Estimating causal effects using neural autoregressive density estimators and do-calculus

Sergio Garrido, Stanislav Borysov, Jeppe Rich, Francisco Pereira

Technical University of Denmark DTU | MLSM Machine Learning for Smart Mobility group http://mlsm.man.dtu.dk

Paper under review at the Journal of causal inference

What is causality?

causality noun

Save Word

cau·sal·i·ty | \ kȯ-ˈza-lə-tē \

plural **causalities**

Definition of *causality*

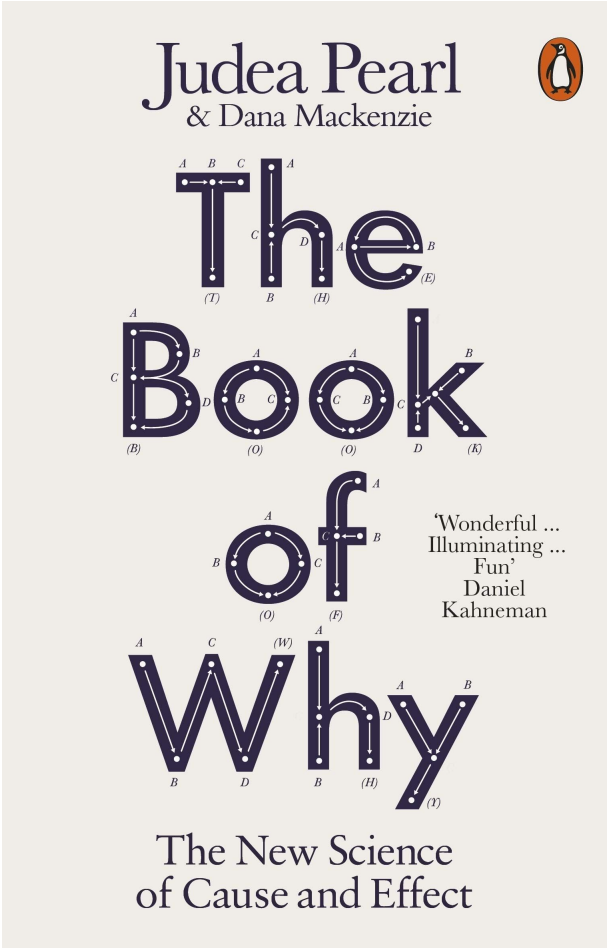
1 : a causal quality or agency

2 : the relation between a cause and its effect or between regularly correlated events or phenomena

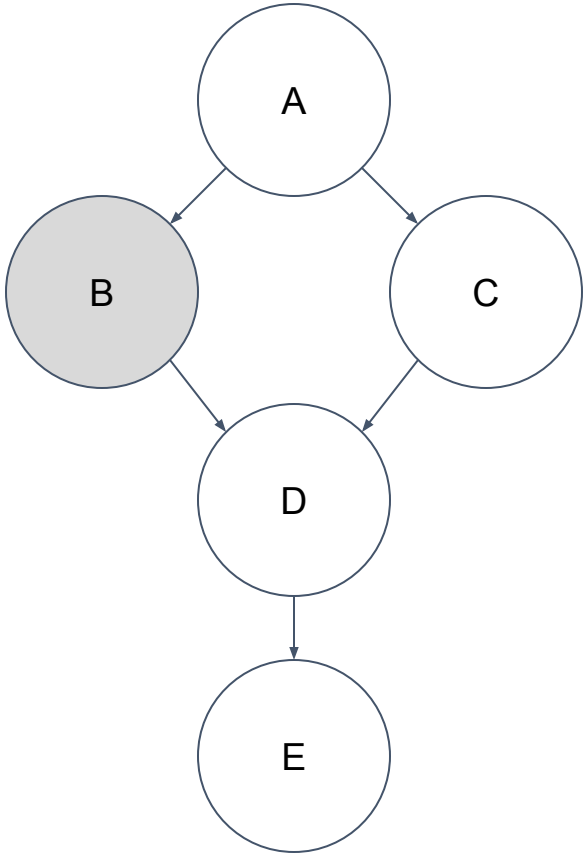
There are different definitions and ways to tackle the causality problem. Rubin's "potential outcomes" and Pearl's "do-calculus" being probably the most famous.

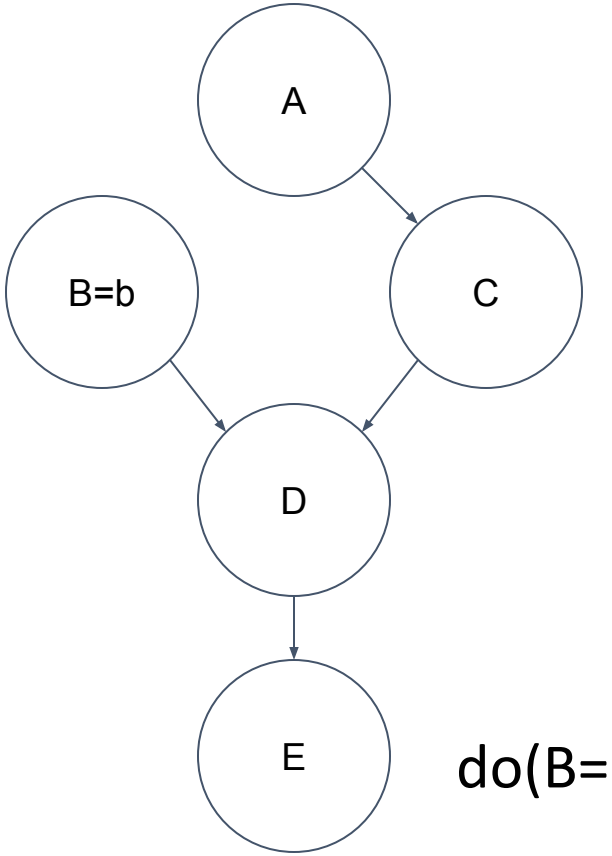
Pearl's causality

- Based on graphical models and structural equation models (SEM).
- The basic idea is that there is a "true" causal graph. (Markov equivalence class?).
- Conditioning (not causal, observation association, prediction), Intervening (doing) and Counterfact (imagining what would have been if).

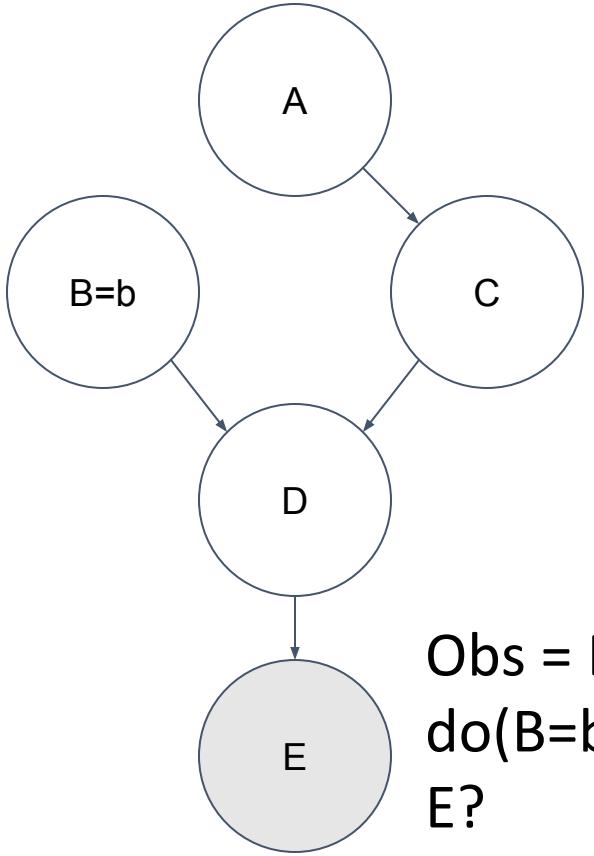


The three operations/queries





do(B=b)



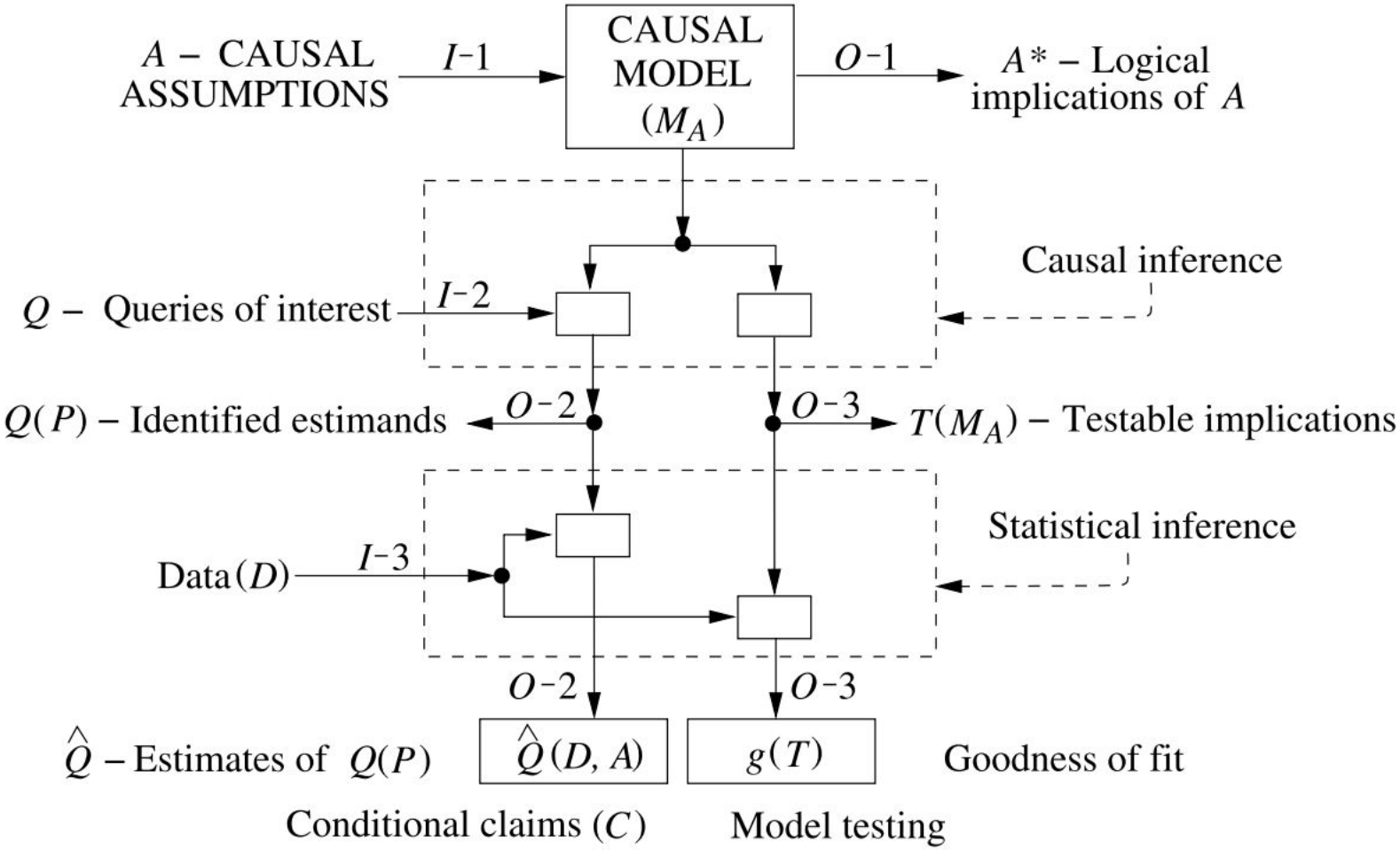
Obs = E, do(B=b), E?

Experimental vs. Observational

Independent Causal Mechanisms

$$p(x) = \prod_i p(x_i \mid PA(x_i))$$

Causality as an inference machine



DTU MLSM Machine Learning for Smart Mobility

Problem... and solution

Pearl (2009) does not offer a way to estimate the ICM + linear assumption does not hold!

We can use ANY flexible approximator to estimate the ICM.

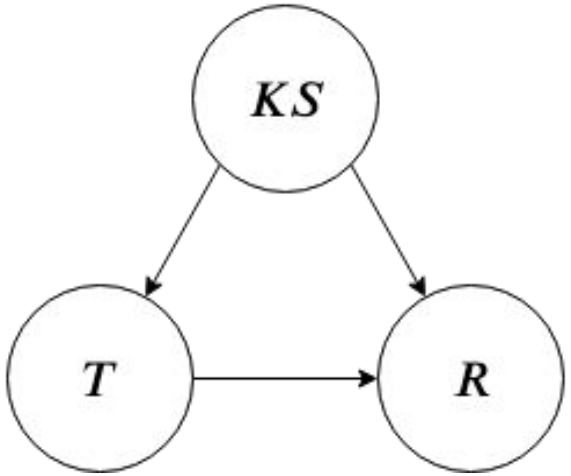
Assumptions

Any inference (be it causal or not) requires assumptions. Causal inference requires causal assumptions which are “another level” of assumptions. For this paper we assume:

- We know the **causal graph.**
- We know the **distribution** that any variable on the graph follows (this is relaxed later).
- The causal graph is **acyclical** (although I have the feeling that this need not be the case).

Questions?

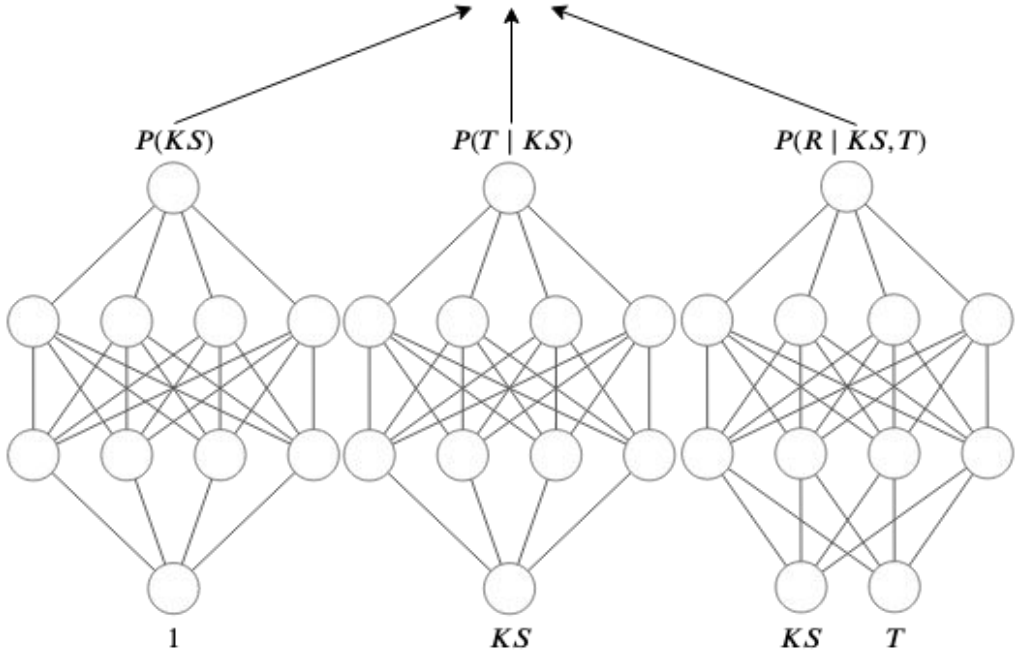
Neural Autoregressive Density Estimator



$$\prod_j^J P(X_j \mid PA(X_j))$$

$$P(KS,T,R) = P(KS \mid \mathrm{PA}(KS))P(T \mid \mathrm{PA}(T))P(R \mid \mathrm{PA}(R))$$

$$= P(KS)P(T \mid KS)P(R \mid KS,T)$$



$$p(x) = \prod_i p(x_i \mid PA(x_i))$$

DTU MLSM
Machine Learning for Smart Mobility

There is a whole family of these models!

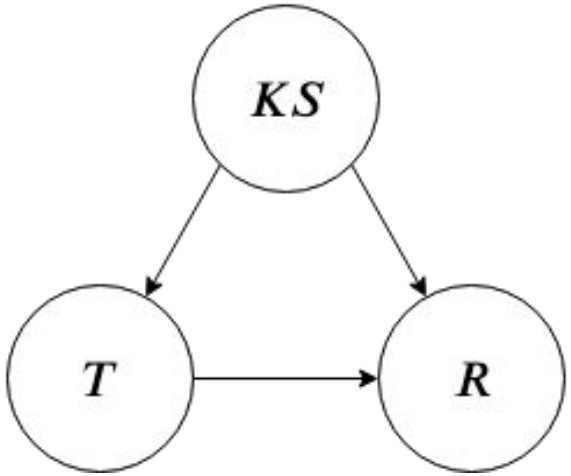
- Bengio, Y. and Bengio, S. (1999). Modeling high-dimensional discrete data with multi-layer neural networks.
- Larochelle, H. and Murray, I. (2011). The neural autoregressive distribution estimator.
- Uria, B., Murray, I., and Larochelle, H. (2013). Rnade: The real-valued neural autoregressive density-estimator.
- Germain, M., Gregor, K., Murray, I., and Larochelle, H. (2015). Made: Masked autoencoder for distribution estimation.
- ...

Why people don’t use them that much?

- Slooooooooow…

Back to causality!

if: observational data;
then: do-calculus;



	Treatment type	
Size	A	B
Small	**93%** (81/87)	87% (234/270)
Large	**73%** (192/263)	69% (55/80)
Total	78% (273/350)	**83%** (289/350)

$$P(R=1 \mid do(T=A)) = \sum_{ks} P(R=1 \mid T=A, KS=ks) P(KS=ks)$$

Questions?

Simulation studies

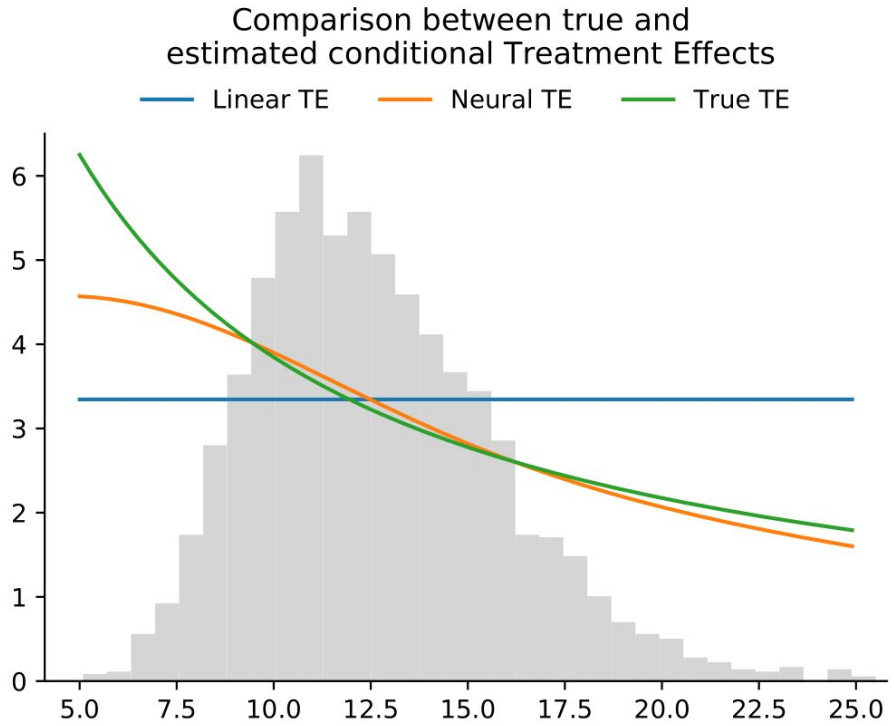
1. All binary variables simulated from the “classic” problem.
2. Continuous (R)ecovery variable .
3. Continuous (R)ecovery variable and (K)idney (S)tone.
4. Non linear effects.
5. A different problem setting (The front-door adjustment).

Do we really need to know the distribution of each variable?

Continuous confounding ATE (eq. (6))		
	Gamma confounding	Log-normal confounding
1 sample	3.77	4.15
5 samples	3.77	4.15
25 samples	3.77	4.15
50 samples	3.77	4.15
Total analytical ATE		4

Non-linear case

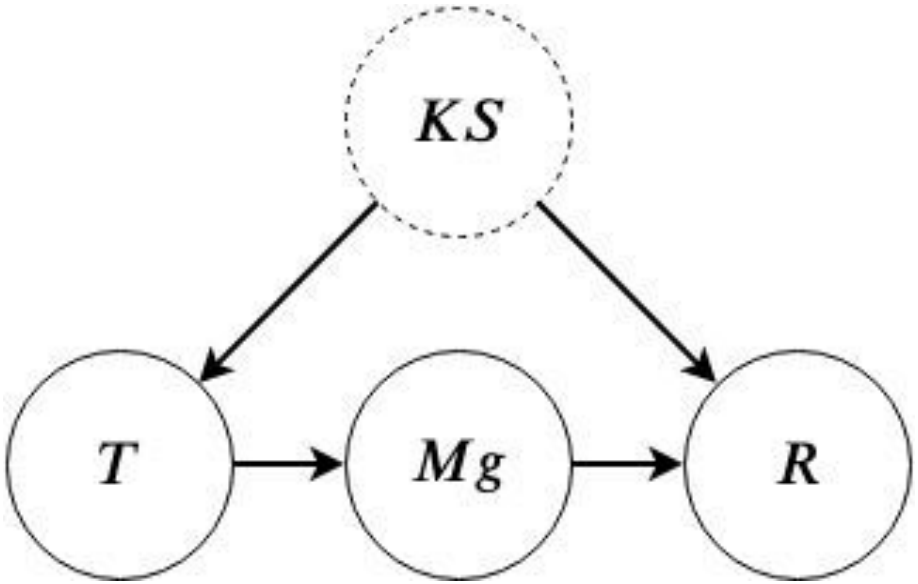
1: **for** each individual i **do**
2: Draw $KS \sim \text{Log-Normal}(2.5, 0.25)$
3: Draw $P(T = A) = \frac{1}{1+exp((KS-\mu)/10)}$
4: Draw $R \sim N\left(\frac{50T}{KS+3}, 1\right)$
5: **end for**



What is the difference between this approach and any other causal NN?

- Shi, C., Blei, D., & Veitch, V. (2019). Adapting neural networks for the estimation of treatment effects.
- Johansson, F., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference.
- Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep IV: A flexible approach for counterfactual prediction
- ...

A (slightly) more challenging problem

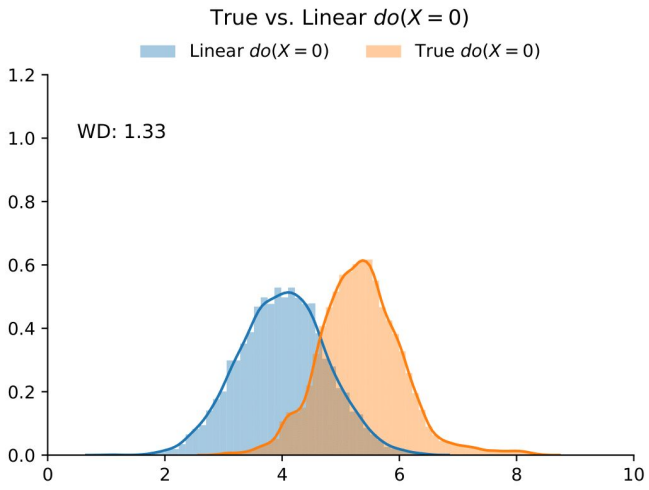


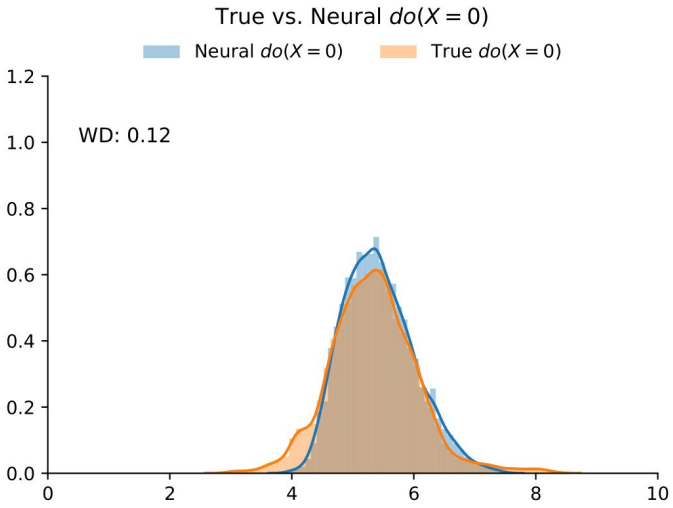
$$P(R \mid do(T = \hat{t})) = \int_{mg} P(Mg = mg \mid T = \hat{t}) \int_{t'} P(R \mid Mg = mg, T = t')P(T = t').$$

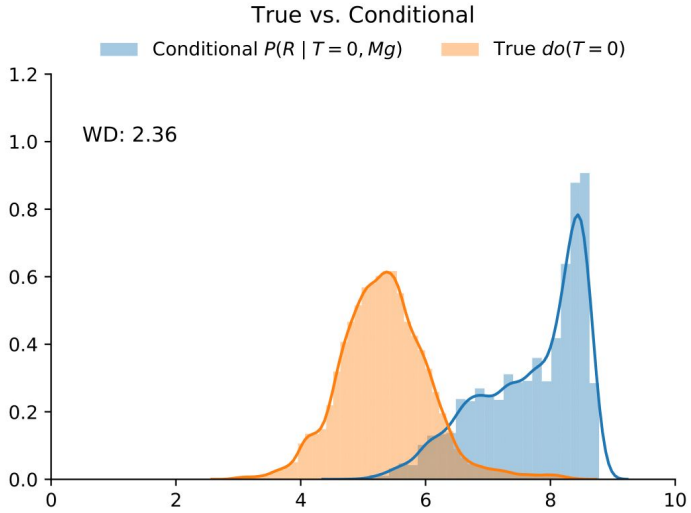
Simulation study 5

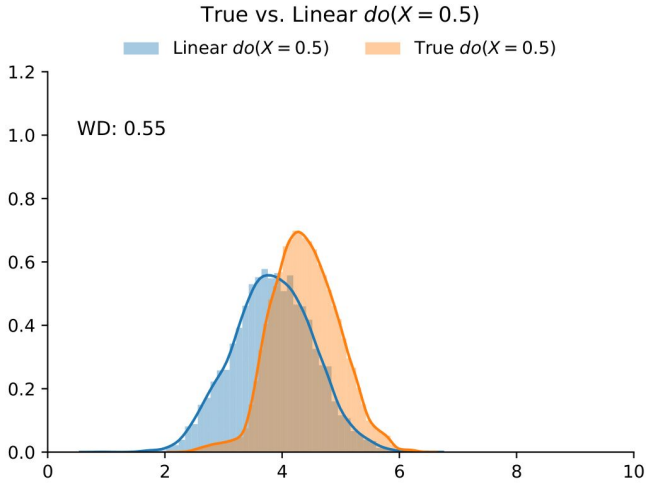
for each individual i **do**
    Draw $KS \sim N(0, 1)$
    Draw $T \sim N(\sin(KS), 0.1)$
    Draw $Mg \sim N(1 + T^2, 0.1)$
    Draw $R \sim N\left(\sin(KS^2) + \frac{5}{Mg}, 0.1\right)$
end for

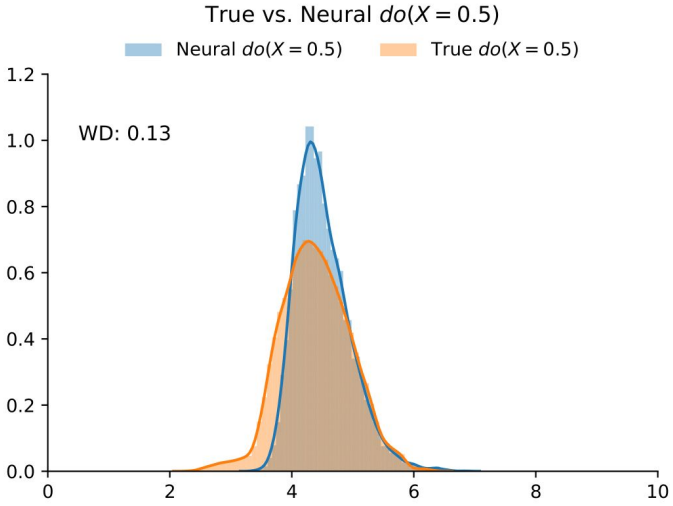
Results

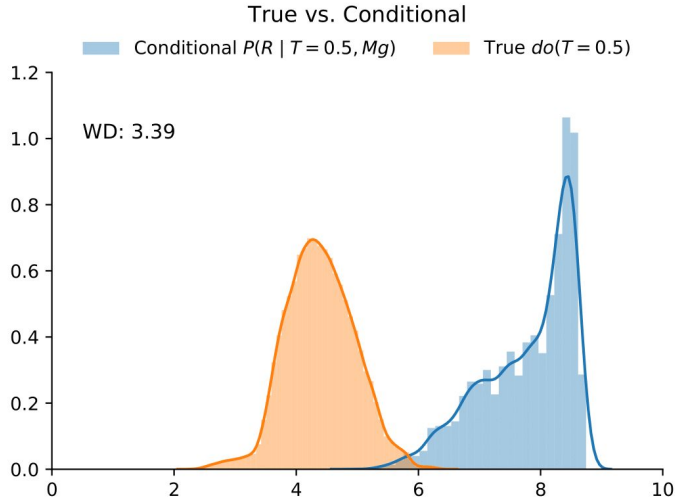

True vs. Linear $do(X = 0)$
Linear $do(X = 0)$
True $do(X = 0)$
WD: 1.33


True vs. Neural $do(X = 0)$
Neural $do(X = 0)$
True $do(X = 0)$
WD: 0.12


True vs. Conditional
Conditional $P(R \mid T = 0, Mg)$
True $do(T = 0)$
WD: 2.36


True vs. Linear $do(X = 0.5)$
Linear $do(X = 0.5)$
True $do(X = 0.5)$
WD: 0.55


True vs. Neural $do(X = 0.5)$
Neural $do(X = 0.5)$
True $do(X = 0.5)$
WD: 0.13


True vs. Conditional
Conditional $P(R \mid T = 0.5, Mg)$
True $do(T = 0.5)$
WD: 3.39

Next steps

- The “logical” next step is to extend this model to work with structural causal models. Not anymore…
- Explore the data/computation efficiency of the model as opposed to a purely conditional model.
- Extend to other types of data. The counterfactual people already worked with images!

Questions?