{\theta}^{(-j)}(X_i^{(-j)}), \hat{\nu}^{(-j)}(X_i^{(-j)})}(X_i)$$

$$\psi_i^{(-j)}(O_k) = \psi^{(-j)}_{\hat{\theta}^{(-j)}(X_i^{(-j)}), \hat{\nu}^{(-j)}(X_i^{(-j)})}(O_k)$$

Theorem : Approximate VIs converge to original VI

Theorem (Nakamura, 2023)

Under some regularity conditions, including assumptions in Athey et al.,(2019), estimation error of $\widehat{\mathrm{AVI}}^{(P)}(X^{(j)}) - \widehat{\mathrm{VI}}^{(P)}(X^{(j)})$ goes to zero, that is:

$$\widehat{\mathrm{AVI}}^{(P)}(X^{(j)}) - \widehat{\mathrm{VI}}^{(P)}(X^{(j)}) \xrightarrow{P} 0 \tag{16}$$

and

$$\widehat{\mathrm{AVI}}^{(NU)}(X^{(j)}) - \widehat{\mathrm{VI}}^{(NU)}(X^{(j)}) \xrightarrow{P} 0 \tag{17}$$

This result follows from next Lemma.

Lemma 1 : Approximate mean squared error of GRF

Lemma

Assuming that GRF satisfies the assumption of Athey et al., (2019), let $\hat{\theta}(x)$ be the solution of GRF for the point x for the score function $\psi_{\theta,\nu}(O)$, subsample size s, regularization parameter $\omega < 0.2$, random partition parameter $\pi > 0$, then

$$\left\|\hat{\theta}(x) - \theta(x)\right\|_2^2 = \left(\sum_{i=1}^{n} \alpha_i(x) \xi^T \hat{V}_{\theta,\nu}(x)^{-1} \psi_{\theta(x),\nu(x)}(O_i)\right)^2 \tag{18}$$

$$+ \mathcal{O}_P\left\{\max\left(\frac{s^{1-\pi\cdot\frac{\log((1-\omega)^{-1})}{\log(\omega^{-1})}}}{n}, \left(\frac{s}{n}\right)^{\frac{4}{3}}\right)\right\} \tag{19}$$

holds, where

$$\hat{V}_{\theta,\nu}(x) := \sum_{i=1}^{n} \alpha_i(x) \nabla M_{\theta,\nu}(x), \quad M_{\theta,\nu}(x) = \mathrm{E}[\psi_{\theta,\nu}(O)|X_i = x]$$

Proposing Approximate Variable Importance

- From the theorem, two types of VIs (11) and (13) can be approximated by only through the score function $\psi_{\theta,\nu}(O_i)$ without unobserved true parameters $\theta(x)$.
- Each of AVI are generalizations of variable importance for random forest.
- The following theorem shows $\widehat{AVI}$ does not always converges to total Sobol index $ST^{(j)}$.

Theorem : Convergence of Permutation type AVI

Theorem (Nakamura, 2023)

$$\widehat{\mathrm{AVI}}^{(P)}(X^{(j)}) \xrightarrow{L_1} \mathrm{E}\left[\left\{\theta(X) - \mathrm{E}[\theta(X_{\pi_j})|X^{(-j)}]\right\}^2\right]$$

and right term can be decomposed to total Sobol index and noise,

$$\mathrm{E}\left[\left\{\theta(X) - \mathrm{E}[\theta(X_{\pi_j})|X^{(-j)}]\right\}^2\right] = \mathrm{Var}(\theta(X)) \times \mathrm{ST}^{(j)} + \delta \qquad (20)$$

where δ can be expressed as

$$\delta = \mathrm{E}\left[\left\{\mathrm{E}[\theta(X)|X^{(-j)}] - \mathrm{E}[\theta(X_{\pi_j})|X^{(-j)}]\right\}^2\right]$$

This theorem follows from Lemma 2.

Lemma 2 : Approximation of MSE by OOB-GRF estimator

Lemma

Under the assumption of Athey et al. (2019), for fixed sample size n and subsample size s_n, and the number of Gradient trees constituting the GRF are B, for $i \in 1, 2, ..., n$, the following relation holds.

$$\left\| \mathrm{E}\left[\left\{ \hat{\theta}^{OOB}_{B,s_n,n}(X_i) - \theta(X_i) \right\}^2 \right] - \mathrm{E}\left[\{\theta_{B,s_n,n}(X) - \theta(X)\}^2 \right] \right\| = O\left(\frac{1}{B}\right)$$

where $\hat{\theta}^{OOB}_{B,s_n,n}(X_i)$ is OOB estimator and $\theta_{B,s_n,n}(X)$ is theoretical random forest.

This lemma shows that the average prediction error of a random forest for test point $x \in \mathcal{X}$ can be approximated by the out-of-bag estimates. Furthermore, as the number of trees increases, then the difference of them are vanished.

Permutation type AVI convergence under some settings

Corollary 1: X is independent

If covariates X are independent then,

$$\widehat{\mathrm{AVI}}^{(P)}(X^{(j)}) \xrightarrow{L_1} \mathrm{Var}(\theta(X)) \times \mathrm{ST}^{(j)}$$

Corollary 2: $\theta(x)$ is additive

If $\theta(X)$ is additive, that is $\theta(X) = \sum_j \theta_j(X^{(j)})$

$$\widehat{\mathrm{AVI}}^{(P)}(X^{(j)}) \xrightarrow{L_1} \mathrm{Var}(\theta(X)) \times \mathrm{ST}^{(j)}_{mg}$$

where $ST^{(j)}$ is total Sobol index and $ST^{(j)}_{mg}$ is marginal total sobol index which is defined as:

$$ST^{(j)} = \frac{\mathrm{E}[\mathrm{Var}\{\theta(X)|X^{(-j)}]\}}{\mathrm{Var}(\theta(X))}, \quad ST^{(j)}_{mg} = \frac{\mathrm{E}[\mathrm{Var}(\theta(X_{\pi_j})|X^{(-j)}]}{\mathrm{Var}(\theta(X))}$$

(Remark) What do $ST^{(j)}$ and $ST_{mg}^{(j)}$ mean ?

$ST^{(j)}$ (Total Sobol index)

- This quantity depends on joint distribution of $X = (X^{(j)}, X^{(-j)})$.
- If there are some highly correlated variables with $X^{(j)}$ in $X^{(j)}$, $ST^{(j)}$ becomes small.

$ST_{mg}^{(j)}$ (Marginal total Sobol index)

- This quantity depends on product of distribution

$$f_{X^{(-j)}}(x^{(-j)}) \times f_{X^{(j)}}(x^{(j)})$$

- Even if there are some highly correlated variables with $X^{(j)}$ in $X^{(j)}$, $ST_{mg}^{(j)}$ does not change.

Application to causal inference and simulation

Settings

- As an application of variable importance for GRF, we consider to estimate conditional average causal effect (CATE).
- Here we use R-loss (Nie and Wager, 2021) as a score function that satisfies Neyman orthogonality (Chernozhukov et al., 2018).
- Let $\{ Y_i, A_i, X_i \}$ are observations, where
- $A_i \in \{0, 1\}$: treatment variable
- $Y_i \in \mathbb{R}$: outcome variable
- $X_i \in \mathbb{R}^n$: feature variables
- $Y_{i,a=1}$, $Y_{i,a=0}$: potential outcomes

R-loss

Under the strongly ignorable assignment

$$(Y_{i,a=1}, Y_{i,a=0}) \perp\!\!\!\perp A_i \mid X_i$$

and the propensity score $\pi(x) := \mathrm{E}[A_i | X_i = x]$ satisfies

$$0 < \pi(x) < 1 \quad \text{forall} \quad x \in \mathcal{X}$$

Consider to estimate CATE which is defined as:

$$\tau(x) = \mathrm{E}\left[Y_{a=1} - Y_{a=0} \mid X_i = x \right]$$

R-loss for $\tau(x)$ is given by

$$\psi_{\tau(x)}(O_i) = (A_i - \pi(X_i)) \left\{ (Y_i - m(X_i)) - (A_i - \pi(X_i)) \tau(X_i) \right\} \quad (21)$$

In practice, $m(x)$ and $\pi(x)$ are estimated by cross-fitting.

Setting

Simulation is performed under following setting.

- Feature variables are generated from

$$X^{(j)} \overset{i.i.d.}{\sim} N(0,1) \quad (j = 1, 2, ..., 20)$$

- Treatment variable is generated by

$$P(A = 1|X) = 0.6 \cdot 1\{X_1 > 0\} + 0.4 \cdot 1\{X_1 \leq 0\}$$

- Outcome variable Y is generated by

$$Y = A \cdot \tau(X) + \mu(X) + \varepsilon$$

where $\varepsilon_i \sim N(0,1)$,

$$\tau(X) = 2\min(X^{(1)}, 0) - 2(X^{(2)})^2 + X^{(3)}(1 - X^{(4)})^2$$

and

$$\mu(X) = X^{(5)} + \min(X^{(6)}, 0)$$

Setting

- Data are generated with a sample size of $n = 1000$
- Estimate the causal effect $\tau(x)$ using a GRF with the R-Learner score (21)
- Compute variable importances by $\widehat{\text{AVI}}^{(P)}$ and $\widehat{\text{AVI}}^{(NU)}$, respectively.

Estimation Results

	Permutation	Noise-up
X_1	0.2859	0.1761
X_2	2.3114	2.0635
X_3	2.6034	1.7526
X_4	1.1118	0.7170
X_5	0.0029	-0.0812
X_6	0.0029	-0.0829
X_7	0.0034	-0.0789
X_8	0.0057	-0.0811
X_9	0.0028	-0.0792
X_{10}	0.0005	-0.0839

	Permutation	Noise-up
X_{11}	0.0023	-0.0798
X_{12}	0.0009	-0.0830
X_{13}	0.0023	-0.0835
X_{14}	0.0028	-0.0795
X_{15}	0.0032	-0.0795
X_{16}	0.0034	-0.0804
X_{17}	0.0020	-0.0833
X_{18}	0.0035	-0.0775
X_{19}	0.0045	-0.0821
X_{20}	0.0034	-0.0800

Table 1: Variable importance for conditional causal effects calculated for Permutation and Noise-up

Results

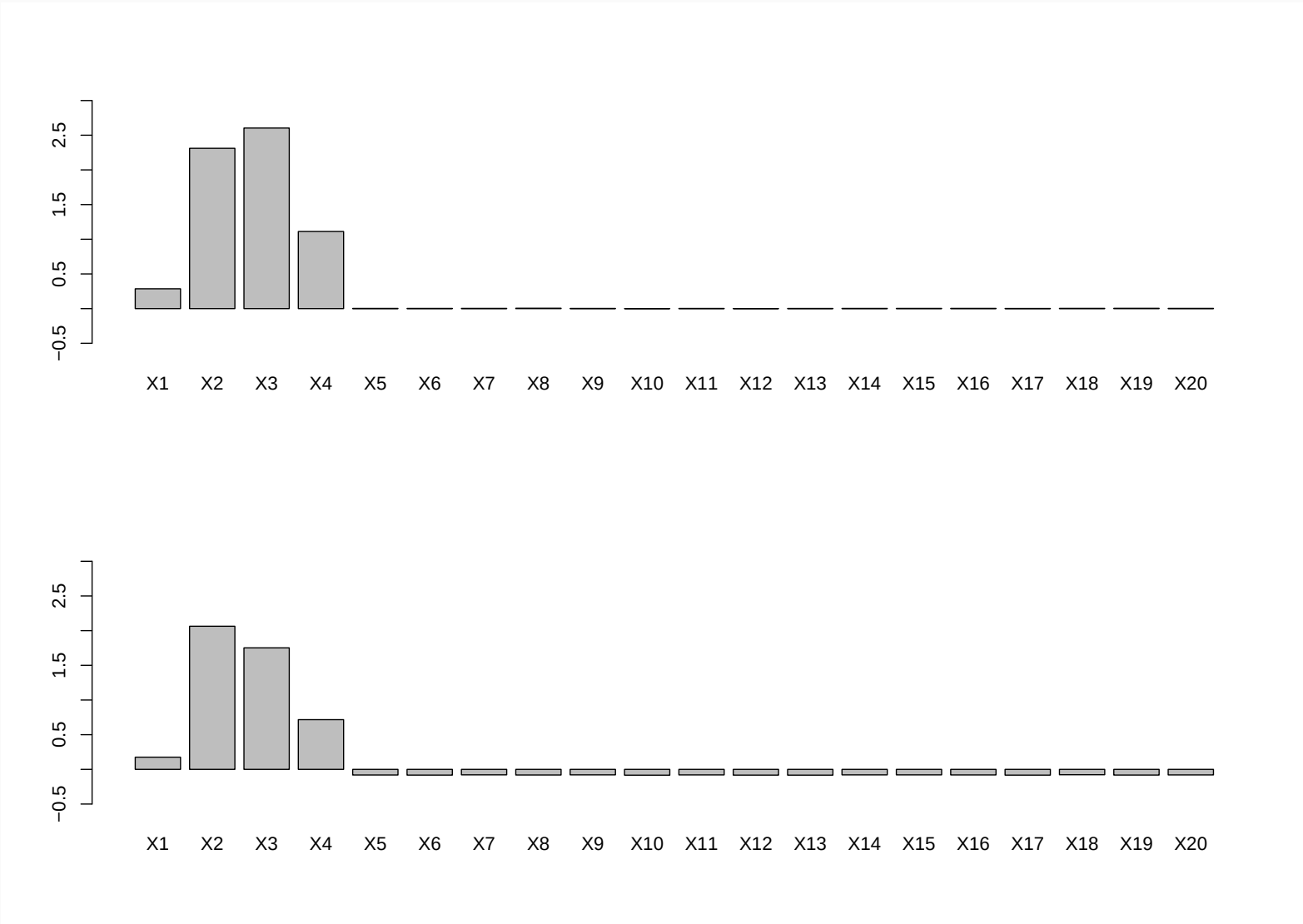


Figure 2: Variable importance using Permutation (upper figure) and Noise-up (lower figure)

Results

- The results of the simulation are shown in Table 1 and Figure 2.
- From these results, it can be seen that the variables $X^{(1)}, ..., X^{(4)}$ included in the conditional causal effect are estimated to have a larger variable importance than other variables not included, regardless of the method used.
- On the other hand, while Permutation estimates $X^{(3)}$ as having the highest variable importance, Noise-up estimates $X^{(2)}$ as the highest, showing a difference.
- The variable importance of Noise-up, although relatively small for the variables $5 \leq j \leq 20$ that are desired to be sufficiently close to 0, is not close enough to 0.

Summary and Future Work

Summary of presentation

- We define a new variable importance measures for generalized random forests and propose approximate variable importance that is not depends on unobserved ground-truth $\theta(x)$
- We show the two types of Approximate variable importance measure $\widehat{\mathrm{AVI}}^{(P)}$ and $\widehat{\mathrm{AVI}}^{(NU)}$ converge to $\widehat{\mathrm{VI}}^{(P)}$ and $\widehat{\mathrm{VI}}^{(NU)}$.
- As a specific application, we proposed variable importance for conditional causal effects and We demonstrated the effectiveness of the proposed method through simulation.

Future Work

- It is known that Permutation MDA for random forest does not work well when there is correlation between features or when the functional parameter includes interactions. Also, Noise-up MDA has a problem in terms of accuracy.
- We plan to resolve these issues by extending Projected CART by B'enard et al (2022) to gradient tree.
- Implementing the algorithm in C++, and developing R packages and Python modules.

References

References (in part)

- Susan Athey, Julie Tibshirani, and Stefan Wager. ”Generalized random forests”. The Annals of Statistics, 47(2):1148 - 1178, 2019
- Clément Bénard et al. ”Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA”. Biometrika, 109(4):881-900, 2022.
- Victor Chernozhukov et al. ”Double/debiased machine learning for treatment and structural parameters”. The Econometrics Journal, 21(1):C1-C68, 2018.
- Vincent Dorie et al. ”Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition”. Statistical Science, 34(1):43 - 68, 2019.
- Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1):217-240, 2011.
- Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika, 108(2):299-319, 2021.

Thank you for Listening !

Have a nice day !

Appendix

Gradient Trees

The tree estimating $\theta(x)$, which is defined as the solution of the estimation equation, is called a *Gradient Tree* (Athey et al., 2019). A Gradient Tree is a recursive partitioning algorithm that divides nodes by focusing on the heterogeneity of $\theta(x)$.

1. Labeling step: Using the data of parent node P, we estimate $\hat{\theta}_P$ and $\hat{\nu}_P$.

$$(\hat{\theta}_P, \hat{\nu}_P) \in \text{argmin}_{\theta,\nu} \left\{ \left\| \sum_{i:X_i \in P} \psi_{\theta,\nu}(O_i)) \right\|_2 \right\} \tag{22}$$

   We then define Γ_P as the consistent estimator for the derivative of the score function, $\nabla \mathrm{E}[\psi_{\hat{\theta}_P,\hat{\nu}_P} | X_i \in P]$. For example,

$$\Gamma_P = \frac{1}{|i : X_i \in P|} \sum_{\{i:X_i \in P\}} \nabla \psi_{\hat{\theta}_P,\hat{\nu}_P}(O_i). \tag{23}$$

   Using these, we construct pseudo-outcomes.

$$\rho_i = -\xi^T \Gamma_P^{-1} \psi_{\hat{\theta}_P,\hat{\nu}_P}(O_i) \in \mathbb{R} \tag{24}$$

Gradient Trees

2. Regression step: We partition the pseudo-outcome ρ_i just like CART does. That is, we divide the parent node P into child nodes C_1 and C_2 by using the variable X as a criterion, in order to maximize the following criterion.

$$\Delta(C_1, C_2) = \sum_{j=1}^{2} \frac{1}{|i : X_i \in C_j|} \left(\sum_{i: X_i \in C_j} \rho_i \right)^2 \tag{25}$$

Athey, Tibshirani, and Wager (2019) have shown that maximizing the evaluation function Δ is asymptotically equivalent to minimizing the following error.

$$\sum_{j=1,2} \Pr\left[X \in C_j | X \in P\right] \mathrm{E}\left[\left(\hat{\theta}_{C_j} - \theta(X)\right)^2 | X \in C_j\right] \tag{26}$$

Here, $\hat{\theta}_{C_j}$ is the solution of the estimation equation in child node C_j.

Projected gradient trees

Projected gradient trees

Projected gradient trees are a extension of Projected CART algorighm proposed by Bénard et al.,(2021).

- $A_n(X)$ is the cell of the original gradient tree partition where X falls.
- $A_n^{(-j)}(X^{(-j)})$ is the projected partition

We respectively denote associate projected gradient tree and projected out-of bags forest-weights as

$$T_b^{(-j)}(X^{(j)}) \quad \text{and} \quad \alpha_i^{(-j,OOB)}(x_{i'}^{(-j)}),$$

respectively defined as following slides

Projected gradient trees

$$\alpha_i^{(-j,OOB)}(X_{i'}^{(-j)}) = \frac{1}{|\Lambda_{i'}|} \sum_{b \in \Lambda_{i'}} \alpha_{bi}^{(-j)}(X_{i'}^{(-j)}) \tag{27}$$

$$\alpha_{bi}^{(-j)}(X_{i'}^{(-j)}) = \frac{1\{X_i^{(-j)} \in L_b(X_{i'}^{(-j)})\}}{|L_b(X_{i'}^{(-j)})|} \tag{28}$$

Projected gradient trees

Define a projected out-of-bags generalized estimator for $(\theta^{(-j,OOB)}(X_i^{(-j)}), \nu^{(-j,OOB)}(X_i^{(-j)}))$ as:

Def: Projected out-of-bags GRF Estimator

$$\left(\hat{\theta}^{(OOB,-j)}(X_{i'}^{(-j)}), \hat{\nu}^{(-j,OOB)}(X_{i'}^{(-j)})\right) \in \underset{\theta,\nu}{\operatorname{argmin}} \left\{ \left\| \sum_{i=1, i\neq i'}^{n} \alpha_i^{(-j,OOB)}(X_{i'}^{(-j)}) \psi_{\theta,\nu}(O_i) \right\|_2 \right\} \tag{29}$$