Causal Inference Tutorial

Rahul Singh

Original: July 23, 2019; Updated: September 10, 2020

The goal of this tutorial is to introduce central concepts, algorithms, and techniques of causal inference for a machine learning audience.

There are three sections.

1. Causal frameworks. I present the three most common languages for expressing causal assumptions: potential outcomes, DAGs, and moment restrictions. These are three self-contained axiomatic frameworks, with trade-offs in interpretability, expressiveness, and tractability. Roughly speaking, these frameworks were developed by statisticians and epidemiologists; computer scientists; and economists, respectively. A causal parameter is identified if it can be expressed as a function of data at the population level; it must be the case that two parameter values are not observationally equivalent. As a running example, I present the relevant assumptions for identification of average treatment effect using each of the three languages.
2. Reduced form toolkit. The previous section discusses causal assumptions at the population level. How are these concepts brought to data? I introduce a toolkit for causal inference in observational data that requires relatively few assumptions about the data generating process–specifically, it involves assumptions about treatment assignment and adherence. This toolkit is called the reduced form toolkit in econometrics because it deals with consequences of equilibrium rather than equilibrium itself. I refer to some ways in which this toolkit is being reimagined with ML estimators.
3. De-biased machine learning. This section presents a different paradigm for combining ML and causal inference: delegate prediction tasks to black-box ML estimators, and create an appropriate harness around the ML estimators for valid causal inference. I present the DML algorithm, and I give references for both its econometric theory and its statistical learning theory.

At the end of each section, I list further topics, canonical references, and relevant courses at MIT. Since this tutorial summarizes a great deal of material across disciplines, I only refer to works that summarize a topic (rather than all the papers in the intellectual history of that topic) or that present a machine learning extension. Comprehensive citations will be found therein.

Contents

1 Causal frameworks

1.1 Potential outcomes

1.1.1 Framework

Let's fix the following notation. Y is the outcome of interest, D is the treatment, and X are control variables. For a given individual, we observe (Y, D, X). I suppress individual index i for readability. In the potential outcomes framework, a given individual also possesses latent variables $\{Y_d\}$ where $d \in \mathcal{D} = supp(D)$. For fixed treatment value d, Y_d is the potential outcome an individual would experience had they received treatment d. The outcome we actually observe, Y, is equal to the potential outcome Y_d iff $D = d$.

Philosophically, in the potential outcomes framework we believe an individual contains many, latent selves. For example, consider the treatment variable indicating attendance at this tutorial: $D = 1$ means attendance and $D = 0$ means non-attendance. Let Y be research productivity. There exist two potential Anish's, each with a different productivity: Y_1 is Anish's productivity had he attended; and Y_0 is Anish's productivity had he not attended. Since in reality $D = 1$ for Anish, his actual productivity Y will coincide with Y_1, and we will never observe Anish's Y_0. Formally, with binary treatment,

$$Y = DY_1 + (1 - D)Y_0$$

The individual treatment effect of the tutorial on Anish's productivity is $Y_1 - Y_0$. But we have a missing data problem: we only ever observe Y_1 or Y_0 for a given individual. For this reason, we consider a population average treatment effect instead:

$$\theta_0 = \mathbb{E}[Y_1 - Y_0]$$

We now have a causal language to ask: for the average PhD, what is the productivity effect of attending the tutorial vs. not?

1.1.2 Treatment effect

How can we formalize the above discussion into axioms that identify the average treatment effect? There are three key axioms

1. Consistency: if $D = d$ then $Y = Y_d$. Recall our story about Anish's productivity. This assumption has more substance than it seems. It rules out interference across units: potential outcome of unit i is unaffected by treatment of unit j. This condition is also called the stable unit treatment value assumption. It would be violated if we were studying vaccines, for example. Or if Devavrat's attendance at the tutorial affects Anish's potential productivity.

2. Positivity: if $\mathbb{P}(X = x) > 0$ then $\mathbb{P}(D = d|x) \in (0, 1)$. There are no PhD characteristics (# years, subfield, etc.) that deterministically pin down treatment status. This condition is also called common support.

3. Conditional exchangeability: $\{Y_d\} \perp\!\!\!\perp D|X$. Here we formalize the notion that, conditional on controls X, treatment assignment D is as good as random; the observational data are quasi-experimental. This condition is also called selection on observables. If we are interested in the effect of lab D (e.g. CSAIL vs. LIDS) on research productivity Y, conditional exchangeability could be satisfied taking controls X to be the set of labs to which an individual receives admission. Conditional on a choice set, it is plausibly random which lab a PhD chooses.

Finally, we see our first identification argument. Under the three assumptions above

$$\begin{aligned}\theta_0 &= \mathbb{E}[Y_1 - Y_0] \\ &= \int \{\mathbb{E}[Y_1|x] - \mathbb{E}[Y_0|x]\} d\mathbb{P}(x) \\ &= \int \{\mathbb{E}[Y_1|D = 1, x] - \mathbb{E}[Y_0|D = 0, x]\} d\mathbb{P}(x) \\ &= \int \{\mathbb{E}[Y|D = 1, x] - \mathbb{E}[Y|D = 0, x]\} d\mathbb{P}(x)\end{aligned}$$

Note that we have expressed θ_0 as a function of data at the population level; ATE is identified.

1.2 DAGs

1.2.1 Framework

I review the concept of directed acyclic graphs (DAGs), introduce causal DAGs, and finally present single world intervention graphs (SWIGs) as a way to relate causal DAGs with potential outcomes.

A DAG is a graph whose vertices are random variables $V = (V_0, ..., V_J)$ with directed edges and no directed cycles. Denote by $pa(V_j)$ the set of vertices from which there is a direct edge into V_j. A DAG represents the joint density of V iff $\mathbb{P}(V) = \Pi_{j=1}^{J} \mathbb{P}(V_j|pa(V_j))$. A given vertex V_j may or may not be observed.

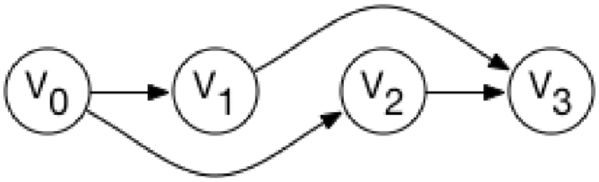

Figure 1: DAG [23]

D-separation is a topological property of a DAG. If there are no active paths from A to B when $C_1, ..., C_K$ are observed (shaded), then $A \perp\!\!\!\perp_d B|\{C_1, ..., C_K\}$. The rules governing whether a path is active are summarized by Bayes ball. In the figure below, an undirected path is active if a Bayes ball traveling along it never encounters a stop symbol.

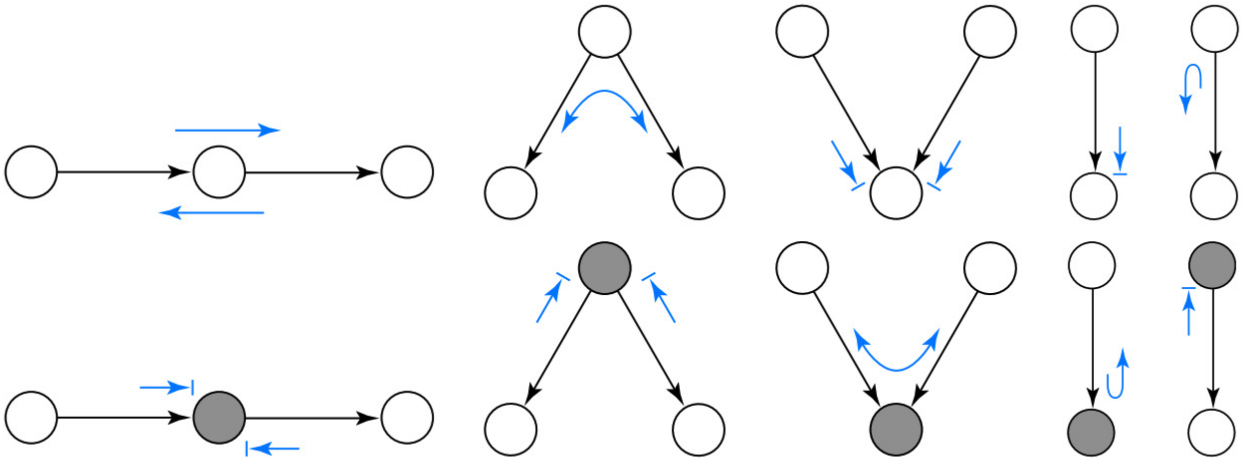

Figure 2: The 10 rules of Bayes ball [21]

D-separation implies statistical independence. However, d-connection does not imply statistical dependence. The faithfulness assumption patches this discrepancy.

A causal DAG is a DAG associated to a causal model. A lack of an arrow from V_j to V_k means the absence of direct causal effect of V_j on V_k. Any variable is a cause of all its descendants. Temporally, a cause must take place before an effect.

1.2.2 Treatment effect

Can we formalize the above discussion into axioms that identify the average treatment effect? There are two key conditions

1. Faithfulness: d-connection implies statistical dependence. Bayes ball becomes sufficient for analysis of independence.
2. Back-door criterion: outcome Y and treatment D are d-separated conditional on the measured controls X, in a graph in which the arrows out of D are removed. The back-door criterion may be satisfied if
   (a) No common causes of Y and D
   (b) No unmeasured confounding. Though there are common causes of Y and D, there are sufficiently many and well-placed controls X such that conditioning on X blocks all backdoor paths in the modified graph described above

As stated, the backdoor criterion is difficult to verify at a glance. Single world intervention graphs (SWIGs) at once make the backdoor criterion easier to verify and formalize the connection between causal DAGs and potential outcomes. The procedure to convert a DAG to a SWIG is called node-splitting. The procedure is as follows.

Topologically order the DAG from left to right. Consider the intervention of setting $D = d$ (in Pearl's language, $do(D = d)$). Decompose the vertex corresponding to D into a two vertices: a left side, and a right side. The left side, labeled D, inherits all the arrows that previously led into the original D. The right side, labeled $D = d$, inherits all the arrows that previously led out of the original D. The left side and right side are not connected. All children of the right side are now potential outcomes indexed by d.

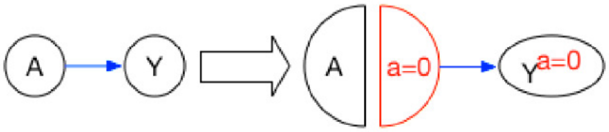

Figure 3: SWIG [23]

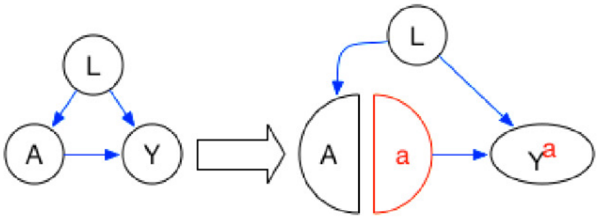

Figure 4: SWIG with confounding [23]

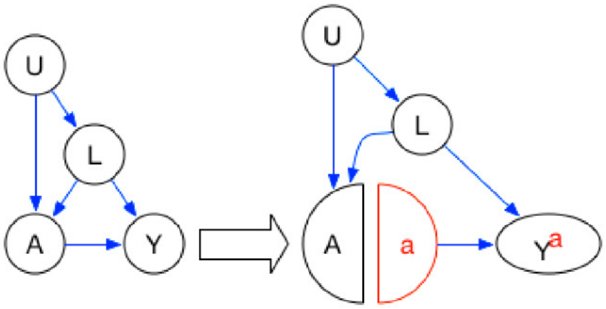

Figure 5: SWIG with confounding [23]

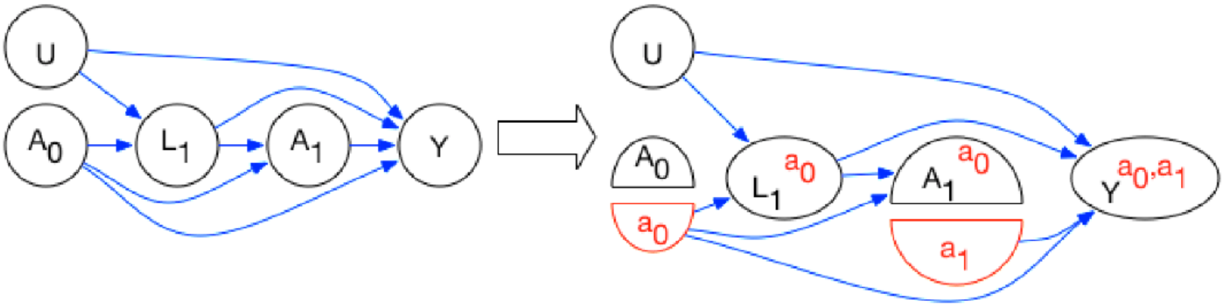

Figure 6: SWIG with multiple treatments [24]

Once we have converted a DAG to a SWIG, we can verify conditional exchangeability via d-separation.

1.3 Moment restrictions

1.3.1 Framework

We have seen the potential outcome and causal DAG frameworks, which are close to equivalent. The former is quite interpretable, while the latter is quite expressive. Identification in both frameworks relies on conditional independences, which are rather strong assumptions. Most of econometrics relies on moment restrictions instead: assumptions about the error terms in a structural model. We will see how moment restrictions are quite tractable; functional form assumptions permit us to relax conditional independence to conditional mean independence and even just orthogonality of error terms.

A moment restriction is of the form

$$\mathbb{E}[\psi(W,\theta)] = 0 \text{ iff } \theta = \theta_0$$

where W is the concatenation of all variables in an observation, and θ_0 is the causal parameter of interest. ψ is called the moment function. Estimation is then possible via generalized method of moments (GMM), which can be viewed as a unifying framework for method of moments, maximum likelihood, linear regression, nonlinear regression, and more.

As a toy example, consider the linear regression model. $W = (Y, D, X)$, and we assume

$$Y = \theta_0 D + X'\beta + \epsilon, \quad \mathbb{E}\left[\epsilon \begin{bmatrix} D \\ X \end{bmatrix}\right] = 0$$

Here the moment restriction is

$$0 = \mathbb{E}[\psi(W,\theta_0)] = \mathbb{E}\left[\epsilon \begin{bmatrix} D \\ X \end{bmatrix}\right] = \left[(Y - \theta_0 D - X'\beta) \begin{bmatrix} D \\ X \end{bmatrix}\right]$$

We have traded in strong functional form assumptions to weaken our independence requirement to orthogonality. θ_0 is identified as the ATE. This is a popular empirical strategy in economics papers.

1.3.2 Treatment effect

It may be the case that moment function ψ depends on additional nuisance parameters. Indeed, this is the case for nonparametric estimation of ATE. In the exposition below, I show how the potential outcome story about ATE can be reformulated as a moment restriction. This approach previews de-biased machine learning (DML).

Define the conditional mean function

$$\gamma_0(D, X) = \mathbb{E}[Y|D, X]$$

We have seen that under consistency, positivity, and conditional exchangeability, ATE is identified and

$$\theta_0 = \int \{\mathbb{E}[Y|D=1,x] - \mathbb{E}[Y|D=0,x]\} d\mathbb{P}(x) = \mathbb{E}[\gamma_0(D=1,X) - \gamma_0(D=0,X)]$$

In other words, we have moment restriction

$$0 = \mathbb{E}[\psi(W,\theta_0,\gamma_0)] = \mathbb{E}[\gamma_0(D=1,X) - \gamma_0(D=0,X) - \theta_0]$$

where $W = (Y, D, X)$ and the moment function ψ has nuisance parameter γ_0. DML will explain how to deal with γ_0 and other nuisance parameters in a sophisticated manner.

1.4 Further topics

1.4.1 G-formula

What if the intervention of interest is a regime of treatments that depend on patient status over time?

1.4.2 Front door criterion

What if treatment and outcome are confounded, but they are mediated by only one variable that is not confounded?

1.4.3 Instrumental variables

What if there is imperfect compliance, so that (randomized) treatment assignment and actual treatment are different variables?

1.4.4 Structural equation modeling

What if we know a great deal about how the data are generated and want to learn structural parameters that allow us to simulate events that have never happened, e.g. a merger?

1.4.5 Panel data

What if we observe a cohort of individuals over time?

1.4.6 Mediation

Can we discern the mechanism through which treatment affects outcome?

1.4.7 Causal discovery

Can we actually learn the edges of the causal DAG from data?

1.5 Resources

1.5.1 Literature

Readable introductions to potential outcomes are in [7, 17]. The classic reference for causal DAGs is [22]. SWIGs are covered in [17] as well. The classic reference for the moment restriction framework is [20].

1.5.2 Courses

In 14.387a, and starting this year in 14.381, Josh Angrist teaches the potential outcomes framework. In 6.438, Guy Bresler teaches graphical models. In 6.244, Caroline Uhler gives a more advanced treatment with applications to causal discovery. Harvard School of Public Health offers a one-week summer course (free for students) covering SWIGs and other tools developed by epidemiologists. In 14.385, Alberto Abadie and Whitney Newey teach the moment restriction framework.

2 Reduced form toolkit

2.1 Sub-classification

Recall that under consistency, positivity, and conditional exhangeability,

$$\theta_0 = \int \{\mathbb{E}[Y|D=1,x] - \mathbb{E}[Y|D=0,x]\} d\mathbb{P}(x)$$

The idea of sub-classification is to approximate each component of this expression.

Assume control variable $X \in \mathcal{X} = supp(X) = \{x^1, ..., x^K\}$. X may be discrete or discretized into K cells. Define

$$\begin{aligned} n_d^k &= \sum_{i=1}^n 1_{X_i = x^k, D_i = d} \\ \bar{Y}_d^k &= \frac{1}{n_d^k} \sum_{i: X_i = x^k, D_i = d} Y_i \end{aligned}$$

n_d^k is the number of observations in cell k with treatment value d. $\bar{Y}_d^k$ is the mean outcome among observations in cell k with treatment value d. Then the subclassification estimator is the sample analogue to θ_0:

$$\hat{\theta} = \sum_{k=1}^K [\bar{Y}_1^k - \bar{Y}_0^k] \frac{n_0^k + n_1^k}{n}$$

Clearly this estimator is infeasible with even moderately dimensional control X, but it is a reasonable starting point.

2.2 Matching

The central idea of matching is to impute the unobserved potential outcome for each observation using nearest neighbors of the opposite treatment status. If observation i is treated, find the M nearest neighbors $j_1(i), ..., j_M(i)$ to i among untreated observations. Likewise, if observation i is untreated, find the M nearest neighbors $j_1(i), ..., j_M(i)$ to i among treated observations. Observations i and j are close if $\|X_i - X_j\|$ is small, where the norm is often taken to be Mahanalobis distance. The matching estimator is then

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n [2D_i - 1] \left\{ Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} \right\}$$

Further work includes a bias correction term.

2.3 Propensity score

The propensity score is defined as $\pi(x) = \mathbb{P}(D = 1|X = x)$. I present three important roles that the propensity score plays in causal inference:

1. rephrasing conditional exchangeability
2. dimension reduction in matching
3. new class of estimators based on Riesz representation

The first role is classical. The second is practical. The third is a more modern interpretation, previewing DML.

I now show that conditional exchangeability can be rephrased in terms of the propensity score. Formally,

$$\{Y_d\} \perp\!\!\!\perp D|X \implies \{Y_d\} \perp\!\!\!\perp D|\pi(X), \quad D \perp\!\!\!\perp X|\pi(X)$$

Assume $\{Y_d\} \perp\!\!\!\perp D|X$.

$$\begin{aligned}
\mathbb{P}(D=1|Y_1,Y_0,\pi(X)) &= \mathbb{E}[D|Y_1,Y_0,\pi(X)] \\
&= \mathbb{E}_{X,Y_1,Y_0}[\mathbb{E}[D|X,Y_1,Y_0]|Y_1,Y_0,\pi(X)] \\
&= \mathbb{E}_{X,Y_1,Y_0}[\mathbb{E}[D|X]|Y_1,Y_0,\pi(X)] \\
&= \mathbb{E}_{X,Y_1,Y_0}[\pi(X)|Y_1,Y_0,\pi(X)] \\
&= \pi(X)
\end{aligned}$$

Similarly

$$\begin{aligned}
\mathbb{P}(D=1|\pi(X)) &= \mathbb{E}[D|\pi(X)] \\
&= \mathbb{E}_X[\mathbb{E}[D|X]|\pi(X)] \\
&= \mathbb{E}_X[\pi(X)|\pi(X)] \\
&= \pi(X)
\end{aligned}$$

so $\mathbb{P}(D=1|Y_1,Y_0,\pi(X)) = \mathbb{P}(D=1|\pi(X))$, i.e. $\{Y_d\} \perp\!\!\!\perp D|\pi(X)$. To see the second result, write

$$\mathbb{P}(D=1|X,\pi(X)) = \mathbb{P}(D=1|X) = \pi(X)$$

hence $D \perp\!\!\!\perp X|\pi(X)$.

This result is practical in the sense that we can now perform sub-classification or matching using the scalar propensity score $\pi(X)$ instead of the possibly high dimensional control variable X. It is also practical in giving us a way to test a consequence of conditional exchangeability: we expect the balancing property to hold:

$$\mathbb{P}(X|D=1,\pi(X)) = \mathbb{P}(X|D=0,\pi(X))$$

Finally, I turn to another important class of reduced form estimators that make use of the propensity score. Observe that

$$\begin{aligned}
\mathbb{E}\left[Y\frac{D}{\pi(X)}\middle|X\right] &= \mathbb{E}\left[Y\frac{1}{\pi(X)}\middle|D=1,X\right]\mathbb{P}(D=1|X) \\
&= \mathbb{E}\left[Y\frac{1}{\pi(X)}\middle|D=1,X\right]\pi(X) \\
&= \mathbb{E}\left[Y|D=1,X\right]
\end{aligned}$$

and likewise

$$\mathbb{E}\left[Y\frac{1-D}{1-\pi(X)}\middle|X\right] = \mathbb{E}\left[Y|D=0,X\right]$$

In summary, we can write ATE as

$$\begin{aligned}
\theta_0 &= \int \{\mathbb{E}[Y|D=1,x] - \mathbb{E}[Y|D=0,x]\}d\mathbb{P}(x) \\
&= \int \left\{\mathbb{E}\left[Y\frac{D}{\pi(X)}\middle|X\right] - \mathbb{E}\left[Y\frac{1-D}{1-\pi(X)}\middle|X\right]\right\}d\mathbb{P}(x) \\
&= \mathbb{E}\left[Y\frac{D}{\pi(X)}\right] - \mathbb{E}\left[Y\frac{1-D}{1-\pi(X)}\right]
\end{aligned}$$

This expression suggests, by the analogy principle, a Horvitz-Thompson estimator of the ATE.

Written in terms of $\gamma_0(D,X) = \mathbb{E}[Y|D,X]$,

$$\theta_0 = \mathbb{E}[\gamma_0(D=1,X) - \gamma_0(D=0,X)] = \mathbb{E}\left[\gamma_0(D,X)\left\{\frac{D}{\pi(X)} - \frac{1-D}{1-\pi(X)}\right\}\right]$$

The middle expression formulates ATE as a functional–a scalar-valued summary–of an underlying regression γ_0. The last expression formulates ATE as an inner product between γ_0 and the an object we will call the Riesz representer α_0

$$\alpha_0(D,X) = \left\{\frac{D}{\pi(X)} - \frac{1-D}{1-\pi(X)}\right\}$$

Our ability to write ATE in both ways is no coincidence; by Riesz representation theorem, we are guaranteed that any continuous linear functional can be represented as an inner product with that functional's unique Riesz representer. Written more generally, for functional m and its unique Riesz representer α_0

$$\theta_0 = \mathbb{E}[m(W,\gamma_0)] = \mathbb{E}[\alpha_0(D,X)\gamma_0(D,X)]$$

and indeed

$$\mathbb{E}[m(W,\gamma)] = \mathbb{E}[\alpha_0(D,X)\gamma(D,X)], \quad \forall \gamma \in L^2 \text{ i.e. } \mathbb{E}[\gamma(D,X)]^2 < \infty$$

2.4 Further topics

2.4.1 Synthetic control

Can we estimate the treatment effect on a treated unit by estimating its potential outcome had it not received treatment as a convex combination of untreated neighbors (w.r.t. X)?

2.4.2 Regression discontinuity

What if treatment is a deterministic and discontinuous function of a control variable, e.g. $D = 1_{X \geq x_0}$ for threshold value x_0?

2.4.3 Two stage least squares

How can we bring instrumental variable identification (recall the imperfect compliance story) to data?

2.5 Resources

2.5.1 Literature

See [26] for sub-classification, [2] for matching, and [25] for propensity score. The canonical reference for synthetic control is [1].

The theory of regression discontinuity design is [15] with popular user guide [19]. In [10], the authors propose a Gaussian process regression generalization.

See [6] for a potential outcomes analysis of instrumental variables, and [5] for its linear implementation as two-stage least squares (2SLS). In [16], the authors consider a neural net generalization. In [27], the authors consider an RKHS generalization with finite sample excess risk guarantees.

There is a recent literature bringing the technology of random forests to some of the methods described above [31, 9]. There is another literature bringing the technology of matrix completion to these methods [8, 4, 3].

2.5.2 Courses

In 14.385, Alberto Abadie and Whitney Newey teach the reduced form toolkit. I'm the TA for that course.

3 De-biased machine learning

3.1 Motivation

In causal inference, we are often interested in a parameter defined as a functional of some underlying regression. We have already seen how ATE is one such example. Other common examples are average treatment effect on the treated, average policy effect, and average consumer surplus. In this discussion, I restrict attention to DML applied to functionals. This is the original setting of DML, called semi-parametric estimation. DML has been extended to a wide variety of settings since.

I briefly recap the ATE formalism we have seen before in the context of a concrete example. Consider the following simple design. The horizontal axis is $\mathcal{X}$ and the vertical axis is $\mathcal{Y}$. Observations (grey dots) are generated from $\gamma_0(D = 1, X) = \mathbb{E}[Y|D = 1, X]$ (the orange curve) if treated and from $\gamma_0(D = 0, X) = \mathbb{E}[Y|D = 0, X]$ (the blue curve) if untreated. The fact that observations are scattered reflects sampling from the respective curves with noise. Note that $\gamma_0(D = 1, X) - \gamma_0(D = 0, X) = 1$ everywhere, so in particular the ATE is $\theta_0 = 1$ by construction. Given access only to the grey dots $W = (Y, D, X)$ and not the data generating curves, we wish to estimate the ATE with true value $\theta_0 = 1$.

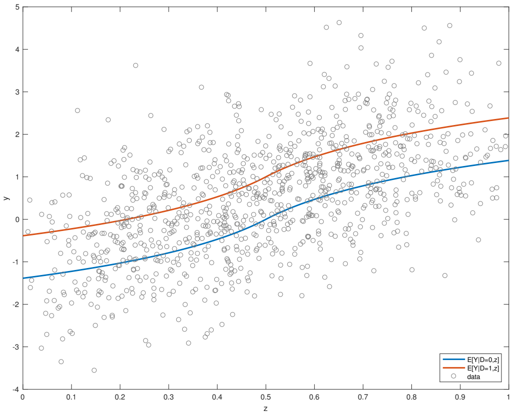

Figure 7: CEF for treated vs. untreated. ATE is $\theta_0 = 1$

Recall that in the ATE example,

$$\theta_0 = \mathbb{E}[m(W, \gamma_0)] = \mathbb{E}[\gamma_0(D = 1, X) - \gamma_0(D = 0, X)]$$

A naive approach, suggested earlier, is as follows.

1. learn CEF with some ML estimator $\hat{\gamma}$
2. plug into the functional

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} m(W_i, \hat{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}[\hat{\gamma}(D = 1, X_i) - \hat{\gamma}(D = 0, X_i)]$$

This is equivalent to method of moments with moment function

$$\psi(w, \theta, \gamma) = m(w, \gamma) - \theta$$

and nuisance parameter γ. Let's see how the naive approach performs in 100 simulations.

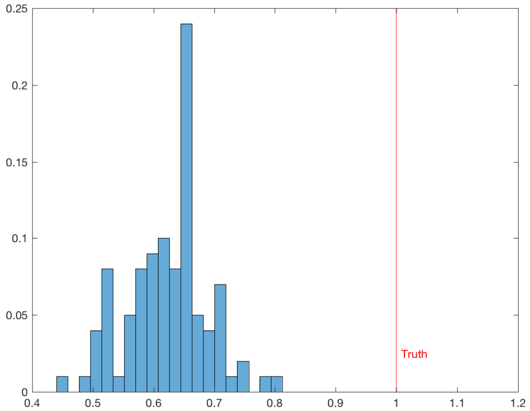

Figure 8: Naive approach $\hat{\theta}$

This phenomenon, in which ML estimation of nuisance parameters badly biases estimation of causal parameters, motivates a more sophisticated approach: de-biased machine learning.

3.2 Influence adjustment and double robustness

Previously we used

$$\psi(w, \theta, \gamma) = m(w, \gamma) - \theta$$

Now use

$$\psi(w, \theta, \gamma, \alpha) = m(w, \gamma) - \theta + \alpha(d, x)[y - \gamma(d, x)]$$

which has nuisance parameters (γ, α). As our prior discussion foreshadowed, the new nuisance parameter is the Riesz representer. The adjustment term is the product of Riesz representer α and regression residual $y - \gamma(x)$.

In what sense is this new moment function more sophisticated? I prove the following properties:

1. $\mathbb{E}[\psi(W, \theta_0, \gamma, \alpha_0)] = 0$, $\forall \gamma \in L^2$. The argument is as follows.

$$\begin{aligned}
\mathbb{E}[\psi(W, \theta_0, \gamma, \alpha_0)] &= \mathbb{E}[m(W, \gamma) - \theta_0 + \alpha_0(D, X)[Y - \gamma(D, X)]] \\
&= \mathbb{E}[m(W, \gamma) - \theta_0 + \alpha_0(D, X)[\gamma_0(D, X) - \gamma(D, X)]] \\
&= \mathbb{E}[m(W, \gamma) - \theta_0 + m(W, \gamma_0) - m(W, \gamma)]] \\
&= \mathbb{E}[m(W, \gamma) - m(W, \gamma_0) + m(W, \gamma_0) - m(W, \gamma)]]
\end{aligned}$$

2. $\mathbb{E}[\psi(W, \theta_0, \gamma_0, \alpha)] = 0$, $\forall \alpha \in L^2$. The argument is as follows.

$$\begin{aligned}
\mathbb{E}[\psi(W, \theta_0, \gamma_0, \alpha)] &= \mathbb{E}[m(W, \gamma_0) - \theta_0 + \alpha(D, X)[Y - \gamma_0(D, X)]] \\
&= \mathbb{E}[m(W, \gamma_0) - \theta_0 + \alpha(D, X)[\gamma_0(D, X) - \gamma_0(D, X)]] \\
&= \mathbb{E}[m(W, \gamma_0) - m(W, \gamma_0) + \alpha(D, X)[\gamma_0(D, X) - \gamma_0(D, X)]]
\end{aligned}$$

Our new ψ it is not only a valid moment function, but also doubly robust: the moment function remains valid even for incorrect values of γ or α. Estimation of θ is robust to first stage estimation error of either nuisance parameter. Double robustness is a strong property, and much DML analysis requires the weaker condition of local robustness. For general moment function $\psi(w, \theta, \eta)$ with nuisance parameter η, local robustness is the property that

$$\partial_\eta \mathbb{E}\psi(W, \theta_0, \eta)|_{\eta=\eta_0} = 0$$

This condition is also called Neyman orthogonality.

It turns out there is a rich theory that delivers a formula for finding locally robust moment functions. Astonishingly, the locally robust moment function for a semi-parametric parameter θ_0 is the limit of the Gateaux derivative of the parameter θ_0 with respect to a smooth deviation away from the true distribution $\mathbb{P}$ as the deviation approaches a point mass.

3.3 Algorithm

I finally present the DML algorithm. It involves the doubly robust moment function and sample splitting.

Partition observations into L distinct subsets $\{I_\ell\}_{\ell=1}^L$

1. learn CEF and Riesz representer
   (a) estimate $\hat{\gamma}_\ell$ from observations not in I_ℓ
   (b) estimate $\hat{\alpha}_\ell$ from observations not in I_ℓ
      - manually (plug-in)
      - automatically (Lasso, Dantzig selector)
2. DR method of moments

$$\hat{\theta} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{i \in I_\ell} \{m(W_i, \hat{\gamma}_\ell) + \hat{\alpha}_\ell(W_i)[Y_i - \hat{\gamma}_\ell(X_i)]\}$$

Any consistent, black-box ML estimator can be used for $\hat{\gamma}$. Let me clarify the distinction between manual and automatic estimation of $\hat{\alpha}$ in the context of ATE.

First, manual estimation of $\hat{\alpha}$. In ATE, we have analytically derived the functional form

$$\alpha_0(D, X) = \left\{ \frac{D}{\pi(X)} - \frac{1 - D}{1 - \pi(X)} \right\}$$

An analyst, knowing the functional form, may estimate $\hat{\alpha}$ by estimating the propensity score $\hat{\pi}$ and plugging it into the formula above. We may be concerned about plugging-in an estimated probability in a denominator. For more complicated parameters, we may be concerned about deriving the Riesz representer's functional form.

Second, automatic estimation of $\hat{\alpha}$. Recent research shows that α_0 is directly identified from the data W, and so we can use machine learning–e.g. Lasso or Dantzig selector–to estimate $\hat{\alpha}$ directly. In other words, we do not have to estimate its components, or even know the functional form of true α_0. We can automatically de-bias ML with more ML!

Let's return to the simple design and see how DML with automatic Riesz representer estimation performs in 100 simulations. First recall the set-up.

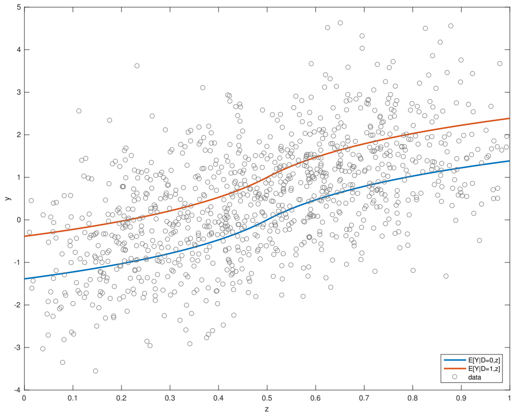

Figure 9: CEF for treated vs. untreated. ATE is $\theta_0 = 1$

Here is the performance of DML with automatic Riesz representer.

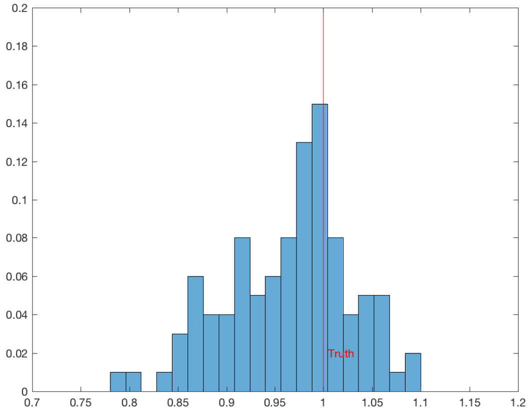

Figure 10: DML with automatic Riesz representer $\hat{\theta}$

Finally, a side-by-side comparison.

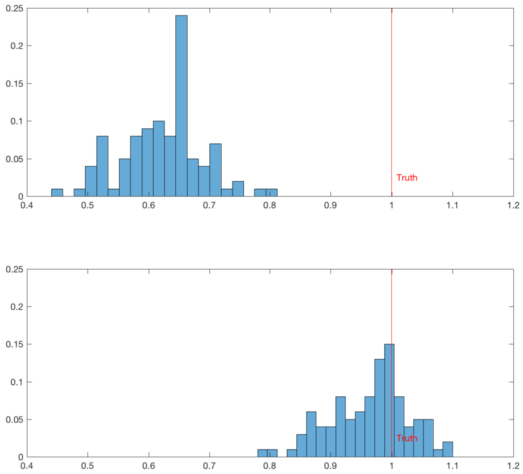

Figure 11: Naive approach vs. DML with automatic Riesz representer

3.4 Further topics

3.4.1 Econometric theory

Is it possible to derive asymptotic confidence intervals for the estimator $\hat{\theta}$?

3.4.2 Statistical learning theory

Is it possible to derive finite sample bounds on excess risk of $\hat{\theta}$?

3.5 Resources

3.5.1 Literature

The core econometric theory of DML is in [11]. The automatic Riesz representer extension of DML is in [12, 13]. The core statistical learning theory of DML is in [14]. See [18] for a nice review of influence function theory in DML, and [30, Chapter 25] for a rigorous statistical background. Finally, see [28, 29] for extensions with instrumental variables.

3.5.2 Courses

In 14.387b, Victor Chernozhukov introduces DML more thoroughly. In 14.386, Whitney Newey teaches some influence function theory (which delivers locally robust moment functions).

Conclusion

Causal inference spans statistics, epidemiology, computer science, and economics. There are three languages to express causal assumptions and conclusions: potential outcomes, causal DAGs, and moment restrictions. Recent research has begun to reimagine the reduced form toolkit with ML techniques such as matrix completion, random forests, neural networks, and RKHS methods. Another paradigm for bridging causal inference and ML is DML, a meta-algorithm around black-box ML estimators that guarantees valid causal inference. My email address is `rahul.singh@mit.edu` and I would love to talk more!

References

[1] Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

[2] Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

[3] Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *arXiv:1902.10920*, 2019.

[4] Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852, 2018.

[5] Joshua D Angrist and Guido W Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995.

[6] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

[7] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press, 2008.

[8] Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research, 2018.

[9] Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[10] Zach Branson, Maxime Rischard, Luke Bornn, and Luke W Miratrix. A nonparametric Bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, 202:14–30, 2019.

[11] Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K Newey, and James M Robins. Locally robust semiparametric estimation. *arXiv:1608.00033*, 2016.

[12] Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Double/de-biased machine learning of global and local parameters using regularized Riesz representers. *arXiv:1802.08667*, 2018.

[13] Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Learning L2 continuous regression functionals via regularized Riesz representers. *arXiv:1809.05224*, 2018.

[14] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv:1901.09036*, 2019.

[15] Jinyong Hahn, Petra Todd, and Wilbert Van der Klaauw. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209, 2001.

[16] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423, 2017.

[17] Miguel A Hernan and James M Robins. *Causal Inference*. CRC Press, 2010.

[18] Hidehiko Ichimura and Whitney K Newey. The influence function of semiparametric estimators. *arXiv:1508.01378*, 2015.

[19] David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355, 2010.

[20] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

[21] Mark Paskin. A short course on graphical models, 2003.

[22] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[23] James Robins. Graphical models for causal inference, 2017.

[24] James Robins. Time-varying exposures and the g-formula, 2018.

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

[26] Paul R Rosenbaum and Donald B Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

[27] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *arXiv:1906.00232*, 2019.

[28] Rahul Singh and Liyang Sun. De-biased machine learning for compliers. *arXiv:1909.05244*, 2019.

[29] Vasilis Syrgkanis, Victor Lei, Miruna Oprescu, Maggie Hei, Keith Battocchi, and Greg Lewis. Machine learning estimation of heterogeneous treatment effects with instruments. *arXiv:1905.10176*, 2019.

[30] Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

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