Causal Machine Learning

Lukáš Lafférs

Matej Bel University, Dept. of Mathematics

ROBUST 2022

[...an icebreaker joke here...]

This presentation

Introduction to Double Machine Learning framework

Three applications

Machine learning and causality

ML is (mostly) about prediction.

Machine learning and causality

ML is (mostly) about prediction.

Prediction is nice, but economists often care more about the **underlying mechanism** more.

Machine learning and causality

ML is (mostly) about prediction.

Prediction is nice, but economists often care more about the **underlying mechanism** more.

While ML gives us many great prediction tools, we are often interested in a **certain variable of interest**.

Machine learning and causality

ML is (mostly) about prediction.

Prediction is nice, but economists often care more about the **underlying mechanism** more.

While ML gives us many great prediction tools, we are often interested in a **certain variable of interest**.

Having a lot of information we need to cope with high dimensionality of covariates.

Job-seeker went through a training/course. Did it help?

Job-seeker went through a training/course. Did it help?

We know a lot about these job-seekers (say 300 variables).

Job-seeker went through a training/course. Did it help?

We know a lot about these job-seekers (say 300 variables).

But sample size is small.

Job-seeker went through a training/course. Did it help?

We know a lot about these job-seekers (say 300 variables).

But sample size is small.

We may try LASSO, but it will give us biased estimates.

Can we make use of the great predictive capabilities of ML algorithms for improving the estimation of parameters of interest?

This is an area of active research. Here we will discuss one important paper on **DOUBLE MACHINE LEARNING**

Victor, Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J.: "Double/debiased machine learning for treatment and structural parameters." The Econometrics Journal 21.1 (2018): C1-C68.

Double machine learning

Victor, Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. : ”Double/debiased machine learning for treatment and structural parameters.” The Econometrics Journal 21.1 (2018): C1-C68.

Double Machine Learning framework

Example: Consider the following partially linear model. θ is the parameter of interest.

$$\begin{aligned} Y &= \theta D + g(X) + U, \qquad & E[U|D,X] = 0 \\ D &= m(X) + V, \qquad & E[V|X] = 0 \end{aligned}$$

Double Machine Learning framework

Example: Consider the following partially linear model. θ is the parameter of interest.

$$\begin{aligned} Y &= \theta D + g(X) + U, \qquad && E[U|D,X] = 0 \\ D &= m(X) + V, && E[V|X] = 0 \end{aligned}$$

Split the data into two parts

- Use the first one to get $\hat{g}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{\theta}$ from regressing $Y - \hat{g}(X)$ on D

Double Machine Learning framework

Example: Consider the following partially linear model. θ is the parameter of interest.

$$
\begin{aligned}
Y &= \theta D + g(X) + U, \qquad & E[U|D,X] = 0 \\
D &= m(X) + V, \qquad & E[V|X] = 0
\end{aligned}
$$

Split the data into two parts

- Use the first one to get $\hat{g}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{\theta}$ from regressing $Y - \hat{g}(X)$ on D

$$
\hat{\theta}_1 = \left(\frac{1}{n} \sum_i D_i^2 \right)^{-1} \frac{1}{n} \sum_i D_i (Y_i - \hat{g}(X_i))
$$

Double Machine Learning framework

Example: Consider the following partially linear model. θ is the parameter of interest.

$$\begin{aligned} Y &= \theta D + g(X) + U, & E[U|D,X] &= 0 \\ D &= m(X) + V, & E[V|X] &= 0 \end{aligned}$$

Split the data into two parts

- Use the first one to get $\hat{g}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{\theta}$ from regressing $Y - \hat{g}(X)$ on D

$$\hat{\theta}_1 = \left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{n}\sum_i D_i(Y_i - \hat{g}(X_i))$$

$\hat{\theta}_1$ is based on $E[\psi_1] = 0$ where $\psi_1 = D(Y - g(X) - \theta D)$

Double Machine Learning framework

How does this naive estimator behave?

$$\sqrt{n}(\hat{\theta}_1 - \theta) = \underbrace{\left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_i D_i U_i}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{\left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_i D_i (g(X_i) - \hat{g}(X_i))}_{\text{In general divergent.}}$$

Double Machine Learning framework

How does this naive estimator behave?

$$\sqrt{n}(\hat{\theta}_1 - \theta) = \underbrace{\left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_i D_i U_i}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{\left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_i D_i(g(X_i) - \hat{g}(X_i))}_{\text{In general divergent.}}$$

Why?

$$\left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_i D_i(g(X_i) - \hat{g}(X_i)) = \left(E[D_i^2]\right)^{-1} \frac{1}{\sqrt{n}}\sum_i \underbrace{m_i(X_i)}_{\neq 0} \underbrace{(g(X_i) - \hat{g}(X_i))}_{\text{more slowly than } \sqrt{n}} + \underbrace{o_P(1)}_{\to_P 0}$$

Double Machine Learning framework

How does this naive estimator behave?

$$\sqrt{n}(\hat{\theta}_1 - \theta) = \underbrace{\left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_i D_i U_i}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{\left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_i D_i (g(X_i) - \hat{g}(X_i))}_{\text{In general divergent.}}$$

Why?

$$\left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_i D_i (g(X_i) - \hat{g}(X_i)) = \left(E[D_i^2]\right)^{-1} \frac{1}{\sqrt{n}}\sum_i \underbrace{m_i(X_i)}_{\neq 0} \underbrace{(g(X_i) - \hat{g}(X_i))}_{\text{more slowly than } \sqrt{n}} + \underbrace{o_P(1)}_{\to_P 0}$$

So it leads to a regularization bias.

Double Machine Learning framework

Now we do something else.

Double Machine Learning framework

Now we do something else.

Instead of $\psi_1 = D(Y - g(X) - \theta D)$ we will base our estimation on different moment conditions:

Double Machine Learning framework

Now we do something else.

Instead of $\psi_1 = D(Y - g(X) - \theta D)$ we will base our estimation on different moment conditions:

$$\psi_2 = V(Y - g(X) - \theta D) = (D - m(X)) \cdot (Y - g(X) - \theta D)$$
$$\psi_3 = V(Y - g(X) - \theta V) = (D - m(X)) \cdot (Y - g(X) - \theta(D - m(X)))$$

Double Machine Learning framework

Now we do something else.

Instead of $\psi_1 = D(Y - g(X) - \theta D)$ we will base our estimation on different moment conditions:

$$\psi_2 = V(Y - g(X) - \theta D) = (D - m(X)) \cdot (Y - g(X) - \theta D)$$
$$\psi_3 = V(Y - g(X) - \theta V) = (D - m(X)) \cdot (Y - g(X) - \theta (D - m(X)))$$

These moment conditions are somewhat more ”clever” as the problematic regularization bias part will converge to zero under mild conditions.

$\hat{\theta}_2$ based on ψ_2

Split the data into two parts

- Use the first one to get $\hat{g}$ and $\hat{m}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{V} = D - \hat{m}(X)$ and use this to get $\hat{\theta}_2$ .....

$\hat{\theta}_2$ based on ψ_2

Split the data into two parts

- Use the first one to get $\hat{g}$ and $\hat{m}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{V} = D - \hat{m}(X)$ and use this to get $\hat{\theta}_2$ .....

$$\sqrt{n}(\hat{\theta}_2 - \theta) = \underbrace{a^*}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

$\hat{\theta}_2$ based on ψ_2

Split the data into two parts

- Use the first one to get $\hat{g}$ and $\hat{m}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{V} = D - \hat{m}(X)$ and use this to get $\hat{\theta}_2$ .....

$$\sqrt{n}(\hat{\theta}_2 - \theta) = \underbrace{a^*}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

- Regularization bias : $b^* = \left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}} \sum_i (m(X_i) - \hat{m}(X_i))(g(X_i) - \hat{g}(X_i))$

$\hat{\theta}_2$ based on ψ_2

Split the data into two parts

- Use the first one to get $\hat{g}$ and $\hat{m}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{V} = D - \hat{m}(X)$ and use this to get $\hat{\theta}_2$ .....

$$\sqrt{n}(\hat{\theta}_2 - \theta) = \underbrace{a^*}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

- Regularization bias : $b^* = \left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}} \sum_i (m(X_i) - \hat{m}(X_i))(g(X_i) - \hat{g}(X_i))$
- Overfitting bias: Sample splitting takes care of this.

Regularization bias : $b^* = \left(\frac{1}{n}\sum_i D_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_i (m(X_i) - \hat{m}(X_i)) \cdot (g(X_i) - \hat{g}(X_i))$

Regularization bias : $b^* = \left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}} \sum_i (m(X_i) - \hat{m}(X_i)) \cdot (g(X_i) - \hat{g}(X_i))$

$\hat{g}$ and $\hat{m}$ no longer need to converge at the rate $n^{-1/2}$

Regularization bias : $b^* = \left(\frac{1}{n}\sum_i D_i^2\right)^{-1} \frac{1}{\sqrt{n}} \sum_i (m(X_i) - \hat{m}(X_i)) \cdot (g(X_i) - \hat{g}(X_i))$

$\hat{g}$ and $\hat{m}$ no longer need to converge at the rate $n^{-1/2}$

It is sufficient if they both converge at the rate $n^{-1/4}$ and this is much much easier for the ML algorithms.

$\hat{\theta}_3$ based on ψ_3

Split the data into two parts

- Use the first one to get $\hat{g}$ and $\hat{m}$ by some ML algorithm (LASSO, RF)

$\hat{\theta}_3$ based on ψ_3

Split the data into two parts

- Use the first one to get $\hat{g}$ and $\hat{m}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{V} = D - \hat{m}(X)$ and $\hat{W} = Y - \hat{m}(X)$ and use this to get $\hat{\theta}_3$ via regressing $\hat{W}$ on $\hat{V}$

$\hat{\theta}_3$ based on ψ_3

Split the data into two parts

- Use the first one to get $\hat{g}$ and $\hat{m}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{V} = D - \hat{m}(X)$ and $\hat{W} = Y - \hat{m}(X)$ and use this to get $\hat{\theta}_3$ via regressing $\hat{W}$ on $\hat{V}$

This is, in fact orthogonalization.

We project both D and Y onto space spanned by X. By means of Frisch-Waugh-Lowell theorem we recover the coefficient of D.

$\hat{\theta}_3$ based on ψ_3

Split the data into two parts

- Use the first one to get $\hat{g}$ and $\hat{m}$ by some ML algorithm (LASSO, RF)
- Use the second portion of data to get $\hat{V} = D - \hat{m}(X)$ and $\hat{W} = Y - \hat{m}(X)$ and use this to get $\hat{\theta}_3$ via regressing $\hat{W}$ on $\hat{V}$

This is, in fact orthogonalization.

We project both D and Y onto space spanned by X. By means of Frisch-Waugh-Lowell theorem we recover the coefficient of D.

Similar decomposition can be shown. Regularization bias also includes cross product $(m(X_i) - \hat{m}(X_i)) \cdot (g(X_i) - \hat{g}(X_i))$

What makes ϕ_2 and ϕ_3 different from ϕ_1 ???

What makes ϕ_2 and ϕ_3 different from ϕ_1 ???

Regularization bias vanishes under mild conditions.

What makes ϕ_2 and ϕ_3 different from ϕ_1 ???

Regularization bias vanishes under mild conditions.

In other words, ϕ_2 and ϕ_3 are both locally insensitive to some mild perturbations of $\hat{m}, \hat{g}$ around m, g.

Neyman-orthogonality

This local insensitiveness has a name: **Neyman-orthogonality**.

Neyman-orthogonality

This local insensitiveness has a name: **Neyman-orthogonality**.

- ψ is a moment condtion
- θ is the parameter of interest (target parameter), θ_0 is the true one
- η is the nuisance parameter, η_0 is the true one
- W denotes data

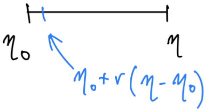
Neyman-orthogonality

This local insensitiveness has a name: **Neyman-orthogonality**.

- ψ is a moment condtion
- θ is the parameter of interest (target parameter), θ_0 is the true one
- η is the nuisance parameter, η_0 is the true one
- W denotes data

(r is small)

In the neigborhood of η_0, ψ does not change much



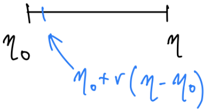
Neyman-orthogonality

This local insensitiveness has a name: **Neyman-orthogonality**.

- ψ is a moment condtion
- θ is the parameter of interest (target parameter), θ_0 is the true one
- η is the nuisance parameter, η_0 is the true one
- W denotes data

(r is small)

In the neigborhood of η_0, Ψ does not change much



$$\frac{\partial}{\partial r} E[\psi(W; \theta_0, \eta_0 + r(\eta - \eta_0))]\Big|_{r=0} = 0$$

Neyman-orthogonality of ψ_2

We will verify that ψ_2 satisfy the Neyman-orthogonality condition, while ψ_1 does not.

Notation

- $\eta = (m, g)$ is the vector of nuisance parameters, $\theta_0 = (m_0, g_0)$ is the true one
- $\eta_r = \eta_0 + r(\eta - \eta_0)$.

Neyman-orthogonality of ψ_2

$$
\begin{aligned}
\psi_2(W;\theta_0,\eta_r) &= (D-m_0(X)-r(m(X)-m_0(X)))\cdot(Y-g_0(X)-r(g(X)-g_0(X))-D\theta_0) \\
&= (D-m_0(X))\cdot(Y-g_0(X)-D\theta_0)+ \\
&\quad -r(D-m_0(X))\cdot(g(X)-g_0(X)) \\
&\quad -r(m(X)-m_0(X))\cdot(Y-g_0(X)-D\theta_0) \\
&\quad +r^2(m(X)-m_0(X))\cdot(g(X)-g_0(X))
\end{aligned}
$$

Neyman-orthogonality of ψ_2

$$
\begin{aligned}
\psi_2(W;\theta_0,\eta_r) &= (D - m_0(X) - r(m(X) - m_0(X)))\cdot(Y - g_0(X) - r(g(X) - g_0(X)) - D\theta_0)\\
&= (D - m_0(X))\cdot(Y - g_0(X) - D\theta_0) + \\
&\quad -r(D - m_0(X))\cdot(g(X) - g_0(X))\\
&\quad -r(m(X) - m_0(X))\cdot(Y - g_0(X) - D\theta_0)\\
&\quad +r^2(m(X) - m_0(X))\cdot(g(X) - g_0(X))\\
\frac{\partial}{\partial r}E[\psi_2(W;\theta_0,\eta_r)] &= -E[(D - m_0(X))\cdot(g(X) - g_0(X))]\\
&\quad -E[(m(X) - m_0(X))\cdot(Y - g_0(X) - D\theta_0)]\\
&\quad +2\cdot r\cdot E[(m(X) - m_0(X))\cdot(g(X) - g_0(X))]
\end{aligned}
$$

Neyman-orthogonality of ψ_2

$$\begin{aligned}
\psi_2(W;\theta_0,\eta_r) &= (D - m_0(X) - r(m(X) - m_0(X))) \cdot (Y - g_0(X) - r(g(X) - g_0(X)) - D\theta_0) \\
&= (D - m_0(X)) \cdot (Y - g_0(X) - D\theta_0) + \\
&\quad - r(D - m_0(X)) \cdot (g(X) - g_0(X)) \\
&\quad - r(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0) \\
&\quad + r^2(m(X) - m_0(X)) \cdot (g(X) - g_0(X)) \\
\frac{\partial}{\partial r} E[\psi_2(W;\theta_0,\eta_r)] &= -E[(D - m_0(X)) \cdot (g(X) - g_0(X))] \\
&\quad - E[(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0)] \\
&\quad + 2 \cdot r \cdot E[(m(X) - m_0(X)) \cdot (g(X) - g_0(X))] \\
\left.\frac{\partial}{\partial r} E[\psi_2(W;\theta_0,\eta_r)]\right|_{r=0} &= -E[(D - m_0(X)) \cdot (g(X) - g_0(X))] \\
&\quad - E[(m(X) - m_0(X)) \cdot (Y - g_0(X) - D\theta_0)]
\end{aligned}$$

Neyman-orthogonality of ψ_2

$$
\begin{aligned}
\frac{\partial}{\partial r} E[\psi(W;\theta_0,\eta_r)]\Big|_{r=0} &= -E[(D-m_0(X))\cdot(g(X)-g_0(X))] \\
&\quad -E[(m(X)-m_0(X))\cdot(Y-g_0(X)-D\theta_0)] \\
&= 0
\end{aligned}
$$

Neyman-orthogonality of ψ_2

$$
\begin{aligned}
\frac{\partial}{\partial r} E[\psi(W;\theta_0,\eta_r)]\Big|_{r=0} &= -E[(D - m_0(X))\cdot(g(X) - g_0(X))] \\
&\quad -E[(m(X) - m_0(X))\cdot(Y - g_0(X) - D\theta_0)] \\
&= 0
\end{aligned}
$$

because

Neyman-orthogonality of ψ_2

$$
\begin{aligned}
\frac{\partial}{\partial r} E[\psi(W;\theta_0,\eta_r)]\Big|_{r=0} &= -E[(D-m_0(X))\cdot(g(X)-g_0(X))] \\
&\quad -E[(m(X)-m_0(X))\cdot(Y-g_0(X)-D\theta_0)] \\
&= 0
\end{aligned}
$$

because

$$
E[(D-m_0(X))\cdot(g(x)-g_0(X))] \quad = \quad E[(g(X)-g_0(X))\cdot \underbrace{E[D-m_0(X)|X]}_{E[V|X]=0}] = 0
$$

Neyman-orthogonality of ψ_2

$$
\begin{aligned}
\frac{\partial}{\partial r} E[\psi(W;\theta_0,\eta_r)]\Big|_{r=0} &= -E[(D-m_0(X))\cdot(g(X)-g_0(X))] \\
&\quad -E[(m(X)-m_0(X))\cdot(Y-g_0(X)-D\theta_0)] \\
&= 0
\end{aligned}
$$

because

$$
\begin{aligned}
E[(D-m_0(X))\cdot(g(x)-g_0(X))] &= E[(g(X)-g_0(X))\cdot \underbrace{E[D-m_0(X)|X]}_{E[V|X]=0}] = 0 \\
E[(m(X)-m_0(X))\cdot(Y-g_0(X)-D\theta_0)] &= E[(m(X)-m_0(X))\cdot \underbrace{E[Y-g_0(X)-D\theta_0|X,D]}_{E[U|X,D]=0}] = 0
\end{aligned}
$$

and hence ψ_2 is indeed Neyman-orthogonal.

Neyman-orthogonality of ψ_3

Aimilarly as ψ_2 but the derivation is a bit longer.

Neyman-orthogonality of ψ_1 ???

$$\psi_1(W;\theta_0,\eta_r) \quad = \quad D\cdot(Y - g_0(X) - r(g(X) - g_0(X)) - D\theta_0)$$

Neyman-orthogonality of ψ_1 ???

$$\begin{aligned}
\psi_1(W;\theta_0,\eta_r) &= D\cdot(Y - g_0(X) - r(g(X) - g_0(X)) - D\theta_0) \\
\frac{\partial}{\partial r}E[\psi_2(W;\theta_0,\eta_r)] &= -E[D\cdot(g(X) - g_0(X))]
\end{aligned}$$

Neyman-orthogonality of ψ_1 ???

$$
\begin{aligned}
\psi_1(W;\theta_0,\eta_r) &= D\cdot(Y - g_0(X) - r(g(X)-g_0(X)) - D\theta_0) \\
\frac{\partial}{\partial r} E[\psi_2(W;\theta_0,\eta_r)] &= -E[D\cdot(g(X)-g_0(X))] \\
\left.\frac{\partial}{\partial r} E[\psi(W;\theta_0,\eta_r)]\right|_{r=0} &= -E[D\cdot(g(X)-g_0(X))]
\end{aligned}
$$

Neyman-orthogonality of ψ_1 ???

$$
\begin{aligned}
\psi_1(W;\theta_0,\eta_r) &= D\cdot(Y - g_0(X) - r(g(X)-g_0(X)) - D\theta_0) \\
\frac{\partial}{\partial r}E[\psi_2(W;\theta_0,\eta_r)] &= -E[D\cdot(g(X)-g_0(X))] \\
\left.\frac{\partial}{\partial r}E[\psi(W;\theta_0,\eta_r)]\right|_{r=0} &= -E[D\cdot(g(X)-g_0(X))] \\
&\neq 0
\end{aligned}
$$

There is nothing we could do to use $E[U|X,D]=0$ and $E[V|X]=0$ to make this term equal to zero.

Overfitting bias

$$\sqrt{n}(\hat{\theta}_2 - \theta) = \underbrace{a^*}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

Overfitting bias

$$\sqrt{n}(\hat{\theta}_2 - \theta) = \underbrace{a^*}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

Overfitting bias may arise from the fact that the same data is used for both estimation of nuisance functions and target parameter.

Overfitting bias

$$\sqrt{n}(\hat{\theta}_2 - \theta) = \underbrace{a^*}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

Overfitting bias may arise from the fact that the same data is used for both estimation of nuisance functions and target parameter.

We can **split the data**.

Overfitting bias

$$\sqrt{n}(\hat{\theta}_2 - \theta) = \underbrace{a^*}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

Overfitting bias may arise from the fact that the same data is used for both estimation of nuisance functions and target parameter.

We can **split the data**.

$\rightarrow$ But then we loose many observations.

Overfitting bias

$$\sqrt{n}(\hat{\theta}_2 - \theta) = \underbrace{a^*}_{\text{Nicely behaved, approx. Gaussian}} + \underbrace{b^*}_{\text{Regularization bias}} + \underbrace{c^*}_{\text{Overfitting bias}}$$

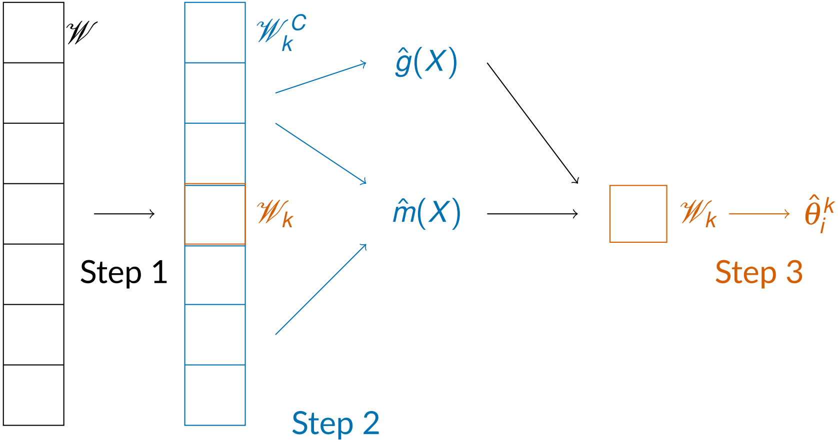
Overfitting bias may arise from the fact that the same data is used for both estimation of nuisance functions and target parameter.

We can **split the data.**

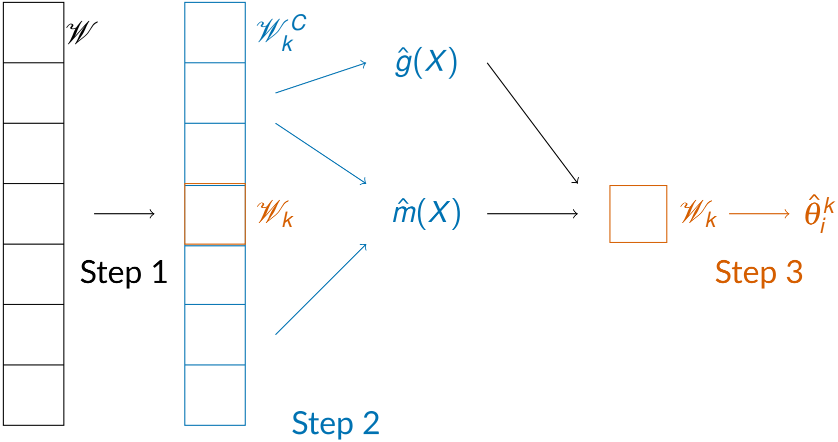
$\rightarrow$ But then we loose many observations.

How to fix this? **Swap the roles** of the two data parts and then average across them!

Sample splitting for dealing with overfitting bias



Sample splitting for dealing with overfitting bias



$$\hat{\theta} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \hat{\theta}_i^k$$

Step 4

DML wrap-up (1)

We saw three: $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$.

DML wrap-up (1)

We saw three: $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$.

Based on: ψ_1, ψ_2 and ψ_3.

DML wrap-up (1)

We saw three: $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$.

Based on: ψ_1, ψ_2 and ψ_3.

While ψ_1 was locally sensitive to some small changes in the η, the other two ψ_2 and ψ_3 were not.

DML wrap-up (1)

We saw three: $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$.

Based on: ψ_1, ψ_2 and ψ_3.

While ψ_1 was locally sensitive to some small changes in the η, the other two ψ_2 and ψ_3 were not.

This allows us to get rid of the regularization bias.

DML wrap-up (1)

We saw three: $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$.

Based on: ψ_1, ψ_2 and ψ_3.

While ψ_1 was locally sensitive to some small changes in the η, the other two ψ_2 and ψ_3 were not.

This allows us to get rid of the regularization bias.

Sample-splitting removes the overfitting bias.

DML wrap-up (2)

- Estimator $\hat{\theta}$ based on Neyman-orthogonal moment function ψ
- Apply sample splitting
- Nuisance parameter estimators are ”good enough” (e.g. converge at rate at least $n^{-1/4}$ - so that the regularization bias vanishes)

DML wrap-up (2)

- Estimator $\hat{\theta}$ based on Neyman-orthogonal moment function ψ
- Apply sample splitting
- Nuisance parameter estimators are "good enough" (e.g. converge at rate at least $n^{-1/4}$ - so that the regularization bias vanishes)

We get that (Theorem 1 in Chernozhukov et al. 2019)

$$\sqrt{n}(\hat{\theta} - \theta) \rightarrow N(0, \sigma^2)$$

Asymptotically normally distributed estimator that is $\sqrt{n}$ consistent.

DML wrap-up (3)

DML provides a framework for developing estimators that:

DML wrap-up (3)

DML provides a framework for developing estimators that:

- can handle high-dimensional data

DML wrap-up (3)

DML provides a framework for developing estimators that:

- can handle high-dimensional data
- make use of predictive powers of ML

DML wrap-up (3)

DML provides a framework for developing estimators that:

- can handle high-dimensional data
- make use of predictive powers of ML
- are well behaved under mild conditions

Heterogeneity of effects

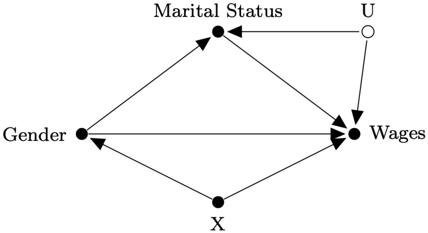
Use X_i to predict estimated effect $\hat{\Delta}_i$

Different samples for:
(i) estimation of $\hat{\Delta}_i$ using DML
(ii) association between X_i and $\hat{\Delta}_i$

Wager, Stefan, and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." Journal of the American Statistical Association 113.523 (2018): 1228-1242.

Limitations - Kitchen sink regression

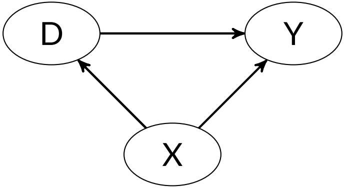
[proper source should be cited here]



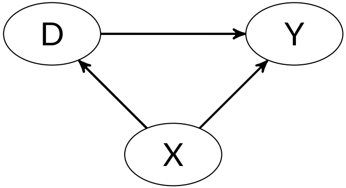
Hünermund, Beyers and Caspi (2021)

Hünermund, Paul, Beyers Louw, and Itamar Caspi. ”Double Machine Learning and Bad Controls–A Cautionary Tale.” arXiv preprint arXiv:2108.11294 (2021).

DML and treatment effects



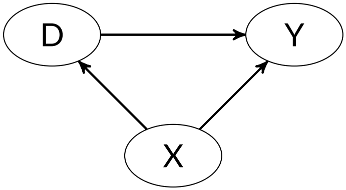
DML and treatment effects



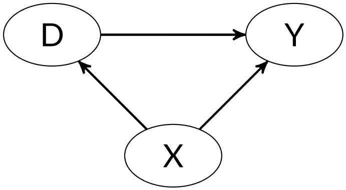
Notation:

- $Y(d)$: (Potential) outcome as function of treatment d
- Y - observed outcome
- D - observed treatment
- X - observed covariates

DML and treatment effects



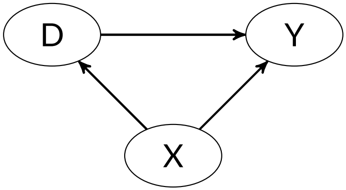
DML and treatment effects



Object of interest:

$$\Delta = E[Y(1) - Y(0)]$$

DML and treatment effects



Object of interest:

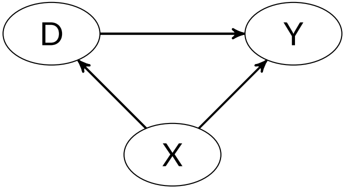
$$\Delta = E[Y(1) - Y(0)]$$

Indentifying assumptions:

1) Conditional independence of D:
$\{Y(1), Y(0)\} \perp D \mid X$

2) Common support:
$\Pr(D = d | X = x) > 0$

DML and treatment effects

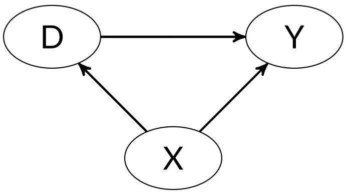


Moment function:

$$\psi(W;\theta_0,\eta) = \frac{I\{D=d\}\cdot[Y_2-\mu(d,X)]}{p(X)}+\mu(d,X)-\theta_0.$$

$$E\left[\psi(W;\theta_0,\eta)\right] = E\left[Y(d)\right]-\theta_0=0$$

DML and treatment effects

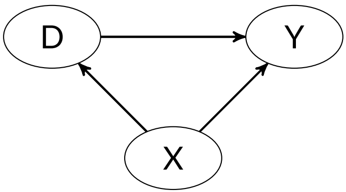


Moment function:

$$
\begin{aligned}
\psi(W;\theta_0,\eta) &= \frac{I\{D=d\}\cdot[Y_2-\mu(d,X)]}{p(X)}+\mu(d,X)-\theta_0. \\
E\Big[\psi(W;\theta_0,\eta)\Big] &= E\Big[Y(d)\Big]-\theta_0=0
\end{aligned}
$$

Data: $W=(Y,D,X)$

DML and treatment effects



Moment function:

$$\begin{aligned}\psi(W;\theta_0,\eta) &= \frac{I\{D=d\}\cdot[Y_2-\mu(d,X)]}{p(X)}+\mu(d,X)-\theta_0.\\ E\left[\psi(W;\theta_0,\eta)\right] &= E\left[Y(d)\right]-\theta_0=0\end{aligned}$$

Data: $W=(Y,D,X)$

Nuisance functions: $\eta=(p,\mu)$

- $p(X)\equiv\Pr(D=d|X)$
- $\mu(D,X)\equiv E[Y|D,X]$

also Doubly robust estimator.

Bang, Heejung, and James M. Robins. "Doubly robust estimation in missing data and causal inference models." Biometrics 61.4 (2005): 962-973.

So far, none of this was my work.

DML applications

There are **many**:

Double/debiased **machine learning** for treatment and structural parameters
V Chernozhukov, D Chetverikov, M Demirer, E Duflo… - 2018 - academic.oup.com
… To estimate η 0 , we consider the use of statistical or **machine learning** (ML) methods, which are … We call the resulting set of methods **double** or debiased ML (DML). We verify that DML …
☆ Save   Cite   Cited by 1198   Related articles   All 22 versions   Web of Science: 279

CITATIONS	VIEWS	ALTMETRIC
387	46,917	99

More metrics information

Most read

Double/debiased machine learning for treatment and structural parameters

DML applications

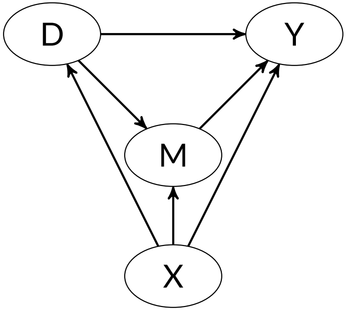
- mediation analysis (with H. Farbmacher, M. Huber, H. Langen, M. Spindler)
- dynamic treatment effects (with H. Bodory, M. Huber)
- sample selection models (with M. Bia, M. Huber)

First application

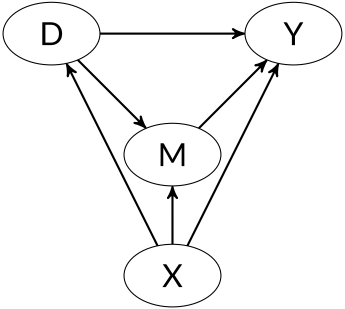
DML and mediation analysis

Causal mediation analysis with double machine learning (Econometrics Journal, 2022, 25 (2), 277—300, with Helmut Farbmacher, Martin Huber, Henrika Langen and Martin Spindler)

DML and mediation analysis



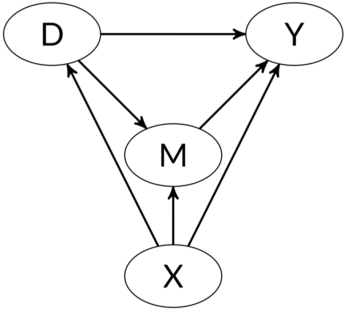
DML and mediation analysis



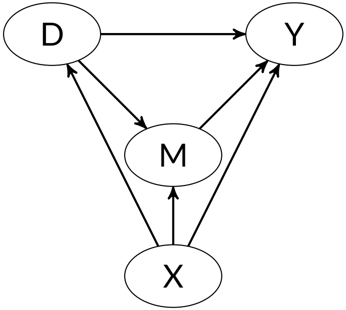
Notation:

- $M(d)$: (Potential) mediator under treatment $d \in \{0, 1\}$
- $Y(d, m)$: (Potential) outcome as function of treatment d and mediator m
- Y - observed outcome
- D - observed treatment
- M - observed mediator
- X - observed covariates

DML and mediation analysis



DML and mediation analysis

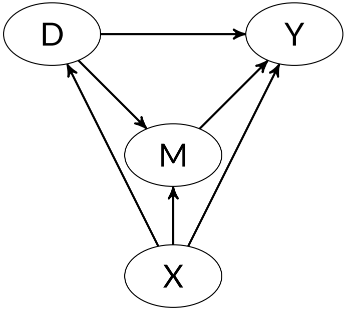


Objects of interest:

$$\delta(d) = E[Y(d, M(1)) - Y(d, M(0))]$$

$$\theta(d) = E[Y(1, M(d)) - Y(0, M(d))]$$

DML and mediation analysis



Objects of interest:

$$\delta(d) = E[Y(d, M(1)) - Y(d, M(0))]$$

$$\theta(d) = E[Y(1, M(d)) - Y(0, M(d))]$$

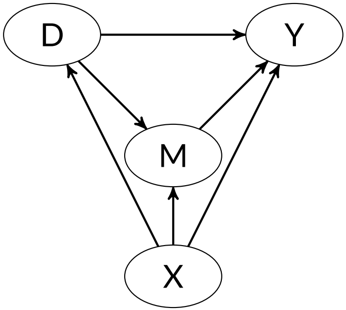
Indentifying assumptions:

1) Conditional independence of D:
$\{Y(d', m),\ M(d)\} \perp D \mid X$

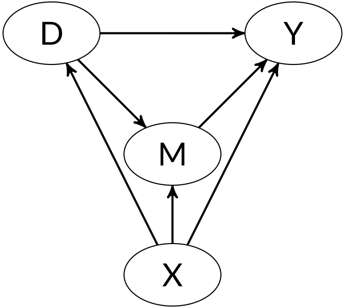
2) Conditional independence of M:
$Y(d', m) \perp M \mid D = d,\ X = x$

3) Common support:
$\Pr(D = d | M = m, X = x) > 0$

DML and mediation analysis



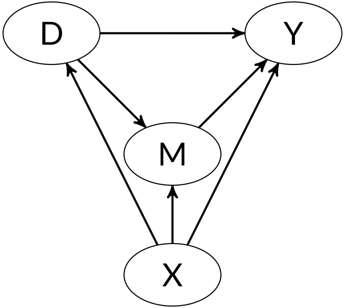
DML and mediation analysis



Moment function:

$$
\begin{aligned}
\psi(W;\theta_0,\eta) &= \frac{I\{D=d\}(1-p_d(M,X))}{p_{dm}(M,X)\cdot 1-p_d(X)}\cdot[Y-\mu(d,M,X)] \\
&+ \frac{I\{D=1-d\}}{1-p_d(X)}\cdot\Big[\mu(d,M,X)-\omega(1-d,X)\Big] \\
&+ E\Big[\mu(d,M,X)\Big|D=1-d,X\Big]-\theta_0. \\
E\Big[\psi(W;\theta_0,\eta)\Big] &= E\Big[Y(d,M(1-d))\Big]-\theta_0=0
\end{aligned}
$$

DML and mediation analysis

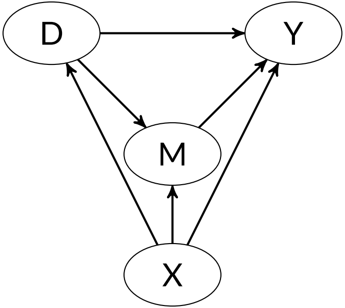


Moment function:

$$
\begin{aligned}
\psi(W;\theta_0,\eta) &= \frac{I\{D=d\}(1-p_d(M,X))}{p_{dm}(M,X)\cdot 1-p_d(X)}\cdot[Y-\mu(d,M,X)] \\
&+ \frac{I\{D=1-d\}}{1-p_d(X)}\cdot\Big[\mu(d,M,X)-\omega(1-d,X)\Big] \\
&+ E\Big[\mu(d,M,X)\Big|D=1-d,X\Big]-\theta_0. \\
E\Big[\psi(W;\theta_0,\eta)\Big] &= E\Big[Y(d,M(1-d))\Big]-\theta_0=0
\end{aligned}
$$

Data: $W=(Y,D,M,X)$

DML and mediation analysis



Moment function:

$$
\begin{aligned}
\psi(W;\theta_0,\eta) &= \frac{I\{D=d\}(1-p_d(M,X))}{p_{dm}(M,X)\cdot 1-p_d(X)}\cdot[Y-\mu(d,M,X)] \\
&+ \frac{I\{D=1-d\}}{1-p_d(X)}\cdot\Big[\mu(d,M,X)-\omega(1-d,X)\Big] \\
&+ E\Big[\mu(d,M,X)\Big|D=1-d,X\Big]-\theta_0. \\
E\Big[\psi(W;\theta_0,\eta)\Big] &= E\Big[Y(d,M(1-d))\Big]-\theta_0=0
\end{aligned}
$$

Data: $W=(Y,D,M,X)$

Nuisance functions: $\eta=(p_d,p_{dm},\mu,\omega)$

- $p_d(X)=Pr(D=d|X)$
- $p_{dm}(M,X)=Pr(D=d|M,X)$
- $\mu(D,M,X)=E(Y|D,M,X)$
- $\omega(1-d,X)=E[\mu(d,M,X)|D=1-d,X]$

DML and mediation analysis: application

Application to NLSY1997:

- National Longitudinal Survey of Youth 1997; representative survey of 8,984 individuals born in the years 1980-84 in the U.S.
- D: Health insurance coverage at 2006 interview.
- M: Routine check-up between 2006 and 2007 interview.
- Y: Self-reported general health at 2008 interview (1=excellent; 5=poor).
- X: 770 control variables, 601 of which are dummies (incl. 252 dummies for missing values) measured in or prior to 2005.

Application

Results:

		direct		indirect	
	$\hat{\Delta}$	$\hat{\theta}(1)$	$\hat{\theta}(0)$	$\hat{\delta}(1)$	$\hat{\delta}(0)$
	Modified score using Bayes' rule				
effect	-0.05	-0.07	-0.05	0.00	0.02
se	0.03	0.03	0.03	0.01	0.01
p-value	0.10	0.03	0.10	0.89	0.07

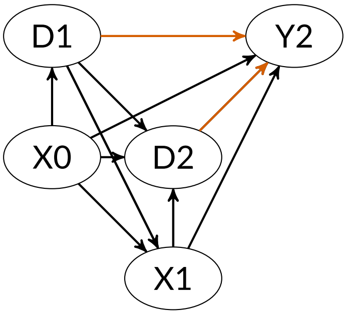
- Health insurance coverage appears to moderately improve general health in the short run among young adults in the U.S. through mechanisms other than routine checkups.

Second application

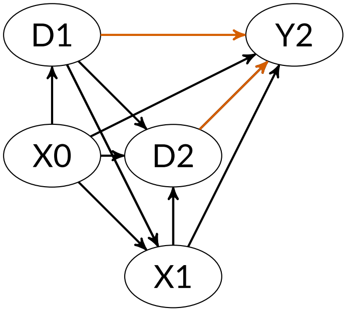
DML and dynamic treatment effects

Evaluating (weighted) dynamic treatment effects by double machine learning (forthcoming in Econometrics Journal with Hugo Bodory and Martin Huber)

DML and dynamic treatment effects



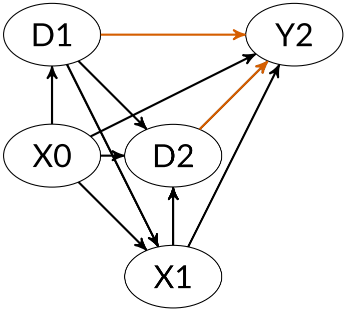
DML and dynamic treatment effects



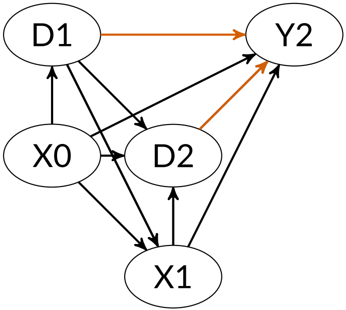
Notation:

- D_t, Y_t, X_t: Treatment, outcome, covariates in period $t \in \{0, 1, 2\}$
- $d_1, d_2 \in \{0, 1, ..., Q\}$, Q is the number of non-zero treatments
- Treatment sequence $\underline{D}_2 \equiv (D_1, D_2)$ and $\underline{d}_2 \equiv (d_1, d_2)$
- $Y_2(\underline{d}_2)$: (Potential) outcome in period 2 under sequence $\underline{d}_2$
- Covariate sequence $\underline{X}_1 \equiv (X_0, X_1)$

DML and dynamic treatment effects



DML and dynamic treatment effects

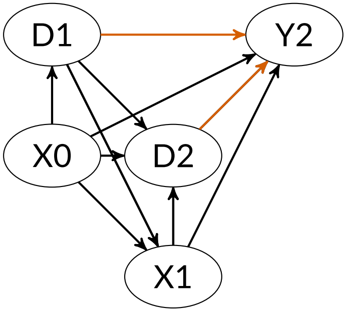


Objects of interest:

$$E[Y(\underline{d}_2)] - E[Y(\underline{d}_2^*)]$$

Indentifying assumptions:

DML and dynamic treatment effects



Objects of interest:

$$E[Y(\underline{d}_2)] - E[Y(\underline{d}_2^*)]$$

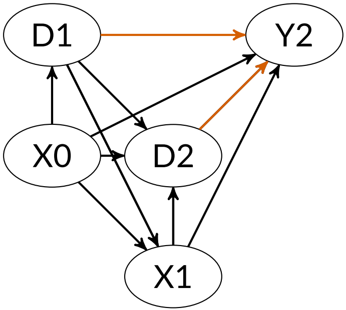
Indentifying assumptions:

1) Conditional ind. of the first treatment:
$Y_2(\underline{d}_2) \perp D_1 | X_0$, for $\underline{d}_2 \in \{0, 1, ..., Q\}^2$

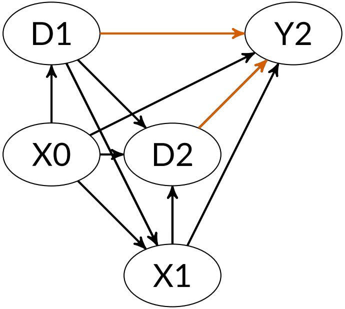
2) Conditional ind. of the second treatment:
$Y_2(\underline{d}_2) \perp D_2 | D_1, X_0, X_1$, for $\underline{d}_2 \in \{0, 1, ..., Q\}^2$.

3) Common support:
$\Pr(D_1 = d_1 | X_0) > 0$, $\Pr(D_2 = d_2 | D_1, \underline{X}_1) > 0$

DML and dynamic treatment effects



DML and dynamic treatment effects

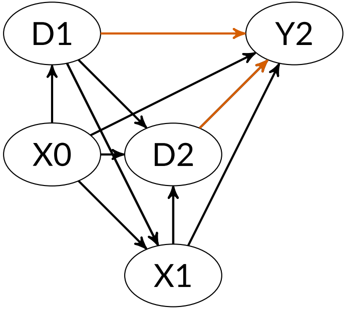


Moment function:

$$\psi(W;\theta_0,\eta) = \frac{I\{D_1=d_1\}\cdot I\{D_2=d_2\}\cdot[Y_2-\mu^{Y_2}(\underline{d}_2,\underline{X}_1)]}{p^{d_1}(X_0)\cdot p^{d_2}(d_1,\underline{X}_1)} + \frac{I\{D_1=d_1\}\cdot[\mu^{Y_2}(\underline{d}_2,\underline{X}_1)-\nu^{Y_2}(\underline{d}_2,X_0)]}{p^{d_1}(X_0)} + \nu^{Y_2}(\underline{d}_2,X_0)-\theta_0.$$

$$E\Big[\psi(W;\theta_0,\eta)\Big] = E\Big[Y_2(\underline{d}_2)\Big]-\theta_0=0$$

DML and dynamic treatment effects

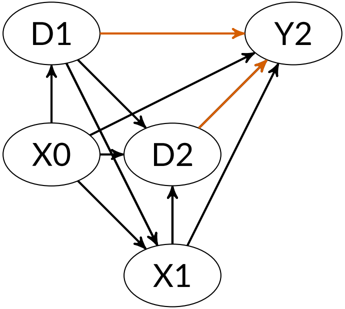


Moment function:

$$
\begin{aligned}
\psi(W;\theta_0,\eta) &= \frac{I\{D_1=d_1\}\cdot I\{D_2=d_2\}\cdot[Y_2-\mu^{Y_2}(\underline{d}_2,\underline{X}_1)]}{p^{d_1}(X_0)\cdot p^{d_2}(d_1,\underline{X}_1)} \\
&+ \frac{I\{D_1=d_1\}\cdot[\mu^{Y_2}(\underline{d}_2,\underline{X}_1)-\nu^{Y_2}(\underline{d}_2,X_0)]}{p^{d_1}(X_0)} + \nu^{Y_2}(\underline{d}_2,X_0)-\theta_0. \\
E\Big[\psi(W;\theta_0,\eta)\Big] &= E\Big[Y_2(\underline{d}_2)\Big]-\theta_0=0
\end{aligned}
$$

Data: $W=(Y_2,D_1,D_2,X_0,X_1)$

DML and dynamic treatment effects



Moment function:

$$\begin{aligned}
\psi(W;\theta_0,\eta) &= \frac{I\{D_1=d_1\}\cdot I\{D_2=d_2\}\cdot[Y_2-\mu^{Y_2}(\underline{d}_2,\underline{X}_1)]}{p^{d_1}(X_0)\cdot p^{d_2}(d_1,\underline{X}_1)} \\
&+ \frac{I\{D_1=d_1\}\cdot[\mu^{Y_2}(\underline{d}_2,\underline{X}_1)-\nu^{Y_2}(\underline{d}_2,X_0)]}{p^{d_1}(X_0)} + \nu^{Y_2}(\underline{d}_2,X_0)-\theta_0. \\
E\Big[\psi(W;\theta_0,\eta)\Big] &= E\Big[Y_2(\underline{d}_2)\Big]-\theta_0=0
\end{aligned}$$

Data: $W=(Y_2,D_1,D_2,X_0,X_1)$

Nuisance functions: $\eta=(p^{d_1},p^{d_2},\mu^{Y_2},\nu^{Y_2})$

- $p^{d_1}(X_0)\equiv\Pr(D_1=d_1|X_0)$
- $p^{d_2}(D_1,\underline{X}_1)\equiv\Pr(D_2=d_2|D_1,\underline{X}_1)$
- $\mu^{Y_2}(\underline{D}_2,\underline{X}_1)\equiv E[Y_2|\underline{D}_2,X_0,X_1]$
- $\nu^{Y_2}(\underline{D}_2,X_0)\equiv E[E[Y_2|\underline{D}_2,X_0,X_1]|D_1,X_0]$,

DML and dynamic treatment effects: Simulation study

Data generating process:

$$
\begin{aligned}
Y_2 &= D_1 + D_2 + X_0'\beta_{X_0} + X_1'\beta_{X_1} + U, \\
D_2 &= I\{0.3D_1 + X_0'\beta_{X_0} + X_1'\beta_{X_1} + V > 0\}, \quad D_1 = I\{X_0'\beta_{X_0} + W > 0\}, \\
X_1 &\sim N(0, \Sigma_1), \quad X_0 \sim N(0, \Sigma_0), \quad U, V, W \sim N(0,1), \text{ independently of each other.}
\end{aligned}
$$

- i-th element in β_{X_0} and β_{X_1} corresponds to $0.4/i^4$ for $i = 1, \ldots, p$.
- Σ_0 and Σ_1 are defined by setting the covariance of the ith and jth covariate in X_0 or X_1 to $\Sigma_{b,ij} = 0.5^{|i-j|}$, with $b \in \{0,1\}$.

DML and dynamic treatment effects: Simulation study

covar-iates	sample size	true effect	absolute bias	standard deviation	average SE	RMSE	coverage in %
		ATE: $\hat{\Delta}(\underline{d}_2, \underline{d}_2^*)$ (all)					
50	2500	2	0.027	0.07	0.069	0.075	91.6
50	10000	2	0.007	0.035	0.034	0.036	94.4
100	2500	2	0.04	0.072	0.069	0.083	88.7
100	10000	2	0.011	0.035	0.034	0.037	94.4
500	2500	2	0.063	0.07	0.068	0.094	83.4
500	10000	2	0.019	0.035	0.034	0.04	90.0

DML and dynamic treatment effects: Simulation study

covar-iates	sample size	true effect	absolute bias	standard deviation	average SE	RMSE	coverage in %
		ATE: $\hat{\Delta}(\underline{d}_2, \underline{d}_2^*)$ (all)					
50	2500	2	0.027	0.07	0.069	0.075	91.6
50	10000	2	0.007	0.035	0.034	0.036	94.4
100	2500	2	0.04	0.072	0.069	0.083	88.7
100	10000	2	0.011	0.035	0.034	0.037	94.4
500	2500	2	0.063	0.07	0.068	0.094	83.4
500	10000	2	0.019	0.035	0.034	0.04	90.0

·4

·4

·4

÷2

÷2

÷2

DML and dynamic treatment effects: Application

Application to Job Corps experimental study:

- Sample comes from the Job Corps experimental study conducted in mid-90's, see Schochet et all (2008): 11313 young individuals with completed interviews four years after randomization (6828 assigned to Job Corps, 4485 randomized out).
- Outcome is employment four years after randomization.
- Treatment sequences are based on participation in academic or vocational training in the first or second year after randomization among those randomized in.

DML and dynamic treatment effects: Application

	Dynamic treatments		Job Corps	Observations
code	year 1	year 2		
00	no educ/train	no educ/train	no	4485
11	no educ/train	no educ/train	yes	320
12	no educ/train	acad educ	yes	43
13	no educ/train	voc train	yes	42
21	acad educ	no educ/train	yes	1328
22	acad educ	acad educ	yes	341
23	acad educ	voc train	yes	183
31	voc train	no educ/train	yes	1279
32	voc train	acad educ	yes	109
33	voc train	voc train	yes	573
missings				2610

DML and dynamic treatment effects: Application

- 1188 raw characteristics (**socio-economic characteristics**, pre-treatment education and training, labor market histories, job search activities, welfare receipt, health, crime...).

Table: Regressors

Type	X_0	X_1
raw variables		
dummy	295	575
categorical	53	13
numeric	26	226
total	374	814
modified for data analysis		
dummy	883	1201
numeric	26	226
total	909	1427

DML and dynamic treatment effects: Application

Results (outcome: employment after 4 years):

3= vocational 2=academic

10% ↑ employment after 4years

d_2	d_2^*	$\hat{E}[Y_2(d_2^*)\|S=1]$	$\hat{\Delta}(d_2, d_2^*, S=1)$	SE	p-value	observations	trimmed
33	22	0.76	0.1	0.06	0.11	3783	507
33	21	0.82	0.05	0.03	0.07	3783	43
33	11	0.81	0.08	0.03	0.02	2346	22

1= no training

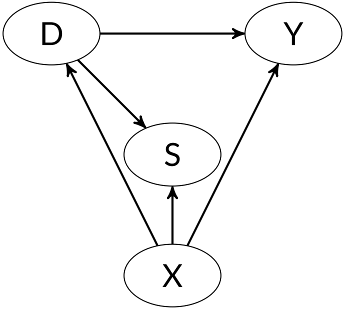
- ATE is estimated in the subsample with first treatment entering one of the treatment sequences compared.
- Random forests and 3-fold cross-validation.

Third application

DML and sample selection models

Double machine learning for sample selection models (arXiv:2012.00745 with Michela Bia and Martin Huber, revision requested)

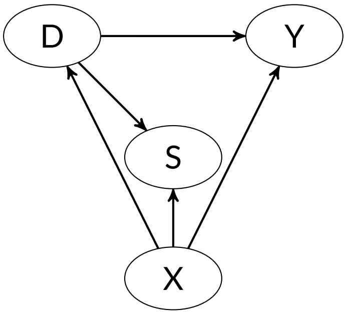
DML and sample selection models



Notation

- $Y(d)$: (Potential) outcome under treatment $d \in \{0, 1, ..., Q\}$.
- D: Treatment.
- Y: Outcome.
- S: Selection indicator.
- X: Covariates.

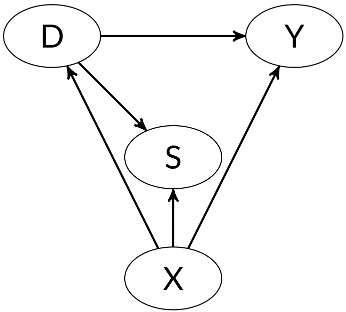
DML and sample selection models



Object of interest:

$$E[Y(d)] - E[Y(d^*)]$$

DML and sample selection models



Object of interest:

$$E[Y(d)] - E[Y(d^*)]$$

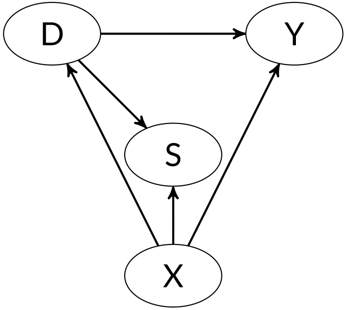
Indentifying assumptions:

1) Conditional independence of the treatment):
$Y(d) \perp D | X = x$

2) Conditional independence of selection:
$Y \perp S | D = d, X = x$

3) Common support:
(a) $\Pr(D = d | X = x) > 0$ and (b) $\Pr(S = 1 | D = d, X = x) > 0$

DML and sample selection models

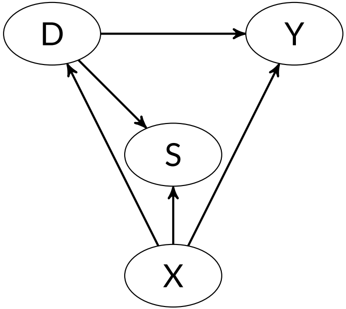


Moment function:

$$\psi(W;\theta_0,\eta) = \frac{I\{D=d\}\cdot S\cdot[Y-\mu(d,1,X)]}{p_d(X)\cdot\pi(d,X)} + \mu(d,1,X) - \theta_0.$$

$$E\Big[\psi(W;\theta_0,\eta)\Big] = E\Big[Y(d)\Big] - \theta_0 = 0$$

DML and sample selection models



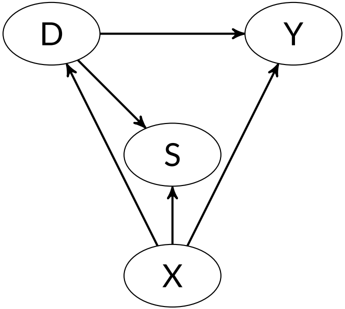
Moment function:

$$\psi(W;\theta_0,\eta) = \frac{I\{D=d\}\cdot S\cdot[Y-\mu(d,1,X)]}{p_d(X)\cdot\pi(d,X)} + \mu(d,1,X) - \theta_0.$$

$$E\Big[\psi(W;\theta_0,\eta)\Big] = E\Big[Y(d)\Big] - \theta_0 = 0$$

Data: $W = (Y.S, S, D, X)$

DML and sample selection models



Moment function:

$$\psi(W;\theta_0,\eta) = \frac{I\{D=d\}\cdot S\cdot[Y-\mu(d,1,X)]}{p_d(X)\cdot\pi(d,X)} + \mu(d,1,X) - \theta_0.$$

$$E\Big[\psi(W;\theta_0,\eta)\Big] = E\Big[Y(d)\Big] - \theta_0 = 0$$

Data: $W = (Y.S, S, D, X)$

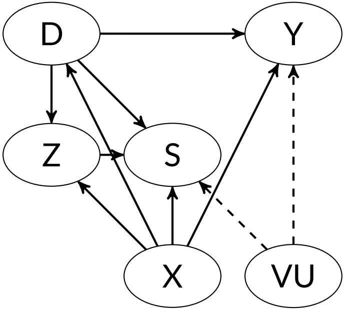
Nuisance functions: $\eta = (p^d, \pi, \mu)$

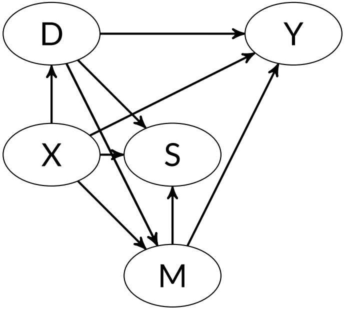
- $p^d(X) = \Pr(D = d|X)$
- $\pi(D, X) = \Pr(S = 1|D, X)$
- $\mu(D, S, X) = E[Y|D, S, X]$

Reg. conditions Algorithm

DML and sample selection models: other frameworks

Two additional setups not considered here.





DML and sample selection models: Simulation

Data generating process:

$$
\begin{aligned}
Y &= D + X'\beta + U \text{ with } Y \text{ being observed if } S = 1, \\
S &= I\{D + X'\beta + V > 0\}, \\
D &= I\{X'\beta + W > 0\}, \\
X &\sim N(0, \sigma^2_X), \quad (U, V) \sim N(0, \sigma^2_{U,V}), \quad W \sim N(0, 1).
\end{aligned}
$$

- ith element in the coefficient vector β is set to $0.4/i^2$ for $i = 1, ..., p$.
- σ^2_X is defined based on setting the covariance of the ith and jth covariate in X to $0.5^{|i-j|}$.
- $\sigma^2_{U,V} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

DML and sample selection models: Simulation

	true	bias	sd	RMSE	meanSE	coverage
n=2000						
DML MAR	1.000	0.003	0.060	0.060	0.063	0.939
n=8000						
DML MAR	1.000	0.012	0.031	0.033	0.034	0.934

DML and sample selection models: Application

Job Corps again.

- Outcome Y is **hourly wage** in last week of first year or four years after randomization, observed conditional on employment S.
- Treatment D is participation in academic or vocational **training** in the first year after randomization among those randomized in.

Evaluation sample:

Table: Treatment distribution

treatment	observations
randomized out of JC	1698
controls (no training)	200
academic training	830
vocational training	843

DML and sample selection models: Application

Table: ATE estimates

$D = 1$	$D = 0$	ATE	se	p-value
Theorem 1 (MAR)				
academic	no training	-0.170	0.253	0.501
vocational	no training	-0.519	0.405	0.199
Theorem 3 (IV)				
academic	no training	-0.192	0.174	0.705
vocational	no training	-0.537	0.404	0.199
Theorem 4 (sequential)				
academic	no training	0.170	0.117	0.147
vocational	no training	0.442	0.096	0.000

We observe small longer-term wage gains in terms of hourly wage.

Recapitulation

DML is a useful framework for estimation under high-dimensional setting.

Recapitulation

DML is a useful framework for estimation under high-dimensional setting.

It can automatically select among many covariates and avoid both regularization bias (via Neyman-orthogonal score) and overfitting bias (via cross-fitting) and provide root-n consistent and asymptotically normal estimator.

Recapitulation

DML is a useful framework for estimation under high-dimensional setting.

It can automatically select among many covariates and avoid both regularization bias (via Neyman-orthogonal score) and overfitting bias (via cross-fitting) and provide root-n consistent and asymptotically normal estimator.

I have shown a few instances where DML appears to be empirically relevant and useful.
(implemented in `causalweight` R package (Bodory and Huber 2018))

Thank you for your attention!

References

- Double machine learning framework: Chernozhukov, Victor, et al. ”Double/debiased machine learning for treatment and structural parameters.” The Econometrics Journal 21.1 (2018): C1-C68.
- Somewhat accessible intro to DML: https://towardsdatascience.com/double-machine-learning-for-causal-inference-78e0c6111f9d
- DML video by one of the authors of DML https://www.youtube.com/watch?v=eHOjmyoPCFU
- DoubleML package in R https://cran.r-project.org/web/packages/DoubleML/DoubleML.pdf
- Bach, Philipp, et al. ”DoubleML–An Object-Oriented Implementation of Double Machine Learning in R.” arXiv preprint arXiv:2103.09603 (2021).
- Bang, Heejung, and James M. Robins. ”Doubly robust estimation in missing data and causal inference models.” Biometrics 61.4 (2005): 962-973.
- Wager, Stefan, and Susan Athey. ”Estimation and inference of heterogeneous treatment effects using random forests.” Journal of the American Statistical Association 113.523 (2018): 1228-1242.
- Hünermund, Paul, Beyers Louw, and Itamar Caspi. ”Double Machine Learning and Bad Controls–A Cautionary Tale.” arXiv preprint arXiv:2108.11294 (2021).
- Farbmacher, Helmut, et al. ”Causal mediation analysis with double machine learning.” The Econometrics Journal 25.2 (2022): 277-300.
- Bodory, Hugo, Martin Huber, and Lukáš Lafférs. ”Evaluating (weighted) dynamic treatment effects by double machine learning.” forthcoming in The Econometrics Journal (2022).
- Bia, Michela, Martin Huber, and Lukáš Lafférs. ”Double machine learning for sample selection models.” arXiv preprint arXiv:2012.00745 (2020).
- Bodory, Hugo, and Martin Huber. ”The causalweight package for causal inference in R.” (2018).

Some additional materials:

Double machine learning

Algorithm 1: Estimation of $E[Y(d)]$

- Let $\mathscr{W} = \{W_i | 1 \leq i \leq n\}$ with $W_i = (Y_i \cdot S_i, D_i, S_i, X_i)$ for all i denote the set of observations in an i.i.d. sample of size n.

1. Split $\mathscr{W}$ in K subsamples. For each subsample k, let n_k denote its size, $\mathscr{W}_k$ the set of observations in the sample and $\mathscr{W}_k^C$ the complement set of all observations not in k.
2. For each k, use $\mathscr{W}_k^C$ to estimate the model parameters of the plug-ins $\mu(D, S=1, X)$, $p_d(X)$, $\pi(D, X)$ in order to predict these plug-ins in $\mathscr{W}_k$, where the predictions are denoted by $\hat{\mu}^k(D, 1, X)$, $\hat{p}_d^k(X)$, and $\hat{\pi}^k(D, X)$.
3. For each k, obtain an estimate of the score function (see ψ_d in (**??**)) for each observation i in $\mathscr{W}_k$, denoted by $\hat{\psi}_{d,i}^k$:

$$\hat{\psi}_{d,i}^k = \frac{I\{D_i = d\} \cdot S_i \cdot [Y_i - \hat{\mu}^k(d, 1, X_i)]}{\hat{p}_d^k(X_i) \cdot \hat{\pi}^k(d, X_i)} + \hat{\mu}^k(d, 1, X_i). \tag{1}$$

4. Average the estimated scores $\hat{\psi}_{d,i}^k$ over all observations across all K subsamples to obtain an estimate of $\Psi_{d0} = E[Y(d)]$ in the total sample, denoted by $\hat{\Psi}_d = 1/n \sum_{k=1}^K \sum_{i=1}^{n_k} \hat{\psi}_{d,i}^k$.

Back

Double machine learning (2)

Regularity conditions and root-n consistency:

Assumption 10 (regularity conditions and quality of plug-in parameter estimates):
For all probability laws $P \in \mathscr{P}$, where $\mathscr{P}$ is the set of all possible probability laws the following conditions hold for the random vector (Y, D, S, X) for $d \in \{0, 1, ..., Q\}$:

(a) $\|Y\|_q \leq C$, $\left\|E[Y^2|D=d, S=1, X]\right\|_\infty \leq C^2$,

(b) $\Pr(\varepsilon \leq p_{d0}(X) \leq 1-\varepsilon) = 1$, $\Pr(\varepsilon \leq \pi_0(d, X)) = 1$,

(c) $\|Y - \mu_0(d,1,X)\|_2 = E\left[(Y - \mu_0(d,1,X))^2\right]^{\frac{1}{2}} \geq c$

(d) Given a random subset I of $[n]$ of size $n_k = n/K$, the nuisance parameter estimator $\hat{\eta}_0 = \hat{\eta}_0((W_i)_{i \in I^C})$ satisfies the following conditions. With P-probability no less than $1 - \Delta_n$:

$$
\begin{aligned}
&\|\hat{\eta}_0 - \eta_0\|_q \leq C, \quad \|\hat{\eta}_0 - \eta_0\|_2 \leq \delta_n, \\
&\|\hat{p}_{d0}(X) - 1/2\|_\infty \leq 1/2 - \varepsilon, \quad \|\hat{\pi}_0(D,X) - 1/2\|_\infty \leq 1/2 - \varepsilon, \\
&\|\hat{\mu}_0(D,S,X) - \mu_0(D,S,X)\|_2 \times \|\hat{p}_{d0}(X) - p_0(X)\|_2 \leq \delta_n n^{-1/2}, \\
&\|\hat{\mu}_0(D,S,X) - \mu_0(D,S,X)\|_2 \times \|\hat{\pi}_0(D,X) - \pi_0(D,X)\|_2 \leq \delta_n n^{-1/2}.
\end{aligned}
$$