Estimating Price Elasticities with Text and Images

Australian National University

Research School of Economics: Computational Economics Series

Tuesday, 11 June 2024

Sven Klaassen
University of Hamburg
Economic AI

Jan Teichert-Kluge
University of Hamburg

Philipp Bach
University of Hamburg
Economic AI

Victor Chernozhukov
Massachusetts Institute of Technology

Martin Spindler
University of Hamburg
Economic AI

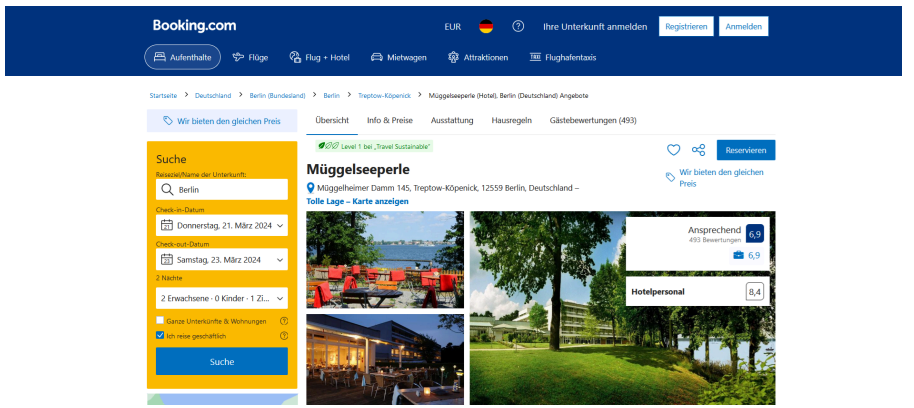
Motivation

Motivation

- Causal Inference mainly relies on tabular data

	A	B	C	D	E	F	G	H	I	J	K
1	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Y
2	0,26222194	0,73438599	0,44667203	0,97138377	0,081064	0,28441456	0,30204598	0,42916989	0,45202869	0,48458485	0,69688241
3	0,02086748	0,76377164	0,82984213	0,44114913	0,6791843	0,10440879	0,06871173	0,65049866	0,38145381	0,37883627	0,68724919
4	0,80214042	0,60108851	0,3924631	0,64680441	0,52447716	0,71441139	0,92058036	0,50487157	0,0639613	0,72018174	0,70586344
5	0,69860016	0,48544746	0,02697306	0,56612716	0,64030572	0,38122836	0,77357736	0,7665354	0,56153968	0,40036439	0,96259876
6	0,76871281	0,52693951	0,23152259	0,99343286	0,19512117	0,52281687	0,74631891	0,51990814	0,8916446	0,13854693	0,15014547
7	0,80559982	0,49248903	0,52094313	0,83491804	0,19224862	0,51690107	0,99896405	0,21315395	0,13844972	0,5972433	0,30108879
8	0,14274622	0,68082916	0,21375762	0,12917741	0,101114	0,35962408	0,37101036	0,88481715	0,0917219	0,81144147	0,23748577
9	0,5565959	0,27805181	0,38191987	0,96746964	0,88262331	0,64204409	0,61391038	0,14798767	0,75991882	0,71364563	0,26831957
10	0,86680082	0,98940603	0,08032276	0,51284266	0,85067788	0,31879029	0,14212956	0,58889047	0,85003262	0,40984155	0,90320049
11	0,49294831	0,9930191	0,32376346	0,76486908	0,71150197	0,49307446	0,00732655	0,1366301	0,95413681	0,8192993	0,41659136
12	0,83367503	0,99203204	0,94500561	0,38442651	0,08267135	0,44319059	0,50597368	0,33789173	0,71150674	0,71723848	0,82498573
13	0,30901589	0,45442227	0,39182182	0,94058761	0,84662039	0,13237356	0,39203009	0,23863556	0,66904622	0,66739594	0,05703788
14	0,84854041	0,07220964	0,60564451	0,11235493	0,11423158	0,19425556	0,10525554	0,69400453	0,24273427	0,74521656	0,1934563
15	0,41204349	0,37890159	0,21172493	0,47214042	0,28972671	0,28045945	0,26144942	0,50182126	0,56848988	0,34171099	0,55046509
16	0,8270131	0,7550563	0,68080658	0,99718432	0,54748336	0,07330951	0,96389321	0,43492645	0,32444209	0,37929049	0,54724228
17	0,65968289	0,42809072	0,50651859	0,85136778	0,67725248	0,80776288	0,80154682	0,41961695	0,03794024	0,10550978	0,18250098
18	0,00250478	0,10976101	0,58665193	0,59539311	0,06932994	0,44532097	0,38697059	0,86107306	0,32445673	0,03897245	0,4593923
19	0,1646558	0,11563336	0,54131562	0,70212986	0,78935386	0,65759136	0,96083497	0,18585388	0,02245136	0,08388247	0,9684134

- In a lot of applications additional unstructured data is available



- We consider multimodal data as **confounders**
- Applications in Marketing, Medicine / Health, Finance, ….
  - Price elasticity of demand
  - Estimation of treatment effects conditioning on X-ray images

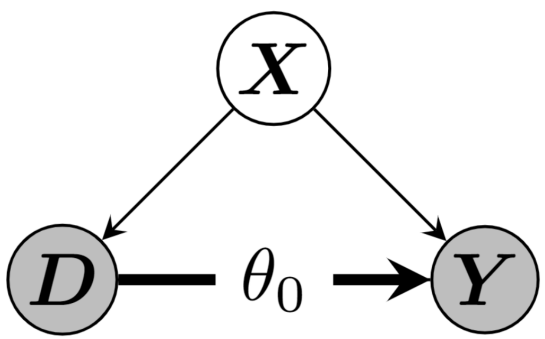
Motivating Example - PLR

Partially linear regression model (PLR)

$$
\begin{aligned}
Y &= \theta_0 D + g_0(X) + \varepsilon, & \mathbb{E}[\varepsilon | X, D] &= 0 \\
D &= m_0(X) + \vartheta, & \mathbb{E}[\vartheta | X] &= 0
\end{aligned}
$$

with

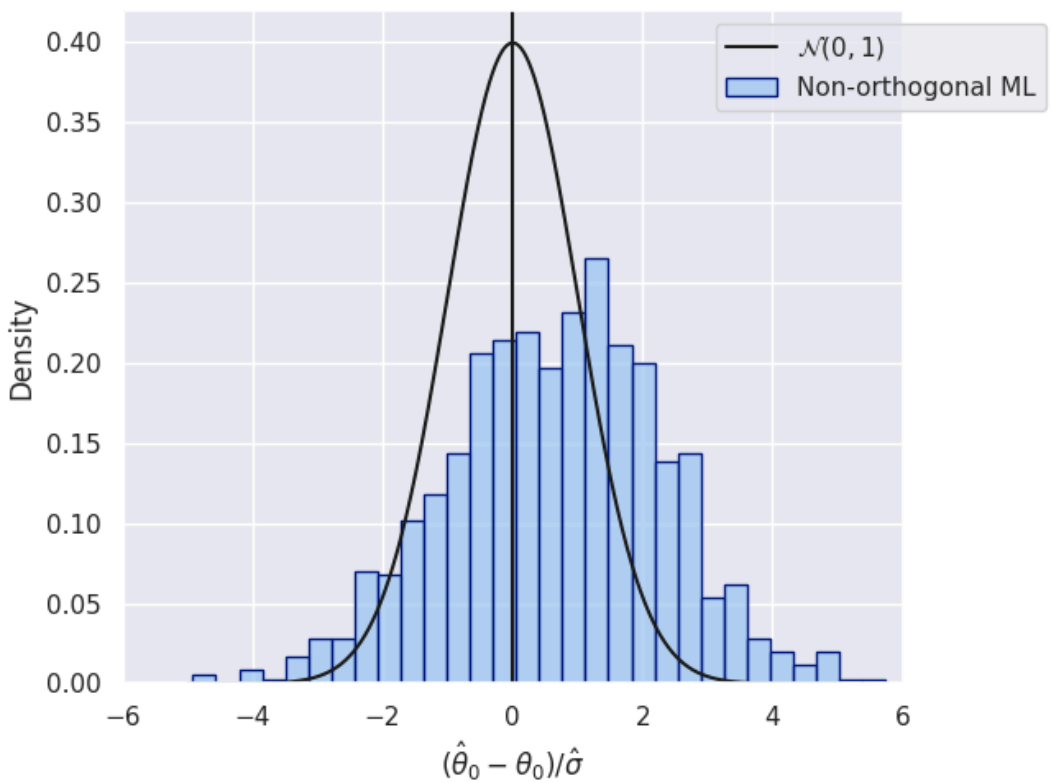
- Y - outcome variable
- D - policy/treatment variable
- $X = (X_1, \ldots)^T$ - vector of additional controls
- ε, ϑ - stochastic errors



DAG for PLR Model

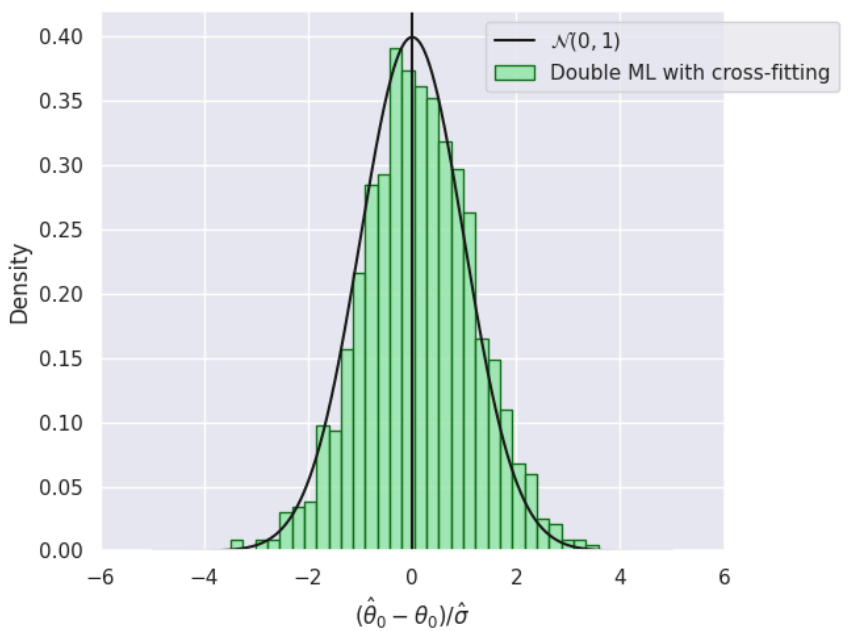
Motivating Example - Regularization Bias

- What if we simply plug-in ML predictions $\hat{g}(X)$ for $g_0(X)$ into $Y = \theta_0 D + g_0(X) + \varepsilon$?



Motivating Example - Orthogonalization

- *Frisch-Waugh-Lovell* style approach: θ_0 can be consistently estimated by **partialling out** X, i.e,
  1. Predict Y and D by $\mathbb{E}[Y|X]$ and $\mathbb{E}[D|X]$, obtained using ML methods
  2. Residualize $\tilde{Y} = Y - \mathbb{E}[Y|X]$ and $\tilde{D} = D - \mathbb{E}[D|X]$
  3. Regress $\tilde{Y}$ on $\tilde{D}$ to obtain $\hat{\theta}$



Double/Debiased Machine Learning (DML)

- Use machine learning methods to fit

$$l_0(X) := \mathbb{E}[Y|X]$$
$$m_0(X) := \mathbb{E}[D|X]$$

- Construct the orthogonalized score (either cross-fitting or split the sample) for a fitted nuisance learner $\hat{\eta} := (\hat{l}, \hat{m})$

$$\psi(W, \theta, \hat{\eta}) := \Big(Y - \hat{l}(X) - \theta\big(D - \hat{m}(X)\big)\Big)\Big(D - \hat{m}(X)\Big)$$

- Compute the estimate as the solution of

$$0 = \frac{1}{n}\sum_{i=1}^{n} \psi(W, \hat{\theta}, \hat{\eta})$$

Double/Debiased Machine Learning (DML)

- **Neyman Orthogonality**

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

  ensures that the moment condition identifying θ_0 is **insensitive to small pertubations** of the nuisance function η around η_0

- Assumptions for the nuisance elements (see Chernozhukov et al. (2018))
  $\|\hat{m} - m_0\|_{P,2} \times \left(\|\hat{m} - m_0\|_{P,2} + \|\hat{\ell} - \ell_0\|_{P,2}\right) \leq \delta_N N^{-1/2}$
- Under some regularity conditions the estimator $\hat{\theta}$ concentrates in $1/\sqrt{n}$-neighborhood of θ_0 and

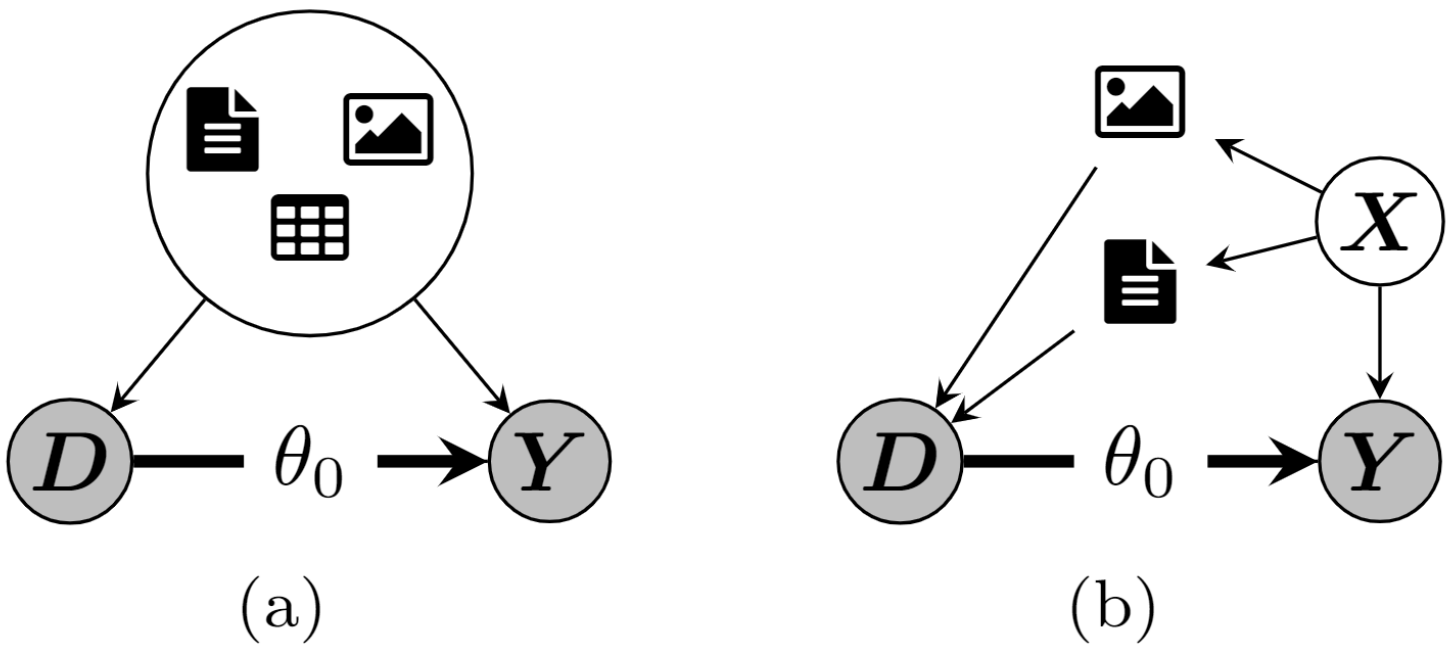
$$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathcal{N}(0, \sigma^2)$$

$\Rightarrow$ simple approach to control for confounders, *if they can be used for ML models!*

DoubleML Deep

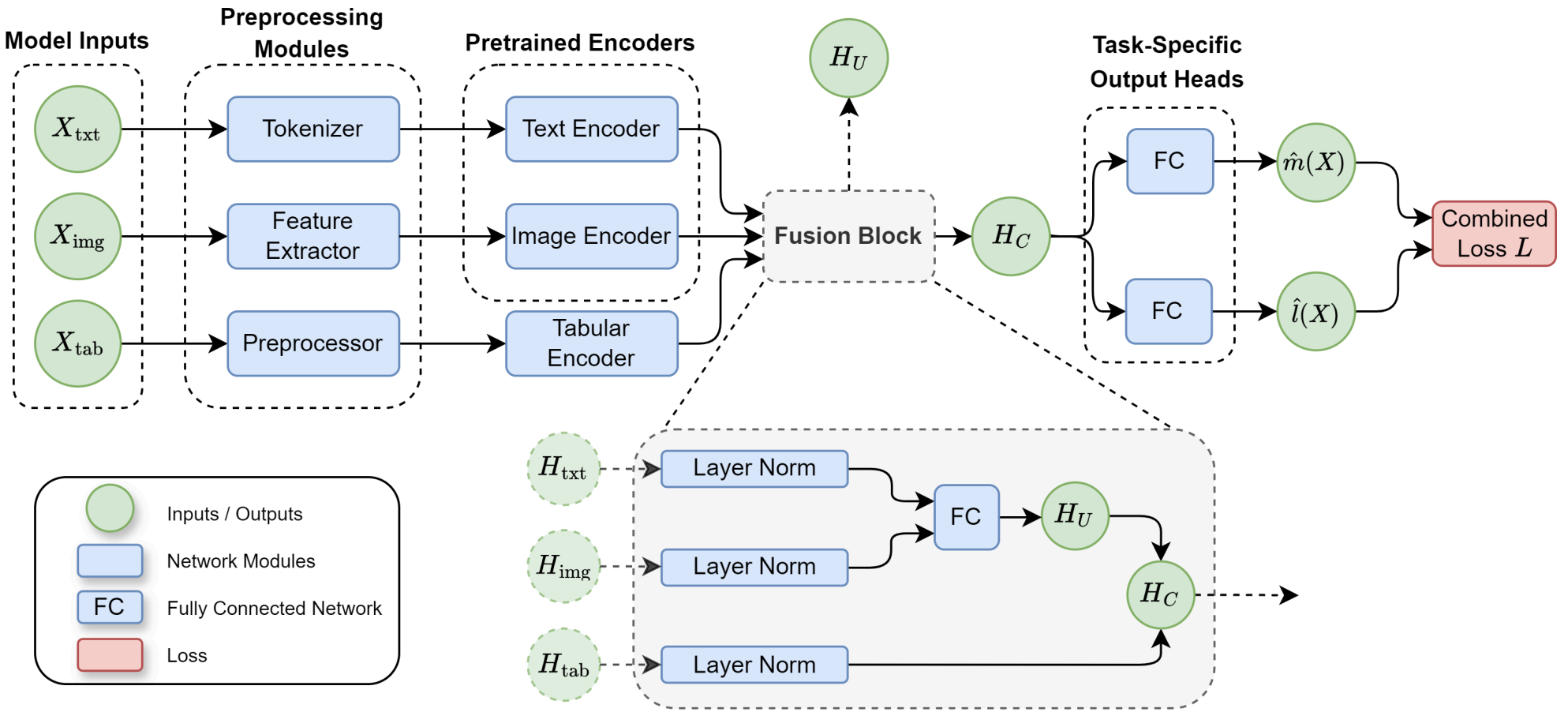
DoubleML Deep - Motivation

- Use multimodal data (text and images) *additionally* to conventional tabular features



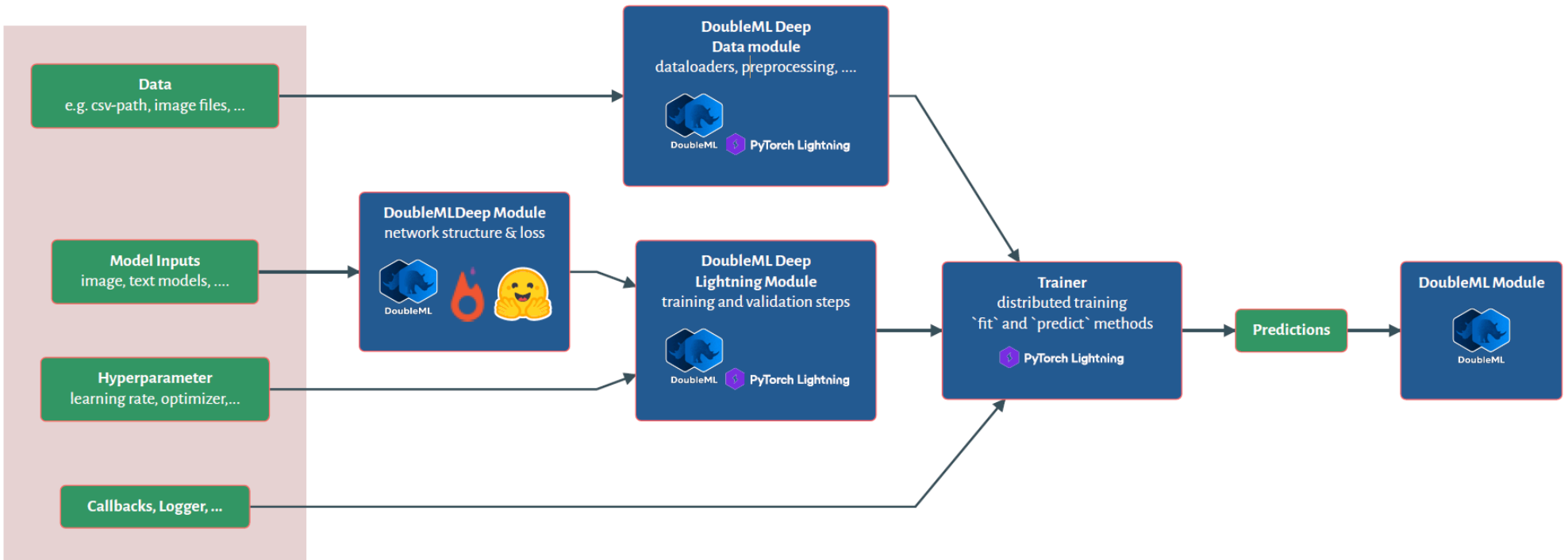
Examples of directed acyclic graphs (DAGs) with image and text confounding. (a) Direct confounding via image, text and tabular data. (b) Treatment decision is driven by text and images. All backdoor paths are blocked by conditioning on both image and text data.

DoubleML Deep - Module Structure



High-Level PLR Model Architecture. Both nuisance components are trained simultaneously with a combined loss.

DoubleML Deep - Workflow



High-Level Workflow of *DoubleMLDeep*

DoubleML Deep - More Details

- Use pretrained models for unstructured data e.g. Bert and Beit
- Basic models for tabular data seem sufficient
- Monitor the nuisance losses
  - $\|Y - \hat{l}(X)\|_{P,2} \leq \|l_0(X) - \hat{l}(X)\|_{P,2} + \sigma_\epsilon$
  - $\|D - \hat{m}(X)\|_{P,2} \leq \|m_0(X) - \hat{m}(X)\|_{P,2} + \sigma_\vartheta$

Semi-Synthetic Data

Original Datasets

- We generate a semi-synthetic dataset with a known treatment effect parameter
- To generate credible confounding wil be based on the **labels** or **outcomes** of the corresponding supervised learning task
- We use the following datasets:

Modality	**Dataset**	**Target** $\tilde{X}$	**Control** X
Tabular	DIAMONDS Wickham (2016)	$\log(\text{Price})$	Carat, Cut, Color, Clarity, ...
Text	IMDB Maas et al. (2011)	Sentiment	Review Text
Image	CIFAR-10 Krizhevsky (2009)	Label	Image

Data Generating Process

We generate a semi-synthetic dataset according to the underlying PLR model

$$
\begin{aligned}
Y &= \theta_0 D + \tilde{g}_0(\tilde{X}) + \varepsilon, \\
D &= \tilde{m}_0(\tilde{X}) + \vartheta,
\end{aligned}
$$

where $\tilde{X} = (\tilde{X}_{\text{tab}}, \tilde{X}_{\text{txt}}, \tilde{X}_{\text{img}})$ with the following additive structure

$$
\begin{aligned}
\tilde{g}_0(\tilde{X}) &= \sum_{\text{mod} \in \{\text{tab,txt,img}\}} \tilde{g}_{\text{mod}}(\tilde{X}_{\text{mod}}) \\
\tilde{m}_0(\tilde{X}) &= \sum_{\text{mod} \in \{\text{tab,txt,img}\}} \tilde{m}_{\text{mod}}(\tilde{X}_{\text{mod}})
\end{aligned}
$$

and $\varepsilon, \vartheta \sim \mathcal{N}(0, 1)$.

Data Generating Process

- The effect on the outcome Y is generated via a standardized version of target variable to balance the confounding impact of all modalities:

$$\tilde{g}_{\text{mod}}(\tilde{X}_{\text{mod}}) = \frac{\tilde{X}_{\text{mod}} - \mathbb{E}[\tilde{X}_{\text{mod}}]}{\sigma_{\tilde{X}_{\text{mod}}}}, \quad \text{mod} \in \{\text{tab}, \text{txt}, \text{img}\}$$

- Further, to ensure a strong confounding, the impact on the treatment D is defined via:

$$\tilde{m}_{\text{mod}}(\tilde{X}_{\text{mod}}) = -\tilde{g}_{\text{mod}}(\tilde{X}_{\text{mod}}), \quad \text{mod} \in \{\text{tab}, \text{txt}, \text{img}\}$$

- The treatment effect is set to $\theta_0 = 0.5$ with $n = 50.000$ samples in the dataset
- Both $\tilde{g}_0(X)$ and $\tilde{m}_0(X)$ are rescaled to ensure a signal-to-noise ratio of 2 for Y and D (given unit variances of the error terms)

Challenges

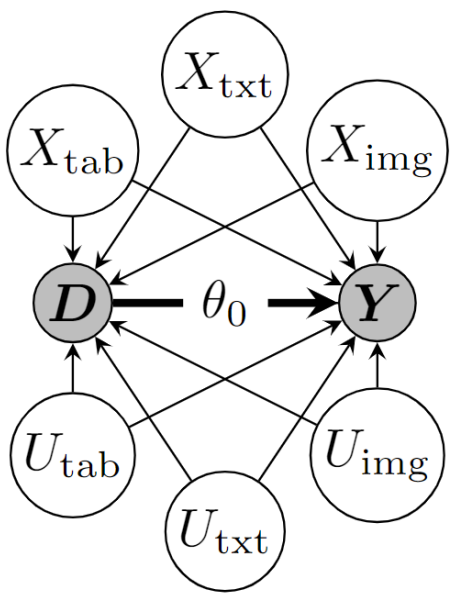
- Dependency on the target of the modality $\tilde{X}_{\text{mod}}$
  - Might not be fully explained by the corresponding features X_{mod}
  - E.g. price of diamonds can not be fully explained by the carat, cut, color, etc.
- Consequently, the estimate $\hat{\theta}$ might only be able to account for the part of confounding which can be explained by the input features:

$$\tilde{X}_{\text{mod}} = \mathbb{E}[\tilde{X}_{\text{mod}} | X_{\text{mod}}] + U_{\text{mod}},$$

  where U_{mod} can not be controlled for

Challenges

- Due to the negative sign and the additive structure, the confounding effect will ensure that higher outcomes Y occur with lower treatment values D, creating a **negative bias**
- The independence of all three original datasets and the additive negative confounding results in a negative bias even if we only control for a subset of confounding factors



DAG for the semi-synthetic dataset. The confounding via the features $X = (X_{\text{tab}}, X_{\text{txt}}, X_{\text{img}})$ can be adjusted for, whereas the unexplained/noise parts $U = (U_{\text{tab}}, U_{\text{txt}}, U_{\text{img}})$ are unobserved.

Bounds for the estimation of θ_0

- Since all modalities contribute a negative bias, the semi-synthetic dataset can be used as a benchmark with an oracle upper bound of an effect estimate of

$$\theta_0 = 0.5$$

- For the lower bound, one can not account for confounding at all and evaluate a basic ordinary least squares model with outcome Y on the treatment variable D (excluding all confounding variables)

$$\hat{\theta}_{\text{OLS}} = -0.4594$$

Bounds for the estimation of η_0

- To access the predictive performance of the nuisance models, we can rely on oracle predictions of

$$\begin{aligned}\tilde{m}_0(\tilde{X}) &:= \mathbb{E}[D|\tilde{X}] \\ \tilde{l}_0(\tilde{X}) &:= \mathbb{E}[Y|\tilde{X}] = \theta_0 \tilde{m}_0(\tilde{X}) + \tilde{g}_0(\tilde{X})\end{aligned}$$

- Evaluating the oracle predictions $\tilde{m}_0(\tilde{X})$ and $\tilde{l}_0(\tilde{X})$ results in the following upper bounds for the performance of the nuisance estimators

$$\begin{aligned}R^2(D, \tilde{m}_0(\tilde{X})) &= 0.6713 \\ R^2(Y, \tilde{l}_0(\tilde{X})) &= 0.5845\end{aligned}$$

  on the whole dataset of $n = 50.000$ observations

Bounds for the estimation of η_0

- In order to compare the predictive performance of different models, a relative r^2-score with respect to the upper bound is defined as

$$0 \leq r^2(D, \hat{m}) := \frac{R^2(D, \hat{m}(X))}{R^2(D, \tilde{m}_0(\tilde{X}))} \leq 1$$

$$0 \leq r^2(Y, \hat{l}) := \frac{R^2(Y, \hat{l}(X))}{R^2(Y, \tilde{l}_0(\tilde{X}))} \leq 1$$

Models

- **Baseline Model**
  - Standard DML approach, relying only on tabular data X_{tab}
  - Based on the LightGBM package for estimation of the nuisance elements
- **Deep Model**
  - Uses the out-of-sample predictions of $\hat{m}(X)$ and $\hat{l}(X)$ generated from the model
  - Based on the RoBERTa Model for text modality, a VIT Model for images and a SAINT model for the tabular data
- **Embedding Model**
  - Generated embedding H_E is used together with the tabular features X_{tab} as input for LightGBM
  - Based on the RoBERTa Model for text modality, a VIT Model for images and a SAINT model for the tabular data

Simulation Results

Results of Simulation Study. Reported: mean ± sd. over five random train-test splits

	Baseline	**Embedding**	**Deep**
$r^2(Y, \hat{l}_0)$	0.31 ± 0.01	0.87 ± 0.02	$\mathbf{0.90 \pm 0.01}$
$r^2(D, \hat{m}_0)$	0.31 ± 0.01	0.87 ± 0.02	$\mathbf{0.90 \pm 0.01}$
$\hat{\theta}$	-0.32 ± 0.01	$\mathbf{0.28 \pm 0.01}$	0.27 ± 0.01

Higher = better (best in bold)

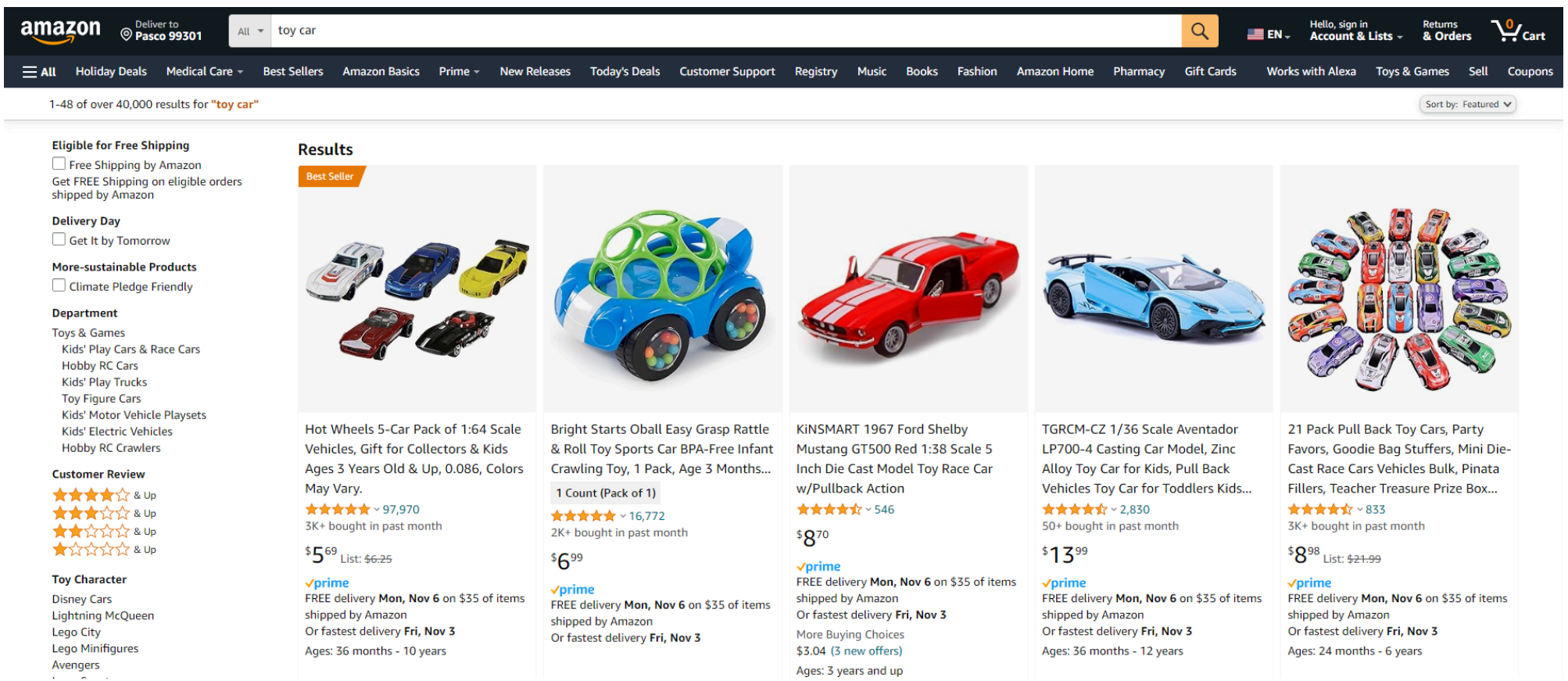
Simulation Results - Performance of $\hat{\theta}$

Boxplots of $\hat{\theta}$. The Embedding Model and Deep Model have similar estimates. This indicates a stable and information-rich embedding H_E, which provides a high explanatory contribution independent of the subsequent ML method for predicting Y and D. θ_0 represents the upper bound.

Application

Application: Estimation of Price Elasticity

- Understanding price elasticity is crucial for economic analysis and business decisions
  - It influences strategies, pricing, and market dynamics
- Specifically in online marketplaces, unstructured data is available



Amazon Toys Dataset

- Public Data from the Amazon sales platform from category toys and games for subcategories vehicles, cars and trucks.
- Using the Sales Rank as a proxy for quantities, as shown in Bajari et al. (2023).

Variable	Description
Sales Rank	Sales rank as weighted mean for the last 30 days
Price	Price as weighted mean for the last 30 days
Text	Combination of Title, Category, Description, etc.
Image	Image of the product

Amazon Toys Dataset

Continous Variables

- Reviews: Rating
- Reviews: Review Count
- New Offer Count: Current
- Count of retrieved live offers: New, FBA
- Count of retrieved live offers: New, FBM

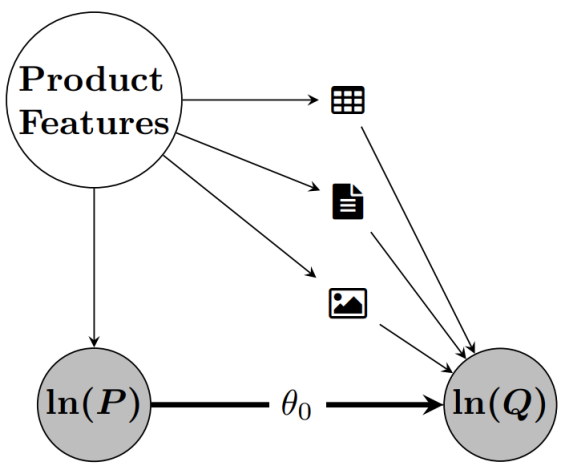
Categorical Variables

- Lightning Deals: Upcoming Deal
- Buy Box: Is FBA

Model

- Run a simple log-log regression model
- Images and Text block backdoor path from price to demand (or sales rank)

$$\ln(Q) = \theta_0 \ln(P) + g_0(X) + \epsilon$$



$\Rightarrow$ The causal parameter θ_0 can be interpreted as price elasticity of demand!

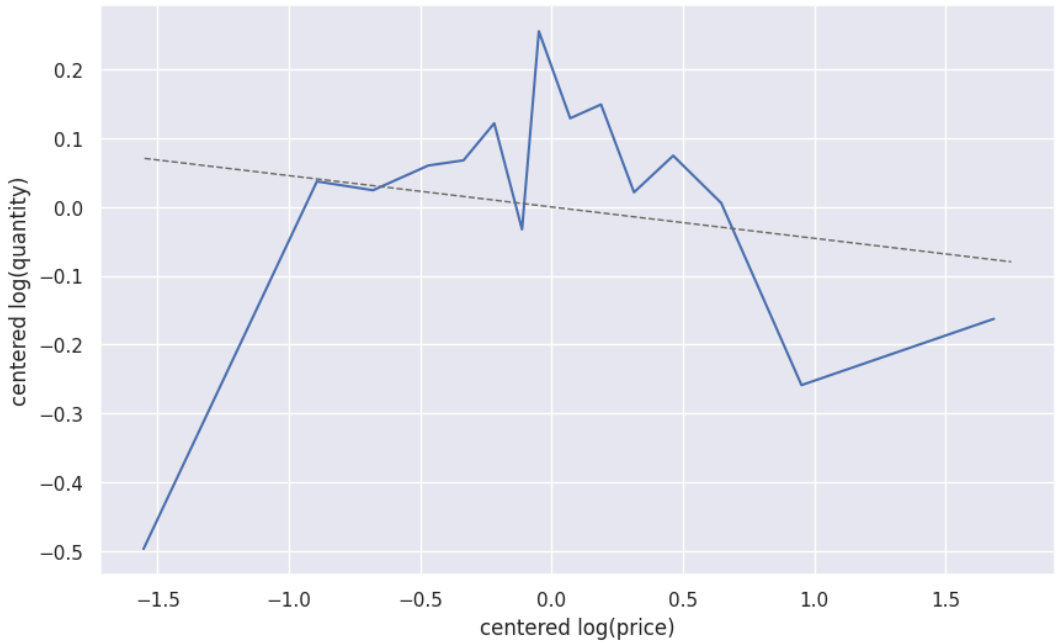
Baseline OLS Model

- Baseline estimate with tabular covariates X_{tab}
- OLS:

$$\ln(Q) = \theta_0 \ln(P) + \beta^T X_{\text{tab}} + \epsilon$$

- $R^2 = 0.330$

2.5 %	$\hat{\theta}$	97.5 %
-0.072	-0.046	-0.019

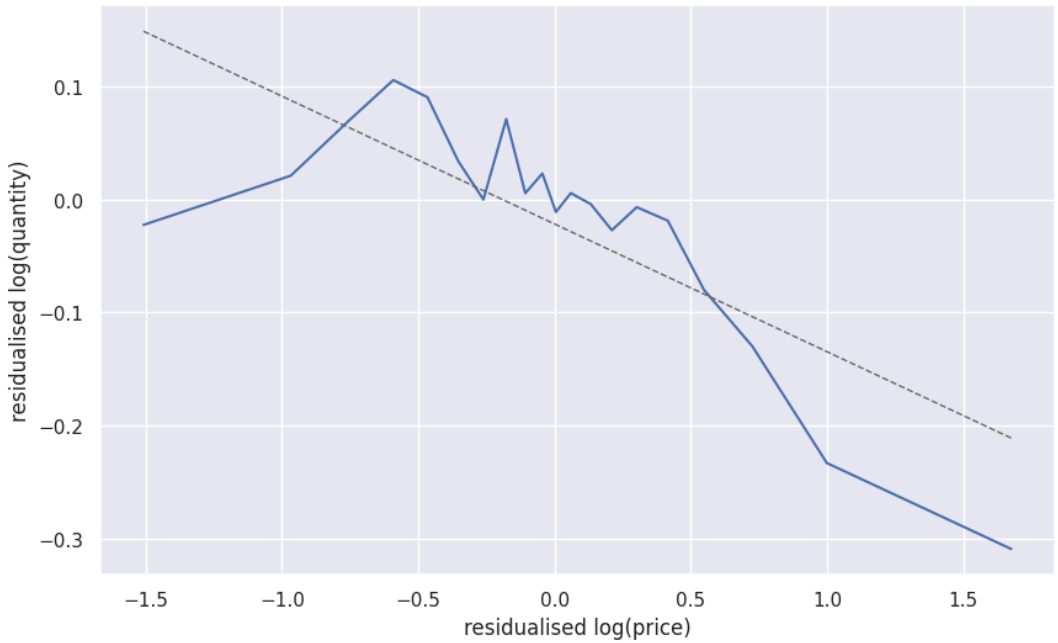


Centered logarithmic negative sales rank over centered logarithmic price as binned plot

Baseline DML Model

- Estimate with tabular data as covariates X_{tab} in `DoubleMLPLR` model with `RandomForest` regressors
  - $l_0(X_{\text{tab}}) := \mathbb{E}[\ln(Q)|X_{\text{tab}}]$
    - $R^2_{l_0} = 0.5986$
  - $m_0(X_{\text{tab}}) := \mathbb{E}[\ln(P)|X_{\text{tab}}]$
    - $R^2_{m_0} = 0.1884$

2.5 %	θ	**97.5 %**
-0.132	-0.1098	-0.080



Residualised logarithmic negative sales rank over residualised logarithmic price as binned plot

Deep Learning Models

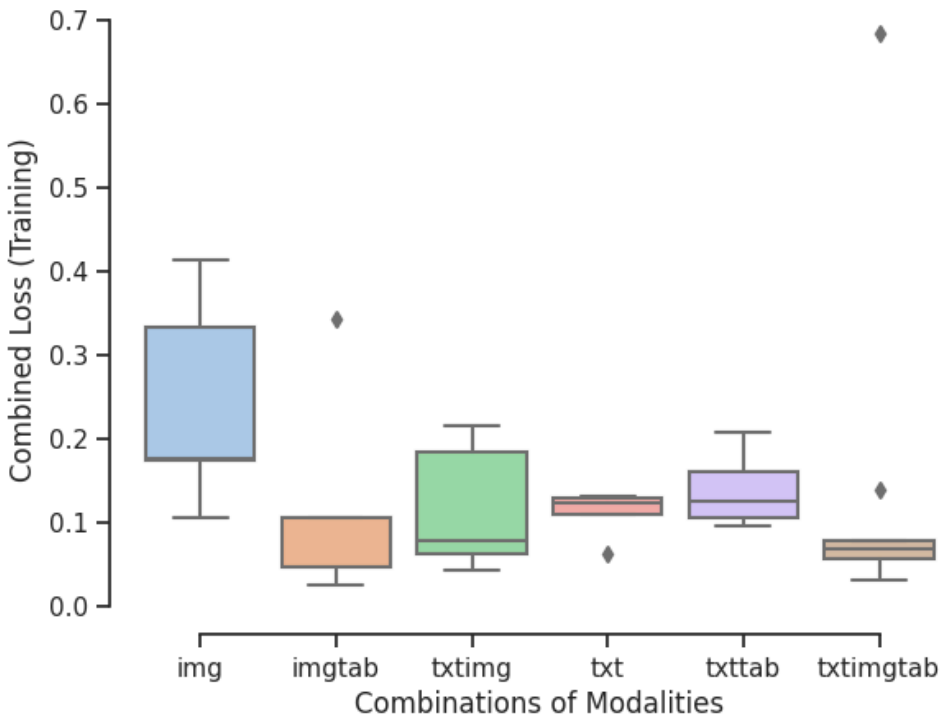
- In this study, a variety of input data combinations were used to train `DoubleMLPLRDeep` models [1]:

Key	Used Confounders
`img`	$X = (X_{img})$
`imgtab`	$X = (X_{img}, X_{tab})$
`txtimg`	$X = (X_{txt}, X_{img})$
`txt`	$X = (X_{txt})$
`txttab`	$X = (X_{txt}, X_{tab})$
`txtimgtab`	$X = (X_{txt}, X_{img}, X_{tab})$

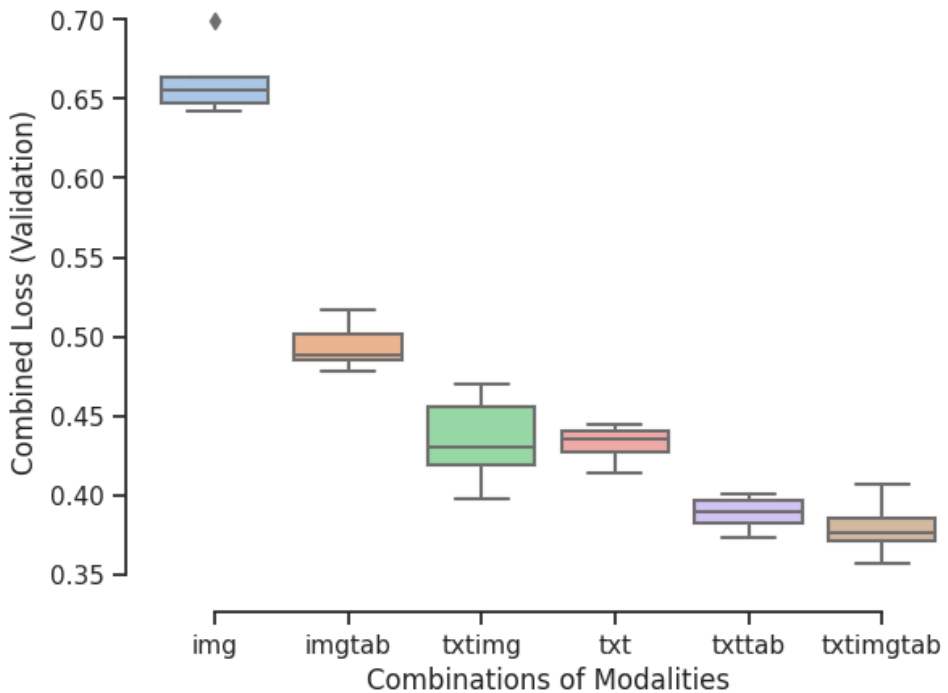
$\Rightarrow$ The aim is to emphasize the benefits of utilizing unstructured data.

Deep Learning Models: RMSE-Scores

- Combined RMSE Score (Training)

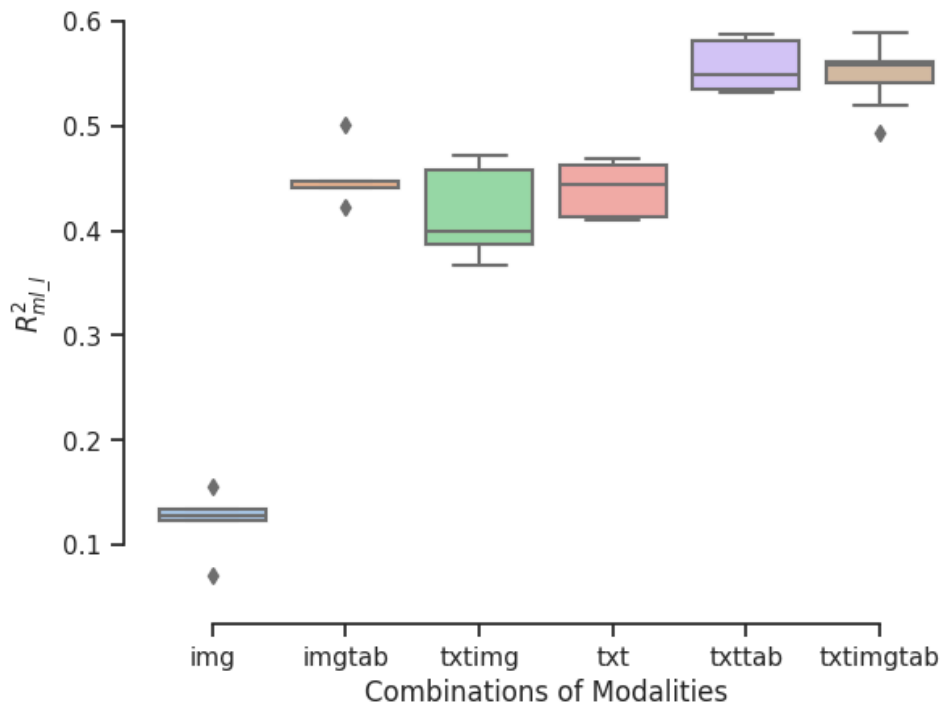


- Combined RMSE Score (Validation)

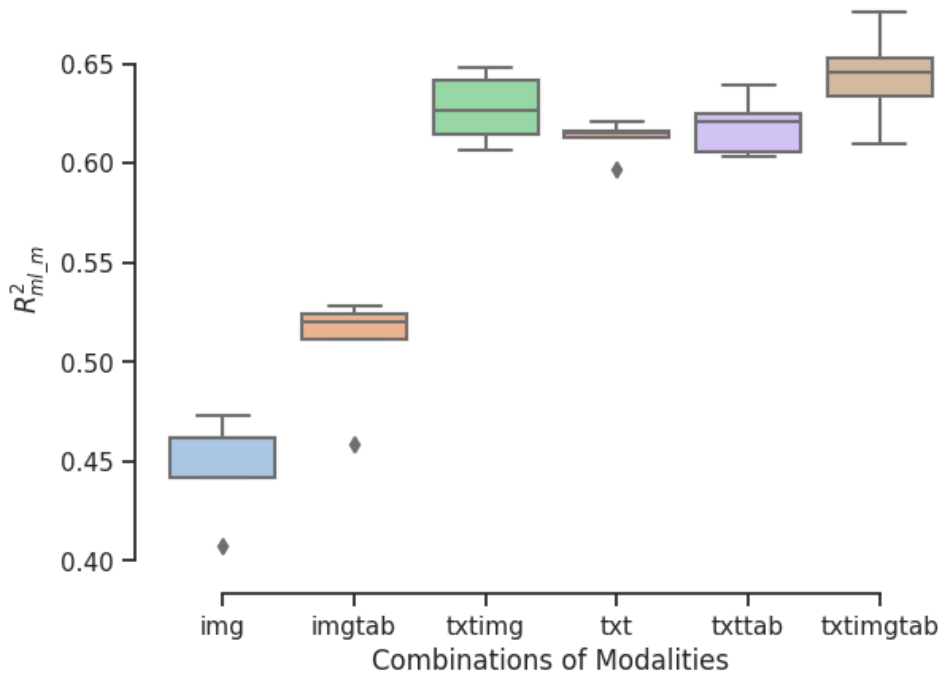


Deep Learning Models: R^2-Scores

- R^2 of log(Quantity) on Validation Set

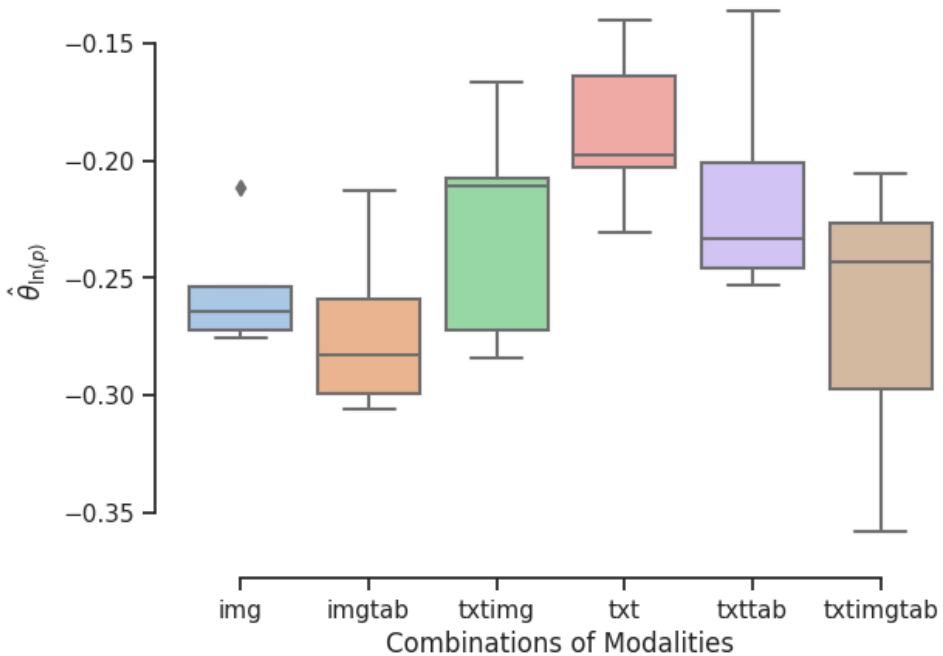


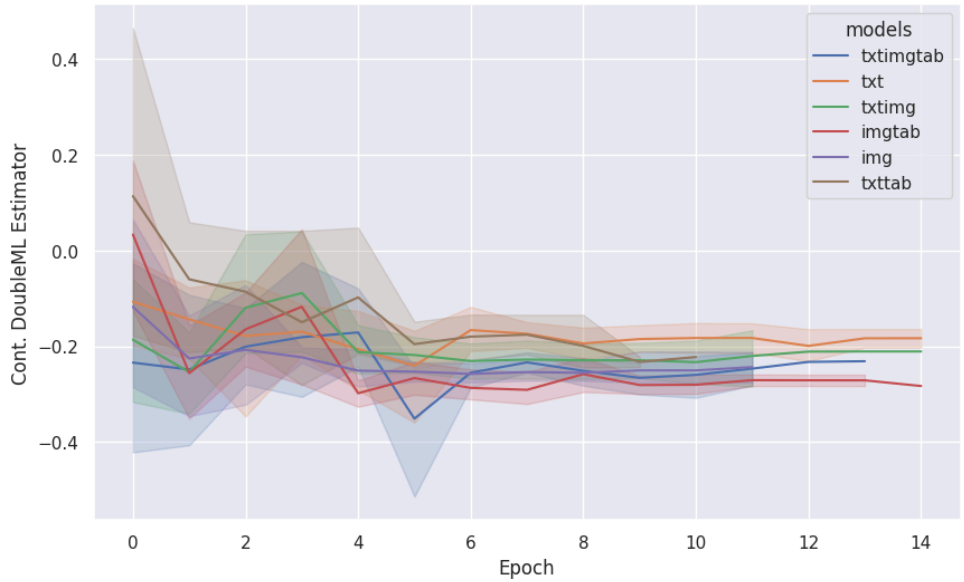
- R^2 of log(Price) on Validation Set



(First) Results

- Using different combinations of tabular, image and text data we obtain the following estimates



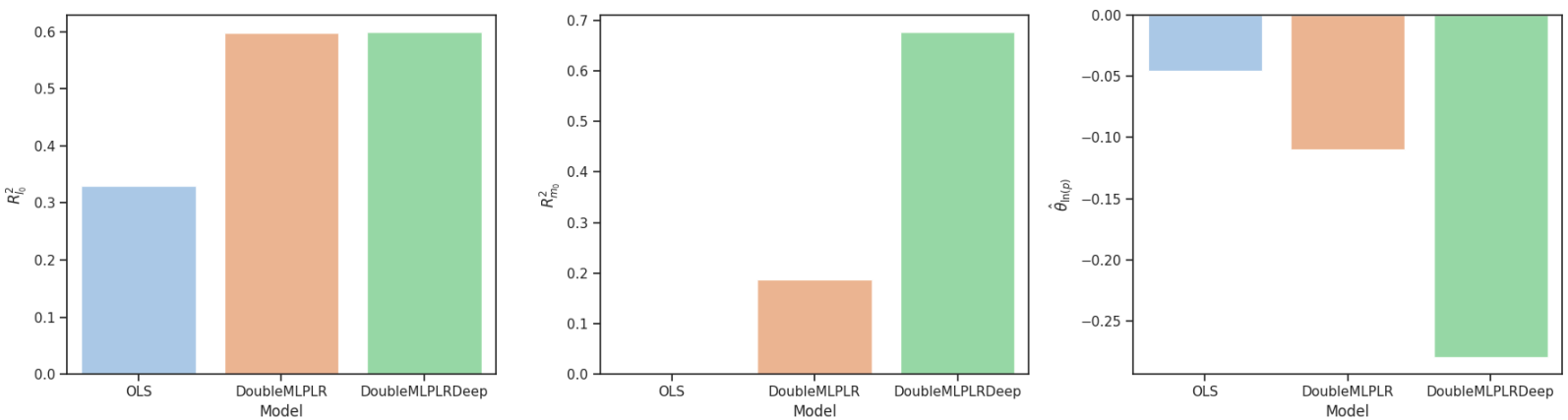


Continous Estimates from `DoubleMLPLR` on Validation Set during the Training Process

Comparison to Baseline Estimates

Model	Covariates	$R^2_{l_0}$	$R^2_{m_0}$	$\hat{\theta}_0$
`OLS`	$X = X_{tab}$	0.3300	-	-0.0455
`DoubleMLPLR`	$X = X_{tab}$	0.5986	0.1884	-0.1098
`DoubleMLPLRDeep`	$X = (X_{tab}, X_{txt}, X_{img})$	0.5990	0.6765	-0.2794

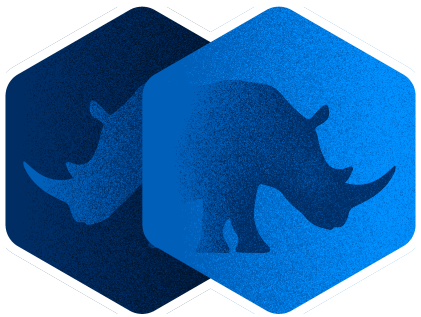
1



More on Double Machine Learning

Papers & Book

- CausalML Book
- R package- with a nontechnical introduction to DML: Bach et al. (2021)
- Python package: Bach et al. (2022)



Software implementation:

- https://github.com/DoubleML/doubleml-for-py
- https://github.com/DoubleML/doubleml-for-r
- Docu and examples: docs.doubleml.org

Thank you!

GitHub Repository

If you like our package, you can leave us a ⭐ on GitHub

Contact

In case you have questions or comments, feel free to contact me

jan.teichertkluge@uni-hamburg.de

References

References

Bach, Philipp, Victor Chernozhukov, Malte S Kurz, and Martin Spindler. 2021. "DoubleML – An Object-Oriented Implementation of Double Machine Learning in R." https://arxiv.org/abs/2103.09603.

———. 2022. "DoubleML-an Object-Oriented Implementation of Double Machine Learning in Python." *Journal of Machine Learning Research* 23: 53–51.

Bajari, Patrick, Zhihao Cen, Victor Chernozhukov, Manoj Manukonda, Suhas Vijaykumar, Jin Wang, Ramon Huerta, et al. 2023. "Hedonic Prices and Quality Adjusted Price Indices Powered by AI." https://arxiv.org/abs/2305.00044.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21 (1): C1–68. https://onlinelibrary.wiley.com/doi/abs/10.1111/ectj.12097.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems* 30: 3146–54.

Krizhevsky, Alex. 2009. "Learning Multiple Layers of Features from Tiny Images."

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. "Learning Word Vectors for Sentiment Analysis." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–50. Portland, Oregon, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/P11-1015.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Appendix

Amazon Toys Dataset

Data Examples

Kinsmart Set of 4 McLaren 720s Toy | ...

Pixar Cars Mack Uncle Lightning...

Neyman Orthogonality

The naive approach minimizes the following MSE

$$min_\theta \mathbb{E}[(Y - D\theta_0 - g_0(X))^2]$$

This implies the following moment equation

$$\mathbb{E}[\underbrace{(Y - D\theta_0 - g_0(X))D}_{=:\psi(W,\theta_0,\eta)}] = 0$$

Whereas for the partialling-out approach minimizes

$$min_\theta \mathbb{E}\left[\left(Y - \mathbb{E}[Y|X] - (D - \mathbb{E}[D|X])\theta\right)^2\right]$$

which implies

$$\mathbb{E}\Big[\underbrace{\left(Y - \mathbb{E}[Y|X] - (D - \mathbb{E}[D|X])\theta\right)(D - \mathbb{E}[D|X])}_{=:\psi(W,\theta_0,\eta)}\Big] = 0$$

Neyman Orthogonality

Naive approach

$$\psi(W, \theta_0, \eta) = (Y - D\theta_0 - g_0(X))D$$

Regression adjustment score

$$\begin{aligned} \eta &= g(X), \\ \eta_0 &= g_0(X). \end{aligned}$$

FWL partialling out

$$\begin{aligned} \psi(W, \theta_0, \eta_0) = &\Big((Y - E[Y|X]) \\ &- (D - E[D|X])\theta_0\Big) \\ &(D - E[D|X]) \end{aligned}$$

Neyman-orthogonal score (Frisch-Waugh-Lovell)

$$\begin{aligned} \eta &= (\ell(X), m(X)), \\ \eta_0 &= (\ell_0(X), m_0(X)), \\ &= (\mathbb{E}[Y \mid X], \mathbb{E}[D \mid X]). \end{aligned}$$

DML Key Ingredients

1. Neyman Orthogonality

- Inference is based on a moment condition that satisfies the **Neyman orthogonality condition** $\psi(W; \theta, \eta)$

$$E[\psi(W; \theta_0, \eta_0)] = 0,$$

- where $W := (Y, D, X, Z)$ and with θ_0 being the unique solution that obeys the **Neyman orthogonality condition**

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

- ∂_η denotes the pathwise (Gateaux) derivative operator

DML Key Ingredients

1. Neyman Orthogonality

- **Neyman orthogonality** ensures that the **moment condition** identifying θ_0 is **insensitive to small pertubations** of the nuisance function η around η_0
- Using a Neyman-orthogonal score **eliminates the first order biases** arising from the replacement of η_0 with a ML estimator $\hat{\eta}_0$
- PLR example: Partialling-out score function

$$\psi(\cdot) = (Y - E[Y|X] - \theta(D - E[D|X]))(D - E[D|X])$$

DML Key Ingredients

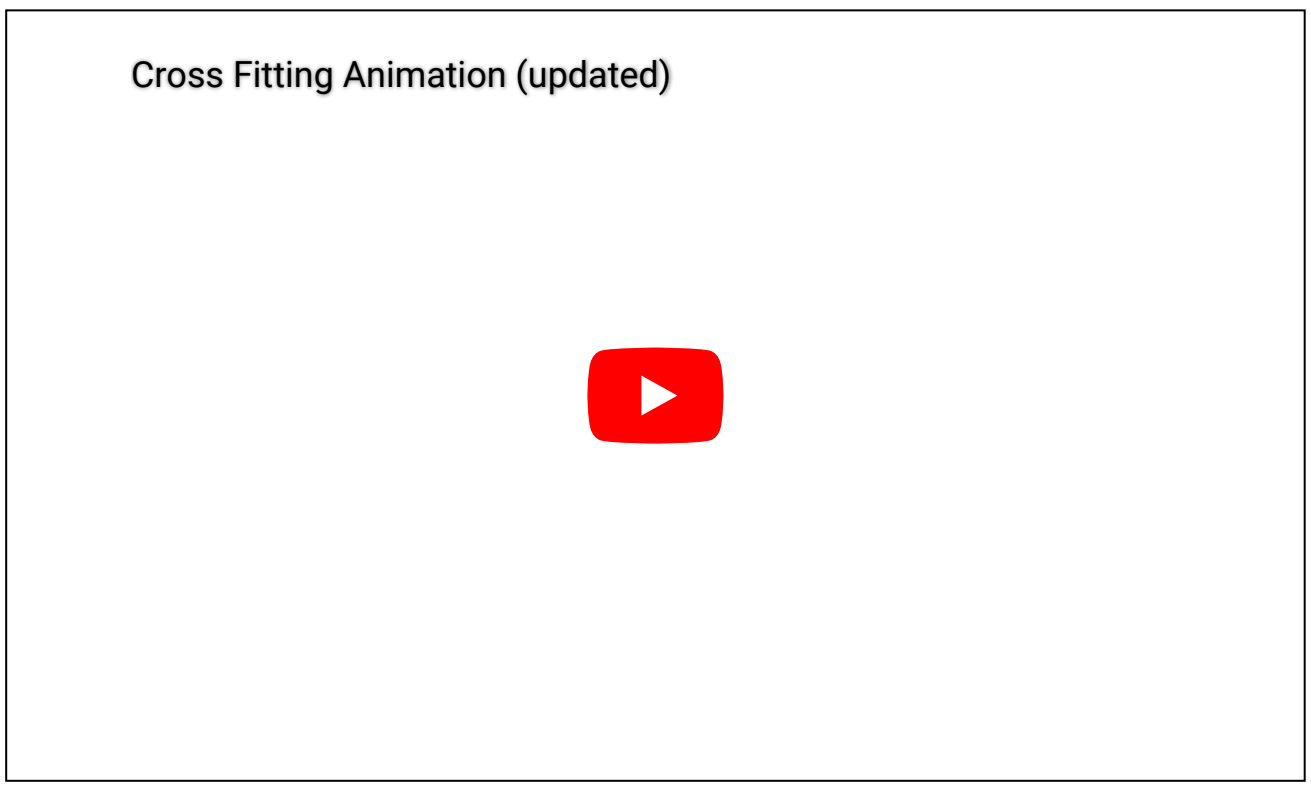
2. High-Quality Machine Learning Estimators

- The nuisance parameters are estimated with high-quality (fast-enough converging) machine learning methods.
- Different structural assumptions on η_0 lead to the use of different machine-learning tools for estimating η_0 Chernozhukov et al. (2018) (Section 3)
- Rate requirements depend on the causal model and orthogonal score, e.g. (see Chernozhukov et al. (2018)),
  - PLR, partialling out:
    $\|\hat{m}_0 - m_0\|_{P,2} \times \left(\|\hat{m}_0 - m_0\|_{P,2} + \|\hat{\ell}_0 - \ell_0\|_{P,2}\right) \leq \delta_N N^{-1/2}$
  - IRM/DR score, ATE: $\|\hat{m}_0 - m_0\|_{P,2} \times \|\hat{\ell}_0 - \ell_0\|_{P,2} \leq \delta_N N^{-1/2}$

DML Key Ingredients

3. Sample Splitting

- To avoid the biases arising from overfitting, a form of **sample splitting** is used at the stage of producing the estimator of the main parameter θ_0.
- Efficiency gains by using cross-fitting (swapping roles of samples for train / hold-out)



DML Key Ingredients

Main result in Chernozhukov et al. (2018)

> There exist regularity conditions, such that the DML estimator $\tilde{\theta}_0$ concentrates in a $1/\sqrt{N}$-neighborhood of θ_0 and the sampling error is approximately
>
> $$\sqrt{N}(\tilde{\theta}_0 - \theta_0) \sim N(0, \sigma^2),$$
>
> with
>
> $$\sigma^2 := J_0^{-2}\mathbb{E}(\psi^2(W; \theta_0, \eta_0)),$$
> $$J_0 = \mathbb{E}(\psi_a(W; \eta_0)).$$

- See this example based on Chernozhukov et al. (2018)