Machine Learning

Marica Valente[1]

June 25, 2024

[1]Assistant Professor, University of Innsbruck (marica.valente@uibk.ac.at)

Outline

Chapter 3: Causal ML

Review of causal inference methods in HD

1. Double ML methods and post-LASSO
    - Belloni and Chernozhukov 13; Chernozhukov et al. 17
2. Causal Trees and Causal Forests
    - Athey and Imbens 16; Wager and Athey 18, Athey et al. 2019

Review of causal inference methods in HD

1. **Double ML methods and post-LASSO**
   - **Belloni and Chernozhukov 13; Chernozhukov et al. 17**
2.
   -

What to do with a great predictor?

Double ML: Prediction in the service of estimation
$\Rightarrow$ Transform $\hat{\beta}$ problems into $\hat{y}$ problems

- Causal analysis: estimate the impact of a low-dimensional parameter, e.g. the effect of a treatment d
- Problem: many other variables x correlate with y AND d
- These variables x are called "confounders"

Causal effects with confounding: Example

- Smoking $(d) \rightarrow$ Lung cancer (y)
- Compare $y_{smokers}$ ("treated" group) to $y_{nonsmokers}$ ("control" group)
- Collect a sample of smokers $(d > 0)$ with/without cancer
- Collect a sample of nonsmokers $(d = 0)$ with/without cancer
- Estimate α using model: $y_i = \alpha d_i + \beta x_i + \epsilon_i$ for each individual i

$\Rightarrow$ What are possible confounders x_i?

Causal effects with confounding: Example

- Smoking (d) $\rightarrow$ Lung cancer (y)
- Compare $y_{smokers}$ ("treated" group) to $y_{nonsmokers}$ ("control" group)
- Collect a sample of smokers ($d > 0$) with/without cancer
- Collect a sample of nonsmokers ($d = 0$) with/without cancer
- Estimate α using model: $y_i = \alpha d_i + \beta x_i + \epsilon_i$ for each individual i

$\Rightarrow$ What are possible confounders x_i?

- Anything that makes smokers differ from nonsmokers and correlates with y
- E.g. age, living in polluted cities, family history of lung cancer

ML methods to adjust for confounding

- Use ML to **predict impact of** x that confound estimation of d on y
- Intuition:
  1. remove the impact of x on y
  2. remove the impact of x on d
  3. then estimate causal effect of d on y
- When confounding is large (many, correlated x), OLS breaks down

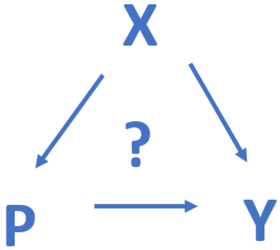
$\Rightarrow$ **Double selection/Residualization** methods to flexibly remove high-dimensional confounding

Residualization: Policy Example

Consider the **causal effect of a Policy (P)** on an outcome (Y), e.g. *the effects of a green tax policy (P) on pollution (Y)*

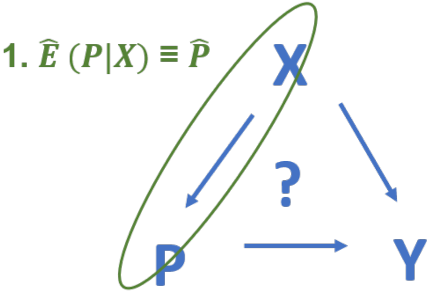
We are interested in the impact of P on Y

Residualization: Policy Example



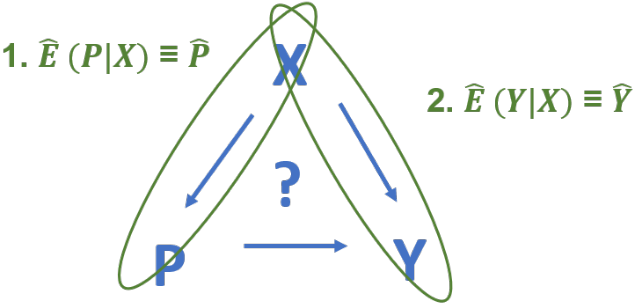
But we have a set of X that may impact both

Residualization: Policy Example

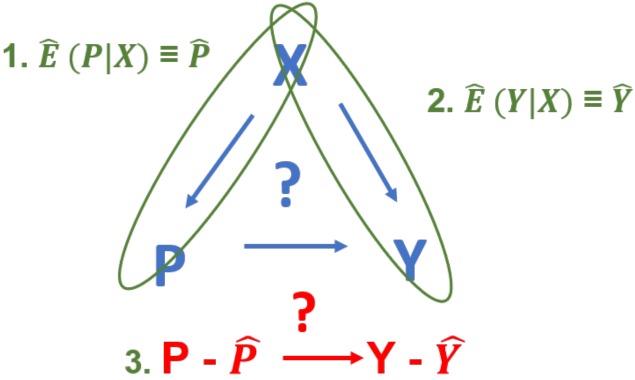


Control for confounding in three stages

Residualization: Policy Example



Residualization: Policy Example



Using LASSO for Prediction (Step 1 and 2)

- LASSO can estimate the mean outcome y given x with nearly the fastest possible rate of convergence given the model complexity, and thus is hard to improve on
- LASSO (or any other method) is **not perfect at model selection** - might include meaningless variables, exclude some relevant regressors
- LASSO biases/shrinks the non-zero coefficient estimates towards 0

$\Rightarrow$ Motivates the use of **Least squares after Lasso, or Post-Lasso**

LASSO biases

- Fit LASSO with (x_1, x_2, x_3) on true model $y = 2x_2 - x_3 + \epsilon$
- **Selection** biases:
  1. $\rho(x_2, x_3)$ large $\rightarrow \hat{\beta}_3 = 0$ (compactification bias)
  2. $\rho(x_1, 2x_2 - x_3)$ large $\rightarrow \hat{\beta}_1 \neq 0$ (expansion bias)
- **Size** biases:
  1. $\hat{\beta}_3 = 0 \rightarrow x_3$ not selected $\rightarrow \hat{\beta}_2$ biased (omitted variable)
  2. $\hat{\beta}_1 \neq 0 \rightarrow$ even if x_2, x_3 selected, $\hat{\beta}_2, \hat{\beta}_3$ biased towards zero (shrinkage)

$\Rightarrow$ In high dimensions, empirical (vs. real) correlations ubiquitous

The Post-LASSO estimator

1. In step one, select the model using LASSO
2. In step two, apply OLS to the selected model

Properties of Post-LASSO (post model selection estimator):

- Performs at least as well as LASSO and has a lower bias (unshrinks β)
- This nice performance occurs even if the LASSO fails in step one, i.e., misses important regressors
- Intuition?

Slides based on Chernozhukov NBER lectures 2013, for video click here.

The Post-LASSO estimator

Intuition behind improved performance of Post-LASSO:

- Intuition: LASSO omits only those components with small coefficients
- Not a big deal if we miss out x that are only weak predictors of outcome and treatment in step 1 and 2
- Why? Small mistakes in step 1 and 2 are going to wash out in step 3 (at least in theory)

$\Rightarrow$ This result was first derived for LS by Belloni and Chernoz. (Bernoulli, 13). Extended to heteroscedastic, non-Gaussian case in Belloni, Chen, Chernoz., Hansen (Econometrica, 12)

Slides based on Chernozhukov NBER lectures 2013, for video click here.

Post-LASSO for inference on target parameter

Consider **inference on the target coefficient** α in the model:

$$y_i = d_i\alpha + x_i'\beta + \epsilon_i,\ \mathbb{E}\left[\epsilon_i x_i\right] = 0, \mathbb{E}\left[\epsilon_i d_i\right] = 0$$

- d_i is target regressor e.g. treatment/policy variable
- In general $\rho(d_i, x_i) \neq 0$, so α cannot be consistently estimated by the regression of y_i on d_i (regularization/shrinkage bias)
- Assuming approximate sparsity, the relationship of d_i to x_i writes:

$$d_i = x_i'\pi^d + \gamma_i^d\ ,\ \mathbb{E}\left[\gamma_i^d x_i\right] = 0$$

$\Rightarrow$ We canNOT use naive estimates of α based simply on applying LASSO and Post-LASSO to the first equation - Why?

The Naive Post-LASSO estimator

1. Select controls terms by running LASSO of y_i on d_i and x_i
2. Estimate α by OLS of y_i on d_i and selected x_i

Caveats:

- Omitted variable bias from estimating $x_i'\beta$ in HD
- Breaks down both theoretically (Leeb and Potscher 09) and practically

$\Rightarrow$ Such a strategy in general does not produce good estimators of α

$\Rightarrow$ Solution: Use residualization/double selection methods for α

Residualization/Double Selection Methods

1. **Double Selection**
   1. Select controls x that predict y by LASSO
   2. Select controls x that predict d by LASSO
   3. Run OLS of y on d and the **union** of controls selected in steps 1 and 2
2. **Partialling Out / Residualization / R-learning (in HD)**
   1. Partial out the x-variables from y
   2. Partial out the x-variables from d
   3. Run OLS on the **residuals**

Cross-fitting

Intuition: Decorrelate model error from estimation error for consistency
$\Rightarrow$ Run Post-LASSO in step 3 on held-out data not used in steps 1 and 2

- Chern. et al. 17 (AER) suggest to implement doubly-robust estimators by "cross-fitting" = k-fold cross-validation
  - Split the data in k folds (parts)
  - Estimate step 1 and 2 on K-1 folds (without using data from k)
  - Estimate causal effect for fold k using estimates in step 1,2
  - Repeat for every fold k=1:K
  - Final causal effect is computed as average of these K estimators

$\Rightarrow$ Estimator is consistent and $\sqrt{n}$-convergent

Pros and cons of cross-fitting

Pros:

- Each ML estimator of steps 1,2 may converge slowly
- "Bad" estimators can be combined

Cons:

- In practice, steps 1, 2 rely on assumptions to produce credible estimates of causal effects
- Prediction of d and y can be imprecise but in practice must be accurate (otherwise researchers are skeptical)

Double Selection in linear models

1. Run the outcome equation: $y_i = x_i' \pi^y + \gamma_i^y$, $\mathbb{E}\left[\gamma_i^y x_i\right] = 0$
2. Run the selection equation: $d_i = x_i' \pi^d + \gamma_i^d$, $\mathbb{E}\left[\gamma_i^d x_i\right] = 0$
3. Run the final outcome equation: $y_i = \alpha x_i + \epsilon_i$, $\mathbb{E}\left[\epsilon_i x_i\right] = 0$

- Three steps: LASSO for steps 1 and 2, Post-LASSO for step 3 with union of variables selected in steps 1 and 2
- Small model selection mistakes will no longer be important under approx. sparsity of 1 and 2
- OLS st.err. valid if 3 is estimated on independent sample from 1 and 2

Pros and cons of Double Selection

Advantages:

- Good statistical properties
- Easy to implement, not computationally heavy

Disadvantages:

- Final outcome model can include controls related to d but not y

$\Rightarrow$ Threat to assumption of approximate sparsity (many x selected)

- Union may contain variables that are highly correlated

$\Rightarrow$ Multicollinearity problems (especially with polynomials!)

Residualization in linear models

1. Remember the selection equation: $d_i = x_i'\pi^d + \gamma_i^d$, $\mathbb{E}\left[\gamma_i^d x_i\right] = 0$
2. Consider the outcome equation: $y_i = x_i'\pi^y + \gamma_i^y$, $\mathbb{E}\left[\gamma_i^y x_i\right] = 0$
3. Consider the regression model: $\gamma_i^y = \alpha\gamma_i^d + \epsilon_i$

- γ_i^y is the residual left after partialling out linear effect of x_i from y_i
- γ_i^d is the residual left after partialling out linear effect of x_i from d_i
- After partialling out, α is coefficient in the reg of γ_i^y on γ_i^d
- This is the so-called Frisch-Waugh-Lovell theorem

Double ML: Summary

Advantages

- Useful for approximately sparse models (most models are not overly complex, few x are useful to explain y)
- Safeguards against specification searches (ad-hoc model selection) and p-hacking (data manipulation)
- Useful for model selection: data-driven and flexible (can specify also non-linear terms and interactions between x)
- Rationalizes why naive Post-LASSO fails (correlation between d, y, x)
- **Use double selection to protect against omitted variable bias**

Double ML with hdm

Partial fit via post-LASSO

```r
rY = rlasso(fmla.y, data = dat)$res
rD = rlasso(fmla.d, data = dat)$res
partial.fit.postlasso = lm(rY ~ rD)
```

Function “rlassoEffect” for double ML methods

```r
PO = rlassoEffect(X[, -1], y, X[, 1], method = "partialling out")
# Does the same as partial.fit.postlasso above
DS = rlassoEffect(X[, -1], y, X[, 1], method = "double selection")
# The two methods are first-order equivalent in both low- and
# high-dimensional settings under regularity conditions
```

Inference on a set of variables of interest (Belloni, Chern., Kato 14)

```r
lasso.e = rlassoEffects(fm, I = ~X1 + X2 + X3 + X50, data = data)
summary(lasso.e)
confint(lasso.e)
plot(lasso.e, main = "Confidence Intervals")
```

Review of causal inference methods in HD

1. Double ML methods and post-LASSO
   - Belloni and Chernozhukov 2013; Chernozhukov et al. 2017

2. Causal Trees and Causal Forests
   - Athey and Imbens 2016; Wager and Athey 2018, Athey et al. 2019

Review of causal inference methods in HD

1.
    -

2. **Causal Trees and Causal Forests**
    - Athey and Imbens 2016; Wager and Athey 2018, Athey et al. 2019

Importance of uncovering heterogeneities

Figure: "When U.S. air force discovered the flaw of averages" by Todd Ross
https://www.thestar.com/news/insight/2016/01/16/when-us-air-force-discovered-the-flaw-of-averages.html

The Story, Part I

- In the late 1940s, the United States Air Force had a serious problem: its pilots could not keep control of their planes (up to 17 deaths per day)
- Pilots had already been pre-selected because they appeared to be average sized
- Military engineers began to wonder if the pilots had gotten bigger over time

Fig. The cockpit problem (U.S. Air Force photo)

- In 1950, the Air Force measured more than 4,000 pilots on many dimensions of size, and then calculated the average for each dimension
- Everyone believed this improved calculation of the average pilot would lead to a better-fitting cockpit and reduce the number of crashes

The Story, Part II

- Gilbert Daniels, a newly hired 23-year-old scientist, had doubts
- "How many pilots really were average?" The average pilot did not exist
- "Average" pilot defined by having most measures within the average range ($\pm$30%)

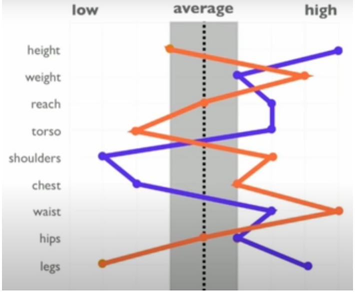


Fig.: Example for 2 pilots (Rose, 2013, "The myth of the averages")

The Story, Part II

- Gilbert Daniels, a newly hired 23-year-old scientist, had doubts
- "How many pilots really were average?" The average pilot did not exist
- "Average" pilot defined by having most measures within the average range ($\pm$30%)

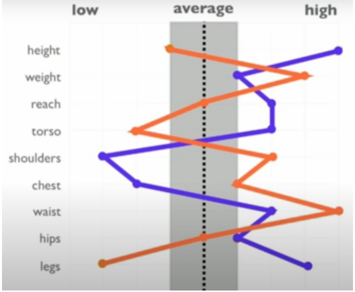


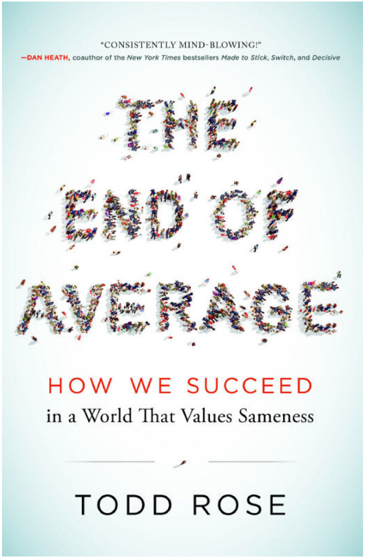
Fig.: Example for 2 pilots (Rose, 2013, "The myth of the averages")

- Out of 4,063 pilots, no single airman fit within the average range on all dimensions
- Less than 3.5% of pilots would be average sized on 3 dimensions

$\Rightarrow$ **Cockpits designed to fit the average pilot would fit no one**

$\Rightarrow$ **Adjustable seats were born. Pilots' performance boomed**

"Systems designed around the average are doomed to fail"



Why heterogeneity matters in science?

Effect heterogeneity: Study how causal effects vary in different subpopulations

Why heterogeneity matters in science?

Effect heterogeneity: Study how causal effects vary in different subpopulations

1. Personalize treatment effects and policy targeting
2. Generalize the causal finding to different populations
3. Better understanding of the causal mechanism
4. Make inference less sensitive to unmeasured confounding

$\Rightarrow$ However, modern applications can easily have tens or hundreds of potential effect modifiers: In this case, it is impractical to consider the subgroups exhaustively

$\Rightarrow$ ML methods come to hand

Using statistics to detect heterogeneity

Traditional approaches:

Using statistics to detect heterogeneity

Traditional approaches:

1. Add interaction terms
2. Stratify the sample

Problems:

1. Which heterogeneity should be pre-specified?
2. P-hacking/data dredging: Report only significant heterogeneity
   "Exploration - which some might term data dredging - is quite different from exogenous selection of a few comparisons. Both have their place. We need to be prepared to deal with either." (Tukey, 1991)
3. Overlook unexpected types of heterogeneity
4. How to stratify continuous variables?
5. Possible interactions $>$ data points likely
6. Spurious heterogeneity: multiple testing problem

Causal forest

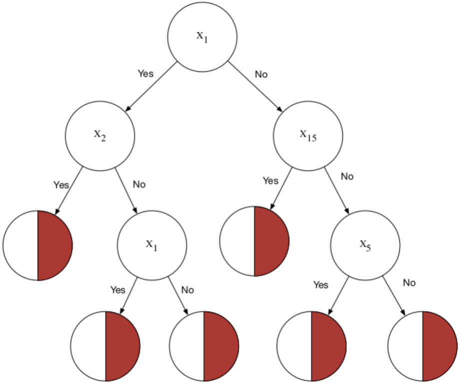
1. Handles large X dimension (failure of standard methods like OLS with interactions, nearest neighbor and kernel matching)
2. Captures possibly complex interactions in data-driven specification
3. *Consistently* estimates *full* distribution of causal effects conditional on x
   $\rightarrow$ estimate a targeting function that maps attributes x to causal effects for each individual

$$\Downarrow$$

Causal forests (Athey et al. 19) $\sim$ **Data-driven** way to estimate **heterogeneous** causal effects

Causal tree

- In each leaf there are treated (red) and untreated (white) obs.
- Causal effect $= \bar{Y}_{red} - \bar{Y}_{white}$
- Each leaf estimates:
  - $\hat{y}(1) = \hat{E}[Y(1)|X] = \bar{Y}(1)$
  - $\hat{y}(0) = \hat{E}[Y(0)|X] = \bar{Y}(0)$
- Causal effect $\hat{\Delta} = \bar{Y}(1) - \bar{Y}(0)$



Causal RF (Wager and Athey 18)

- Causal RF (vs. tree) allows for personalized estimates
- Estimate $\hat{\Delta}$ in each tree with OOB obs. and take their average

Recursive Partitioning for Causal Effects

- Replace y for prediction trees with Δ:

$$\min_{j,s} \quad \sum_{i:x_i \in R_1(j,s)} (\Delta_i - \hat{\Delta}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (\Delta_i - \hat{\Delta}_{R_2})^2$$

But we do NOT observe Δ_i

[2] For derivations, see, e.g., Hitsch and Misra, 2018; Athey and Imbens, 2016; Lundberg, 2017: "A tutorial in high-dimensional causal inference" (link)

Recursive Partitioning for Causal Effects

- Replace y for prediction trees with Δ:

$$\min_{j,s} \sum_{i:x_i \in R_1(j,s)} (\Delta_i - \hat{\Delta}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (\Delta_i - \hat{\Delta}_{R_2})^2$$

  But we do NOT observe Δ_i

  Instead of MIN prediction error (unfeasible):

  SPLIT BY MAX VARIANCE of treatment effects ACROSS LEAVES

- Maximize size (sum of squares) of within-leaf treatment effect as[2]:

$$\max_{j,s} \sum_{i:x_i \in R_1(j,s)} (\hat{\Delta}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (\hat{\Delta}_{R_2})^2$$

[2] For derivations, see, e.g., Hitsch and Misra, 2018; Athey and Imbens, 2016; Lundberg, 2017: "A tutorial in high-dimensional causal inference" (link)

Non-random treatment assignment

- What if treatment is NON-RANDOM? (self-selection)

Covariate imbalance between treated and untreated units

- Confounders are correlated with Y and treatment assignment D

  $\Rightarrow$ Confounding factors induce correlation between Y and D that is NOT indicative of the change in Y due to D (causal effect)

- Need to control for all of them, or for the conditional probability of being treated given these factors (known as propensity score, PS)

Causal RF need adjustment!

Causal forest

Idea:

- Why not running a regression within each leaf?
  $\Rightarrow$ Use double ML methods to estimate Δ in each leaf
- Step 1,2: Predict outcome y and treatment variable d using x
- Residualize outcomes as $y - \hat{y}$ and treatment as $d - \hat{d}$
- Step 3: Predict causal effect by regressing $y - \hat{y}$ on $d - \hat{d}$
- Build causal forest by running step 3 regression in each node

Causal forest

Idea:

- Why not running a regression within each leaf?
  $\Rightarrow$ Use double ML methods to estimate Δ in each leaf
- Step 1,2: Predict outcome y and treatment variable d using x
- Residualize outcomes as $y - \hat{y}$ and treatment as $d - \hat{d}$
- Step 3: Predict causal effect by regressing $y - \hat{y}$ on $d - \hat{d}$
- Build causal forest by running step 3 regression in each node

How to compute uncertainty in estimation?
$\Rightarrow$ Variance of causal effects and CI via bootstrap methods

How to obtain $\hat{y}$ and $\hat{d}$ for step 3?
$\Rightarrow$ Via two separate CART or LASSO, or whatever ML method for prediction

Causal Random Forest example with grf

Build a Causal Random Forest

```
tuned.forest <- causal_forest(X, Y, W, Est. (W = treatm. vector)
                 data=waste_dat,          Dataset to use
                 mtry=sqrt(ncol(X)),      m, higher if X collinear
                 num.trees=1000,          The more, the better
                 min.node.size=10,        Min. nr. obs. per leaf
                 ...)
```

Prediction

```
pred <- predict(tuned.forest)   OOB (or specify test set)
```

Causal effects

```
pred$predictions[W == 1]        Causal effect for treated obs.
mean(pred$predictions[W == 1])  Avg causal effect on treated
sqrt(pred$variance.estimates)   St. errors of causal effects
```