Random Forest and Causal Effect Estimation

Jinhua Wang [1]   Melvyn Weeks [2]

[1] Judge Business School, Cambridge University

[2] Faculty of Economics, Cambridge University

May 23, 2019

Overview

1. Counterfactual - the Fundamental Problem
2. Data [Dehejia and Wahba, 1999]
3. Main Results
   - Trees in the Forest
   - Average Treatment Effect
   - Causal Forest Estimators
   - Heterogeneity
4. Variable Importance and Interactions (Preliminary)
5. Further Research

Counterfactual - the Fundamental Problem

Image we are calculating the impact of Columbus's discovery of America on the American GDP today.

It will be necessary to know what would American GDP have been if Columbus had never discovered the new continent - namely, the **counterfactual**, which is **not** observed.

To find out the counterfactual:

- We can send a man to a parallel universe.
- We can **predict** what would have happened.

Notations

Table: Definitions of Notations Used in This Presentation

	Definition
Symbol	
τ	The treatment effect.
Y_i^0	A scalar outcome of individual i who **did not** receive treatment
Y_i^1	A scalar outcome of individual i who **received** treatment.
X_i	A vector of covariates (characteristics) of individual i.
$W_i=0$	The individual did not receive treatment.
$W_i=1$	The individual received treatment.
$e(X_i)$	The propensity score (the probability of individual i with characteristic X_i receiving treatment).

Random Experiment Data

Unconditional Independence

In a random experiment, the outcome is independent of the treatment conditions.

$$\{Y_i^0, Y_i^1\} \perp\!\!\!\perp W_i \tag{1}$$

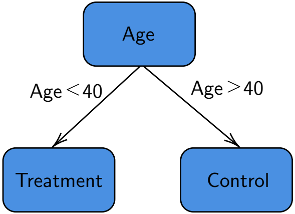
Average Treatment Effect (ATE)

Although it is impossible to recover the counterfactual for each individual, we can still calcualte the *average* counterfactual and infer the *average* treatment effect:

$$\tau = E(Y_i^1 \mid W_i = 1) - E(Y_i^0 \mid W_i = 0) \tag{2}$$

Observational Data and Challenges

Suppose individuals above age 40 are assigned to treatment group, while individuals below age 40 are assigned to control group. In this case, the average treatment effect estimated with equation (2) will be biased, because **age is correlated with income**.



$$\tau_{biased} = E(Y_i^1 \mid W_i = 1) - E(Y_i^0 \mid W_i = 0)$$

Figure: Non-random assignment

Observational Data and Challenges

In a non-random experiment setting, it is necessary to introduce the concept of conditional independence:

Conditional Independence

If the outcome of individual i is independent of the treatment conditional on covariates X_i:

$$\{Y_i^0, Y_i^1\} \perp\!\!\!\perp W_i \mid X_i \tag{3}$$

If assumption (3) holds, the outcome of individual i is also independent of the treatment conditional on the probability of treatment $e(X_i)$. This technique is also called the propensity score matching (PSM) [Rosenbaum and Rubin, 1983] :

$$\{Y_i^0, Y_i^1\} \perp\!\!\!\perp W_i \mid X_i \rightarrow \{Y_i^0, Y_i^1\} \perp\!\!\!\perp W_i \mid e(X_i) \tag{4}$$

Covariates

For example, in our data, X_i is a vector of the following variables:

$$X_i = \begin{bmatrix} re74_i \\ re75_i \\ age_i \\ education_i \\ black_i \\ hispanic_i \\ married_i \end{bmatrix} \tag{5}$$

Conditional independence simply means that the outcome variable Y_i is indepedent of the treatment conditional on X_i.

$e(X_i)$ is traditionally calculated with logit/probit regression which converts X_i into a single scalar ranged between 0 and 1.

Observational Data and Challenges

Conditional Average Treatment Effect (CATE)

[Rubin, 1974] show that if conditional independence holds, the **conditional** average treatment effect can be unbiasedly estimated as:

$$\tau = E[(Y_i^1 \mid W_i = 1) - E(Y_i^0 \mid W_i = 0) \mid X_i] \tag{6}$$

Alternatively with PSM,

$$\tau = E[(Y_i^1 \mid W_i = 1) - (Y_i^0 \mid W_i = 0) \mid e(X_i)] \tag{7}$$

Note that $e(X_i)$ essentially converts a vector X_i of covaraites into a scalar.

UNIVERSITY OF CAMBRIDGE

Observational Data and Challenges

However, several challenges remain.

- The counterfactual is still **not** observed, and PSM is a parametric model to predict the counterfactual.
- The form of the covariates X_i might be high dimensional.
- Probit and Logit models to estimate propensity score $e(X_i)$ asserts strong parametric and additive assumptions.
- **Propensity score estimation is a prediction problem**, but parametric regressions perform poorly on predictions.
- There could be omitted variable bias.

Observational Data and Challenges

For example, a typical logit model would assume **additive** assumptions:

$$\begin{aligned} P(Y_i = 1|X_i) = logit^{-1}(&\beta_0 + \beta_1 re74_i + \beta_2 re75_i + \beta_3 age_i \\ &+ \beta_4 education_i + \beta_5 black_i + \beta_6 hispanic_i + \beta_7 married_i) \end{aligned}$$

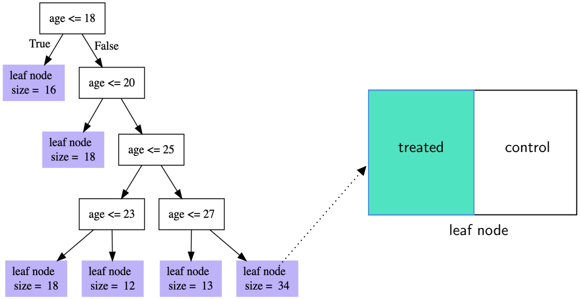
But are the parametric assumptions correct?

- One can simply argue that the income should be in *log* form.
- It is not clear why age is additive to income, education and ethnicity.
- It is not clear which of these variables interact with each others.

Model Selection Bias: What makes it worse is that, many researchers bruteforcely test many parametric forms and select the one that conforms with their expectation for the conclusion. This is a bias that cannot be simply tested with a coefficient t-test.

Tree-based Causal Effect Estimation

The algorithm will place splits on each of the covariates X_i recursively so that the mean-squared-error between the predicted outcome $\hat{Y}_i$ and actual outcome Y_i is minimized.



In a leaf node L of the tree, we can see the individuals in that leaf node as randomly assigned to treatment and control group.

Causal Tree

Under the assumption that the unconfoundedness assumption 3 holds, we can estimate the causal effect with a causal tree.

Causal Effect Estimation

The treatment effect can be consistently estimated as: [Wager and Athey, 2018]:

$$\begin{aligned}\hat{\tau}(X_i) = & \frac{1}{\mid \{i : W_i = 1, X_i \in L\} \mid} \sum_{i:W_i=1,X_i \in L} Y_i \\ & - \frac{1}{\mid \{i : W_i = 0, X_i \in L\} \mid} \sum_{i:W_i=0,X_i \in L} Y_i\end{aligned} \tag{8}$$

Random Forest and Inference

Machine Learning is not a blackbox - it is possible to do inference with it.

Asymptotic Normality of Random Forest

[Wager, 2014] shows that:

- Random forest predictions are **asymptotically normal** under certain conditions.
- The asymptotic variance can be estimated consistently with the **Infinitesimal Jackknife**.

[Wager et al., 2014] further show that:

- The Gaussian confidence interval can be estimated as $\hat{y} \pm z_\alpha \times \hat{\sigma}$
- $\hat{\sigma}$ is the standard error estimated with infinitesimal jackknife
- z_α is the z-score for normal distribution.

Causal Tree

Honest Estimation

- Contrary to the traditional ML, which uses adaptive learning, [Athey and Imbens, 2015] proposed a method called honest estimation, which separates the model construction from the estimation proccess.
- The training set is split into four subsamples: the training subsample, the estimation subsample, the cross-validation subsample and the test subsample.

Causal Forest

Bagging

Because a single tree in the causal forest is unstable in structure, bagging an ensemble of B number of causal trees will help to generate results with **lower variance**.

Treatment Effect

The treatment effect can be consistently estimated as $\hat{\tau}(X_i) = \frac{\sum_{b=1}^{B} \hat{\tau}_b(X_i)}{B}$ [Wager and Athey, 2018]

Inference

The treatment effect $\hat{\tau}(X_i)$ is asymptotically Gaussian distributed and pointwise consistent for the true treatment effect.
[Wager and Athey, 2018].

Data

Now, let's apply the idea of causal forest to a famous dataset - National Supported Work Demonstration Program (NSW) data collected in the mid-1970s.

NSW was a vocational job training program in the mid-1970s that lasted for 12-18 months designed to enhance the candidates' income in the US. Candidates were randomly allocated into treatment and control groups, where training is only provided to the candidates in the treatment group.

The outcome variable Y_i is the income after treatment ($re78$), and the treatment effect measures the income increase of the treated candidates compared to that of the controlled candidates.

NSW Data

Table: Summary Statistics for NSW Sample

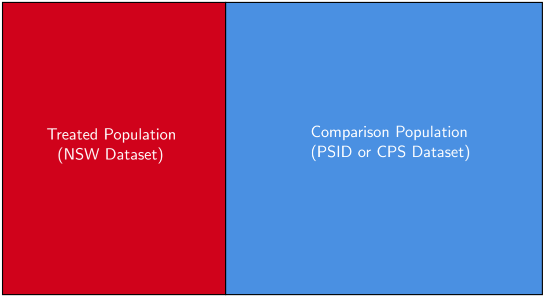
	count	mode	freq	mean	std
treat	445	0	260	**0.42**	
age	445			25.37	7.10
education	445			10.20	1.79
black	445	1	371	0.83	
hispanic	445	0	406	0.09	
married	445	0	370	**0.17**	
nodegree	445	1	348	0.78	
re74	445			2102.27	5363.58
re75	445			1377.14	3150.96
re78	445			5300.76	6631.49

treat, *black*, *hispanic*, *married* and *nodegree* are all binary variables.

[Dehejia and Wahba, 1999]

In 1999, [Dehejia and Wahba, 1999] **replaced** the controlled population in the dataset with non-random experimental data.

By doing so, we can use the NSW random experiment treatment effect as a **benchmark** for causal effect estimation in observational studies.



Treated Population
(NSW Dataset)

Comparison Population
(PSID or CPS Dataset)

Figure: PSID or CPS Synthetic Sample

PSID Data [Dehejia and Wahba, 1999]

Table: Summary Statistics for PSID (Synthetic) Sample

	NSW (Random) Sample			PSID (Synthetic) Sample		
	count	mean	std	count	mean	std
treat	445	**0.42**		2675	**0.07**	
age	445	25.37	7.10	2675	34.23	10.50
education	445	10.20	1.79	2675	11.99	3.05
black	445	**0.83**		2675	**0.29**	
hispanic	445	0.09		2675	0.03	
married	445	**0.17**		2675	**0.82**	
nodegree	445	0.78		2675	0.33	
re74	445	2102.27	5363.58	2675	18230	13722.3
re75	445	1377.14	3150.96	2675	17850.9	13877.8
re78	445	**5300.76**	6631.49	2675	**20502.4**	15632.5

CPS Data [Dehejia and Wahba, 1999]

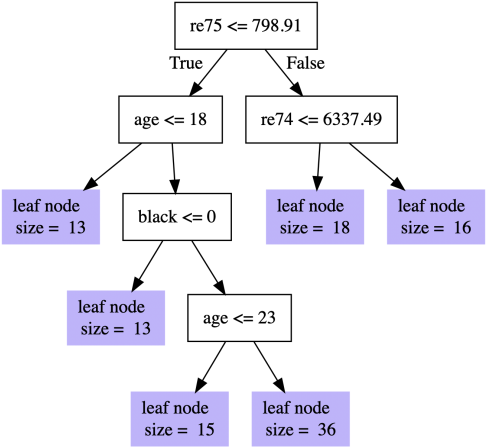
Table: Summary Statistics for CPS (Synthetic) Sample

	NSW (Random) Sample			CPS (Synthetic) Sample		
	count	mean	std	count	mean	std
treat	445	**0.42**		16177	**0.01**	
age	445	25.37	7.10	16177	33.14	11.04
education	445	10.20	1.79	16177	12.01	2.87
black	445	**0.83**		16177	**0.08**	
hispanic	445	0.09		16177	0.07	
married	445	**0.17**		16177	**0.71**	
nodegree	445	0.78		16177	0.30	
re74	445	2102.27	5363.58	16177	13880.5	9613.11
re75	445	1377.14	3150.96	16177	13512.2	9313.21
re78	445	**5300.76**	6631.49	16177	**14749.5**	9671

Main Results (Outline)

- Tree Example in the Causal Forest
- Average Treatment Effects
- Heterogeneity in Conditional Average Treatment Effects
  - Plot of CATE against education
  - Best Linear Fit Test
- Determinants of the Causal Effect
  - Variable Importance Measure
  - Variable Interaction Measure

100th. Tree in the NSW Forest



Average Treatment Effects ([Dehejia and Wahba, 1999])

Table: Average Treatment Effects [Dehejia and Wahba, 1999]

	Treated - Controlled	Stratification	Matching
NSW	1794 (633)		
PSID	-15205 (1154)	1608 (1571)	1691 (2209)
CPS	-8498 (712)	1713 (1115)	1582 (1069)

Causal Forest in Observational Studies

Two essential assumptions need to be satisfied for causal forest.

Unconfoundedness (Conditional Independence)

The outcome of individual i is independent of the treatment conditional on covariates X_i. Note that X_i could be of high-dimensional form.

$$\{Y_i^0, Y_i^1\} \perp\!\!\!\perp W_i \mid X_i \tag{9}$$

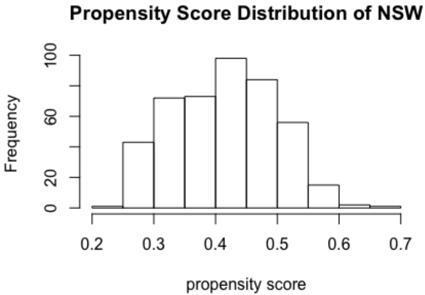
Overlap Assumption

The data must satisfy the overlap assumption to ensure the consistency of results in the causal forest algorithm [Wager and Athey, 2018].

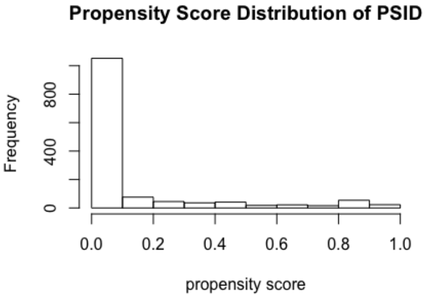
$$\varepsilon < P[treat = 1|X = x] < 1 - \varepsilon, \text{for some } \varepsilon > 0. \tag{10}$$

Causal Forest in Observational Studies

However, the overlap assumption could be hard to satisfy for unbalanced non-random experiment data.



Propensity Score Distribution of NSW



Propensity Score Distribution of PSID

Causal Forest Estimators

In a non-random experiment environment, there are several estimators that can help us reduce the estimation bias.

- Average Treatment Effect for the Treated: Reduce the amount of extrapolation to increase the treatment effect estimation accuracy [Wooldridge, 2012].
- Overlap-weighted Average Treatment Effect:

$$ATE = \frac{\sum_{i=1}^{n} e(X_i)(1 - e(X_i))E[Y(1) - Y(0)|X = X_i]}{\sum_{i=1}^{n} e(X_i)(1 - e(X_i))}$$
$$\text{where } e(x) = P[W_i = 1|X_i = x]$$

- Truncate the comparison sample whose propensity score falls out of the common support of the treated sample:

$$\min e_{treat_i=0}(X_i) \geq \min e_{treat_i=1}(X_i), \text{ for all } i \text{ in the data.}$$

Average Treatment Effects

Table: Average Treatment Effect Estimation Results

	Average Treatment Effects		
	All	Treated	Overlap
Comparison Group			
NSW	**1539.07*****	1785.32***	1604.34***
	(640.7954)	(682.0611)	(645.261)
PSID	−5440.35**	**1396.54****	-1034.79
	(2986.392)	(764.9294)	(966.9764)
PSID (Truncated)	-2685.63	**1163.18***	-1108.75
	(2393.886)	(769.390)	(962.7921)
CPS	−2366.58*	**1560.16****	1058.39*
	(1451.985)	(680.7866)	(698.8165)
CPS (Truncated)	-1157.57	**1874.99*****	1119.26*
	(1237.065)	(659.8012)	(716.6128)

Treatment Effect Heterogeneity

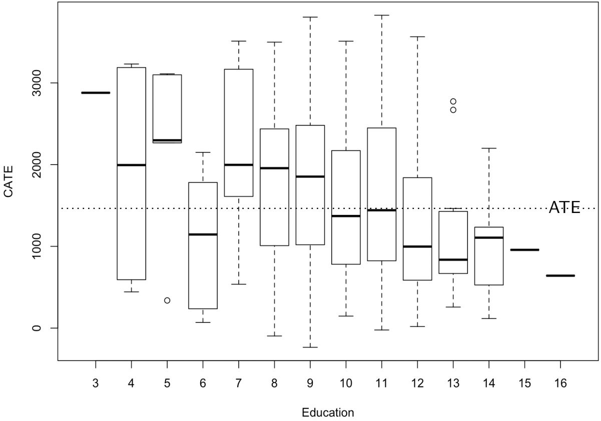


Figure: Predicted Treatment Effect Plotted Against Education (NSW)

Heterogeneity Test

Table: Best Linear Fit Test of Heterogeneity Test

	Best Linear Fit Test of Heterogeneity	
	Mean Forest Prediction	Differential Forest Prediction
Comparison Group		
NSW	1.02***	0.11
	(0.41)	(0.60)
PSID	1.02**	0.91**
	(0.52)	(0.43)
PSID (Truncated)	1.12**	0.99**
	(0.59)	(0.43)
CPS	1.27**	1.01***
	(0.61)	(0.23)
CPS (Truncated)	7.20	1.00***
	(12.33)	(0.25)

Variable Importance Meaure (Preliminary)

Table: Variance Importance of Causal Forest

	Importance Score for Each Sample				
	NSW	PSID	PSID Truncated	CPS	CPS Truncated
Variable					
re74 (Real Income in 1974)	**0.21**	**0.46**	**0.45**	**0.38**	**0.35**
re75 (Real Income in 1975)	0.16	**0.27**	**0.24**	**0.32**	**0.29**
age	**0.31**	**0.14**	**0.16**	**0.17**	**0.22**
education	**0.17**	0.06	0.08	0.07	0.09
black	0.02	0.01	0.01	0.02	0.02
hispanic	< 0.01	0.01	< 0.01	< 0.01	< 0.01
married	0.06	0.03	0.02	0.01	0.01
nodegree	0.05	0.02	0.02	0.01	0.02

Variable Interaction Measure (Preliminary)

Detailed description see `https://github.com/grf-labs/grf/pull/405`.

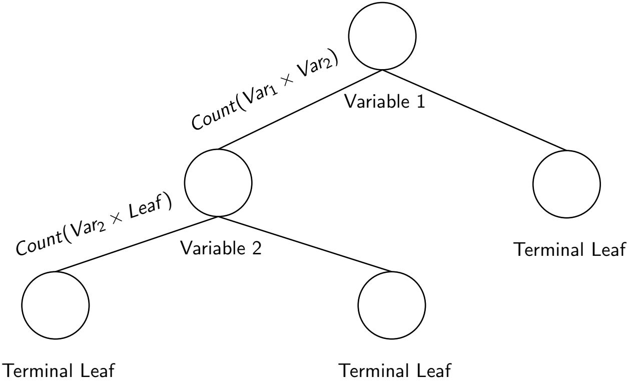


Figure: Interaction Between Variables

Variable Interaction Measure (Preliminary)

Table: NSW Variable Interaction Measure

NSW Dataset			
Variable 1	Variable 2	Interaction Count	Interaction Frequency
terminal leaf	age	59313	0.2765
terminal leaf	education	28008	0.1306
age	education	18706	0.0872
terminal leaf	re75	18294	0.0853
age	age	15427	0.0719
terminal leaf	re74	11881	0.0554
re75	age	10978	0.0512
re74	age	7106	0.0331
re75	education	6171	0.0288
terminal leaf	nodegree	4492	0.0209

Further Research

- Variable Interaction Measure
- Partial Treatment Effect conditional on a convariate

References I

Athey, S. and Imbens, G. (2015).

Recursive partitioning for heterogeneous causal effects.

Dehejia, R. H. and Wahba, S. (1999).

Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs.

Journal of the American Statistical Association, 94(448):1053–1062.

Rosenbaum, P. R. and Rubin, D. B. (1983).

The central role of the propensity score in observational studies for causal effects.

Biometrika, 70(1):41–55.

Rubin, D. B. (1974).

Estimating causal effects of treatments in randomized and nonrandomized studies.

Journal of educational Psychology, 66(5):688.

Wager, S. (2014).

Asymptotic theory for random forests.

References II

Wager, S. and Athey, S. (2018).
Estimation and inference of heterogeneous treatment effects using random forests.
Journal of the American Statistical Association, 113(523):1228–1242.

Wager, S., Hastie, T., and Efron, B. (2014).
Confidence intervals for random forests: The jackknife and the infinitesimal jackknife.
The Journal of Machine Learning Research, 15(1):1625–1651.

Wooldridge, J. (2012).
Treatment effect estimation with unconfounded assignment.
In *American Accounting Association/Financial Accounting and Reporting Section Workshop*.

We **keep moving forward**, opening up new doors and doing new things, because we're curious . . . and curiosity keeps leading us down new paths.

- Walt Disney