EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

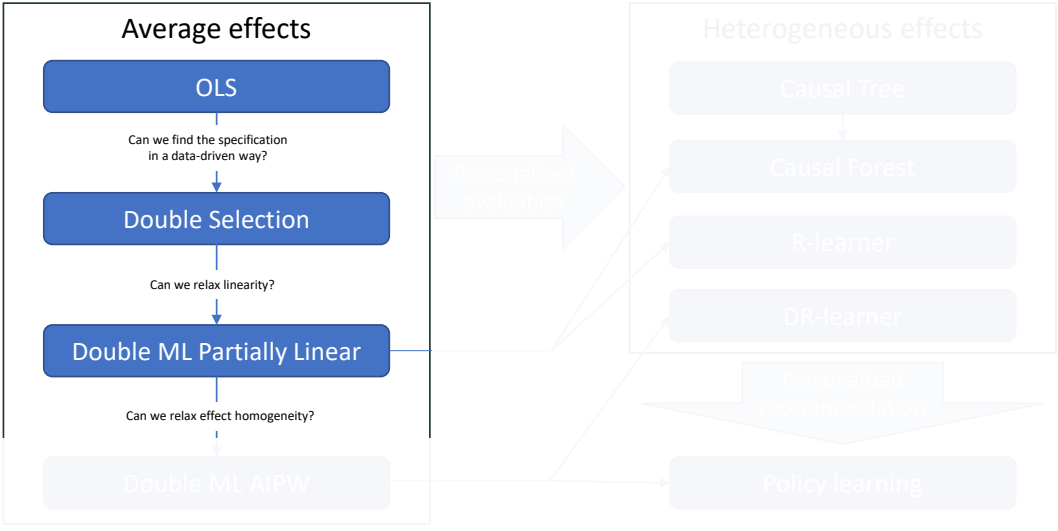
Causal Machine Learning

Average treatment effect estimation: AIPW-Double ML

Michael Knaus

WiSe 25/26

Current state of affairs



Plan of this morning

Can we allow for heterogeneous treatment effects?

1. Nonparametric identification

2. Augmented IPW

3. Consolidation

Nonparametric identification

Beyond effect homogeneity

So far, we assumed that the outcome model in the SCM (and therefore also POs) are (partially) linear $\Rightarrow$ effect heterogeneity restricted to a minimum

This resulted in relatively easy recipes for binary and continuous treatments

This is convenient

However, assuming homogeneous effects is not innocent:

- We might estimate effects for a strange/unintuitive target population if there is actually effect heterogeneity (e.g. Słoczyński, 2022)
- We might miss that we have no comparable units in treated and control group and heavily rely on extrapolation

Let's see how we can relax effect homogeneity for a binary treatment

Recall target parameters

Average target parameters without restricting effect heterogeneity:

- Average potential outcome (APO): $\gamma_w := \mathbb{E}[Y(w)]$
  - What is the expected outcome if everybody receives treatment w?
- Average Treatment Effect (ATE): $\tau_{ATE} := \mathbb{E}[Y(1) - Y(0)] = \gamma_1 - \gamma_0$
  - What is the expected treatment effect in the population?

Definition ✅

Nonparametric identification (1/4)

Identifying Assumption 2 (Strong Ignorability)

(a) $Y(w) \perp\!\!\!\perp W \mid X$ for all $w \in \{0, 1\}$

(b) $0 < \mathbb{P}[W = 1 \mid X = x] =: e(x) < 1$

IA2a is identical to having a valid adjustment set (IA1) of last week but here we focus on binary treatments

IA2b is called common support/overlap/positivity assumption

IA2b is required b/c we do not impose an outcome model that allows to extrapolate the counterfactual into regions where everybody receives the same treatment

This set of assumption allows for arbitrary effect heterogeneity

Non-parametric identification (2/4)

Note that the target parameters are just different aggregations of the Conditional Average Potential Outcome (CAPO) $\mathbb{E}[Y(w) \mid X]$:

- $\gamma_0 := \mathbb{E}[Y(0)] \overset{LIE}{=} \mathbb{E}[\mathbb{E}[Y(0) \mid X]]$
- $\gamma_1 := \mathbb{E}[Y(1)] \overset{LIE}{=} \mathbb{E}[\mathbb{E}[Y(1) \mid X]]$
- $\tau_{ATE} := \mathbb{E}[Y(1) - Y(0)] \overset{LIE}{=} \mathbb{E}[\mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X]]$

$\Rightarrow$ It suffices to show that the CAPO is identified

Non-parametric identification (3/4)

Three common ways to identify the CAPO under IA2:

$$\mathbb{E}[Y(w) \mid X = x] = \mathbb{E}[Y \mid W = w, X = x] =: m(w, x) \tag{1}$$

$$= \mathbb{E}\left[\frac{\mathbb{1}[W = w]Y}{e_w(x)} \middle| X = x\right] \tag{2}$$

$$= \mathbb{E}\left[m(w, x) + \frac{\mathbb{1}[W = w](Y - m(w, x))}{e_w(x)} \;\middle|\; X = x\right] \tag{3}$$

where $e_w(x) := P[W = w \mid X = x]$

(1) motivates estimation via regression adjustment, (2) motivates inverse probability weighting (IPW), and (3) motivates the doubly robust/augmented IPW (AIPW) estimator

Non-parametric identification (4/4)

From an identification perspective, we can plug-in any of the identified estimands

For example:

$$\begin{aligned}\gamma_w := \mathbb{E}\left[Y(w)\right] \overset{LIE}{=} \mathbb{E}\left[\mathbb{E}[Y(w) \mid X]\right] &= \mathbb{E}\left[m(w, X)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}[W = w]Y}{e_w(X)}\middle| X\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[m(w, X) + \frac{\mathbb{1}[W = w](Y - m(w, X))}{e_w(X)} \;\middle|\; X\right]\right]\end{aligned}$$

$\Rightarrow$ identification ✅

Identification of Conditional Average Potential Outcomes - RA

Show for (1)

$$\mathbb{E}[Y(w) \mid X = x] \overset{IA2}{=} \mathbb{E}[Y(w) \mid W = w, X = x] \tag{4}$$

$$\overset{Consistency}{=} \mathbb{E}[Y \mid W = w, X = x] \tag{5}$$

Remark: Note that the common support condition A2b is required, although it might not be obvious. It ensures that we do not condition on an event with zero probability density.

Identification of Conditional Average Potential Outcomes - IPW (option 1)

There are multiple ways to show IPW based identification

I show you three ways to go from $\mathbb{E}[Y \mid W = w, X = x]$ established in (5) to $\mathbb{E}\left[\frac{\mathbb{1}[W = w]Y}{e_w(x)}\middle| X = x\right]$

Note that there is nothing causal here, it is just the transformation of one statical quantity into another form using statistics

A very compact derivation of IPW is to define $D(w) := \mathbb{1}[W = w]$ to note that

$$\mathbb{E}[D(w)Y \mid X = x] \overset{LIE}{=} e_w(x)\,\mathbb{E}[\overbrace{D(w)}^{=1} Y \mid W = w, X = x] + \cancel{(1 - e_w(x))\,\mathbb{E}[\overbrace{D(w)}^{=0} Y \mid W \neq w, X = x]}$$

$$= e_w(x)\,\mathbb{E}[Y \mid W = w, X = x] \qquad \mid \div e_w(x)$$

$$\mathbb{E}\left[\frac{D(w)Y}{e_w(x)}\middle| X = x\right] = \mathbb{E}[Y \mid W = w, X = x]$$

Identification of Conditional Average Potential Outcomes - IPW (option 2)

Less compact but maybe more accessible

$$\mathbb{E}[Y \mid W = w, X = x] = \mathbb{E}[\overbrace{D(w)}^{=1} Y \mid W = w, X = x]$$

$$= \mathbb{E}\left[D(w)Y \frac{e_w(x)}{e_w(x)} \middle| W = w, X = x\right] + \underbrace{(1 - e_w(x))\, \mathbb{E}[\overbrace{D(w)}^{=0} Y \mid W \neq w, X = x]}_{=0} / e_w(x)$$

$$= \frac{\overbrace{e_w(x)\, \mathbb{E}\left[D(w)Y | W = w, X = x\right] + (1 - e_w(x))\, \mathbb{E}[D(w)Y \mid W \neq w, X = x]}^{\overset{LIE}{=} \mathbb{E}[D(w)Y|X=x]}}{e_w(x)}$$

$$= \frac{\mathbb{E}\left[D(w)Y \mid X = x\right]}{e_w(x)} = \mathbb{E}\left[\frac{D(w)Y}{e_w(x)} \middle| X = x\right]$$

Identification of Conditional Average Potential Outcomes - IPW (option 3)

The least hacky reverse engineered / most statistic option:

$$
\begin{aligned}
\mathbb{E}[Y \mid W = w, X = x] &= \int y\, f(y \mid W = w, X = x)\, dy && \text{(def of conditional expectation)} \\
&= \int y\, \frac{f(y, W = w, X = x)}{f(W = w, X = x)}\, dy && \text{(Bayes' rule)} \\
&= \int y\, \frac{f(y, W = w \mid X = x)\cancel{f(X = x)}}{f(W = w \mid X = x)\cancel{f(X = x)}}\, dy && \text{(chain rule)} \\
&= \int y\, \frac{f(y, W = w \mid X = x)}{e_w(x)}\, dy && \text{(by definition)} \\
&= \iint y\, \frac{\mathbb{1}[W = w]}{e_w(x)}\, f(y, W \mid X = x)\, dy\, dW && \text{(indicator trick; integrate over discrete } W\text{)} \\
&= \mathbb{E}\left[\frac{\mathbf{1}(W = w)}{\Pr(W = w \mid X)}\, Y \,\middle|\, X = x \right] && \text{(law of the unconscious statistician)}
\end{aligned}
$$

Identification of Conditional Average Potential Outcomes - AIPW (1/2)

Define $D(w) := \mathbb{1}[W = w]$ for compactness and show for (3)

$$
\begin{aligned}
\mathbb{E}[Y(w) \mid X = x] &= \mathbb{E}\left[m(w,x) + \frac{D(w)(Y - m(w,x))}{e_w(x)} \middle| X = x \right] \\
&= \mathbb{E}\left[Y(w) - Y(w) + m(w,x) + \frac{D(w)(Y - m(w,x))}{e_w(x)} \middle| X = x \right] \\
&\overset{\text{Cons.}}{=} \mathbb{E}\left[Y(w) - Y(w) + m(w,x) + \frac{D(w)(Y(w) - m(w,x))}{e_w(x)} \middle| X = x \right] \\
&= \mathbb{E}[Y(w) \mid X = x] + \underbrace{\mathbb{E}\left[(Y(w) - m(w,x)) \left(\frac{D(w) - e_w(x)}{e_w(x)} \right) \middle| X = x \right]}_{\text{needs to be 0}}
\end{aligned} \tag{6}
$$

Remark: Note that the common support condition IA2b is required for $m(w,x)$ to be defined and to not divide by zero as it ensures $e_w(x) > 0$

Identification of Conditional Average Potential Outcomes - AIPW (2/2)

Show that the second part of (6) is zero

$$
\begin{aligned}
&\mathbb{E}\left[(Y(w) - m(w,x))\left(\frac{D(w) - e_w(x)}{e_w(x)}\right)\middle| X = x\right] \\
&\overset{IA2a}{=} \mathbb{E}\left[(Y(w) - m(w,x))|X = x\right]\mathbb{E}\left[\left(\frac{D(w) - e_w(x)}{e_w(x)}\right)\middle| X = x\right] \\
&= \left(\mathbb{E}\left[Y(w) \mid X = x\right] - m(w,x)\right)\left(\frac{\mathbb{E}\left[D(w) \mid X = x\right] - e_w(x)}{e_w(x)}\right) \\
&\overset{IA3,\text{Cons.}}{=} \left(\mathbb{E}\left[Y \mid W = w, X = x\right] - m(w,x)\right)\left(\frac{\mathbb{E}\left[D(w) \mid X = x\right] - e_w(x)}{e_w(x)}\right) \\
&= \underbrace{\left(m(w,x) - m(w,x)\right)}_{=0}\underbrace{\left(\frac{e_w(x) - e_w(x)}{e_w(x)}\right)}_{=0} = 0 \quad \square
\end{aligned}
$$

Remark: Note that only one of the nuisance parameters needs to be correct $\Rightarrow$ doubly robust identification

Augmented IPW

From identification to estimation

The identification results suggest the following estimators for APO:

$$\text{Eq. (1): } \hat{\gamma}_w^{RA} = \frac{1}{N} \sum_i \hat{m}(w, X_i) \tag{7}$$

$$\text{Eq. (2): } \hat{\gamma}_w^{IPW} = \frac{1}{N} \sum_i \frac{\mathbb{1}[W_i = w] Y_i}{\hat{e}_w(X_i)} \tag{8}$$

$$\text{Eq. (3): } \hat{\gamma}_w^{AIPW} = \frac{1}{N} \sum_i \left(\hat{m}(w, X_i) + \frac{\mathbb{1}[W_i = w](Y_i - \hat{m}(w, X_i))}{\hat{e}_w(X_i)} \right) \tag{9}$$

where we estimated the nuisance parameters $\hat{m}(w, x)$ and $\hat{e}_w(x)$ in a first step

Only one strategy works with ML

Estimators using parametrically estimated nuisance parameters are common and work for all three strategies

BUT model selection problem remains

$\Rightarrow$ Supervised ML could be helpful again

However, estimators based on only one nuisance parameter inherit the slow convergence rates of the ML method $\Rightarrow$ no standard $\sqrt{N}$-based inference

Chernozhukov et al. (2018) show that the estimator in (9) is consistent, asymptotically normal and semiparametrically efficient if nuisance parameters are high-quality and cross-fitted predictions (like last week)

AIPW Double ML: procedure

AIPW Double ML proceeds as follows:

1. Form cross-fitted predictions of $\hat{m}(w, X)$ and $\hat{e}_w(X)$ using ML methods
2. Create a pseudo-outcome

$$\tilde{Y}_{\gamma_w} = \underbrace{\hat{m}(w, X)}_{\text{outcome prediction}} + \underbrace{\frac{\mathbb{1}[W = w](Y - \hat{m}(w, X))}{\hat{e}_w(X)}}_{\text{weighted residual}} \tag{10}$$

3. Estimate APO as mean of the pseudo-outcome $\hat{\gamma}_w^{AIPW} = \frac{1}{N} \sum_i \tilde{Y}_{i,\gamma_w}$
4. Run t-test on the mean for hypothesis testing (no adjustments needed)

This is known as the Doubly Robust or Augmented Inverse Probability Weighting (AIPW) estimator (I prefer AIPW b/c there are more doubly robust estimators than there are more AIPW estimators, but never mind...)

Why does it work?

The "magic" feature is again that the underlying score

$$\frac{1}{N}\sum_{i=1}^{N}\underbrace{\left(\hat{m}(w,X_i)+\frac{\mathbb{1}[W_i=w](Y_i-\hat{m}(w,X_i))}{\hat{e}_w(X_i)}-\hat{\gamma}_w^{AIPW}\right)}_{\psi(Y_i,W_i,\hat{m}(w,X_i),\hat{e}_w(X_i))}=0 \tag{11}$$

$$\iff \quad \hat{\gamma}_w^{AIPW}=\frac{1}{N}\sum_{i=1}^{N}\left(\hat{m}(w,X_i)+\frac{\mathbb{1}[W_i=w](Y_i-\hat{m}(w,X_i))}{\hat{e}_w(X_i)}\right) \tag{12}$$

is Neyman-orthogonal, i.e. small errors in the estimation of nuisance parameters do not distort estimation of the target parameter

Estimators based on regression adjustment or IPW only are not Neyman-orthogonal

Neyman-orthogonality of AIPW (1/5)

Neyman-orthogonality means that the Gateaux derivative with respect to the nuisance parameters is zero in expectation at the true nuisance parameters (NP):

$$\partial_r \mathbb{E}[\psi(\ldots, m + r(\tilde{m} - m), e + r(\tilde{e} - e))]|_{r=0} = 0 \tag{13}$$

where we suppress the dependencies of NPs and denote by, e.g., $\tilde{m}$ any other value of the outcome nuisance than the true value m

This looks scary, but we just need to know how to setup the problem and take standard derivatives (quotient rule)

For simplictly, we get rid of the brackets and the underscore to write the score with true target and nuisance parameters as

$$\psi(Y, W, \gamma_w, m(w, X), e(X)) = m + \frac{\mathbb{1}[W = w]Y}{e} - \frac{\mathbb{1}[W_i = w]m}{e} - \gamma_w$$

Neyman-orthogonality of AIPW (2/5)

Again we use $D(w) := \mathbb{1}[W = w]$ for brevity

First, add perturbations to the true nuisance parameters in the score

$$
\begin{aligned}
&\psi(\ldots, m + r(\tilde{m} - m), e + r(\tilde{e} - e)) \\
&\quad = (m + r(\tilde{m} - m)) + \frac{D(w)Y}{e + r(\tilde{e} - e)} - \frac{D(w)(m + r(\tilde{m} - m))}{e + r(\tilde{e} - e)} - \gamma_w \\
&\quad = \underbrace{(m + r(\tilde{m} - m))}_{(i)} + \underbrace{\frac{D(w)Y}{e + r(\tilde{e} - e)}}_{(ii)} - \underbrace{\frac{D(w)m}{e + r(\tilde{e} - e)}}_{(iii)} - \underbrace{\frac{D(w)r(\tilde{m} - m)}{e + r(\tilde{e} - e)}}_{(iv)} - \gamma_w
\end{aligned}
$$

Note that with $r = 0$, we are back to the original score

With $r \neq 0$ the nuisance parameters are distorted

Next, take the derivative wrt r

Neyman-orthogonality of AIPW (3/5)

Second, take the derivative wrt r

$$\begin{aligned}
&\partial_r \psi(\ldots, m + r(\tilde{m} - m), e + r(\tilde{e} - e)) \\
&\quad = \underbrace{(\tilde{m} - m)}_{\partial_r(i)} - \underbrace{\frac{D(w)Y(\tilde{e} - e)}{(e + r(\tilde{e} - e))^2}}_{\partial_r(ii)} + \underbrace{\frac{D(w)m(\tilde{e} - e)}{(e + r(\tilde{e} - e))^2}}_{\partial_r(iii)} \\
&\qquad - \underbrace{\frac{D(w)(\tilde{m} - m)(e + r(\tilde{e} - e)) - D(w)r(\tilde{m} - m)(\tilde{e} - e)}{(e + r(\tilde{e} - e))^2}}_{\partial_r(iv)}
\end{aligned}$$

Neyman-orthogonality of AIPW (4/5)

Third, evaluate at $r = 0$

$$
\begin{aligned}
&\partial_r \psi(\ldots, m + r(\tilde{m} - m), e + r(\tilde{e} - e))|_{r=0} \\
&\quad = (\tilde{m} - m) - \frac{D(w)Y(\tilde{e} - e)}{(e + 0(\tilde{e} - e))^2} + \frac{D(w)m(\tilde{e} - e)}{(e + 0(\tilde{e} - e))^2} \\
&\qquad - \frac{D(w)(\tilde{m} - m)(e + 0(\tilde{e} - e)) - D(w)0(\tilde{m} - m)(\tilde{e} - e)}{(e + 0(\tilde{e} - e))^2} \\
&\quad = (\tilde{m} - m) - \frac{D(w)Y(\tilde{e} - e)}{e^2} + \frac{D(w)m(\tilde{e} - e)}{e^2} - \frac{D(w)(\tilde{m} - m)e}{e^2}
\end{aligned}
$$

Neyman-orthogonality of AIPW (5/5)

Fourth, take expectation

$$
\begin{aligned}
&\partial_r \, \mathbb{E}[\psi(\ldots, m + r(\tilde{m} - m), e + r(\tilde{e} - e))]|_{r=0} \\
&\quad = \mathbb{E}\left[(\tilde{m} - m) - \frac{D(w)Y(\tilde{e} - e)}{e^2} + \frac{D(w)m(\tilde{e} - e)}{e^2} - \frac{D(w)(\tilde{m} - m)e}{e^2}\right] \\
&\quad \overset{LIE}{=} \mathbb{E}\left[\mathbb{E}\left[(\tilde{m} - m) - \frac{D(w)Y(\tilde{e} - e)}{e^2} + \frac{D(w)m(\tilde{e} - e)}{e^2} - \frac{D(w)(\tilde{m} - m)e}{e^2}\middle| X\right]\right] \\
&\quad = \mathbb{E}\left[(\tilde{m} - m) - \frac{em(\tilde{e} - e)}{e^2} + \frac{em(\tilde{e} - e)}{e^2} - \frac{e(\tilde{m} - m)e}{e^2}\right] = 0
\end{aligned}
$$

because

$$
\begin{aligned}
\mathbb{E}[D(w)Y \mid X] &\overset{LIE}{=} \mathbb{P}[D(w) = 0 \mid X] \, \mathbb{E}[0Y \mid W \neq w, X] + \mathbb{P}[D(w) = 1 \mid X] \, \mathbb{E}[1Y \mid W = w, X] \\
&= \mathbb{P}[D(w) = 1 \mid X] \, \mathbb{E}[Y \mid W = w, X] = em
\end{aligned}
$$

$\Rightarrow$ The Gateaux derivative wrt NP is zero $\Rightarrow$ Neyman-orthogonal score

Average treatment effect (1/2)

Following the same logic we can estimate the ATE as follows:

1. Form cross-fitted predictions of $\hat{m}(1,X)$, $\hat{m}(0,X)$ and $\hat{e}(X)$ using ML methods
2. Create a pseudo-outcome

$$\begin{aligned}\tilde{Y}_{ATE} &= \tilde{Y}_{\gamma_1} - \tilde{Y}_{\gamma_0} \\ &= \underbrace{\hat{m}(1,X) - \hat{m}(0,X)}_{\text{outcome predictions}} + \underbrace{\frac{W(Y-\hat{m}(1,X))}{\hat{e}(X)} - \frac{(1-W)(Y-\hat{m}(0,X))}{1-\hat{e}(X)}}_{\text{weighted residuals}}\end{aligned} \tag{14}$$

   where we use that $\mathbb{1}[W=1] = W$, $\mathbb{1}[W=0] = 1-W$, $e_1(X) = e(X)$, $e_0(X) = 1 - e(X)$
3. Estimate ATE as mean of the pseudo-outcome $\hat{\tau}^{AIPW}_{ATE} = \frac{1}{N}\sum_i \tilde{Y}_{i,ATE}$
4. Run t-test on the mean for hypothesis testing (no adjustments needed)

Average treatment effect (2/2)

As pseudo-outcome (14) is just the difference of the APO pseudo outcomes it inherits the Neyman-orthogonality

$\Rightarrow \hat{\tau}_{ATE}^{AIPW}$ is consistent, asymptotically normal and semiparametrically efficient when nuisance parameters are high-quality and cross-fitted predictions

Estimation ✅

Discussion AIPW Double ML

Advantages:

- No need to impose effect homogeneity $\Rightarrow$ differentiate between ATE and ATT (next lecture)
- Extends naturally to multiple treatments
- Basis for other estimators (see soon)

Disadvantages:

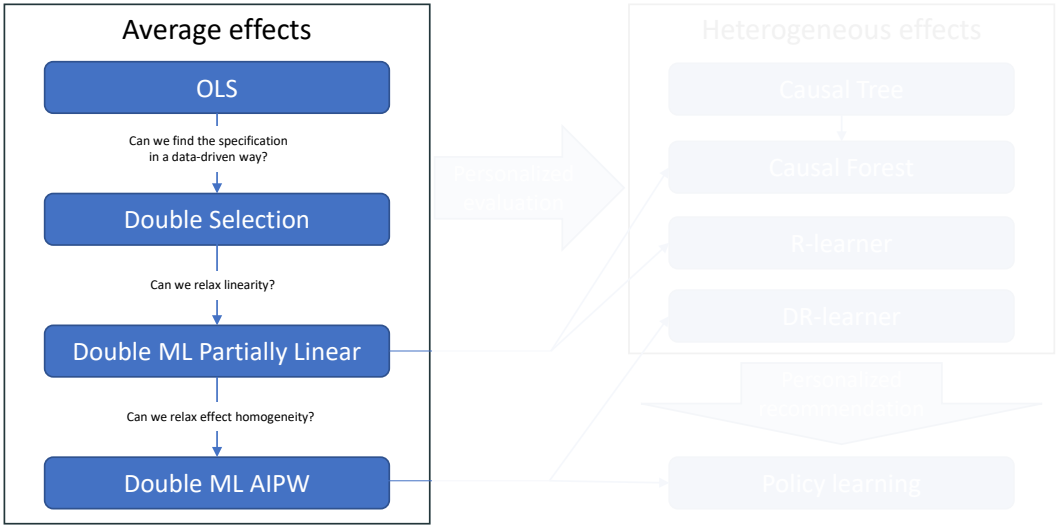
- Looks scary and complicated
- Sensitive to small propensity scores (no/weak overlap)
- Extensions for continuous treatments not trivial (Colangelo & Lee, 2019; Semenova & Chernozhukov, 2021)

Simulation notebook: AIPW Double ML (ATE)

Application notebook: Double ML for average treatment effects

Consolidation

Average effects unlocked



Main take-away

Estimation of average treatment effects can be split into multiple prediction problems

Combining them in the right way allows to use familiar inference procedures

$\Rightarrow$ We can leverage the powerful supervised ML toolbox

$\Rightarrow$ Moves the academic task from hand crafted regression models to the specification of suitable prediction methods

$\Rightarrow$ In the best case this increases

- quality and statistical validity of the estimates
- transparency
- time for researchers to do something more interesting

The recipe of Double ML can be generalized $\Rightarrow$ next week

Ceterum censeo a fancy method alone is not a credible identification strategy

$\Rightarrow$ separate identification and estimation