Machine learning, Bayes and & structural models

Nils Droste

2021 ClimBEco course

introduction

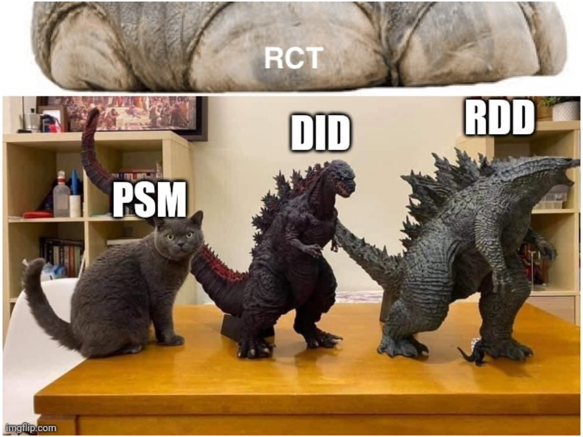
Today, we will talk about

- machine learning techniques
- structural equation models
- Bayesian inference

and their relation to causal inference

introduction

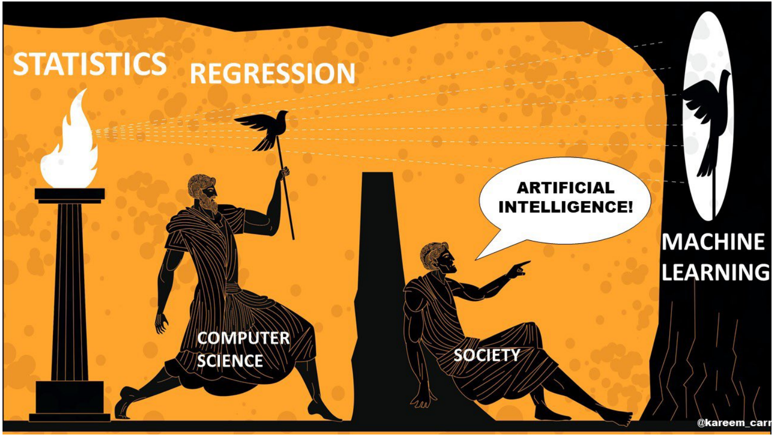
"There is no hierarchy in causal inference"



Moritz Poll2021

introduction

"There is no hierarchy in causal inference"



Kareem Carr 2021

introduction

"There is no hierarchy in causal inference"

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief inY?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing, Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past two years?

The causal hierarchy. Source: Pearl 2019

causal trees

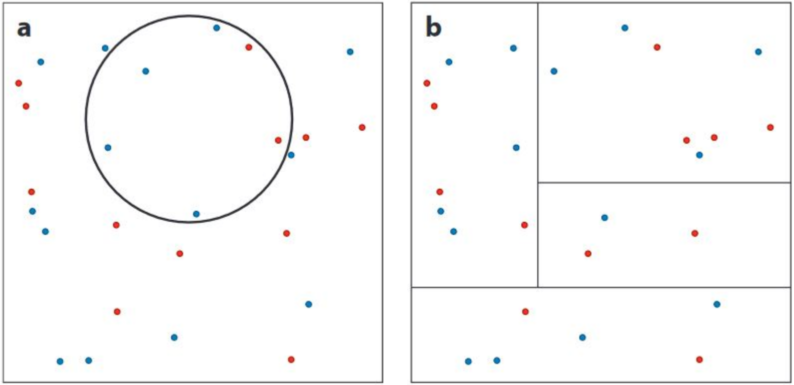
Recall, the conditional average treatment effect (cf. Athey and Imbens 2016)

$$CATE = E(Y_i(1) - Y_i(0)|X_i = x) \tag{1}$$

$\rightarrow$ we could look at average differences between treated and untreated groups within partitions of the observed

causal trees

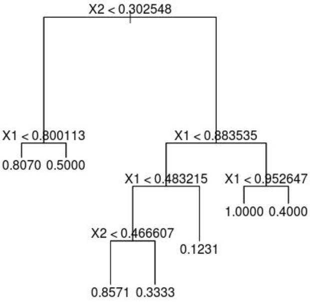
We could use covariates X_i to identify clusters within the observed, e.g. by recursive binary splitting

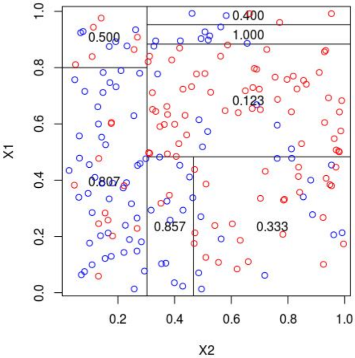


Source: Athey and Imbens 2019

causal trees

We could use covariates X_i to identify clusters within the observed, e.g. by recursive binary splitting

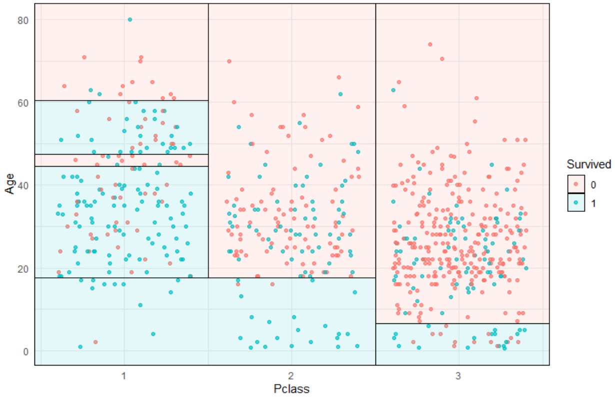




datacamp 2021

causal trees

We could use covariates X_i to identify clusters within the observed, e.g. by recursive binary splitting



McDermott 2021

causal trees

We can use classification and regression trees to estimate

$$\hat{\tau}(x) = \frac{1}{|\{i : D_i = 1, X_i \in L\}|} \sum_{\{i:D_i=1,X_i\in L\}} Y_i - \frac{1}{|\{i : D_i = 0, X_i \in L\}|} \sum_{\{i:D_i=0,X_i\in L\}} Y_i \tag{2}$$

by "recursively splitting the feature space ... into a set of leaves L" (Wager and Athey 2018)

causal trees

We can use classification and regression trees to estimate

$$\hat{\tau}(x) = \frac{1}{|\{i : D_i = 1, X_i \in L\}|} \sum_{\{i:D_i=1,X_i\in L\}} Y_i - \frac{1}{|\{i : D_i = 0, X_i \in L\}|} \sum_{\{i:D_i=0,X_i\in L\}} Y_i \quad (2)$$

by "recursively splitting the feature space ... into a set of leaves L" (Wager and Athey 2018)

and repeat to build an ensemble of trees $\rightarrow$ random forests. This particularly well suited for heterogeneous treatment effects. Athey and Imbens (2016) suggest an *honest* cross validation procedure for robust causal inference, see generalized random forests R package.

honest crossvalidation

This recursive partioning uses multiple optimizations

- within sample (e.g. MSE, possibly with penalties) $\rightarrow$ danger of overfitting to sample
- minimizing prediction error on unseen portions of the data to reduce out-of-sample bias $\rightarrow$ (n-fold) crossvalidation

honest crossvalidation

This recursive partioning uses multiple optimizations

- within sample (e.g. MSE, possibly with penalties) $\rightarrow$ danger of overfitting to sample
- minimizing prediction error on unseen portions of the data to reduce out-of-sample bias $\rightarrow$ (n-fold) crossvalidation

If we do this honestly

- we can compute average treatment effects in the splits
- it promises a valid CATE estimation
- it allows to estimate heterogenous treatment effects

heterogenous treatment effects

Appendix A. Regression tree explanation of CATE

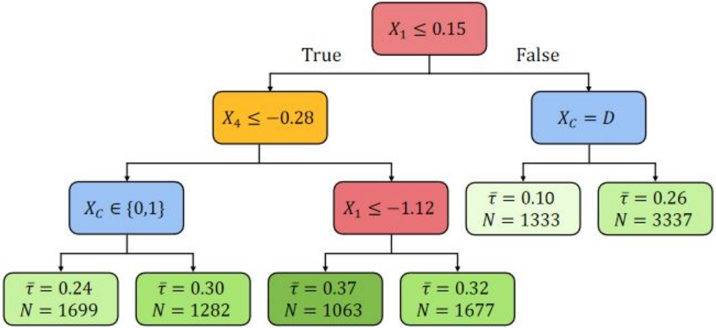


Figure 6: Visualization of a regression tree fit to the imputed CATE values based on the full covariate set.

Source: Johansson 2019

A similar - and earlier - approach

Bayesian Additive Regression Trees (BART), e.g. Hill 2011

- estimate two models
  - one on the treatment assignment process
  - a second on the treatment effect (aka response surface)

A similar - and earlier - approach

Bayesian Additive Regression Trees (BART), e.g. Hill 2011

- estimate two models
  - one on the treatment assignment process
  - a second on the treatment effect (aka response surface)
- seems to perform weel in comparison to linear regression, propensity-score matching, etc. (Hill 2011)

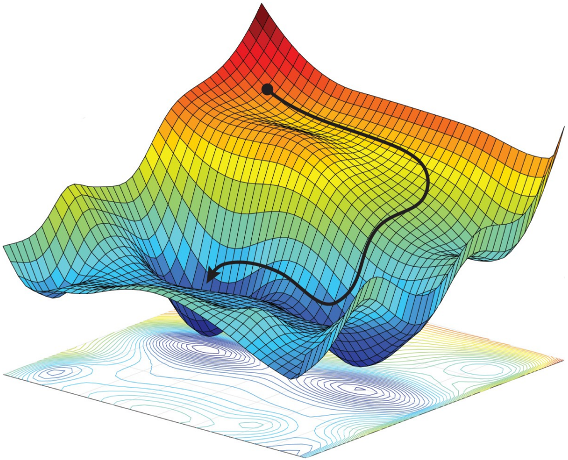
For a suite of latest and different approaches read up on C. Carvalho et al. 2019

A similar - and earlier - approach

Bayesian Additive Regression Trees (BART), e.g. Hill 2011

- estimate two models
  - one on the treatment assignment process
  - a second on the treatment effect (aka response surface)
- seems to perform weel in comparison to linear regression, propensity-score matching, etc. (Hill 2011)
- "*BART is not constrained by prior theories*" (ibid.)
  - not sure this is unproblematic

machine learning estimation



Parmeet Bhatia 2020

estimation techniques

Instead of ordinary least squares

$$b\hat{e}ta^{ols} = \arg\min_{\beta} \sum_{i=1}^{N} (Y - \beta X_i)^2 \quad (3)$$

we can also use machine learning techniques (aka sequential updates / iterative estimation / optimization of parameter values) to estimate

estimation techniques

Instead of ordinary least squares

$$\hat{beta}^{ols} = \arg\min_{\beta} \sum_{i=1}^{N} (Y - \beta X_i)^2 \tag{3}$$

we can also use machine learning techniques (aka sequential updates / iterative estimation / optimization of parameter values) to estimate

$$\hat{beta}^{ml} = \arg\min_{\beta} \sum_{i=1}^{N} (Y - \beta X_i)^2 + \lambda(\|\beta\|_q)^{1/q} \tag{4}$$

estimation techniques

Instead of ordinary least squares

$$b\hat{e}ta^{ols} = \arg\min_{\beta} \sum_{i=1}^{N} (Y - \beta X_i)^2 \tag{3}$$

we can also use machine learning techniques (aka sequential updates / iterative estimation / optimization of parameter values) to estimate

$$b\hat{e}ta^{ml} = \arg\min_{\beta} \sum_{i=1}^{N} (Y - \beta X_i)^2 + \lambda(\|\beta\|_q)^{1/q} \tag{4}$$

with a regularizing penalty term that results in sparse models, $q = 1$ for Lasso, $q = 2$ ridge regression, and $q \to 0$ for best subset regression, or hybrids such as elastic nets (cf. Athey and Imbens 2019) and posterior choice of λ by (n-fold) cross-validation.

estimation techniques

Or we use neural networks, where we transform

$$Z_{ik}^{(1)} = \sum_{j=1}^{K} \beta_{kj}(1) X_{ij} \quad \text{for} \quad k = i, ..., K_1 \tag{5}$$

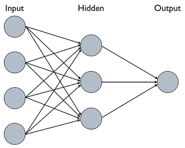
with K covariates (aka features) and latent (aka unobserved / hidden node) variable $Z_j k$, e.g. into a recitified linear function $g(z) = z1_{z>0}$.

estimation techniques

Or we use neural networks, where we transform

$$Z_{ik}^{(1)} = \sum_{j=1}^{K} \beta_{kj}(1) X_{ij} \quad \text{for} \quad k = i, \ldots, K_1 \tag{5}$$

with K covariates (aka features) and latent (aka unobserved / hidden node) variable $Z_j k$, e.g. into a recitified linear function $g(z) = z 1_{z>0}$.
which allows us a to formulat a neural network



$$Y_i = \sum_{k=1}^{K_1} \beta k(2) g[Z_{ik}^2] + \varepsilon_i \tag{6}$$

with a single layer K_1, and a non-linear transformation $g(\cdot)$ (cf. Athey and Imbens 2019; Farrell, Liang and Misra 2021).

matching or synthetic control

Or, we can use machine learning methods for matching or synthetic control (Doudchenko and Imbens 2020)

e.g to estimate weights w and intercept μ for simulating counterfactual

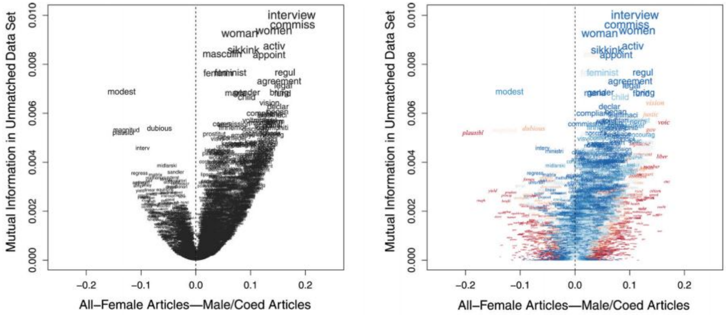
$$\hat{Y}_{t,post}(0) = \mu + \sum_{i=1}^{N} w_i Y_{i,T}^{obs} \tag{7}$$

by

$$(\hat{\mu}^{ols}, \hat{w}^{ols}) = \arg\min_{\mu,w} \sum_{s=1}^{T_0} \left(Y_{0,T_0-s+1}^{obs} - \mu - \sum_{i=1}^{N} w_i \cdot Y_{0,T_0-s+1}^{obs} \right)^2 + \underbrace{\lambda \left(\frac{1-\alpha}{2} \|w\|_2^2 + \alpha \|w\|_1 \right)}_{penaltyfunction} \tag{8}$$

matching

We can also used unsupervised machine learning, such as e.g. latent dirichlet allocation, to use text as data (Gentzkow, Kelly and Taddy 2019) for matching a treated with an untreated population (aka matching on similar content)

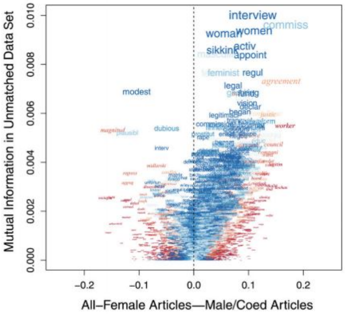


(a) Full Data Set (b) Topic Matched

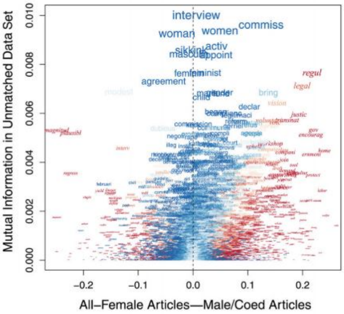
Topical Inverse regression matching (TIRM). Source: Roberts, Stewart and Nielsen 2020

matching

We can also used unsupervised machine learning, such as e.g. latent dirichlet allocation, to use text as data (Gentzkow, Kelly and Taddy 2019) for matching a treated with an untreated population (aka matching on similar content)



(c) Matched on human codes



(d) TIRM

Topical Inverse regression matching (TIRM). Source: Roberts, Stewart and Nielsen 2020

Bayes theorem

$$\mathrm{p}(y|x) = \frac{\mathrm{p}(x|y)\mathrm{p}(y)}{\mathrm{p}(x)} \tag{9}$$

Bayesian inference

Now that we talked about sequential parameter optimization, model ensembles (aka model averaging), and posteriors, ...

Bayesian inference

Now that we talked about sequential parameter optimization, model ensembles (aka model averaging), and posteriors, ...

we are close to talk about a Bayesian approach. Suppose we formulate econometric model

Bayesian inference

Now that we talked about sequential parameter optimization, model ensembles (aka model averaging), and posteriors, ...

we are close to talk about a Bayesian approach. Suppose we formulate econometric model

$$\mathrm{p}(\theta|y) = \frac{\mathrm{p}(y|\theta)\mathrm{p}(\theta)}{\mathrm{p}(y)} \quad \propto \quad \mathrm{p}(y|\theta)\mathrm{p}(\theta) \tag{10}$$

Bayesian inference

Now that we talked about sequential parameter optimization, model ensembles (aka model averaging), and posteriors, ...

we are close to talk about a Bayesian approach. Suppose we formulate econometric model

$$\mathrm{p}(\theta|y) = \frac{\mathrm{p}(y|\theta)\mathrm{p}(\theta)}{\mathrm{p}(y)} \quad \propto \quad \mathrm{p}(y|\theta)\mathrm{p}(\theta) \tag{10}$$

with *prior* parameter probability $\mathrm{p}(\theta)$ and the *posterior* probability (aka conditional probability) $\mathrm{p}(y|\theta)$), the likelihood of observables given the parameter.

Bayesian inference

This framework can be used to estimate literally all types of models that we have been discussing, e.g.

Bayesian inference

This framework can be used to estimate literally all types of models that we have been discussing, e.g.

- Bayesian linear regression (and thus Bayesian 2SLS)

Bayesian inference

This framework can be used to estimate literally all types of models that we have been discussing, e.g.

- Bayesian linear regression (and thus Bayesian 2SLS)
- Panel data (Ning, Ghosal and Thomas 2019)

Bayesian inference

This framework can be used to estimate literally all types of models that we have been discussing, e.g.

- Bayesian linear regression (and thus Bayesian 2SLS)
- Panel data (Ning, Ghosal and Thomas 2019)
- Regression discontinuity (Hinne, Van Gerven and Ambrogioni 2020)
- Ridge and Lasso regressions have their Bayesian interpretation already (cf. Athey and Imbens 2019)

Bayesian inference

This framework can be used to estimate literally all types of models that we have been discussing, e.g.

- Bayesian linear regression (and thus Bayesian 2SLS)
- Panel data (Ning, Ghosal and Thomas 2019)
- Regression discontinuity (Hinne, Van Gerven and Ambrogioni 2020)
- Ridge and Lasso regressions have their Bayesian interpretation already (cf. Athey and Imbens 2019)
- Simulating a synthetic control by Bayesian structural time series analysis (Brodersen et al. 2015)
- Bayesian random forests (Hill 2011; Hahn, Murray and C. M. Carvalho 2020)

It has the merit of being explicit of prior beliefs about distributions (or probalities of parameters) and thus uncertainties

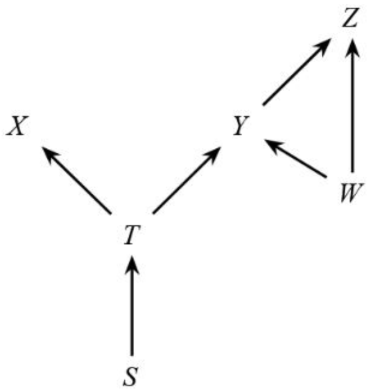
A far greater uncertainty

What do we really now about the world. Image source: Studio Binder 2020

A far greater uncertainty

Black Box Models. Source: Zhao and Hastie 2021

We need a theory of change



Directed Graph Model. Source: Stanford Encyclopedia of Philosophy 2018

structural models

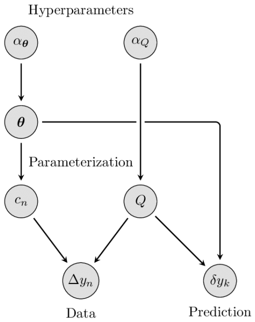
”[G]raphical representations of structural causal models do not require the learner - whether artificial or human - to *impose any distributional or functional-form restrictions* on the underlying causal mechanisms under study.

The approach remains fully nonparametric, a characteristic it shares with the potential outcomes framework. At the same time, however, crucial identification assumptions, suchas ignorability, are derived from the properties of the underlying structural model, rather than being assumed to hold a priori.

Causal graphical models thus combine the accessibility and flexibility of potential outcomes with the preciseness and analytical rigor of structural econometrics” (Hünermund and Bareinboim 2021)

structural causal models (SCM)

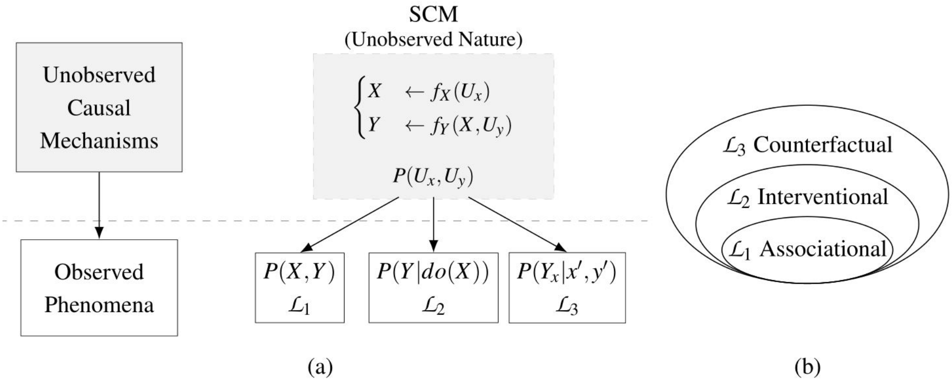
Parameters of a probabilistic structural model can be estimated with a Bayesian network approach ...



Bayesian network. Source: Melendez et al. 2019

... partial least squares, or through double debiased machine learning (cf. citeChernozhukov2018,Jung2021)

The do-calculus



The causal hierarchy. Source: Bareinboim et al. 2020

Wrapping it all up

Wrapping it all up

To estimate a causal effect, we can compare the outcome to its counterfactual.

- potential outcomes: $\tau = E\{Y_i(1)\} - E\{Y_i(0)\}$

Wrapping it all up

To estimate a causal effect, we can compare the outcome to its counterfactual.

- potential outcomes: $\tau = E\{Y_i(1)\} - E\{Y_i(0)\}$
  - RCTs
  - DID & 2WFE
  - IV & synthetic controls
  - causal trees
  - SCM

Wrapping it all up

To estimate a causal effect, we can compare the outcome to its counterfactual.

- potential outcomes: $\tau = E\{Y_i(1)\} - E\{Y_i(0)\}$
  - RCTs
  - DID & 2WFE
  - IV & synthetic controls
  - causal trees
  - SCM
- with frequentist, Bayesian, or machine learning methods

Wrapping it all up

To estimate a causal effect, we can compare the outcome to its counterfactual.

- potential outcomes: $\tau = E\{Y_i(1)\} - E\{Y_i(0)\}$
  - RCTs
  - DID & 2WFE
  - IV & synthetic controls
  - causal trees
  - SCM
- with frequentist, Bayesian, or machine learning methods
- it should be conceptually sound and methodologically robust

Wrapping it all up

To estimate a causal effect, we can compare the outcome to its counterfactual.

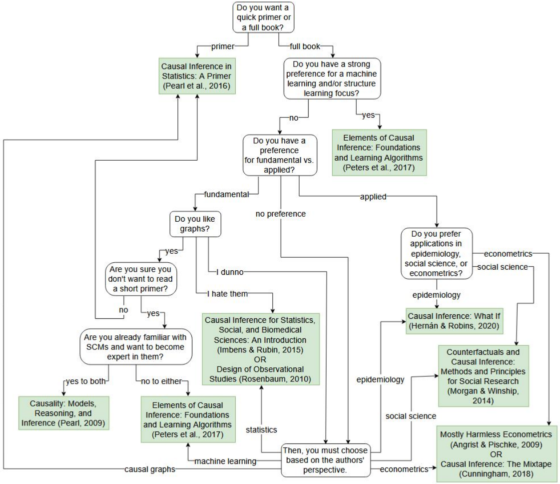
- potential outcomes: $\tau = E\{Y_i(1)\} - E\{Y_i(0)\}$
  - RCTs
  - DID & 2WFE
  - IV & synthetic controls
  - causal trees
  - SCM
- with frequentist, Bayesian, or machine learning methods
- it should be conceptually sound and methodologically robust
- I think the relation with theory should be healthy (inductively so or deductive)

sources

Beyond the coursebooks, see

- James, Witten, Hastie and Tibshirani (2013)
  - introduction to statistical learning (edX)
- Boehmke and Greenwell (2020)
  - Hands-On Machine Learning with R
- Paul Goldsmith-Pinkham 2021
  - applied methods PhD course
- Nick Huntington-Klein 2021
  - The Effect: An Introduction to Research Design and Causality
- Peters, Janzing and Schölkopf 2017
  - Elements of causal inference: foundations and learning algorithms
- Nagarajan, Scutari and Lèbre 2013
  - Bayesian Networks in R

deciding on literature



decision tree, source: Brady Neal 2020

References I

Athey, Susan and Guido W Imbens (2016). 'Recursive partitioning for heterogeneous causal effects'. In: *Proceedings of the National Academy of Sciences of the United States of America* 113.27, pp. 7353–7360. DOI: 10.1073/pnas.1510489113.

— (2019). 'Machine learning methods economists should know about'. In: *Annual Review ofEconomics* 11, pp. 685–725. DOI: 10.1146/annurev-economics-080217-053433.

Bareinboim, Elias et al. (2020). 'On Pearl's Hierarchy and the Foundations of Causal Inference'. In: *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 1–62. URL: https://causalai.net/r60.pdf.

Brodersen, Kay H. et al. (2015). 'Inferring causal impact using Bayesian structural time-series models'. In: *The Annals of Applied Statistics* 9.1, pp. 247–274. DOI: 10.1214/14-AOAS788. URL: http://projecteuclid.org/euclid.aoas/1430226092.

Carvalho, Carlos et al. (2019). 'Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge'. In: *Observational Studies* 5, pp. 21–35.

Doudchenko, Nikolay and Guido W Imbens (2020). 'Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis'. URL: http://arxiv-export-lb.library.cornell.edu/pdf/1610.07748.

Farrell, Max H., Tengyuan Liang and Sanjog Misra (2021). 'Deep Neural Networks for Estimation and Inference'. In: *Econometrica* 89.1, pp. 181–213. DOI: 10.3982/ecta16901.

References II

Gentzkow, Matthew, Bryan Kelly and Matt Taddy (2019). 'Text as data'. In: *Journal of Economic Literature* 57.3, pp. 535–574. DOI: 10.1257/jel.20181020.

Hahn, P. Richard, Jared S. Murray and Carlos M. Carvalho (2020). 'Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion)'. In: *Bayesian Analysis* 15.3, pp. 965–1056. DOI: 10.1214/19-BA1195.

Hill, Jennifer L. (2011). 'Bayesian nonparametric modeling for causal inference'. In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240. DOI: 10.1198/jcgs.2010.08162.

Hinne, Max, Marcel A.J. Van Gerven and Luca Ambrogioni (2020). 'Causal inference using Bayesian non-parametric quasi-experimental design'. URL: https://arxiv.org/pdf/1911.06722.pdf.

Hünermund, Paul and Elias Bareinboim (2021). 'Causal Inference and Data-Fusion in Econometrics'.

James, Gareth et al. (2013). *An introduction to statistical learning*. Springer.

Johansson, Fredrik D. (2019). 'Machine Learning Analysis of Heterogeneity in the Effect of Student Mindset Interventions'. In: *Observational Studies* 5, pp. 71–82. ISSN: 23318422. arXiv: 1811.05975.

Melendez, J. A. et al. (Apr. 2019). 'Quantifying correlated truncation errors in effective field theory'. In: *Physical Review C* 100.4. DOI: 10.1103/PhysRevC.100.044001.

Nagarajan, Radhakrishnan, Marco Scutari and Sophie Lèbre (2013). 'Bayesian networks in R'. In: *Springer* 122, pp. 125–127.

References III

Ning, Bo, Subhashis Ghosal and Jewell Thomas (2019). 'Bayesian method for causal inference in spatially-correlated multivariate time series'. In: *Bayesian Analysis* 14, pp. 1–28. DOI: 10.1214/18-BA1102.

Pearl, Judea (2019). 'The seven tools of causal inference, with reflections on machine learning'. In: *Communications of the ACM* 62.3, pp. 54–60. DOI: 10.1145/3241036.

Peters, Jonas, Dominik Janzing and Bernhard Schölkopf (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press. ISBN: 0262037319.

Roberts, Margaret E., Brandon M. Stewart and Richard A. Nielsen (2020). 'Adjusting for Confounding with Text Matching'. In: *American Journal of Political Science* 64.4, pp. 887–903. DOI: 10.1111/ajps.12526.

Wager, Stefan and Susan Athey (2018). 'Estimation and Inference of Heterogeneous Treatment Effects using Random Forests'. In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242. DOI: 10.1080/01621459.2017.1319839.

Zhao, Qingyuan and Trevor Hastie (2021). 'Causal Interpretations of Black-Box Models'. In: *Journal of Business and Economic Statistics* 39.1, pp. 272–281. DOI: 10.1080/07350015.2019.1624293.