State of the art methods for Causal Inference

IN4143: Data Analysis and Causal Inference

José Arenas R.

Synthetic Control

Introduction to Synthetic Controls

- Well established, but not traditional yet

- Popularized by Abadie, Diamond and Hainmueller (2010, JASA)

- Kind of “diff-in-diff with matching”
  - Same setting as DiD → **Panel data**, some units are treated in some periods
  - It estimates “Adjusted differences”
    - We use pre-treatment data to “adjust for differences between treatment groups”
    - The **causal effect** is equal to the adjusted post-treatment differences between treatment groups
  - Differences with DiD
    - The **pre-treatment adjustment** is made by a matching algorithm (not regression)
    - The **purpose of matching** is to eradicate pre-treatment differences and doesn’t take into account the propensity of being treated

Important aspects of Synthetic Controls

- It relies on **large data availability** before treatment

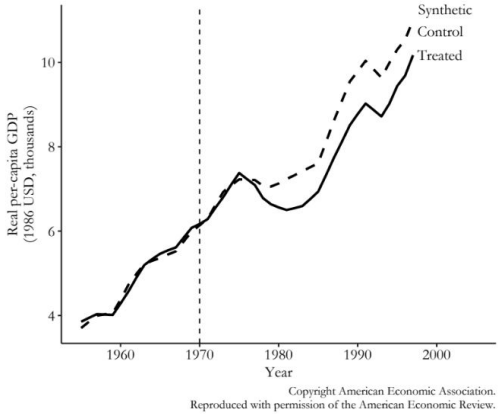
- **After the matching** phase both groups "treated" and "control" must be almost identical before treatment
  - In most cases uses the outcome variable as a matching variable (before treatment)

- The statistical significance is not determined by discovering the sample distribution beforehand, but by making "**randomization inference**"
  - A placebo test for estimating a null distribution with which compare the estimations of the causal effect, by falsifying the treated units → it cycle through every control observation estimating the synthetic control estimates generating a distribution of possible effects. If the original estimates are far into the tails, they are more robust.

Intuition behind Synthetic Controls

- It begins with a group of **treated** observation **and** a "**donor set**" (potential controls)

- Implements a matching algorithm using pre-treatment data, assigning a specific weight to each of the possible control units
  - The temporal trends before the treatment period must be almost identical for treated and controls after the matching phase



Example of Synthetic Controls (Abadie and Gardeazabal 2003)

- Violent conflict in the Basque Region of Spain (late 60s)
  - ETA separatist group
- Effects on economic activity (GDP)
- Donor set ⟶ 17 regions

Example of Synthetic Controls (Abadie and Gardeazabal 2003)

- Violent conflict in the Basque Region of Spain (late 60s)
  - ETA separatist group
- Effects on economic activity (GDP)
- Donor set ⟶ 17 regions

Example of Synthetic Controls (Abadie and Gardeazabal 2003)

TABLE 3—PRE-TERRORISM CHARACTERISTICS, 1960'S

	Basque Country (1)	Spain (2)	"Synthetic" Basque Country (3)
Real per capita GDP[a]	5,285.46	3,633.25	5,270.80
Investment ratio (percentage)[b]	24.65	21.79	21.58
Population density[c]	246.89	66.34	196.28
Sectoral shares (percentage)[d]			
Agriculture, forestry, and fishing	6.84	16.34	6.18
Energy and water	4.11	4.32	2.76
Industry	45.08	26.60	37.64
Construction and engineering	6.15	7.25	6.96
Marketable services	33.75	38.53	41.10
Nonmarketable services	4.07	6.97	5.37
Human capital (percentage)[e]			
Illiterates	3.32	11.66	7.65
Primary or without studies	85.97	80.15	82.33
High school	7.46	5.49	6.92
More than high school	3.26	2.70	3.10

Sources: Authors' computations from Matilde Mas et al. (1998) and Fundación BBV (1999).

[a] 1986 USD, average for 1960–1969.

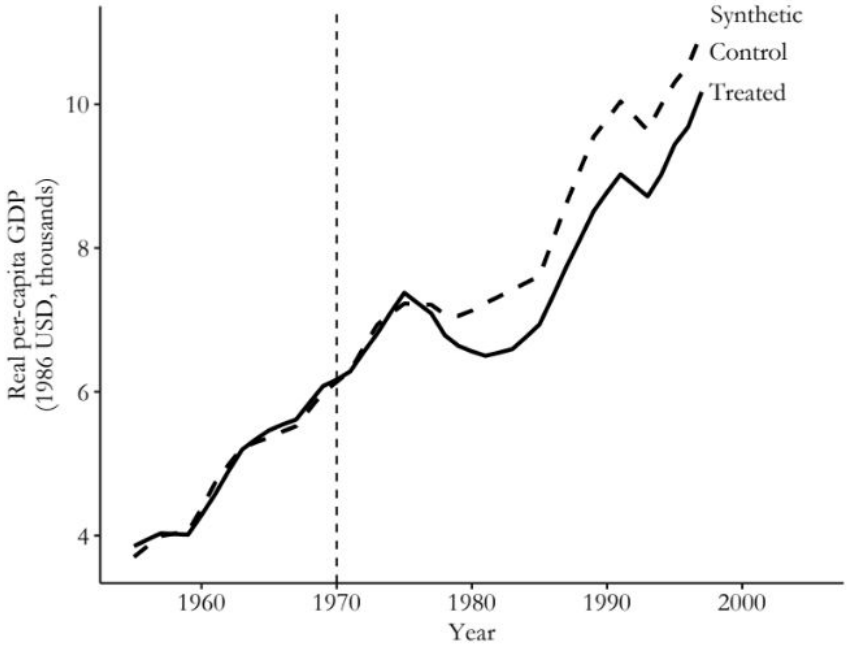
[b] Gross Total Investment/GDP, average for 1964–1969.

[c] Persons per square kilometer, 1969.

[d] Percentages over total production, 1961–1969.

[e] Percentages over working-age population, 1964–1969.

Example of Synthetic Controls (Abadie and Gardeazabal 2003)



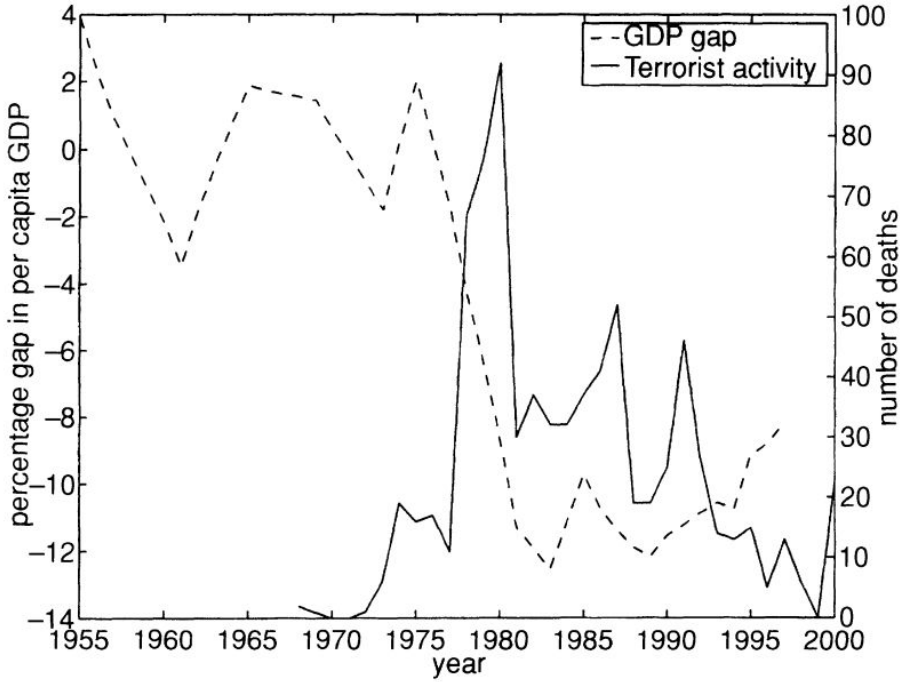


FIGURE 2. TERRORIST ACTIVITY AND ESTIMATED GAP

Example of Synthetic Controls (Abadie and Gardeazabal 2003)

- Placebo test ⟶ Catalonia (“a nonterrorism region”)
  - Very similar regions ⟶ Actually outperformed the synthetic control

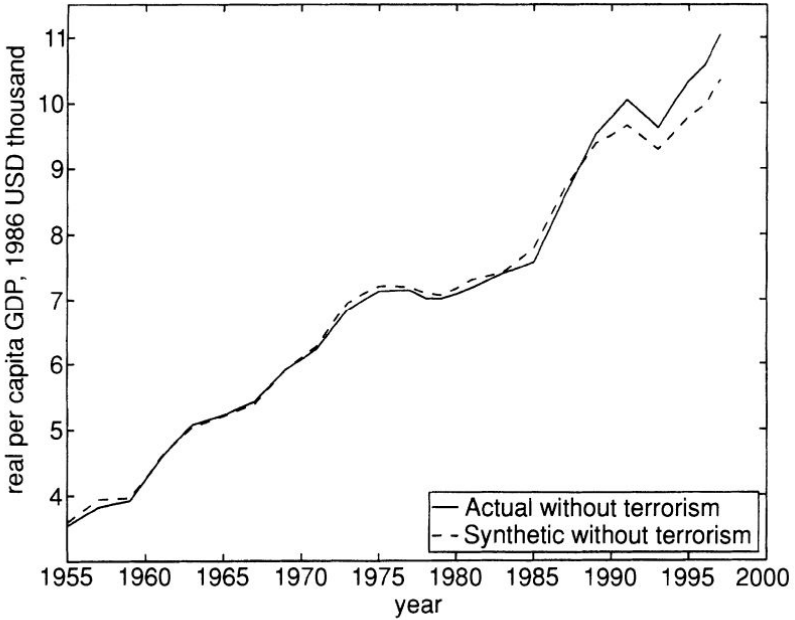
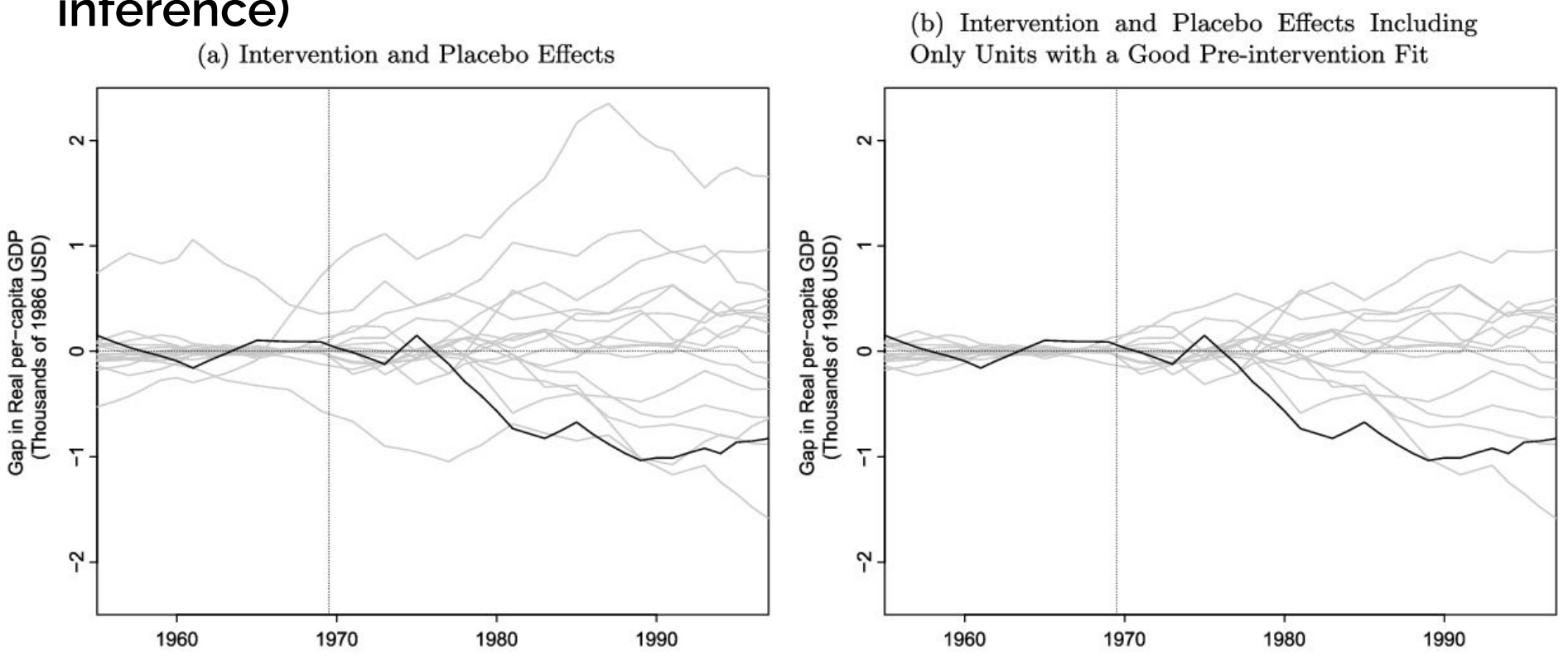


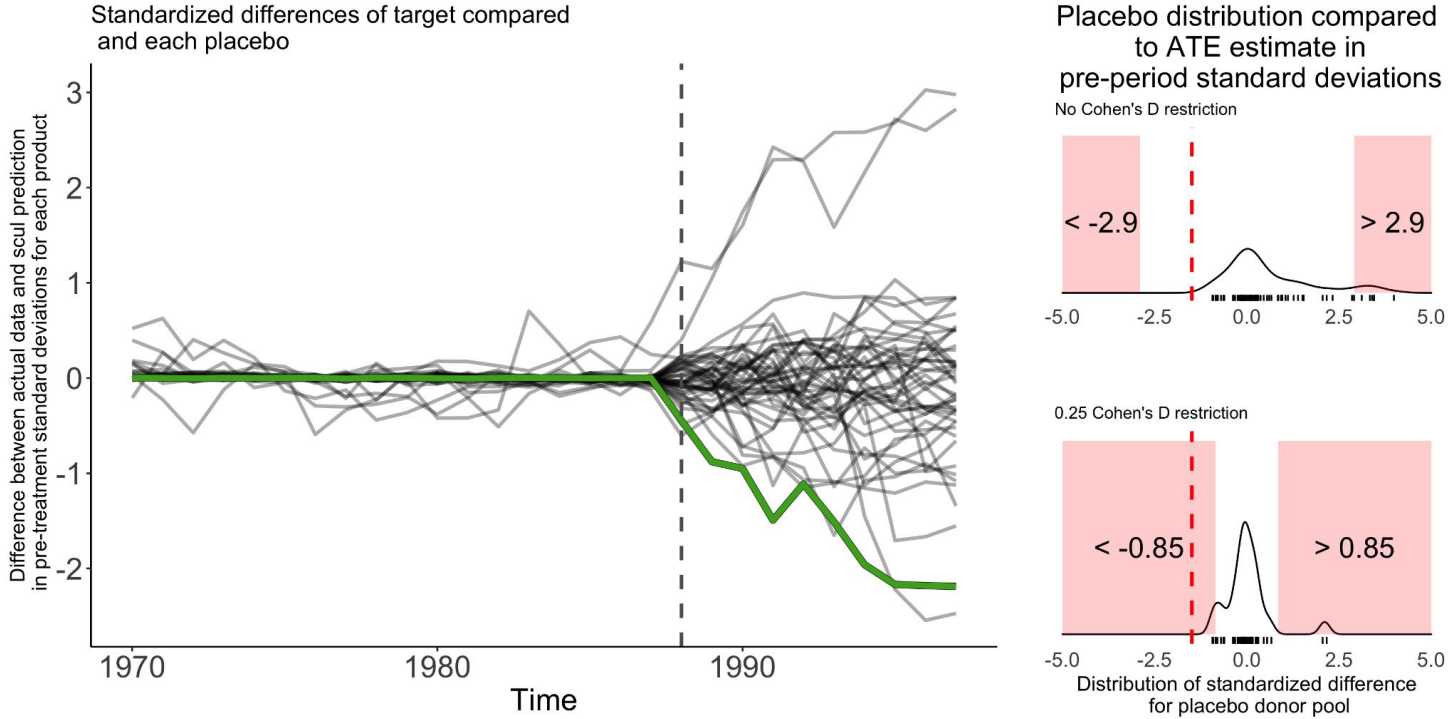
FIGURE 4. A “PLACEBO STUDY,” PER CAPITA GDP FOR CATALONIA

Example of Synthetic Controls (how it looks randomization inference)



Source: Synthetic “Control Method: Inference, Sensitivity Analysis and Confidence Sets (Firpo and Possebom, 2018).”

Example of Synthetic Controls (how it looks randomization inference)



Source: “Tactics for design and inference in synthetic control studies: An applied example using high-dimensional data (Hollingsworth and Wing, 2020).”

Pros and cons of Synthetic Controls (vs. DiD)

"It may well be the most important innovation in the policy evaluation literature in the last fifteen years (Athey and Imbens 2017)"

- **Advantages**
  - It doesn't rely on the parallel trends assumption
    - It "sorta forces" it's own assumptions
    - It makes more discipline the process of finding a valid control group
  - Less sensitive to model specification (not a regression method)
  - Naturally identify dynamic effects without accounting for interaction terms

- **Disadvantages** (Why is not more popular?)
  - It requires a lot of pre-treatment data in order to obtain a high quality matching
  - It tends to overfit (adjust to the noise of the outcome variable)
  - Feels fishy to match based on the outcome variable

Matrix Completion

Introduction to Matrix Completion

- "Using regularized regression to predict missing untreated values, and then comparing the actual treated value to those predictions in order to get an average treatment on the treated" (with ML), kind of

- Debut in Athey et al. (2021)

- Works with panel data, by building **two matrices**: the "treated matrix" and the "untreated matrix"
  - **Treated** matrix ⟶ Value of the outcome variable conditional on getting the treatment
  - **Untreated** matrix ⟶ Value of the outcome variable conditional on not getting the treatment

  **In each cell the values could be 1s, 0s or "?" (if we have a binary outcome)*

Building the matrices in the Matrix Completion method

- If we are looking at the "treated" matrix and the unit didn't receive the treatment in that period, the cell value would be "**?**"

	Untreated				Treated		
Ind.	Time1	Time2	Time3	Ind.	Time1	Time2	Time3
1	1	0	1	1	?	?	?
2	?	1	?	2	0	?	1
3	1	?	1	3	?	0	?
4	?	?	?	4	1	1	0

Table 21.1: Outcomes for Treated and Untreated Individuals

Source: "The Effect: An Introduction to Research Design and Causality (Huntington 2022)"

- Matrix completion is all about filling the "?s" in the "untreated matrix"
  - To have an "untreated" comparison for every treated unit
  - Once we have every "?s" from the untreated matrix imputed we average together the estimated treatment effects over all the treated observations to get an "Average Treatment on the treated"

Key aspects Matrix Completion method

- To fill the “?s” in the untreated cells (in the untreated matrix) we use the “data from the others untreated units”

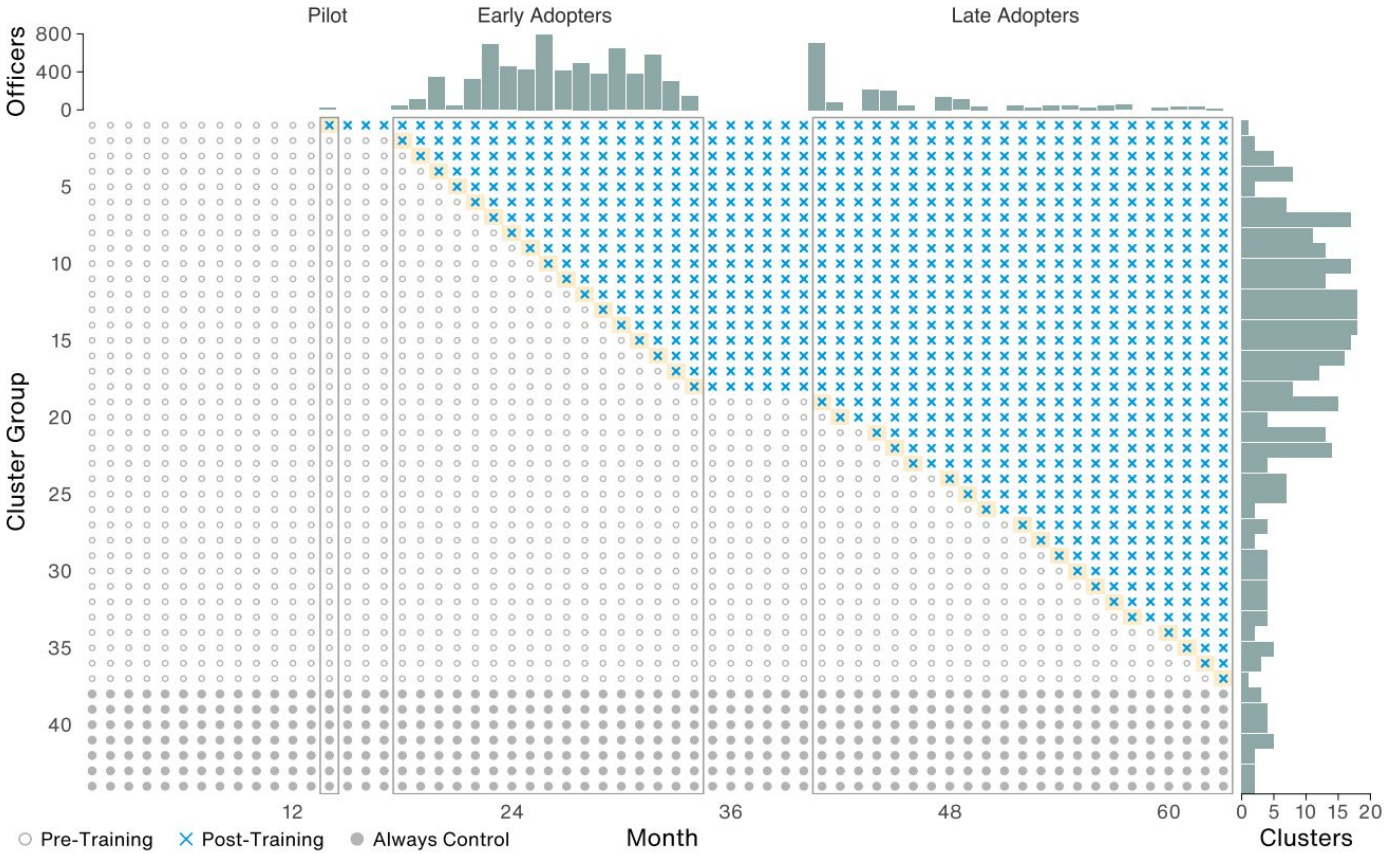
- Determines for each control unit which (and how) other observations are useful for the imputing process by using **regularization** and **matching** weights based on covariates

- **Main assumption** → It requires that the treatment is random conditional on whatever you use to create comparison weights
  - Like regression methods and matching

Example of Matrix Completion (Wood, Tyler, and Papachristos 2020)

- They study → Whether training on procedural justice methods—which encourage respect, neutrality, and transparency in the course of police work—can reduce the use of force.
- Chicago police officers
  - Thousands of policemen took the course
  - Did they receive complaints or “use force” on the job?
  - Staggered adoption

Example of Matrix Completion (Wood, Tyler, and Papachristos 2020)



Example of Matrix Completion (Wood, Tyler, and Papachristos 2020)

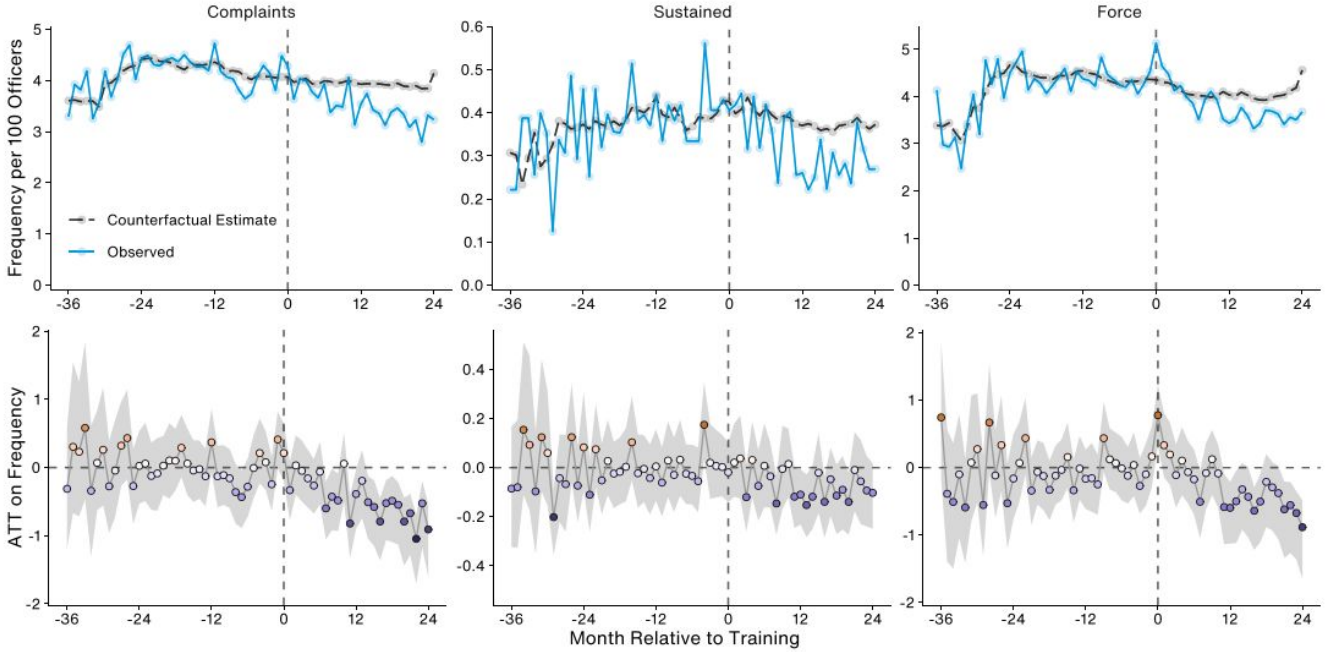


Fig. 2. (*Top*) Observed and counterfactual estimates of complaints, sustained or settled complaints, and use of force per 100 officers per month. Months are recalibrated to be relative to the onset of training. (*Bottom*) The ATT for each month is the estimated counterfactual frequency subtracted from the observed frequency in that month. Monthly ATT estimates are colored according to their value relative to zero. The 95% CIs are computed using 2,000 block bootstrap runs at the cluster level.

Example of Matrix Completion (Wood, Tyler, and Papachristos 2020)

- **Results** → Taking the course reduces de # of complaints by 11% and use of force by almost 8%

Table 1. Average effect of training on complaints received, sustained or settled complaints, and mandatory reports of use of force

	Complaints	Sustained	Force
Cumulative ATT	−11.60	−1.67	−7.45
SE	2.09	0.61	2.33
95% CI	−15.60, −7.45	−2.81, −0.40	−12.40, −3.37
P	<0.001	0.008	0.002
Cluster fixed effects	Yes	Yes	Yes
Month fixed effects	Yes	Yes	Yes
Officers	8,618	8,618	8,618
Months	63	58	63
Clusters	328	328	328
Treated clusters	306	295	306
Always-control clusters	22	33	22
Observations	20,664	19,204	20,664

The cumulative ATT represents the average reduction after 24 mo per 100 trained officers. The 95% CIs are computed using 2,000 block bootstrap runs at the cluster level.

Matrix Completion

- **Final Comment** about Matrix completion
  - Is more of a “general idea” than a specific algorithm
    - Synthetic controls and DiD are specific cases of matrix completion

Causal Discovery

Introduction to Causal Discovery

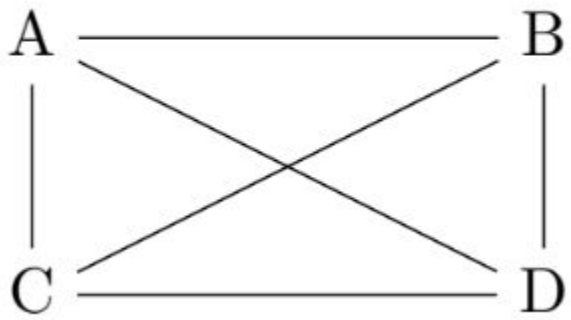
- Is the process of using data to **develop causal diagrams**
  - So far we have seen that we could generate a DAG based on experts knowledge, but can we build them purely from data? → Sometimes
  - How can this be possible? when a given dataset may be consistent with multiple DAGs
    - It depends on the algorithm (and there are many processes and algorithms for Causal Discovery)

- **SGS algorithm** (Spirtes-Glymour-Scheines)
  1. For each pair of variable → Is there an arrow between them?
  2. Once we know where the arrows are → Where are pointing each of the arrows?

  **** The algorithm is more complex than this → it also has ways to deal with unmeasured variables and other approaches to identify the causal relations (e.g., time data)*

SGS Algorithm for Causal Discovery example

- We have four variables (A-B-C-D)
  - We want to determine the causal relations between them (if any)

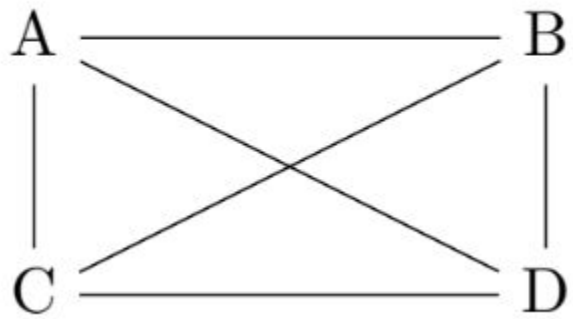


Steps:

1. For each pair of variables → How can we figure out whether there is an arrow between them?
   a. Answer: By checking conditional associations in the data
      i. Is there a correlation between A and B? → if not, delete the line between them
      ii. Does this correlation persists after controlling for C? → if not, delete the line between them

SGS Algorithm for Causal Discovery example

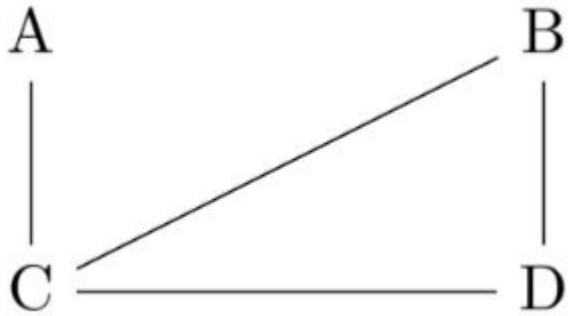
- We have four variables (A-B-C-D)
  - We want to determine the causal relations between them (if any)



- Corr(A,B) = 0 ; Corr(A,B | C) != 0
- Corr(A,C) != 0
- Corr(A,D) = 0
- Corr(B,C) != 0 ; Corr(B,C | D) != 0
- Corr(B,D) != 0
- Corr(C,D) != 0

SGS Algorithm for Causal Discovery example

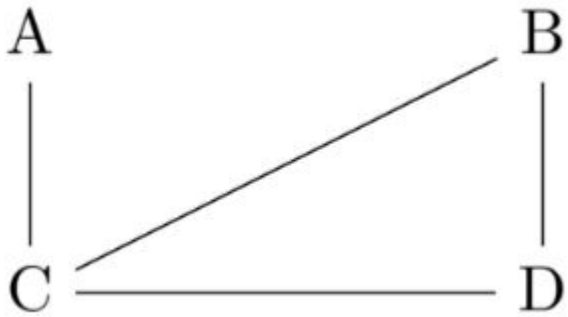
- We have four variables (A-B-C-D)
  - We want to determine the causal relations between them (if any)



- Corr(A,B) = 0 ; Corr(A,B | C) != 0
- Corr(A,C) != 0
- Corr(A,D) = 0
- Corr(B,C) != 0 ; Corr(B,C | D) != 0
- Corr(B,D) != 0
- Corr(C,D) != 0

SGS Algorithm for Causal Discovery example

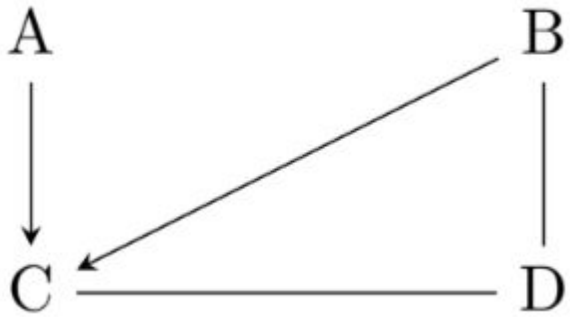
2. Which direction are they pointing at?



- Let's look at the path A - C - B
  - This could be filled in with
    - A→C→B ; A← C ← B ; A ← C → B ; A → C ← B
    - If we are seeing that …
      - There is no correlation between A and B w/out controls
      - There is a correlation between A and B after controlling for C
      - C must be a collider!

SGS Algorithm for Causal Discovery example

2. Which direction are they pointing at?



- Let's look at the path A - C - B
  - This could be filled in with
    - A→C→B ; A← C ← B ; A ← C → B ; A → C ← B
    - If we are seeing that
      - There is no correlation between A and B w/out controls
      - There is a correlation between A and B after controlling for C
      - C must be a collider!
- Let's look at the path C - D - B
  - We get a relationship between B and C whether or not controlling for D
  - We can't say anything else

*we might have run into an "**equivalence class**" → set of DAGs that our data can't distinguish*

Double Machine Learning

Introduction to Double Machine Learning

- DML ⟶ another way to close “back doors”

- The **problem with ML** techniques is that they are usually designed to be good at predicting, but not for statistical or causal inference.
  - So DML finds one step in a standard causal inference design that is inherently a prediction problem

- Which part of the standard causal inference design is a prediction problem?

Introduction to Double Machine Learning

- DML ⟶ another way to close “back doors”

- The problem with ML techniques is that they are usually designed to be good at predicting, but not for statistical or causal inference.
  - So DML finds one step in a standard causal inference design that is inherently a prediction problem

- Which part of the standard causal inference design is a prediction problem?
  - Controlling for covariates! (and estimation of interaction terms)

Controlling for covariates in OLS

$$Y = \beta_0 + \beta_x X + \beta_W W + \beta_Z Z + \varepsilon$$

```
> summary(lm(Y ~ X + W + Z, data = df))

Call:
lm(formula = Y ~ X + W + Z, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7944 -0.5867 -0.1038  0.6188  2.3280

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.008333   0.561452  -0.015   0.9882
X            2.968389   0.088823  33.419  < 2e-16 ***
W            0.182038   0.042884   4.245 5.05e-05 ***
Z           -0.024275   0.009866  -2.460   0.0157 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8867 on 96 degrees of freedom
Multiple R-squared:  0.9234,    Adjusted R-squared:  0.921
F-statistic: 385.6 on 3 and 96 DF,  p-value: < 2.2e-16
```

Controlling for covariates as a prediction task

```
> modelo_Y = lm(Y ~ W + Z, data = df)
> modelo_X = lm(X ~ W + Z, data = df)
>
> df$Yhat = predict(modelo_Y)
> df$Y_resid = df$Y - df$Yhat
>
> df$Xhat = predict(modelo_X)
> df$X_resid = df$X - df$Xhat
>
> summary(lm(Y_resid ~ X_resid, data = df))
```

$X = \alpha_0 + \alpha_W W + \alpha_Z Z + \varepsilon_X$

$Y = \gamma_0 + \gamma_W W + \gamma_Z Z + \varepsilon_Y$

$Y_R = \varepsilon_Y$

$X_R = \varepsilon_X$

$Y_R = \omega_0 + \omega_X X_R + \varepsilon_X^R$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.128e-16  8.776e-02    0.00        1
X_resid     2.968e+00  8.791e-02   33.77   <2e-16 ***
---
```

$\beta_x \approx \omega_X$

we are "removing" the effects of Z and W over X and Y, and then we estimate the relationship after the "cleaning"

Double Machine Learning

- DML consist in executing this "**prediction task**" via any ML technique

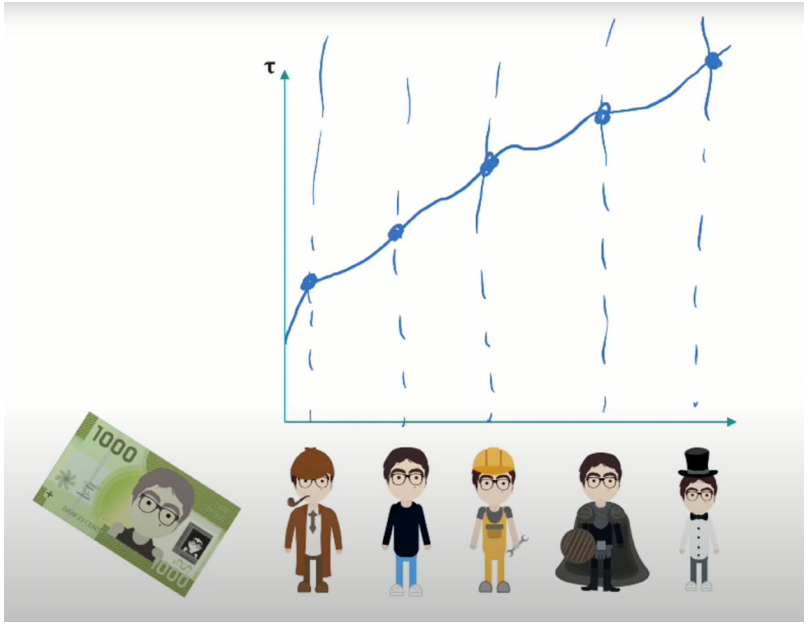
- DML e**mploys sample splitting** to avoid overfitting
  - In its basic form → splits the sample into two halves
    1. Train an ML model with one half, then predicts Y_R and X_R in the other half, and then use this other half to estimate the coefficient of interest
    2. Switch the datasets and repeat the process
    3. Average the estimated effects between the two processes

- Why using DML?
  - Better at high dimensional settings (multiple covariates)
  - Better at capturing nonlinear relations (w/out creating interaction terms)

Modeling Heterogeneous Treatment Effects

Introduction to HTE

- We live in a world of heterogeneous effects
  - We can’t escape h.e. but we can give ourselves some tools to face them



Introduction to HTE

- If we can estimate the **distribution of the effects** we can:
  - Estimate the average treatment effect (as before)
  - We could toss out the average and look at the distribution overall
  - Identify how the effect varies over the sample
    - Standard deviation of the effect
    - For what population the effect is bigger or smaller
      - Poor vs rich; young vs old people, etc

- Under traditional models → generate interaction terms in regressions
  - **Very limited** → You can really only include a couple of interaction terms before your model turns into a poorly-powered impossible-to-interpret slurry

Introduction to HTE

- Particularly active since 2010s
  - The estimation of heterogeneous effects may well end up being ML's most important contribution to causal inference.
  - One problem with HTE → There are so many dimensions they could vary along.
    - How could we check everything? → ML!

  - ML methods → aren't design to identify an effect
    - They assume that you already have a design for identifying an effect.
      - Once you've identified the effect, **these can help you see the entire distribution** of that effect

- **Causal Random Forest**

Introduction to Meta algorithms

- We are used to the idea of fitting a model, and then using it to make a unique prediction of the outcome for each observation.
  - But what if we want to get a unique estimate of the effect of X on Y for each unit?
    - The **regressiony way** is by estimating it through interaction terms (between the treatment and covariates), and then estimate the effect of that person through the sum of those coefficients.
    - The problem with interaction terms is that **regression** simply can't handle a lot of interaction terms, or highly nonlinear interactions.
      - So we end up with a very basic model (using very few interaction terms)

Causal ML: Two non exclusive Approaches

Doubly Robust Estimators Approach

T - Learner

R - Learner

Generalized Random Forest Approach

Causal Trees

Causal Forest

Doubly Robust Approach (Meta-Algorithms)

R - Learner (Nie & Wager 2017)

$$\hat{\tau}(\cdot) = argmin_{\tau} \sum_{i=1}^{n} ((Y_i - \hat{m}^{(-i)}(X_i)) - \tau(X_i)(W_i - \hat{e}^{(-i)}(X_i)))^2 + \Lambda_n(\tau(\cdot))$$

X - Learner (Künzel et al. 2018)

1. Estimar separadamente “Y predicho” para control y tratamiento
2. Imputar Y predicho para control y tratamiento con “modelos cruzados”
3. Estimar CATE con las bases “pooled”
4. Tomar el promedio ponderado de las estimaciones de CATE

$$\hat{\tau}(x) = \hat{g}(x)\hat{\tau}_0(x) + [1 - \hat{g}(x)]\hat{\tau}_1(x)$$

Generalized Random Forest Approach

Causal Tree (Athey & Imbens 2016)

- Partition the data into subpopulations with similar causal effects

Causal Forest (Athey, Tibshirani & Wager 2018)

- Specialization of Generalized Random Forest algorithm
  *R-Learner

Causal Random Forests

Introduction to Causal Random Forests

- CRF ⟶ Morph the task of estimating a unique effect of each unit into a format that works similar to a predicting task in the "ML" way.
  - What is a Random Forest?
    - In RF we split the data with the objective of minimizing the prediction error
    - In CRF we want to see variation in the treatment effect

- So we choose splits based on how different the estimated effect is in each side of the split
  1. For each possible split we estimate the resulting treatment effect in each terminal node
  2. We select the split which maximizes the difference in the treatment effects
  3. Once the split gets small enough, we stop splitting (stopping criterion)
  4. We save the estimation of the treatment effect for every unit in each terminal node
  5. We bootstrap the whole thing again and limit our choice of splitting variables
  6. We estimate the overall treatment effect for every unit by averaging across all the bootstrap samples

"honest split" ⟶ separate "tree growing data" from "treatment effects estimation data"

Generalized Random Forest Approach

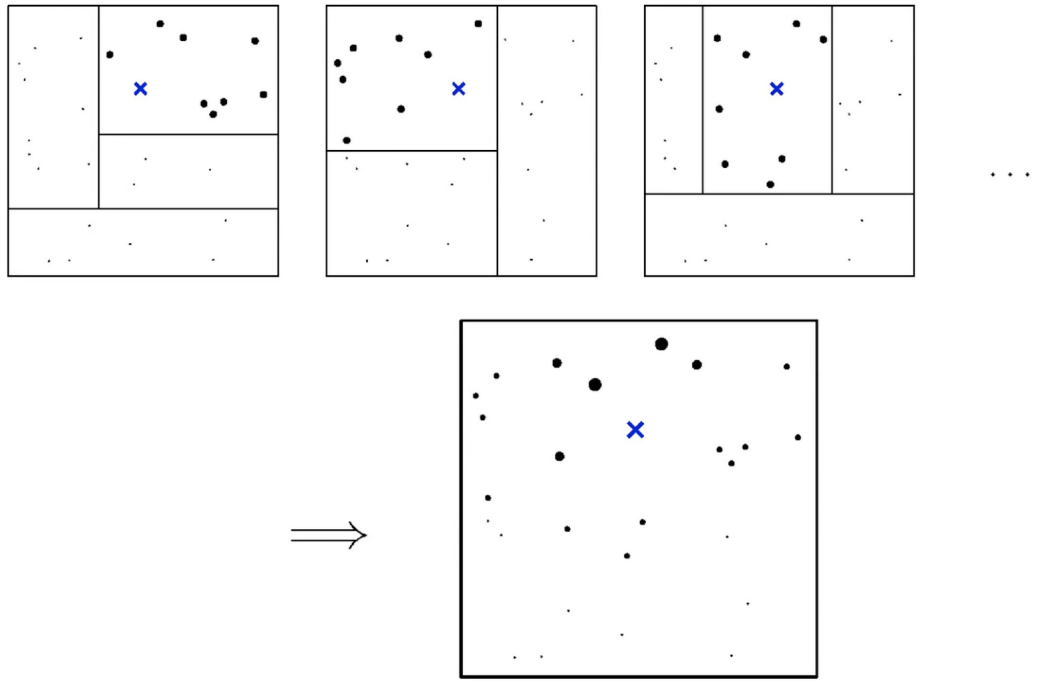
- **“Honest Split”**: Implies that the asymptotic properties of treatment effect estimates within leaves are the same as if the tree partition had been exogenously given
  - P1: Tree Structure
  - P2: Estimation

- Type TOT-H loss function

$$\widehat{\mathrm{EMSE}}_\tau(\mathcal{S}^{\mathrm{tr}}, \Pi) \equiv \frac{1}{N^{\mathrm{tr}}} \sum_{i \in \mathcal{S}^{\mathrm{tr}}} \hat{\tau}^2(X_i; \mathcal{S}^{\mathrm{tr}}, \Pi)$$

$$- \frac{2}{N^{\mathrm{tr}}} \cdot \sum_{\ell \in \Pi} \left(\frac{S^2_{\mathcal{S}^{\mathrm{tr}}_{\mathrm{treat}}}(\ell)}{p} + \frac{S^2_{\mathcal{S}^{\mathrm{tr}}_{\mathrm{control}}}(\ell)}{1-p} \right).$$

The “Kernelized KNN” interpretation



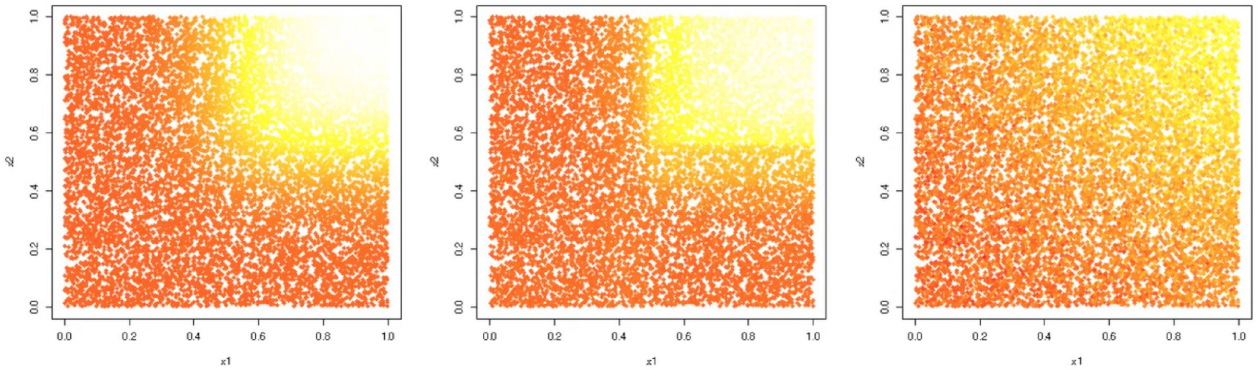
Forests induce a kernel via **averaging tree-based neighborhoods**.
This idea was used by Meinshausen (2006) for quantile regression.

Source: AEA Continue Education (Susan Athey, 2018)

Fit under synthetic data

The plots below depict $\hat{\tau}(x)$ for 10k random test examples, projected into the 2 signal dimensions.

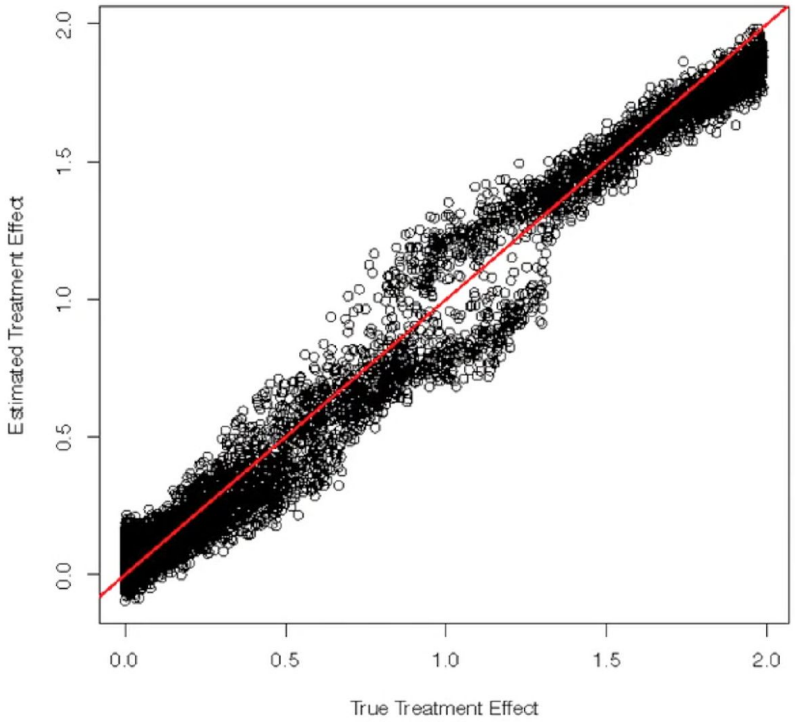
true effect $\tau(x)$ | **causal forest** | **k-NN estimate**



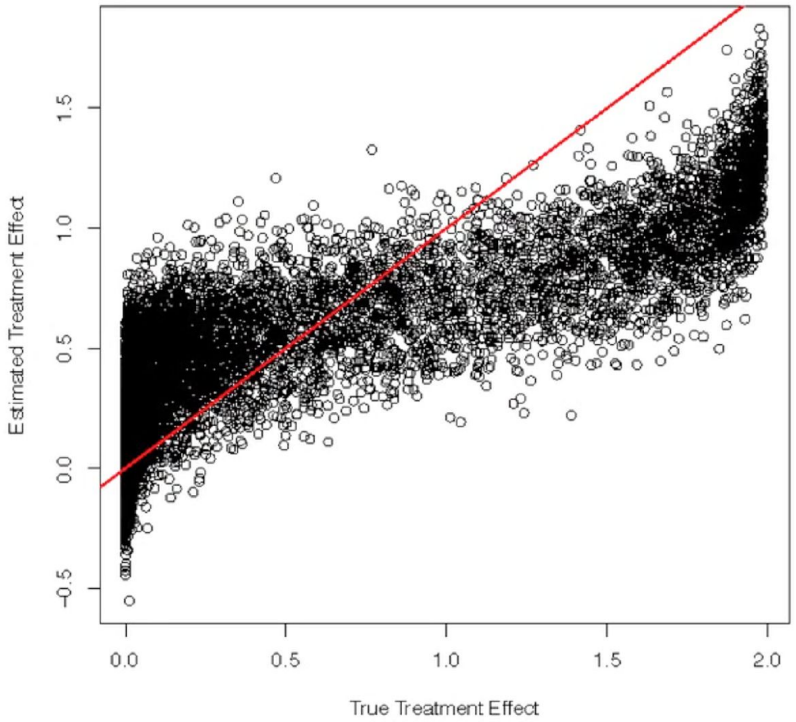
Software: `causalTree` for `R` (Athey, Kong, and Wager, 2015)
available at `github: susanathey/causalTree`

Source: AEA Continue Education (Susan Athey, 2018)

causal forest



k-NN estimate



For $p = 6$, the corresponding MSE ratio for τ is 2.2.

Source: AEA Continue Education (Susan Athey, 2018)

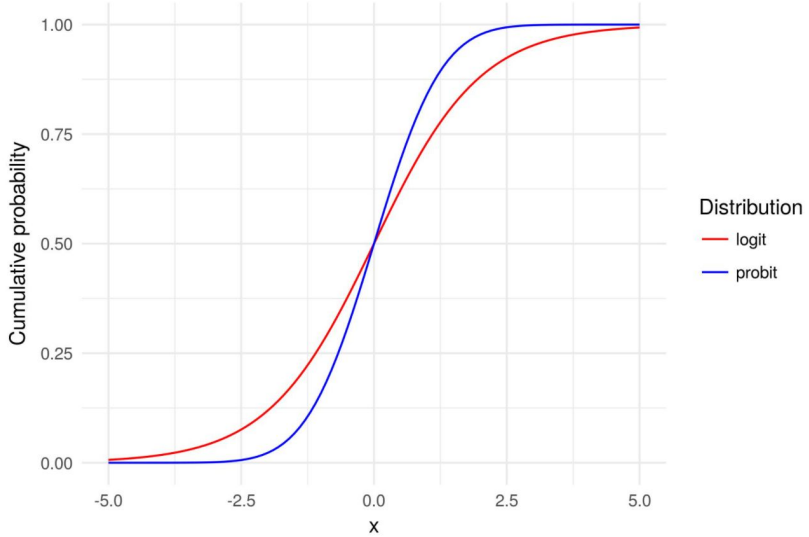
Causal Forest tutorial

R tutorial (by Susan Athey)

Sorted Effects

Introduction to Sorted Effects

- Many **“traditional” nonlinear regression models** naturally have treatment effect heterogeneity
  - Logit and Probit ⟶ since the effect can’t be that big if your probability of having a dependent value of 1 is near 1 or 0, naturally emerges heterogeneity in the treatment effect



Introduction to Sorted Effects

- Many **“traditional” nonlinear regression models** naturally have treatment effect heterogeneity
  - Logit and Probit ⟶ since the effect can’t be that big if your probability of having a dependent value of 1 is near 1 or 0, naturally emerges heterogeneity in the treatment effect

- After obtaining the individual treatment effects, they are **sorted** from the most sensible individual to the treatment to the least
  - We can study the composition of the treatment effect percentiles
  - We can study the distribution of the heterogeneous treatment effects

Introduction to Sorted Effects

- What makes this its own method?
  1. The designers of sorted effects managed to figure out, using bootstrapping, how to **estimate the standard errors** on these individual effects in an accurate way, and how to handle the noisiness in the tails of the effect distribution, both of which can be difficult

  2. They **introduced few methods** for what you can do with the distribution of effects (e.g., comparing who is in different parts of the distribution; who is the most and least affected)

  3. Because this is an approach to treatment effect heterogeneity, **reducing** the **dimensionality** to a **single dimension** ("τ" itself), it allows you to make models with more built-in heterogeneity, easier to interpret, and provide standard errors that account for all the interactions.

Example of Sorted Effects

- Effect of Vitamin E advice over Vitamin E intake
- What if the people who are more likely to follow the advice are the ones that are healthier?

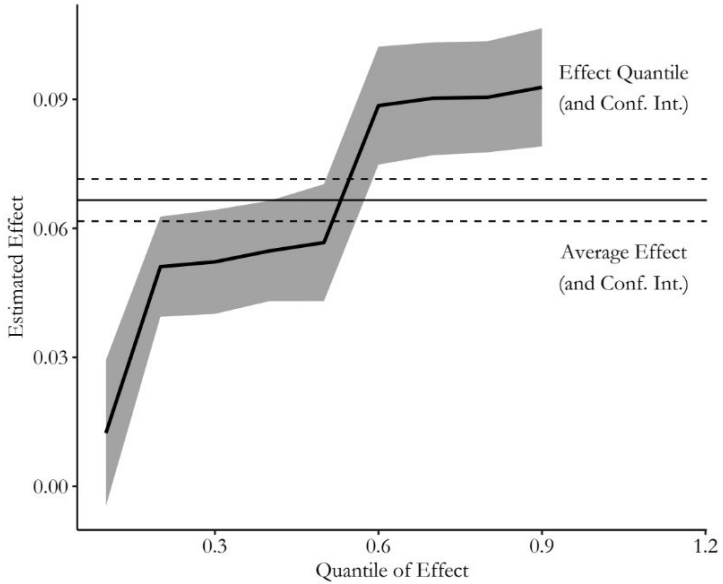


Figure 21.7: Sorted Distribution of Effects of Vitamin E Recommendation on Taking Vitamin E

Example of Sorted Effects

Effect:	10% Least Affected		10% Most Affected	
Variable	Mean	SD	Mean	SD
Smoking	1	0	0	0
Exercise Rating Score	-0.422	0.713	1.446	0.193
Vitamin Behavior Score	-1.985	0.387	1.706	0.105

Table 21.2: Characteristics of Those Least and Most Responsive to Health Recommendation for Vitamin E

Source: “The Effect: An Introduction to Research Design and Causality (Huntington 2022)“

People who followed the advice were healthier before the treatment (in terms of smoking and exercising habits)

- In the 10% most sensible to the treatment group, there are 0 smokers
- In the 10% least sensible to the treatment group, everyone is a smoker

Structural Estimation

Introduction to Structural Estimation

- **First** think about the entire underlying model and figuring out **from that** how to estimate an effect
- Disadvantages
  - Require some **strong assumptions** → they only work if you are right about the underlying data generating process.
    - It sounds risky, but if it's wrong, it is not always clear what other method actually tells you the truth anyway.
- Advantages
  - Doing structural modeling properly would not just use theory to determine the set of alternative pathways to consider, but would **also** see **what kind of statistical model the theory implies**.
    - Structural models estimate theoretical parameters directly
      - This means that it is easier to answer complex causal questions.
      - If you have a structural model, you immediately get access to how any variable in the model causes any other variable. And how that effect differs in different settings.
      - You can ask the model what's going on, you don't have to wait to see more data (just change the parameter values and simulate)
- Estimation methods
  - Maximum likelihood and generalized method of moments.

R packages

- Synthetic Controls → tidysynth and gsynth
- Matrix Completion → gsynth
- Causal Discovery → pcalg
- Causal Random Forest → grf

Any question?