Causal Machine Learning for Political Science[1]

Sam Fuller

Department of Government
Harvard University

March 5th, 2025
Northwestern University

POL 490: Machine Learning in Political Science

[1] Thanks to Jack T. Rametta for collaboration on previous slides and related papers.

Introductions

Sam Fuller (American Politics & Methodology)

- PhD from the University of California, Davis
- Postdoctoral Fellow at the Center for American Political Studies (in the Department of Government) at Harvard University
- **Methods:** ML, surveys, experiments, etc.
- **Substance:** Public opinion, political psychology/behavior, the connection between partisan identity and anti-democratic attitudes (and political violence).

ICPSR Summer Program Topical Workshop & Book

If any of this is of particular interest, Jack T. Rametta and I teach a 40-hr course on causal machine learning each summer. We are also *almost* done with a book on this topic.

Overview

Goals:

- Develop an understanding and familiarity with Causal Machine Learning (CML)
- Enable you to *begin* to confidently apply CML methods in your research

Limitations:

- Only covering cross-sectional approaches
- CML literature = literally a fire hose
- Focusing on tree-based methods
- Focusing on approaches with well developed software

$\hookrightarrow$ **And we only have this class period!**

Outline

Today's presentation proceeds as follows:

1. Motivation
2. Variable importance
3. Feature effects
4. Research contexts
5. DRML
6. Causal forest
7. Example of application

Questions During the Presentation

If you have any questions during the course of the presentation, do not hesitate to ask! I'm happy to clarify during or after the presentation.

Motivation

Background on Causal Machine Learning

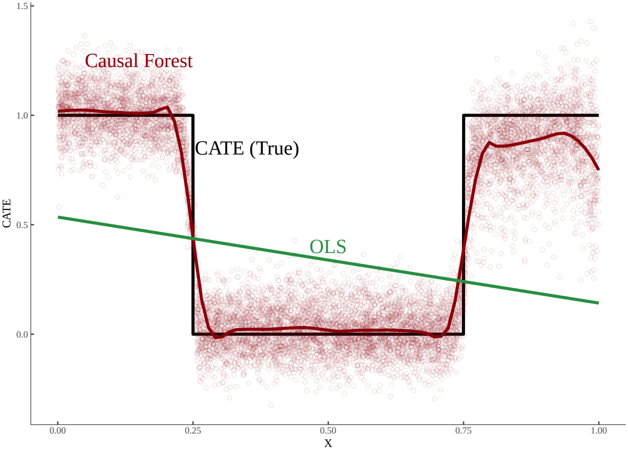
Rapid expansion of ML methods for causal inference:

- Bayesian Additive Regression Trees (Chipman, George, and McCulloch 2010)
- Causal Forest (Wager and Athey 2018)
- Double Machine Learning (Chernozhukov et al. 2023)
- Ratkovic's PLCE and MDEI estimators (Ratkovic and Tingley 2023)

New forest-based ML methods (Montgomery and Olivella 2018) promise theoretical + practical advantages for main effects, HTEs, and beyond.

But old-school, parametric methods still dominate in political science.

Do Social Scientists Really Need ML? Can't We Just Use OLS?



↪ Nonlinearities, interactions, heterogeneity lurking beneath the surface. Researchers are fallible!

Two Cultures No More?

Breiman (2001b)

There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown...

Athey and G. W. Imbens (2019): **No longer true for statistics!**

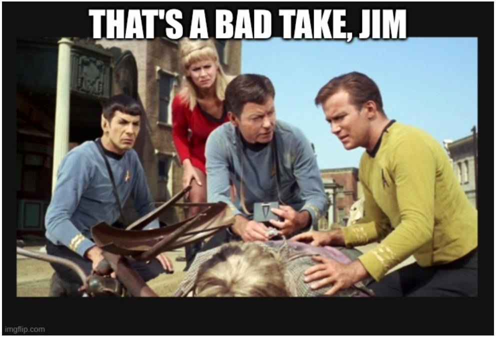
- Though there is much slower adoption of ML in economics, political science, psychology, sociology, etc.

↪ So what's the hold up...?

Two Cultures No More? Not so Fast

One big reason: the causal revolution

↪ Common critique: All this machine learning stuff is just **curve fitting**! We can't use any of this to make causal claims or perform causal tasks.



Two Cultures No More?

In fact, yes we can! Turns out, ML methods like forests can be integrated into the potential outcomes framework (Wager and Athey 2018).

Not only *can we* employ ML for causal inference, ML can be ***better*** than other methods (Chernozhukov et al. 2023) for many causal tasks (ATE estimation, heterogeneous treatment effect detection, etc.).

Blasphemy...

ML is often more honest and reliable for causal inference relative to simple pre-specified regression models.

Figure: Actual Image of Me Presenting CML Research

Advantages of ML for Causal Inference

1. Model specification doesn't fall from the sky!
2. High dimensional confounding and interactions? No problem.
3. Nonlinear function form? No problem.
4. Sample splitting turns out to be good for prediction AND inference
   - Athey and G. Imbens (2016), Ratkovic and Tingley (2023), and Blackwell and Michael P Olson (2022a), etc.!

A Contradiction in Terms? Machine Learning for Causal Inference

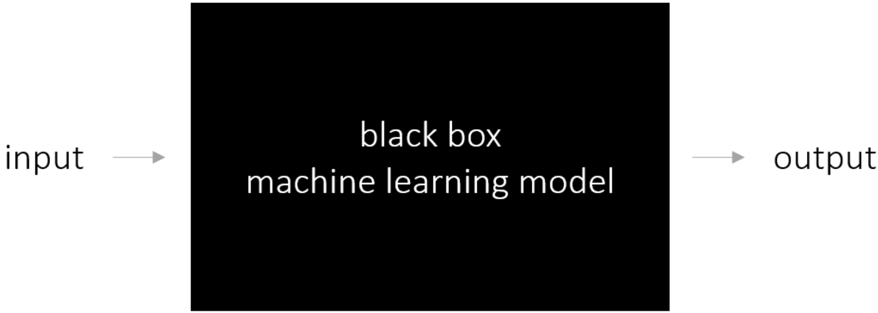
Basic idea that underlies the causal machine learning approach:

- ML algorithms generate better predictions, even if model is black box.
- Many causal inference problems *are* predictions problems. Don't always care what's in the black box, e.g. propensity score estimation.
- CML incorporates ML curve-fitters into estimating equations for causal estimands. Use ML to fit nuisance parameters we wouldn't (and usually shouldn't!) interpret anyway.
- Host of methods to interpret the black box outputs for inference after the fact if desired

$\hookrightarrow$ Not attempting to do inference on simple ML predictions, harnessing the predictions for inference in a structured, rigorous way

Identifying Important Variables: Variable Importance

Extracting Information from the “Black Box”



input → black box machine learning model → output

↪ Unpacking this “black box” is critical for exploratory & causal ML...

Permutation Tests

Attributed to George Box

The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively.

- One of the best and most popular resampling methods for model assessment is the **permutation test**.
- Permutation tests sample from an estimated null distribution of the data (i.e., by randomly permuting the rows and/or columns of a dataset).

Permutation Tests Cont.

- Breiman (2001a) applies this to the problem of estimating feature importance: how does model fit change when we permute a predictor variable (removing all valuable information from it)?
- Provides an intuitive and efficient way to represent the totality of a feature's importance in a black box model.
- Can also repeat over a large number of reshuffles to estimate uncertainty bounds on the feature importance estimates.

Conditional Permutation Tests

- However, we encounter a recurring problem when there are dependencies between features: we generate observations that are outside of the support of the data (or data manifold).
- To address the issue of correlation and estimate partial effects, Strobl et al. (2008) propose a conditional permutation scheme where values of X_S are permuted *conditional* on levels of a correlated variable(s) (see also Debeer and Strobl 2020, and the `permimp` package in `R`).
- **Straightforward idea**: rather than permuting over the marginal distribution $P(X_S)$, instead use the conditional distribution $P(X_S|X_C = x_c)$.

VI: A Fast Moving Research Area

Variable importance is an incredibly fast moving field. Thus, the best variable importance methods will certainly change in the future.

In that vein, there is *a lot* of research on Shapley-related methods See Appendix:

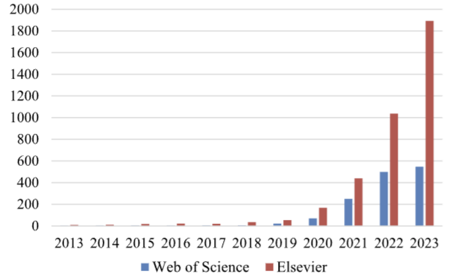


Figure 1 The count of literature with the keywords "Shapley Value, Machine Learning" in the Web of Science and Elsevier databases

Li et al. (2024) has a great review of recent Shapley-related methods.

Identifying Interactions: Variable-Level Friedman's H

If there are no interactions involving x_j, we can decompose the prediction function F into the sum of the partial dependence F_j on x_j and the partial dependence $F_{\setminus j}$ on all other features $\mathbf{x}_{\setminus j}$, i.e.,

$$F(\boldsymbol{x}) = F_j(x_j) + F_{\setminus j}(\mathbf{x}_{\setminus \mathrm{j}}).$$

Correspondingly, Friedman and Popescu (2008)'s statistic of overall interaction strength of x_j is given by:

$$H_j^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{F}(\mathbf{x}_{\mathrm{i}}) - \widehat{F}_j(x_{ij}) - \widehat{F}_{\setminus j}(\mathbf{x}_{\mathrm{i}\setminus \mathrm{j}})\right]^2}{\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{F}(\mathbf{x}_{\mathrm{i}})\right]^2}.$$

Identifying Interactions: Pairwise Friedman's H

Again, if there are no interaction effects between features x_j and x_k, their two-dimensional partial dependence function F_{jk} can be written as the sum of the univariate partial dependencies, i.e.,

$$F_{jk}(x_j, x_k) = F_j(x_j) + F_k(x_k).$$

Correspondingly, Friedman and Popescu (2008)'s statistic of pairwise interaction strength between x_j and x_k is defined as

$$H_{jk}^2 = \frac{A_{jk}}{\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{F}_{jk}(x_{ij}, x_{ik})\right]^2}$$

where

$$A_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{F}_{jk}(x_{ij}, x_{ik}) - \widehat{F}_j(x_{ij}) - \widehat{F}_k(x_{ik})\right]^2.$$

Exploring the Effects of Important Variables: Feature Effects

Calculating Feature/Predictor Effects

Some options for estimating predictor effects:

1. PDP (partial dependence plots)
2. ALE (accumulated local effects)
3. ICE (individual conditional expectations)
4. LOCO (leave-one-covariate-out) or feature ablation
5. Permutation tests
6. Shapley values

See Appendix for details on these other options

Partial Dependence Plots

The partial dependence function estimates the **marginal** effects of a feature(s) (X_S) on the predictions from a machine learning model ($\hat{f}$):

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C) d\mathbb{P}(x_C)$$

where X_C is the set of all other predictor variables (that is, X_S and X_C are complementary).

- Zhao and Hastie (2021) ["Causal Interpretations of Black Box Models"] point out that the PDP is equivalent to Pearl's back-door adjustment that is sufficient to identify the causal effects of X_S on Y **if** X_C satisfies the back-door criterion.

Partial Dependence Plots

The partial dependence function is estimated using the Monte Carlo method by taking samples from the training data:

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$

Hence, the function simply represents the average prediction when we set all observations to a given level of X_S.

Research Contexts

A Motivating *Experimental* Question

Let's assume we have some information treatment and we want to see if it affects vote choice:

information (W) → **vote-choice (Y)**

But we know from a whole bunch of political science and psychological research that information is processed differently by different people. So we might want to see reactions differ by covariates (PID, gender, education, etc.).

So really we're looking at:
information (W) × **covariates (X)** → **vote-choice (Y)**

Conditional treatment effects are our bread and butter in *a lot* of political science research. (More on this later.)

A Motivating *Observational* Question

But let's say we instead want to know how the effect of knowing information about an event, say a speech or a policy-implementation, affects vote-choice. The kinds of people who are exposed to information are very different than those who aren't (e.g., more politically sophisticated, stronger partisanship, etc.).

Then really we're looking at:
$(\textbf{covariates}(\mathbf{X}) \rightarrow \textbf{information}(\mathbf{W})) \times \textbf{covariates}(\mathbf{X}) \rightarrow \textbf{vote-choice}(\mathbf{Y})$

Oh, and those covariates also influence the vote-choice themselves! $X \rightarrow Y$.

This is all to say, generally, we can think of research contexts on two spectra:

1. Are covariates (X) related to the outcome?
2. Are covariates (X) related to treatment assignment?

Research Contexts for DRML

Treatment Assignment (W) ~ Covariates

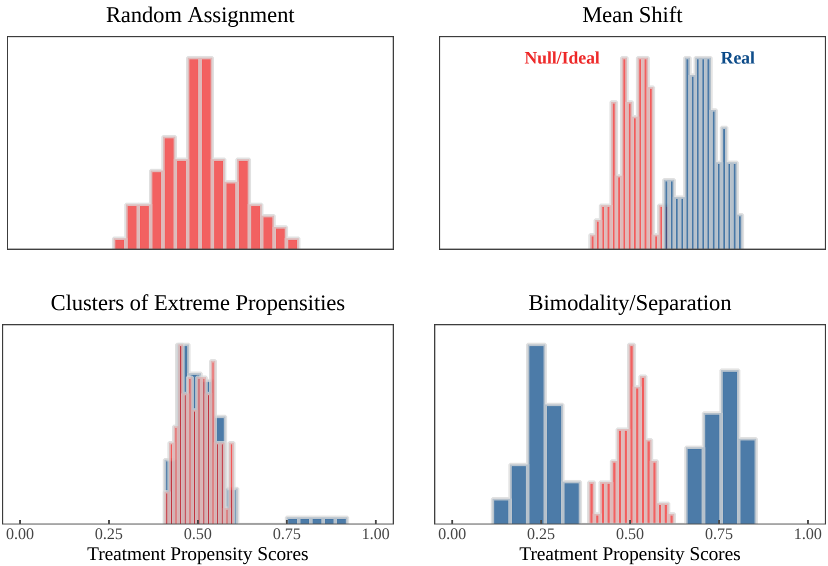
False $\xleftrightarrow{\text{Spectrum}}$ True

Outcome (Y) ~ Covariates

False $\xleftrightarrow{\text{Spectrum}}$ True

	False	True
False	*Cov Relationship* Unrelated to Y Unrelated to W (Balanced Covs) *Research Context* Successful RCT Uninformative Covs	*Cov Relationship* Unrelated to Y Related to W (Imbalanced Covs) *Research Context* Unsuccessful RCT Uninformative Covs
True	*Cov Relationship* Related to Y Unrelated to W (Balanced Covs) *Research Context* Successful RCT Informative Covs	*Cov Relationship* Related to Y Related to W (Imbalanced Covs) *Research Context* Observational Data Unsuccessful RCT Informative Covs

Exogenous Treatment or Selection?



Random Assignment

Mean Shift

Clusters of Extreme Propensities

Bimodality/Separation

The Doubly Robust Estimation Framework

Frist, let's look at classic propensity score adjustments from Rosenbaum and Rubin (1983):

W = Treatment; X = Covariates; Y = Outcome.

Outcome Only: $Y = \beta_0 + \beta' X + \theta W$

Propensity: $\widehat{P} = pr(W = 1|X) = \alpha_0 + \alpha' X$

Where the resulting probabilities $\widehat{P}$ replace the vector of covariates X in our OLS regression:

Propensity Only: $Y = \beta_0 + \zeta \widehat{P} + \theta W$

where ζ represents the effect that treatment propensity $\widehat{P}$ has on our outcome Y.

$\hookrightarrow$ What if we combined the propensity AND outcome models?

The DRML Framework

DRML

$$\begin{aligned} Y &= \theta W + \omega(X) + \epsilon, \quad \mathbb{E}(\epsilon|W,X) = 0 \ \} \text{ Outcome Model} \\ W &= \gamma(X) + V, \qquad \mathbb{E}(V|X) = 0 \ \} \text{ Treatment Model} \end{aligned}$$

- Y is our outcome variable.
- $W = \{0,1\}$ denotes treatment, and θ is the treatment effect.
- The vector $X = (X_1, ..., X_p)$ is composed of all pre-treatment covariates, broadly construed.
- ϵ and V are random errors.

“Weak” DR: If outcome *or* treatment propensity model are correctly specified, then DR estimates are guaranteed consistent.

“Strong” DR: Get both right, estimates consistent, first order equivalent of true treatment effect. Estimates are semi-parametrically efficient, asymptotically normal.

DRML Cont.

In this setup, $\omega(X)$ and $\gamma(X)$ are nuisance parameters. Only need good predictions!

In experimental contexts, DRML with good randomization turns into a fancy regression adjustment (with some distinct benefits).

Depending on the implementation, DRML employs cross-fitting or sample-splitting to guard against overfitting.

$\hookrightarrow$ We will be using a variant of DRML where the models are honest random forests, also known as causal forests (Athey, Tibshirani, and Wager 2019).

Note: Some gains possible from other options but this model is fast, highly predictive, and has nice, known properties.

DRML ATE Estimation: Reducing Noise & Bias

DRML

$$
\begin{aligned}
Y &= \theta W + \omega(X) + \epsilon, \quad \mathbb{E}(\epsilon | W, X) = 0 \,\}\text{Outcome Model} \\
W &= \gamma(X) + V, \quad \mathbb{E}(V | X) = 0 \,\}\text{Treatment Model}
\end{aligned}
$$

Assume we *only* care about estimating ATEs and not CATEs.

In **experimental contexts**:

- Then we still benefit significantly from modeling reducible noise with covariates that inform baseline rates of our outcome.
- Explicitly modeling θ at unit-level can benefit ATE estimates
- Increased precision and lower power requirements for smaller effects.

In **observational contexts**:

- We can significantly reduce bias from nonrandom assignment.
- This is similar to matching and other propensity score methods.
- The overlap estimator (Li, Morgan, and Zaslavsky 2018) can provide credible ATT/ATO estimates with significant nonrandom assignment.

ATE Estimation: DRML vs. SUEs

Treatment Assignment (W) ~ Covariates

False $\xleftrightarrow{\text{Spectrum}}$ True

Outcome (Y) ~ Covariates

False $\xleftrightarrow{\text{Spectrum}}$ True

	False	True
False	*Simulations* Uninformative Covariates *DRML vs. SUEs* Reduces to difference-in-means	*Simulations* Uninformative Covariates *DRML vs. SUEs* Reduces to difference-in-means
True	*Simulations* Standard Experiments *DRML vs. SUEs* Higher Precision Lower Power Reqs.	*Simulations* Systematic Imbalance *DRML vs. SUEs* Higher Consistency Higher Precision Better Coverage Lower Power Reqs.

See Appendix for simulation evidence

Honest Random Forests (Causal Forest) for ATEs

grf & Honest Random Forests

Wager and Athey (2018) introduce generalized random forests tweaking Breiman's RF...

- "Honest" subsample splitting: data used to choose splits not used to populate leaves.
  - In other words, separate data is used to train the model and estimate effects at each split and node.
- Leaves that are unpopulated are "pruned."
- For HTEs, loss function is to explicitly maximize heterogeneity (more later).
  - Has knock-on benefits for ATEs.

Causal Forests

Causal Forests adapt Random Forests to get known statistical properties (consistent estimates, Gaussian asymptotic sampling distribution, and estimable variance to build CIs).

To do this they:

1. Alter the tree-growing algorithm to ensure that trees have bins/leaves L that are small enough such that:
    1. Y_is where $i \in L(x)$ are independently distributed *and*
    2. Combinations of (Y_i, W_i)—where W represents treatment—are as if they come from a randomized experiment.
2. Make sure each tree is trained on a *subsample*, each split contains a minimum number of treatment and control units, and **honesty** is employed in each leaf.
    1. *Within* a tree there is a training/testing split: some portion (`honesty.fraction`) of the data is used to *estimate effects* (test; $\mathcal{I}$) and the other portion is used to determine splits (train; $\mathcal{J}$).
    2. Splits are made by maximizing the variance of the estimated treatment effect conditional on covariates $\hat{\tau}(X_i)$ for $i \in \mathcal{J}$ between splits.

AIPW

$$\hat{\tau}_{\text{AIPW}} =$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\underbrace{W_i\,\frac{Y_i - \hat{\omega}_{(1)}(X_i)}{\hat{\gamma}(X_i)} - (1 - W_i)\frac{Y_i - \hat{\omega}_{(0)}(X_i)}{1 - \hat{\gamma}(X_i)}}_{\text{IPW Component}} + \underbrace{\hat{\omega}_{(1)}(X_i) - \hat{\omega}_{(0)}(X_i)}_{\text{Outcome Component}}\right]$$

AIPW in DRML

To understand how AIPW has doubly robust properties, we decompose Equation 39 into two components, the regression adjustment component D and the residual IPW estimator R as follows:

$$\hat{\tau}_{AIPW} = D + R$$

$$D = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{\omega}_{(1)} \left(X_i \right) - \hat{\omega}_{(0)} \left(X_i \right) \right)$$

$$R = \frac{1}{n} \sum_{i=1}^{n} \left(W_i \frac{Y_i - \hat{\omega}_{(1)}(X_i)}{\hat{\gamma}(X_i)} - (1 - W_i) \frac{Y_i - \hat{\omega}_{(0)}(X_i)}{1 - \hat{\gamma}(X_i)} \right)$$

So... What Does This Mean?

The long and short of it is this:

1. Causal forests have statistically known properties (like OLS) but are far more flexible and powerful at prediction.
   - The overall "box" may still be black, but we now know the properties of CF's ATEs and CATEs!
2. Because they are built to maximize the variance of the treatment effect as predicted by covariates (think demographics) they are phenomenal at:
   1. **Detecting** treatment effect heterogeneity.
   2. **Estimating** conditional average treatment effects.

- There are more assumptions made by CF to recover the statistical guarantees, but many of these are assumed by other models (unconfoundedness) or are directly testable (overlap in treatment propensities).
- Also/however, in cases of complete randomization these assumptions are essentially guaranteed. Hence the beauty of RCTs.

Machine Learning and Treatment Effect Heterogeneity

A New Estimand: Conditional Average Treatment Effects

As we established, estimating individual-level treatment effects = impossible.

But we can estimate conditional effects among observations with similar covariate profiles.

Specifically, we can estimate conditional average treatment effects (CATEs):

$$CATE = \tau(\mathbf{X}) = E\left[Y^{(1)} - Y^{(0)} \mid \mathbf{X} = \mathbf{x}\right]$$

$\hookrightarrow$ **This is a function, not a point estimate!!**

Need:

- SUTVA
- Overlap
- Unconfoundedness
- Colliders also possible!

HTEs in the Social Sciences

In the social sciences, while ATEs from interesting experiments or observational analyses are in and of themselves important, we are very, very often interested in if and how those treatment effects vary by any number of variables or characteristics.

Understanding heterogeneity increases knowledge of subject area, increases understanding the main effect as well.

But there are **numerous** pitfalls in the estimation of conditional average treatment effects (CATEs).

How Should We Estimate Conditional Effects?

The core issue with estimating conditional effects, particularly in experiments, is that there is ***no guarantee of causal inference***.

So how should we estimate conditional effects?

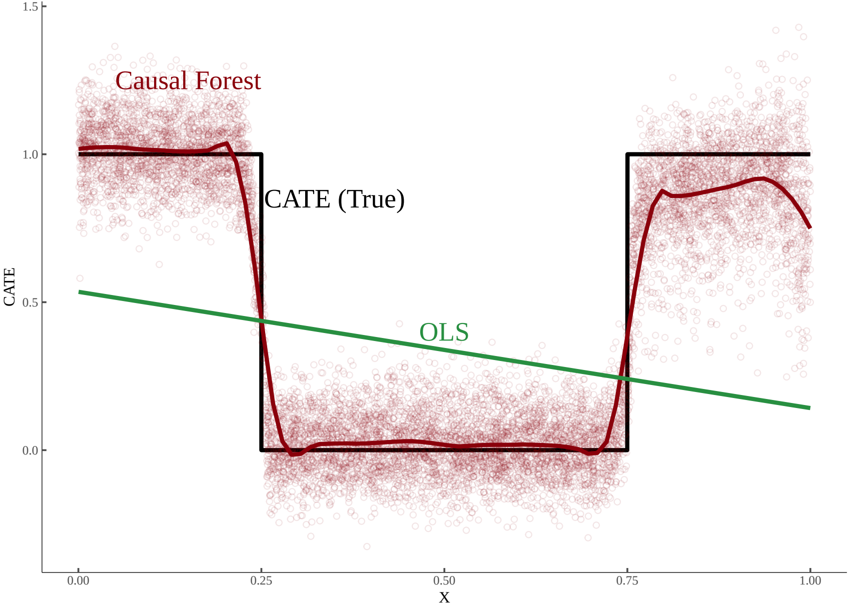
1. Multiple interactions in a simple regression?
    - Could overfit, likely misspecified (Beiser-McGrath and Beiser-McGrath 2020). Need lots of power.
2. Separate single interaction regression models?
    - Could introduce omitted interaction bias! (Blackwell and Michael P. Olson 2022b)
3. Off-the-shelf machine learning methods (RF, boosting, etc.)?
    - Not designed for causal inference, likely overfit, regularization bias (later) (Ratkovic and Tingley 2023).

Pitfalls in the Standard Estimation of CATEs

Additionally, numerous issues can arise when attempting to estimate CATEs with standard methods:

- False-positives
- Researcher's degrees of freedom
- Linear (or simply parametric) assumptions surrounding functional forms
- Difficulty in estimating multi-dimensional interactions
- (Relatedly) the curse of dimensionality

A (Contrived) Example



Learners: Fitting ML Models to DGPs

The basic logic is to fit ML models to the underlying data-generating-process (DGP) to recover (heterogeneous) treatment effects.

In (mostly) chronological order:

- S-Learner
- T-Learner
- R-Learner

↪ We'll focus on R-Learner See Appendix for more information on learners

R-Learner: Intuitive

$X =$ Covariates, $W =$ (Binary) Treatment, $Y =$ Outcome

First, assume we know nothing about treatment and simply model the outcome using our covariates: $Y \sim X$.

Two chunks of error from $\epsilon = Y - \widehat{Y}$, reducible and irreducible. We're interested in the *reducible* component since that's a representation of the effect of treatment (because we ignored that in our first step).

Second, we model treatment assignment $W \sim X$ to determine the weight we place on each observation for training the treatment effect model.

Observations with assignment that is accurately predicted ($\widehat{W} \approx \{0, 1\}$) already have their reducible error explained *through* $Y \sim X$. Whereas observations that are not accurately predicted ($\widehat{W} \approx \frac{n_{W=1}}{n}$) are weighted the most for training the treatment effect model.

Finally, we use ML to model a heterogeneous treatment function, as a function of covariates X, that explains the most of that reducible error (weighted by non-confidence in predicted assignment, $W - \widehat{W}$) across all observations.

R-Learner Benefits

Nie and Wager (2021) demonstrate R-learner:

- Flexible: can use any curve fitting method.(**Must cross-fit!**)
- Convergence depends on complexity of τ, not outcome and propensity.
- Quasi-oracle: same error bounds as oracle w/ knowledge of props and outcome.

$\hookrightarrow$ Can implement via rlearner `R` package, or directly through `grf`.

See Appendix for simulation evidence of R-Learner (`grf`) benefits

Causal Forests for HTEs

Critically: CFs are grown to maximize differences in treatment effects between splits and nodes.

To do this they:

1. Alter the tree-growing algorithm to ensure that trees have bins/leaves L that are small enough such that:
    1. Y_is where $i \in L(x)$ are independently distributed *and*
    2. Combinations of (Y_i, W_i)—where W represents treatment—are as if they come from a randomized experiment.
2. Make sure each tree is trained on a *subsample*, each split contains a minimum number of treatment and control units, and **honesty** is employed in each leaf.
    1. *Within* a tree there is a training/testing split: some portion (`honesty.fraction`) of the data is used to *estimate effects* (test; $\mathcal{I}$) and the other portion is used to determine splits (train; $\mathcal{J}$).
    2. Splits are made by maximizing the variance of the estimated treatment effect conditional on covariates $\hat{\tau}(X_i)$ for $i \in \mathcal{J}$ between splits.

Detecting Heterogeneity & Identifying Important Variables

DRML vs. Traditional: Testing for HTEs

Standard approaches tend to rely on preregistration and interactions in regressions (treatment × conditioning variable).

This approach, as we illustrate in our simulations, can lead to significant bias *and* missing important, unexpected heterogeneity.

Many tests for detecting heterogeneous treatment effects, chronologically:

- The bound test (Athey and Wager 2019)
- Best linear fit test (Chernozhukov et al. 2023)
- Rank-weighted average treatment effect (RATE) tests (Wager 2024a)
- Sequential RATE (Ibid.)

Check omnibus test, then proceed to CATEs IFF you pass HTE tests.

See Appendix for more details on heterogeneity tests

↪ You can use VI measures to identify conditioning variables.

So What Should We Use? The Omnibus Test

There is no single answer or consensus in the literature.

Wager (2024b)'s Sequential RATE test is shown to have the highest power in small sample sizes.

Chernozhukov et al. (2023)'s best linear fit test gives more useful diagnostic information.

↪ Our advice: report multiple statistics, if they disagree, try to assess why. Typically these agree in practice.

Identifying Important Conditioning Variables

We can use standard variable importance measures with causal forest models for identifying important condition variables.

Interpretation of VIs from CF models: **Important** variables are **important** for predicting heterogeneous treatment effects.

Only do this if there is evidence fo heterogeneity (according to the tests described before).

What method should we use?

- Permutation variable importance is good for 1-way importance
- Friedman's H is good for 2-way (interactive) importance

Experimental Analysis Process

1. Correctly code and specify model of interest
2. Run a CF with all pretreatment covariates included in the covariate matrix
    - Tune CF based on calibration test (built into the `grf` package)
3. Check ATEs
4. Run the omnibus heterogeneity test
5. If tests suggest that there is heterogeneity then check:
    - Permutation variable importance
    - Friedman's H
6. Calculate conditional effects using important variables
    - Can calculate group average treatment effects (GATEs)
        - Groups determined by 1-variable
        - Groups determined by multiple variables
    - Can calculate PDPs
        - 1-way
        - 2-way
        - Paneled 3-way (!!!)

See the Appendix for an example of this whole process:

Running Heterogeneity Tests

Using Exploratory & Causal ML in Research

Example Paper

Given the recent research, we had the grand idea to uncover correlates of support for political violence using machine learning.
↪ An intentionally agnostic, exploratory analysis.

Our (Initial) Research Question

What are the correlates/predictors of support for political violence both across datasets and countries?

Motivation

Given recent research on support for political violence (SPV), we had the grand idea to uncover correlates of support for political violence using machine learning.

↪ An intentionally agnostic, exploratory analysis using random forest.

Our (Initial) Research Question

What are the correlates/predictors of support for political violence both across datasets and countries?

Main-Text Datasets

Dataset	Obs	Countries	Focus	SPV Measure
Afrobarometer R5 (2011–13)	51,156	34 African	–	Single Question
Daxecker (2025)	4,783	IN	Identity, Info	Single Question
FIRE Campus Pulse	259,507	US	–	Single Question
Fuller (2022)	11,000	10 Western	Identity	Single Question
Landry (2024)	2,003	US	Identity, Psych	4-Item Scale
Munis (2023)	2,314	US	Identity, System	Single Question
Piazza (2023)	1,889	US	Identity, Psych	Single Question
Voelkel (2024)	35,252	US	Identity, Psych	4-Item Scale
Westwood (2022)	5,927	US	Measurement	Scale & Vignette
WVS Wave 3 (1995–98)	64,181	Global	System	Single Question
WVS Wave 7 (2017–22)	97,220	Global	System	Single Question

Assessing Variable Importance for Prediction

Marginal variable importance is calculated using permutation tests.

Our process:

1. Train the model with all covariates, calculate fit
2. Randomly permute one feature
3. Assess change in prediction error (RMSE, AUC, etc.) w/ permuted feature
4. Repeat with each feature
5. Order by change in prediction error

↪ All predictions are out-of-bag, evaluate by change in RMSE.

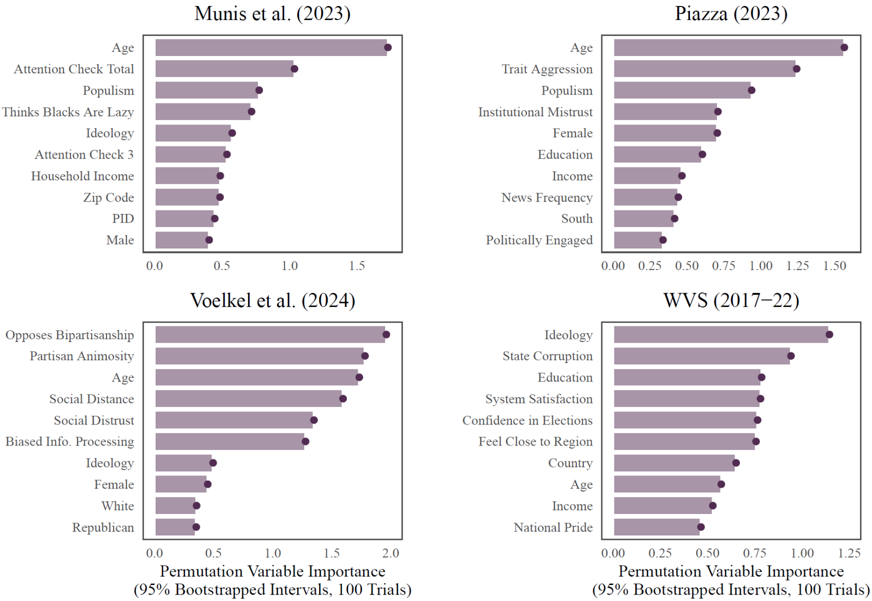
Partial Dependence & Friedman's H Statistic

To test for interactions we employ Friedman's H statistic: how much variance in the predictions depend on a single interaction.

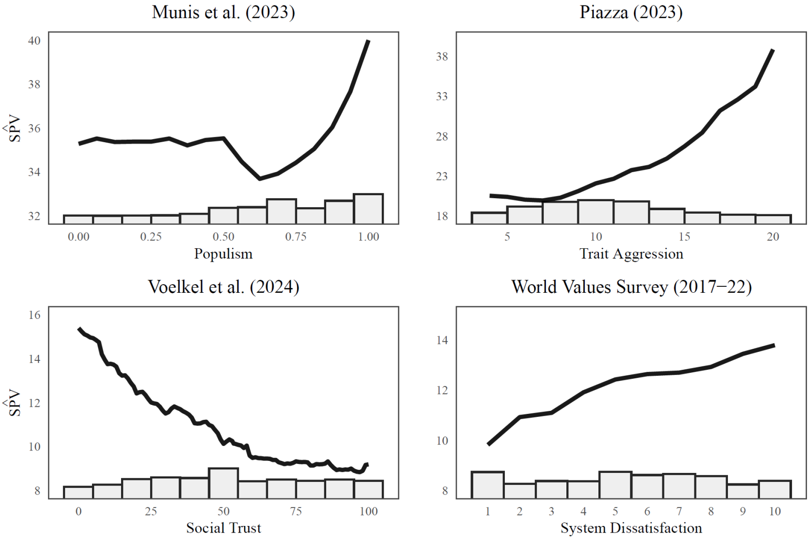
To examine functional relationships, we look at 1D & 2D partial dependence plots (PDPs) *only on important covariates*.

$\hookrightarrow$ Analogous to marginal effects plots and evaluating the statistical significance of an interaction term.

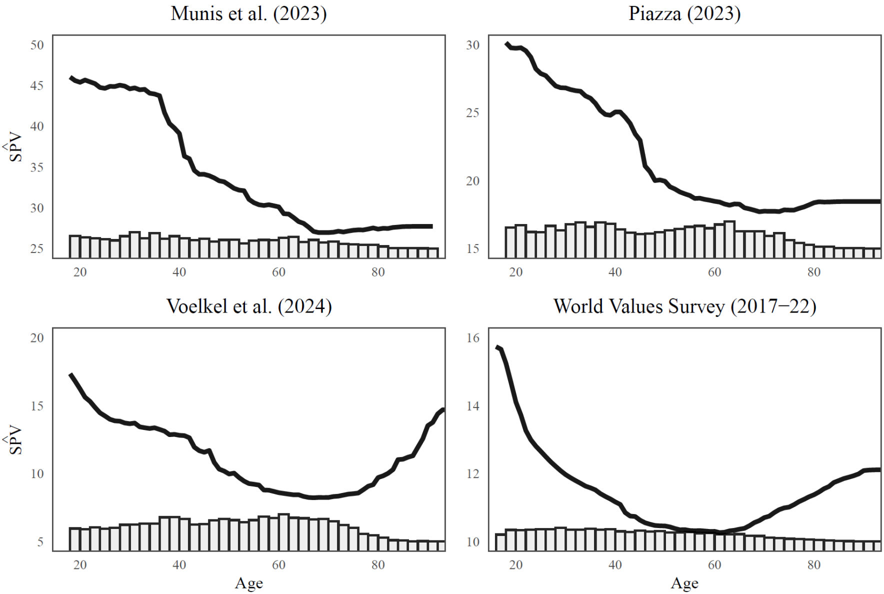
Permutation Variable Importance Results



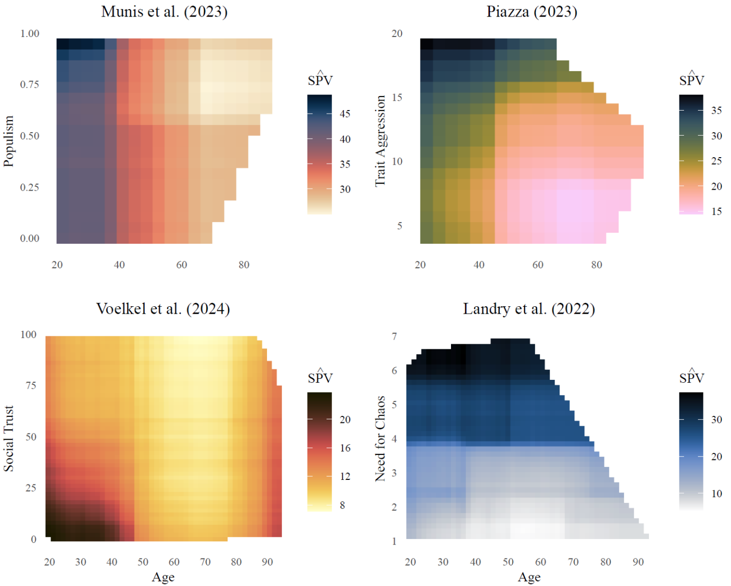
Predicted SPV, Partial Dependence over Variables of Interest



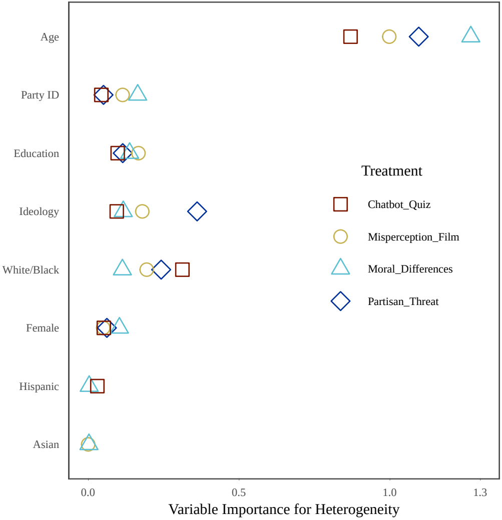
Predicted SPV, Partial Dependence over Age



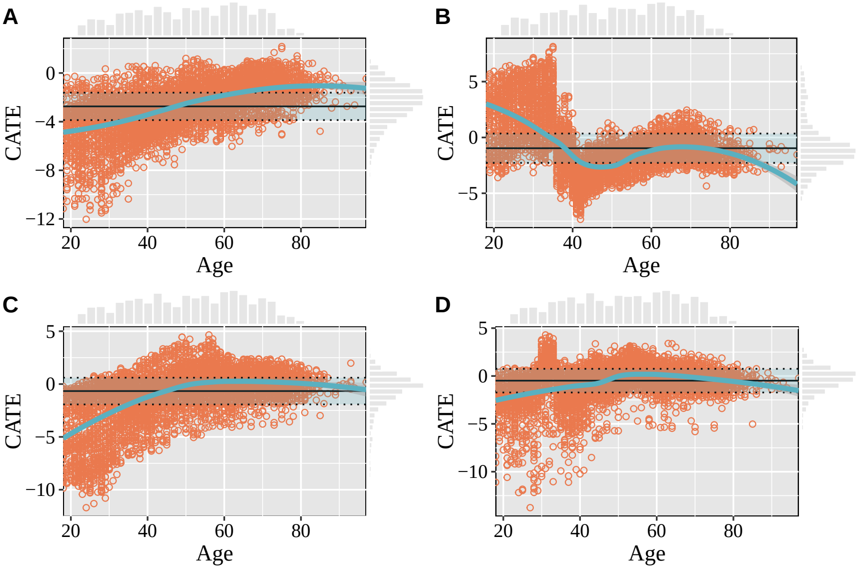
Predicted SPV, Two-Way Interactions with Age



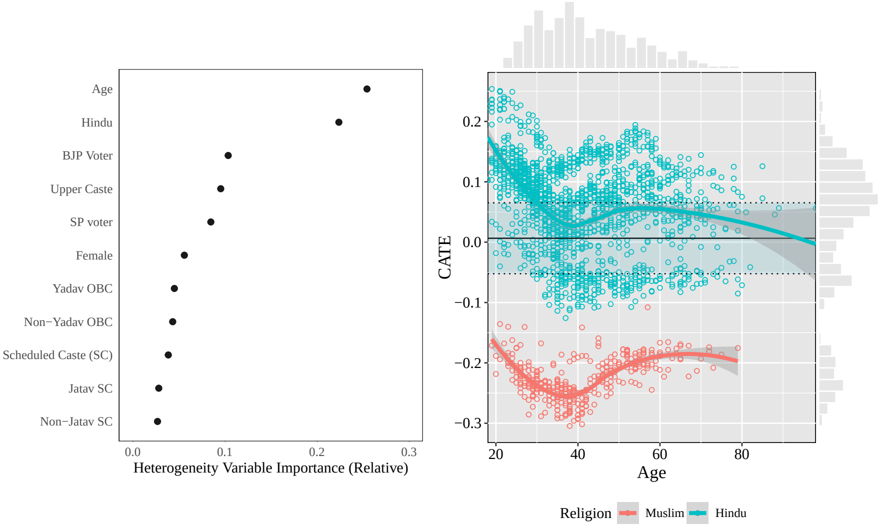
Treatment Effect Conditionality by Pre-Treatment Covariate, Voelkel et al. (US, 2024)



Age × Change in SPV, Voelkel et al. (US, 2024)



VIs and PDP by Age & Religion, Daxecker & Prasad (India, 2025)



Overall Takeaways

Across all datasets we essentially find *the same relationship*...

- Increasing age → decreasing SPV.
- Country, region, and time-period intercept shifts.
- Reinforces previous literature, three/four categories:
  1. Politically salient identities
  2. Psychological characteristics
  3. Attitudes toward the political system
  4. The information environment

Another Example: Fuller, Cerda, and Rametta (2025)

This was the first project I worked on that explicitly leveraged causal forest for an experiment.

That paper has a corresponding preregistration and more straightforward experimental analysis, which may be a good reference if you're interested in running experiments.

Jack and I believe that you should ***always*** use some flavor of causal ML for experiments.

- If you don't, you're leaving money on the table!
- See Appendix for simulation evidence of ATE and CATE benefits!

Appendix

References I

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Athey, Susan, and Guido W Imbens. 2019. "Machine Learning Methods That Economists Should Know About." *Annual Review of Economics* 11:685–725. https://www.annualreviews.org/doi/full/10.1146/annurev-economics-080217-053433.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized Random Forests." *The Annals of Statistics* 47 (2): 1148–1178. https://doi.org/10.1214/18-AOS1709. https://doi.org/10.1214/18-AOS1709.

Athey, Susan, and Stefan Wager. 2019. "Estimating Treatment Effects with Causal Forests: An Application." *Observational Studies* 5 (2): 37–51.

Beiser-McGrath, Janina, and Liam F Beiser-McGrath. 2020. "Problems with products? Control strategies for models with interaction and quadratic effects." *Political Science Research and Methods* 8 (4): 707–730.

Bénard, Clément, and Julie Josse. 2023. "Variable importance for causal forests: breaking down the heterogeneity of treatment effects." *arXiv preprint arXiv:2308.03369.*

Blackwell, Matthew, and Michael P Olson. 2022a. "Reducing Model Misspecification and Bias in the Estimation of Interactions." *Political Analysis* 30 (4): 495–514.

References II

Blackwell, Matthew, and Michael P. Olson. 2022b. “Reducing Model Misspecification and Bias in the Estimation of Interactions.” *Political Analysis* 30 (4): 495–514. https://doi.org/10.1017/pan.2021.19.

Breiman, Leo. 2001a. “Random Forests.” *Machine Learning* 45:5–32.

———. 2001b. “Statistical Modeling: The Two Cultures.” *Statistical Science* 16 (3): 199–231.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2023. “Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments, with an Application to Immunization in India.” arXiv: 1712.04802 [stat.ML]. https://doi.org/10.48550/arXiv.1712.04802.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. “BART: Bayesian Additive Regression Trees.” *The Annals of Applied Statistics* 4 (1): 266–298. https://doi.org/10.1214/09-AOAS285. https://doi.org/10.1214/09-AOAS285.

Debeer, Dries, and Carolin Strobl. 2020. “Conditional permutation importance revisited.” *BMC bioinformatics* 21 (1): 307.

Friedman, Jerome H., and Bogdan E. Popescu. 2008. “Predictive learning via rule ensembles.” *The Annals of Applied Statistics* 2 (3): 916–954. https://doi.org/10.1214/07-AOAS148. https://doi.org/10.1214/07-AOAS148.

Fuller, Sam, Nicolás de la Cerda, and Jack T Rametta. 2025. “Affect, Not Ideology: The Heterogeneous Effects of Partisan Cues on Policy Support.” *Political Behavior,* 1–25.

References III

Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation." *journal of Computational and Graphical Statistics* 24 (1): 44–65.

Hollmann, Noah, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2022. "Tabpfn: A transformer that solves small tabular classification problems in a second." *arXiv preprint arXiv:2207.01848.*

Kennedy, Edward H. 2023. "Towards optimal doubly robust estimation of heterogeneous causal effects." *Electronic Journal of Statistics* 17 (2): 3008–3049.

Kim, Jee-Seon, Xiangyi Liao, and Wen Wei Loh. 2023. "Assessing cross-level interactions in clustered data using cate estimation methods." In *The Annual Meeting of the Psychometric Society,* 87–97. Springer.

Li, Fan, Kari Lock Morgan, and Alan M Zaslavsky. 2018. "Balancing Covariates via Propensity Score Weighting." *Journal of the American Statistical Association* 113 (521): 390–400.

Li, Meng, Hengyang Sun, Yanjun Huang, and Hong Chen. 2024. "Shapley value: from cooperative game to explainable artificial intelligence." *Autonomous Intelligent Systems* 4 (1): 2.

Montgomery, Jacob M, and Santiago Olivella. 2018. "Tree-Based Models for Political Science Data." *American Journal of Political Science* 62 (3): 729–744.

References IV

Nie, Xinkun, and Stefan Wager. 2021. "Quasi-oracle estimation of heterogeneous treatment effects." *Biometrika* 108 (2): 299–319.

Ratkovic, Marc, and Dustin Tingley. 2023. "Estimation and Inference on Nonlinear and Heterogeneous Effects." *The Journal of Politics* 85 (2): 421–435.

Robins, James M, and Andrea Rotnitzky. 2001. "Comment on the Bickel and Kwon article, 'Inference for semiparametric models: Some questions and an answer'." *Statistica Sinica* 11 (January): 920–936.

Rosenbaum, Paul R, and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70 (1): 41–55.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. 2008. "Conditional variable importance for random forests." *BMC bioinformatics* 9 (1): 307.

Štrumbelj, Erik, and Igor Kononenko. 2014. "Explaining prediction models and individual predictions with feature contributions." *Knowledge and information systems* 41 (3): 647–665.

Wager, Stefan. 2024a. "Sequential Validation of Treatment Heterogeneity." *arXiv preprint arXiv:2405.05534.*

———. 2024b. "Sequential Validation of Treatment Heterogeneity." arXiv: 2405.05534 [econ.EM]. https://arxiv.org/abs/2405.05534.

References V

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113 (523): 1228–1242.

Zhao, Qingyuan, and Trevor Hastie. 2021. "Causal interpretations of black-box models." *Journal of Business & Economic Statistics* 39 (1): 272–281.

Shapley Values

Shapley Values

- Shapley values are a classic concept in game theory.
- For a coalition of players who cooperate to earn some collective payoff, Shapley values define the contribution of each player to the coalition.
- They do so by averaging each player's marginal contribution over all possible coalitions.

Shapley Values Cont.

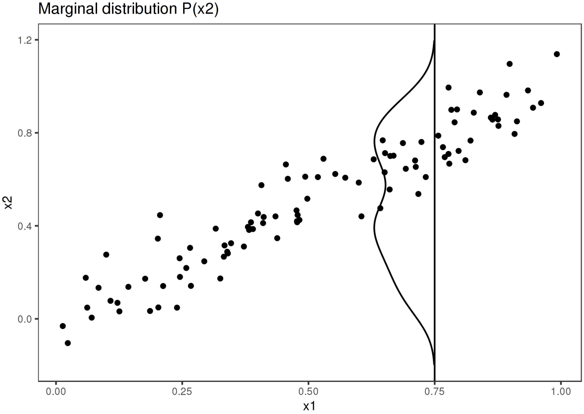
- We can apply this to the problem of estimating predictor importance by treating the features as players and the payoff as the model prediction.
- For a given observation i, compare the predictions with and without a given feature j across all possible coalitions of the other features X_{-j}.
- The mean of these differences is observation i's Shapley value $\phi_(j)$.

Shapley Values Cont.

- Hence, we can interpret the Shapley value ϕ_j as the average marginal contribution of a feature value to the model's prediction across possible coalitions.
  - Note: these are **marginal** Shapley values; there are also **conditional** Shapley values that are estimated by sampling the out-of-coalition variables conditional on X_J (see, e.g., the `shapr` package in `R` and accompanying paper.)
- Shapley values have desirable statistical properties: for example, efficiency (i.e., they correctly capture deviations between expected and actual predictions across individual observations) and nullity (irrelevant features receive zero values).
- Useful for both local and global explanations.
- However, exact solutions usually computationally intractable (for each observation, there are 2^p possible coalitions) and usually requires Monte Carlo sampling (e.g., Štrumbelj and Kononenko 2014, "Explaining Prediction Models and Individual Predictions").

The Problematic Independence Assumption

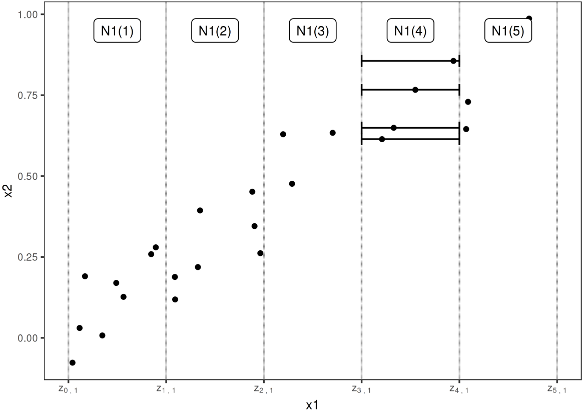
The partial dependence function assumes X_S and X_C are orthogonal (uncorrelated). Otherwise, bogus combinations:



Calculating Predictor Effects

Accumulated Local Effects (ALE)

An alternative to the partial dependence function is the ALE method. ALE calculates the differences in predictions using the conditional distribution of X_C over bins of X_S:



ICE curves

- An issue with PDP and ALE is that they can smooth over heterogeneous effects.
- This motivates the use of **individual conditional expectation** (ICE) curves.
- ICE uses the same calculation as PDP, but shows separate curves for each observation instead of averaging these values.

ICE curves

From Goldstein et al. (2015) [“Peeking Inside the Black Box”]:

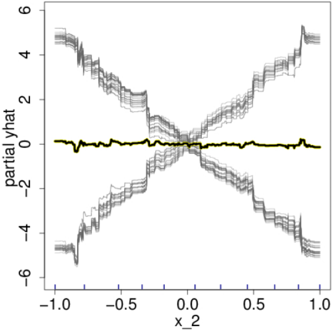


Figure 2: SGB ICE plot for X_2 from 1000 realizations of the data generating process described by Equation 3. We see that the SGB’s fitted values are either approximately linearly increasing or decreasing in X_2.

Surrogate trees

- Straightforward idea: fit a simple, readily interpretable model (such as a single decision tree) to the predictions $\hat{y}$ from a more complex model.
- Determine how well the surrogate tree approximates $\hat{y}$ with the same covariate information using R^2 values.
- Important to note this is not a model of the data, but rather a model (i.e., a simplification) of another model.

Simulation Evidence of CF's ATE Performance

Simulation Overview

For ATEs we evaluate two sets of sims:

- Standard experimental context. RA + simple informative covariates
- High dimensional experimental context. RA + complex informative covariates

Methods to Compare

Competing estimators...

- CF-AIPW: Estimates ATE with CF
- Lin Estimator: Estimates ATE with OLS + regression adjustment (demean+interact)
- Unadjusted Difference in Means

DRML Performance Test Simulation Parameters

Parameter	Possible Settings
Sample Size	100, 500, 1000, 5000
Number of Covariates	5, 10, 20, 50, 100
Heterogeneous Effects	True or False
Error Variance	Low (1), Medium (5), High (10)
Linear Effects	True or False
Interactions	True or False

High Dimensional Simulation Parameters

Parameter	Possible Settings
Treatment Effect	Small (0.5), Medium (1), Large (2)
Signal-to-Noise Ratio	Min. (7), Low (5), Medium (3), High (1)
Sample Size	25, 50, 75, 100, 125, 150, 175, 200, 250, 300, 400, 500, 750, 1000, 2000, 5000

Standard Experimental Context Results

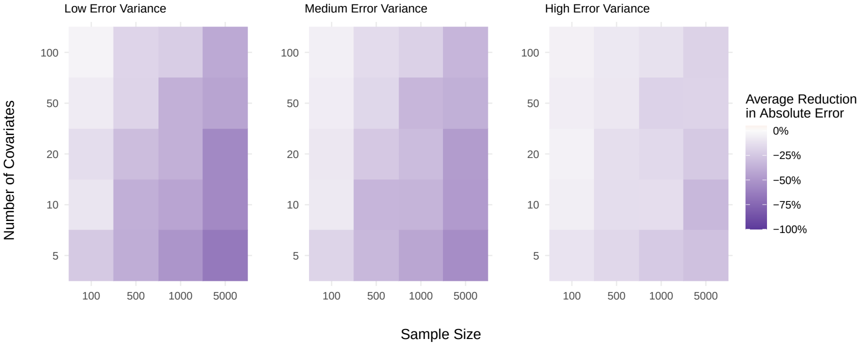
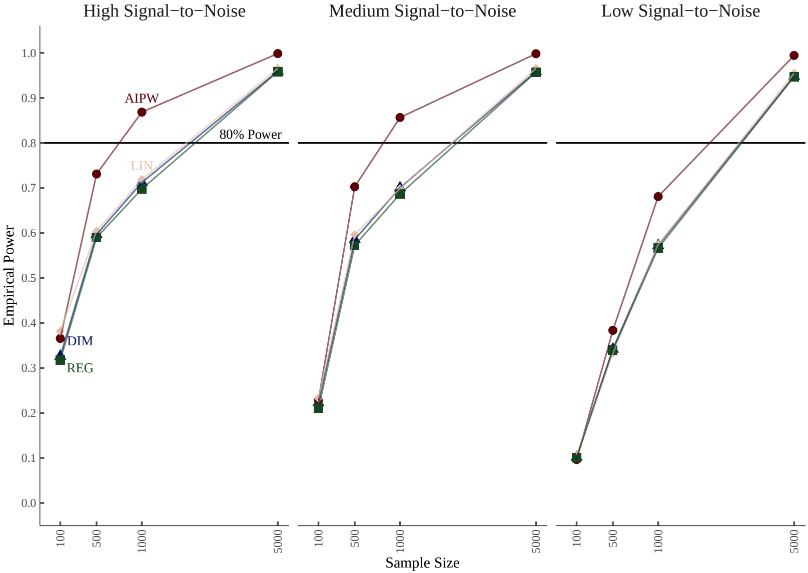


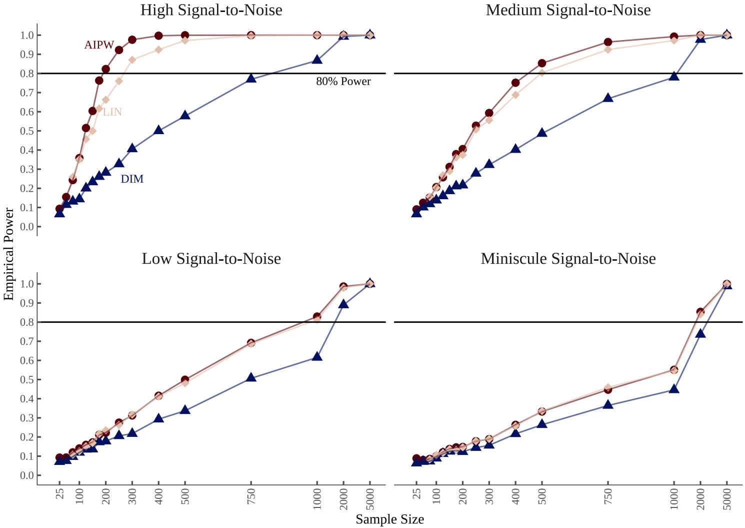
Figure: Average reduction in absolute error of the DRML treatment effect estimates versus unadjusted difference in means, regression-adjusted, and Lin-adjusted effects.

480 permutations x 2,000 seeds = 960,000 simulations covering many DGPs.
$\hookrightarrow$ (*untuned*) DRML wins in $\sim$ 99% **of cases.**

Power Results: Standard Experimental Context



Power Results: High Dimensional Context



↪ DRML leaves no variance reduction on the table.

Long Story Short...

CF performs phenomenally well for recovering ATEs, particularly compared to standard methods like linear regressions and difference-in-means.

This is across:

- Accuracy/bias
- Coverage
- Power

Note: we haven't even complicated the propensity model here!

S-, T-, & R-Learner

ML for CATEs the Wrong Way: S-Learner

- Train a single model on all data with treatment indicator: $\hat{Y}(x, w)$
- Compute two predictions per x:
  - $\hat{Y}(x|w = 1)$ and $\hat{Y}(x|w = 0)$
- Estimate CATE: $\hat{\tau}(x) = \hat{Y}(x|w = 1) - \hat{Y}(x|w = 0)$

$\hookrightarrow$ Treatment is just another covariate here, not considering it separately!

A Better, but still wrong, way: T-Learner

- Fit two separate models:
  - $\hat{Y}_0(x)$ on controls $(W = 0)$
  - $\hat{Y}_1(x)$ on treated $(W = 1)$
- Estimate CATE: $\hat{\tau}(x) = \hat{Y}_1(x) - \hat{Y}_0(x)$

$\hookrightarrow$ This approach misses, need to consider the joint objective of finding CATEs.

R-Learner: In Detail

Nuisance parameters modeled through ML:

$m^*(X) = \text{Outcome Model } (\widehat{Y} = Y \sim X)$
$e^*(X) = \text{Treatment Model } (\widehat{W} = W \sim X)$

Goal in constructing CML model:

$\tau^*(X) = \text{(Heterogeneous) Treatment Effect Model}$

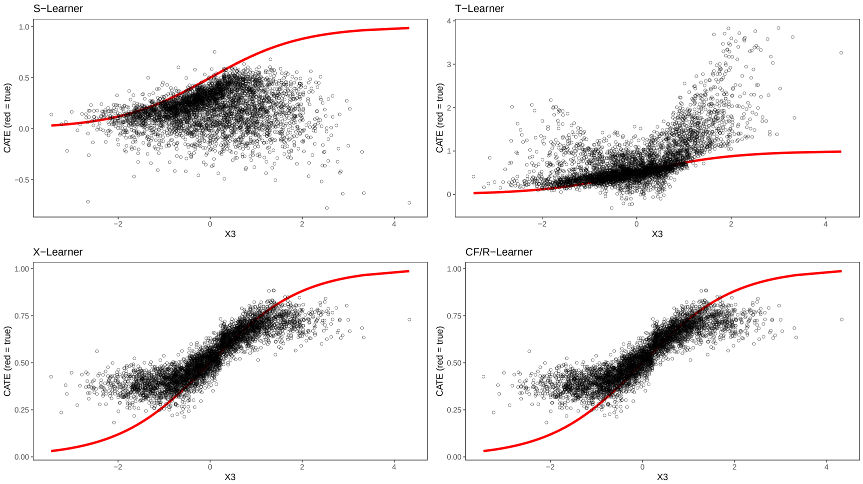
The Math:

$$\tau^*(\cdot) = \text{argmin}_\tau \left\{ E\left(\left[\{Y_i - m^*(X_i)\} - \{W_i - e^*(X_i)\} \tau(X_i) \right]^2 \right) \right\}$$

The Estimation Process:

1. Estimate nuisances: $\hat{\mu}(x) = E[Y|X = x], \quad \hat{\pi}(x) = P(W = 1|X = x)$
2. Compute residuals: $r_Y = Y - \hat{\mu}(x), \quad r_W = W - \hat{\pi}(x)$
3. Solve orthogonalized regression: $\hat{\tau}(\cdot) = \arg\min_\tau \frac{1}{n} \sum_i (r_{Y,i} - r_{W,i}\tau(x_i))^2$
4. Implementation via weighted least squares of r_Y/r_W on X with weights r_W^2.

Example of Regularization Bias: Comparing Learners



DGP via Apoorva Lal, framework from Kunzel 2019.

↪ ML is good for HTEs, **only** when combined with an appropriate loss function!

Simulation Evidence of CF's CATE Performance

Simulation Overview

Table: CATE Simulation Parameters

Parameter	Possible Settings
Signal-to-Noise Ratio (`SNR`)	Minuscule (1), Low (3), Medium (5), High (7)
Sample Size (n)	500, 1000, 2000, 5000
Number of Predictors (p)	5, 10, 15, 20
Heterogeneity Present? (`heterogeneity`)	True, False
Dimension of CATE (`dimension`)	1D, 2D
Nonlinear Effects (`linear`)	True, False

For CATEs we compare CF to saturated OLS in standard experimental context.

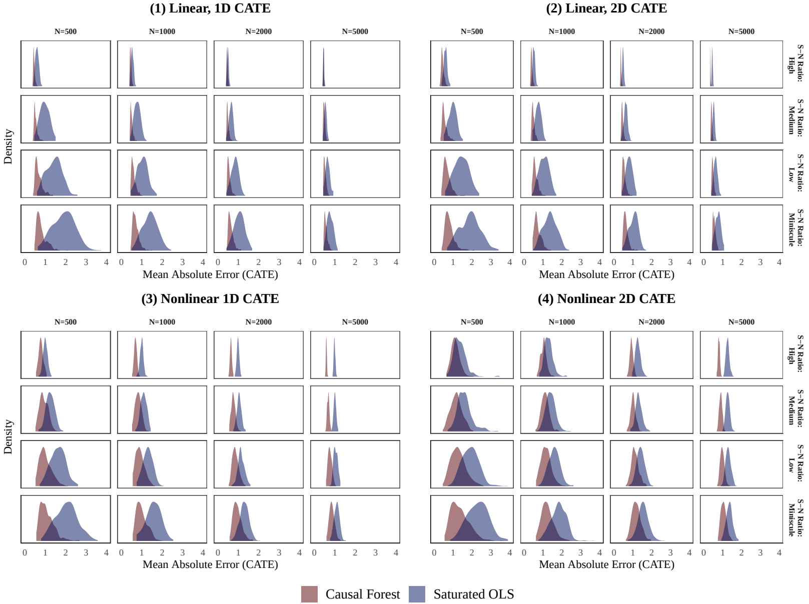
- By dimension (1 vs. 2)
- By linearity (linear, nonlinear)

CF vs. OLS: Heterogeneity Detection False-Positive Rates

	Sample Size							
	500		1000		2000		5000	
S-N Ratio	CF	OLS	CF	OLS	CF	OLS	CF	OLS
High	0.75%	50.0%	1.94%	51.7%	2.69%	53.4%	4.12%	53.5%
Medium	2.94%	45.6%	1.69%	45.0%	4.44%	48.4%	4.75%	46.4%
Low	2.25%	43.9%	1.81%	45.1%	4.31%	46.7%	4.62%	46.8%
Miniscule	2.00%	43.1%	2.69%	43.2%	4.62%	45.6%	5.00%	45.0%

Table: Comparison of False Positive Rates (FPR): No Heterogeneity Present

CF vs. OLS: CATE Accuracy



Long Story Short...

CF outperforms saturated OLS for recovering CATEs, often dramatically so.

This is across:

- Accuracy/bias
- Coverage
- Power

Benefits derive both from ML targeting of CATE & rethinking how HTEs are estimated in the first place (e.g., focus = more agnostic).

In the next module, we'll cover the implications of **DRML for research design and then delve into selection on observables.**

Heterogeneity Tests & Calculating Conditional Effects

Toy Example for HTE Tests

A simple example for us to evaluate different HTE tests:

- Sample Size of 3,000
- 10 Pre-Treatment (X) Covariates, all normally distributed
- Binary treatment with random-assignment
- Normally Distributed Error
- Outcome model is noise

$\hookrightarrow$ CATE $\rightarrow$ linear function of X_1

The Bound Test

The “bound test” or high-vs-low test asks: can our CATE model can differentiate a statistically significant difference between predicted CATEs above and below the median CATE prediction.

1. Estimate the ATE (via AIPW) for observations with CATEs above the median
2. Estimate the ATE (via AIPW) for observations with CATEs below the median
3. Test for a statistically significant different between these two estimates

For our example, this yields a difference of 1.93 with a 95% CI of $[1.19, 2.67]$. Suggests statistically significant HTEs.

$\hookrightarrow$ **This is a heuristic.** Do not rely on this!

Best Linear Fit Test

Chernozhukov et al. (2023)'s best linear fit test assesses model's ability to capture both the average and conditional treatment effects.

Execute a simple linear regression of the average CATE prediction and differential CATE predictions on the residualized outcome variable.

Coefficients of $1 =$ perfect performance. Significance tests interpreted as usual.

$\hookrightarrow$ Insignificant result for differential CATE prediction either no true HTEs or model not capturing. Could still be power, model not converging, etc.

Rank-weighted Average Treatment Effect (RATE) Tests

Rank-weighted average treatment effect (RATE) assesses presence of HTEs via assessment of our CATE model's performance on different samples.

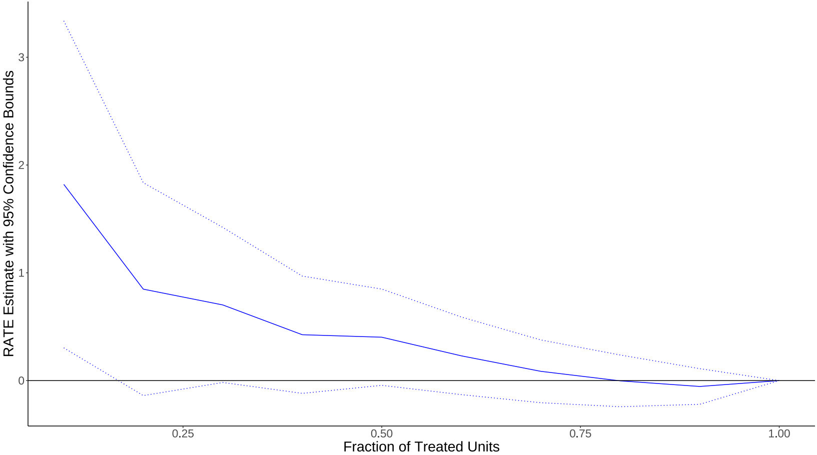
To calculate RATE...

1. CATE predictions arranged in numerical order
2. Generate targeting operator characteristic (TOC) curve, somewhat analogous to the popular receiver operator characteristic (ROC) curves often employed in binary classification settings.
3. RATE is calculated by taking the area under the TOC curve.[2]

$\hookrightarrow$ Intuitively, when treatment effect is constant, RATE approximates zero. Not so if HTEs present.

[2] For a more detailed discussion, see this vignette.

Target Operator Characteristic Curve



Sequential RATE

Could do a t-test to compare RATE estimate to zero, BUT, need to evaluate more robustly.

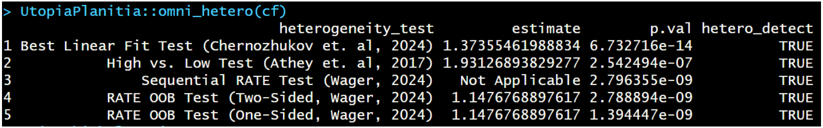
Evaluate RATE on sequential cross-folds to assess performance.

Specifically...

1. Randomly split the data into $k = 1, \ldots, K$ evenly sized folds.
2. For each fold $k = 2, \ldots, K$:
    1. Obtain an estimated CATE function using training data from folds $1, \ldots, k-1$.
    2. Form a RATE estimate using data from the k-th test fold and compute a t-statistic.
3. Aggregate the $K - 1$ different t-statistics and use this as a final test statistic.

$\hookrightarrow$ Implemented in our temp package `UtopiaPlanitia::omni_hetero`.

Sequential RATE and Omnibus HTE Testing



```
> UtopiaPlanitia::omni_hetero(cf)
                             heterogeneity_test         estimate        p.val hetero_detect
1 Best Linear Fit Test (Chernozhukov et. al, 2024) 1.37355461988834 6.732716e-14          TRUE
2          High vs. Low Test (Athey et. al, 2017) 1.93126893829277 2.542494e-07          TRUE
3              Sequential RATE Test (Wager, 2024)   Not Applicable 2.796355e-09          TRUE
4          RATE OOB Test (Two-Sided, Wager, 2024)  1.1476768897617 2.788894e-09          TRUE
5          RATE OOB Test (One-Sided, Wager, 2024)  1.1476768897617 1.394447e-09          TRUE
```

↪ Sequential RATE matches other HTE tests and results based on simple RATE.

Newer Research

To be clear, there are many new methods for CATE estimation using ML. DR-learner (Kennedy 2023), cluster analysis for CATEs (Kim, Liao, and Loh 2023), network/tab-based CATE models (Hollmann et al. 2022), etc.

grf and causal forest are well developed and have mature software. Relative acceptance in social science journals. Probably 80–90% of the benefits with relatively little cost (e.g., you don't need to buy a $3,000 GPU and learn PyTorch).

Identifying Important Variables

The most basic and intuitive VI measure is included in `grf`:

1. Generate a matrix with columns from 1 to max(depth) and rows from 1 to the number of covariates. Each cell is the total number of times each variable appears as a split from any tree at that depth.
   - nrow(matrix) = number of covs; ncol(matrix) = max(depth)
   - Default: Limits to 4 columns (only looks at a maximum depth of 4).
2. Each row (variable) is then divided by its row sum to calculate the depth incidence proportion by tree depth.
3. The initial weights for the matrix are determined by a tuning parameter called "decay" (default = 2): Initial Weight= $\frac{\text{Inc.Prop.}}{\text{Depth}^{\text{Decay}}}$.
4. Relative weights are then calculated by dividing the initial weights by the sum of weights: Initial Weight ÷ sum(Initial Weights).
5. Finally, for each depth the incidence is weighted and the sum of each row (variable) is used as its variable importance.

↪ This results in a *relative* variable importance which can only be interpreted in relation to the other variables in the model.

Bénard and Josse (2023)'s LOCO VI Measure

1. **Local centering** Compute OOB nuisance estimates $\hat{\mu}(X_i) = \mathbb{E}[Y_i \,|\, X_i]$ & $\hat{\pi}(X_i) = \Pr(W_i = 1 \mid X_i)$, then:

$$\tilde{Y}_i = Y_i - \hat{\mu}(X_i), \quad \tilde{W}_i = W_i - \hat{\pi}(X_i).$$

2. **Full model** Fit a causal forest on $(\tilde{Y}_i, \tilde{W}_i, X_i)$ & obtain the CATE predictor $\hat{\tau}(X)$. Its empirical R-loss:

$$\widehat{R}_n(\hat{\tau}) = \frac{1}{n} \sum_{i=1}^{n} \{\tilde{Y}_i - \hat{\tau}(X_i)\tilde{W}_i\}^2$$

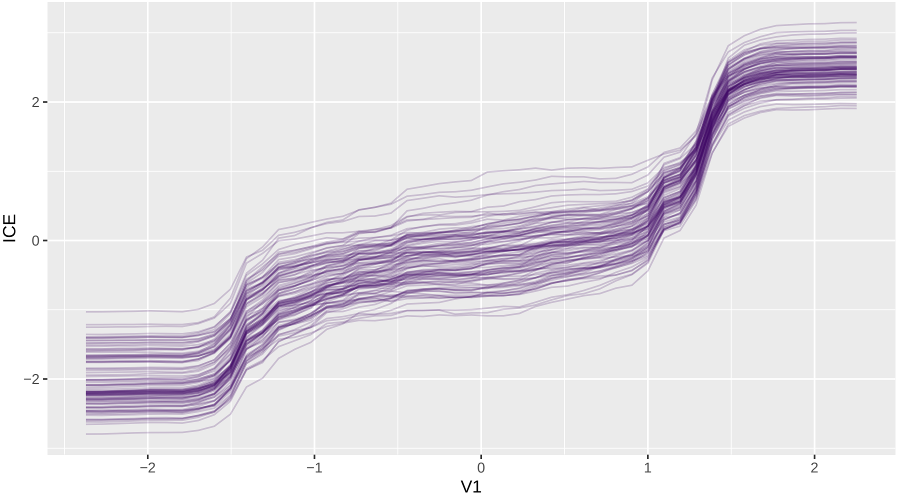
3. **Drop one covariate** Re-train the forest *without* the j^{th} covariate.
4. **Variable–importance score**

$$\mathrm{VI}_j = \widehat{R}_n(\hat{\tau}_{(-j)}) - \widehat{R}_n(\hat{\tau}) \quad \geq 0.$$

$\hookrightarrow$ Larger VI_j = greater degradation in model performance. Code in our temp package: `UtopiaPlanitia::cf_loco`.

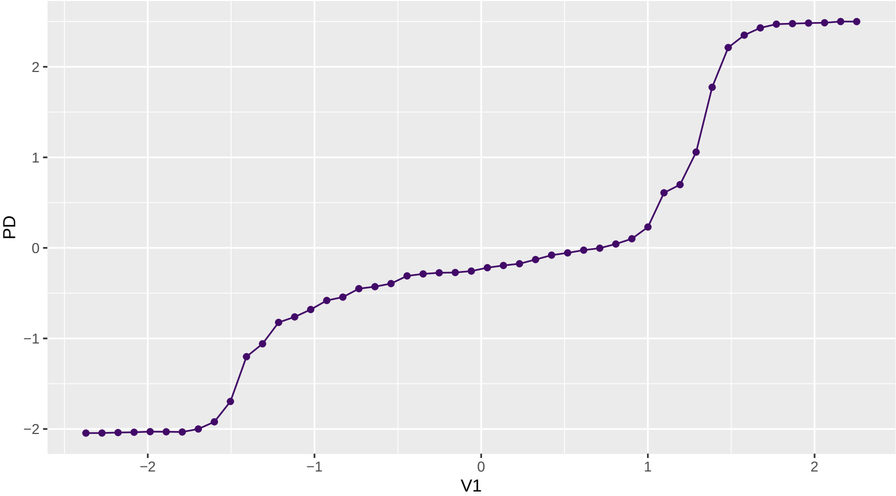
ICE Plot

From our earlier toy example:



1-D Partial Dependence Plots

From our earlier toy example:



Identifying Interactions: Variable-Level Friedman's H

If there are no interactions involving x_j, we can decompose the prediction function F into the sum of the partial dependence F_j on x_j and the partial dependence $F_{\setminus j}$ on all other features $\mathbf{x}_{\setminus j}$, i.e.,

$$F(\boldsymbol{x}) = F_j(x_j) + F_{\setminus j}(\mathbf{x}_{\setminus \mathrm{j}}).$$

Correspondingly, Friedman and Popescu (2008)'s statistic of overall interaction strength of x_j is given by

$$H_j^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{F}(\mathbf{x}_{\mathrm{i}}) - \widehat{F}_j(x_{ij}) - \widehat{F}_{\setminus j}(\mathbf{x}_{\mathrm{i}\setminus \mathrm{j}})\right]^2}{\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{F}(\mathbf{x}_{\mathrm{i}})\right]^2}.$$

Identifying Interactions: Pairwise Friedman's H

Again, if there are no interaction effects between features x_j and x_k, their two-dimensional partial dependence function F_{jk} can be written as the sum of the univariate partial dependencies, i.e.,

$$F_{jk}(x_j, x_k) = F_j(x_j) + F_k(x_k).$$

Correspondingly, Friedman and Popescu (2008)'s statistic of pairwise interaction strength between x_j and x_k is defined as

$$H_{jk}^2 = \frac{A_{jk}}{\frac{1}{n}\sum_{i=1}^{n}\left[\widehat{F}_{jk}(x_{ij}, x_{ik})\right]^2}$$

where

$$A_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{F}_{jk}(x_{ij}, x_{ik}) - \widehat{F}_j(x_{ij}) - \widehat{F}_k(x_{ik})\right]^2.$$

Calculating Friedman’s H

We’ll calculate Friedman’s H-statistic using `hstats`.

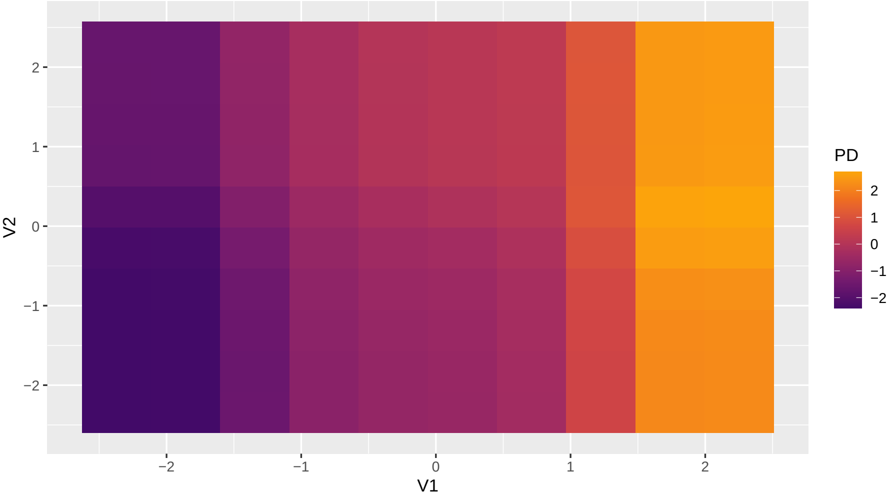
We can plug in `grf` models into `hstats` to calculate Friedman’s H, which is a measure of interaction importance.

Friedman's H Plot

Here we calculate Friedman's H-statistic using `hstats`. Works automatically with `grf`. From our earlier toy example:

2-D Partial Dependence Plots

From our earlier toy example:



Toy Example for with HTEs Present

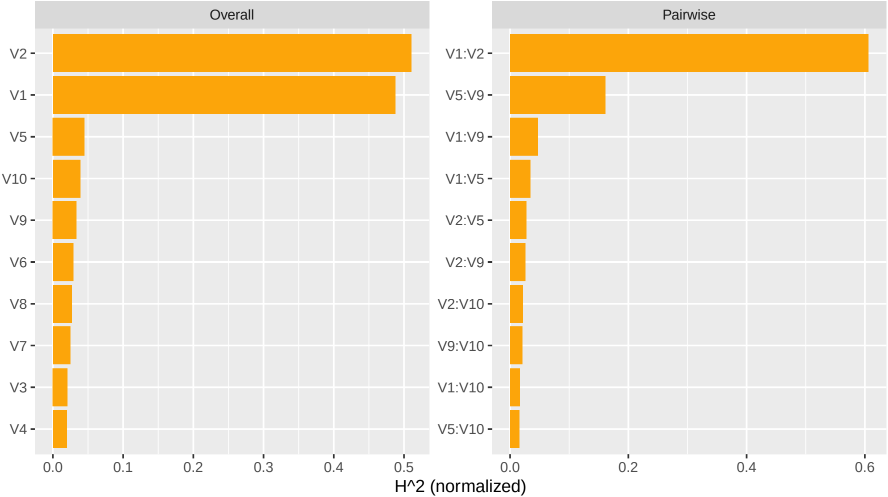
One more example for us to evaluate different HTE tests:

- Sample Size of 3,000
- 10 Pre-Treatment (X) Covariates, all normally distributed
- Binary treatment with random-assignment
- Normally Distributed Error
- Outcome model is noise

$\hookrightarrow$ CATE $\rightarrow (2 * X_2) * X_1$

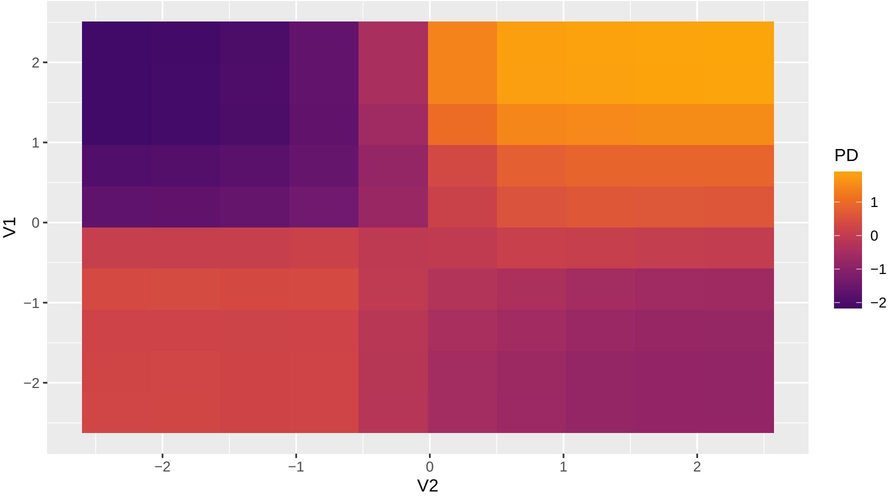
Friedman's H when Interaction Present

From the second toy example.



2-D Partial Dependence Plot

From our second toy example:


V1
V2
PD