Adjustment Sets and Approaches - and limitations / critiques

Wouter van Amsterdam

2024-08-06

Table of contents

Adjustment sets and approaches

How to find adjustment sets?

- adjustment sets:
  - the back-door criterion states that any set Z that blocks all backdoor paths from X to Y is a sufficient adjustment set for causal effect estimation of $P(Y|do(X))$ using the backdoor formula.
  - how do we find these sufficient sets?
  - what if there are multiple?
- adjustment: how to do this?
  - stratification
  - what is regression adjustment?
  - T-learner vs S-learner

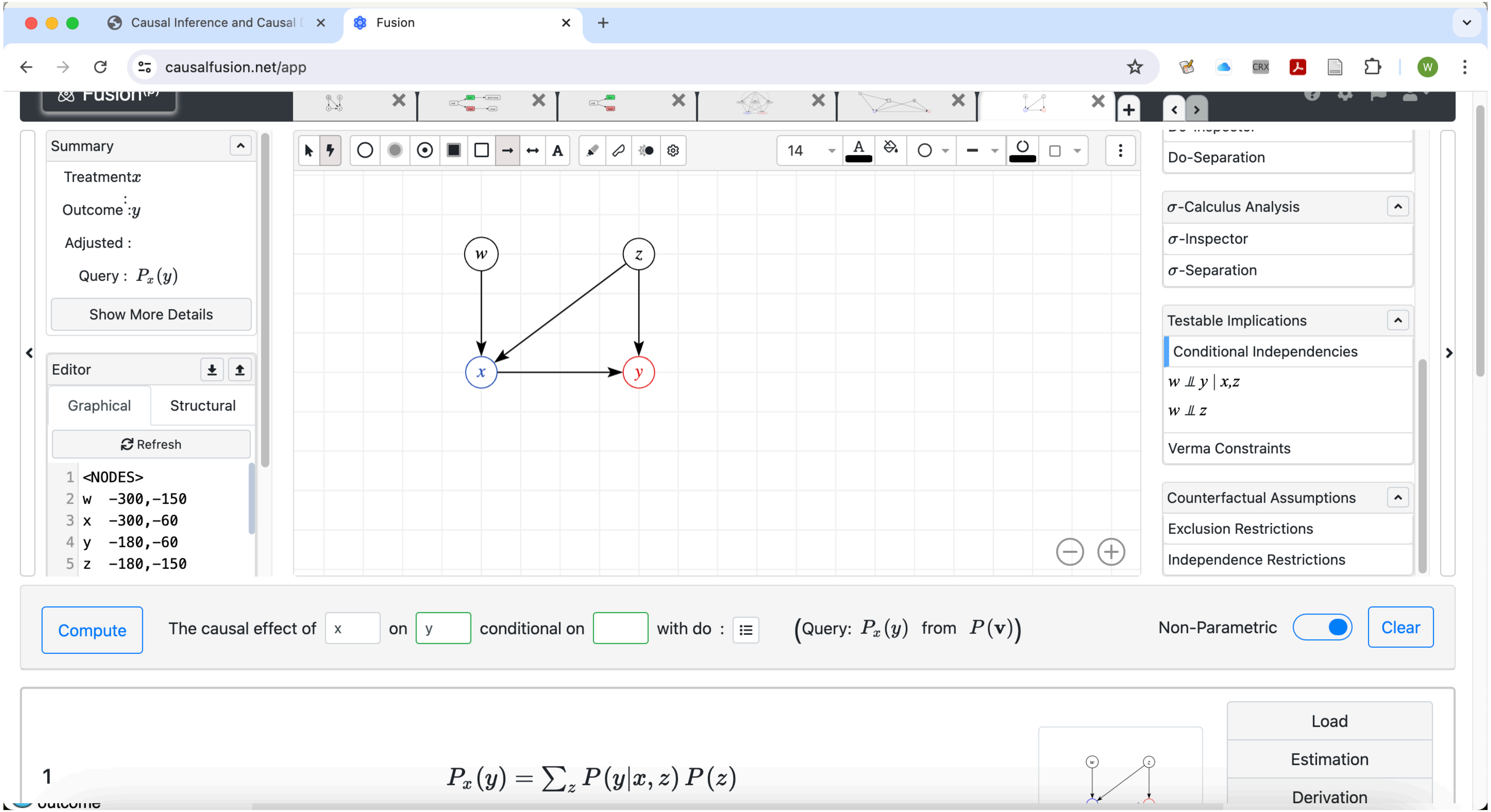
Valid adjustment sets

- in general:
  - PA_{T} (the direct parents of treatment $\mathrm{T}: \mathrm{Z}_1$) are a valid adjustment set
  - PA_{Y} (the direct parents of outcome $\mathrm{Y}: \mathrm{Z}_2$) are a valid adjustment set
- in this case:
  - W is a valid adjustment set

$$
\begin{tikzcd}
 & W \arrow[dl] \arrow[dr] & \\
Z_1 \arrow[d] & & Z_2 \arrow[d] \\
T \arrow[rr] & & Y
\end{tikzcd}
$$

Valid adjustment sets: picking one

- websites like dagitty.net and causalfusion.net provide user-friendly interfaces for creating and exporting DAGs, in addition:
  - valid adjustment sets (if they exist)
  - testable conditional indepdencies



causalfusion.net

dagitty.net

How to do adjustment

What not to do

1. do univariable pre-screening against outcome (and / or treatment)

- this should maybe never be done
- especially not in the context of causal inference

Adjustment formula

$$P(y|do(x)) = \sum_z P(y|x,z)P(z)$$

- entails summing over all possible values of Z
- say Z is 5 categorical variables with each 3 categories, this means $4^5 = 1024$ estimates of:
  - $P(y|x,z)$ for each value of x
- what if Z is continuous?
- in practice, researchers rely on smoothness assumptions (e.g. regression) to estimate $P(Y|x,z)$ with a parametric model
- this assumption *can* be based on substantive causal knowledge, but often seems inspired rather pragmatism or necessity
- misspecification of this estimator leads to biased results (even if you know all the confounders)

Target queries

- up to now we've worked exclusively with $\mathrm{P}(\mathrm{y}|\mathrm{do}(\mathrm{t}))$: the probability of observing outcome y when setting treatment T to t
- this is not typically what is of most interest, say there are two treatment options $\mathrm{T} \in \{0,1\}$ (control and 'treatment')

1. *average treatment effect*

$$\mathrm{ATE} = \mathrm{E}[\mathrm{y}|\mathrm{do}(\mathrm{t}=1)] - \mathrm{E}[\mathrm{y}|\mathrm{do}(\mathrm{t}=0)]$$

2. *conditional average treatment effect*

$$\mathrm{CATE} = \mathrm{E}[\mathrm{y}|\mathrm{do}(\mathrm{t}=1),\mathrm{w}] - \mathrm{E}[\mathrm{y}|\mathrm{do}(\mathrm{t}=0),\mathrm{w}]$$

3. *prediction-under-intervention* $\mathrm{P}(\mathrm{y}|\mathrm{do}(\mathrm{t}),\mathrm{w})$ (more on this on day 4)

- these can be computed from $\mathrm{P}(\mathrm{y}|\mathrm{do}(\mathrm{t}),\mathrm{w})$

The simplest case: linear regression

- assume the following structural causal model (z is confounder, u is exogenous noise):

$$f_y(t, z, u) = \beta_t t + \beta_z z + \beta_u u$$

- then:

$$\text{ATE} = E[Y|\text{do}(t = 1)] - E[Y|\text{do}(t = 0)]$$

- i.e. the ATE collapses to the the regression parameter β_t in a linear regression model of y on t, z

General estimators for the ATE and the CATE (meta-learners)

- denote $\tau(w) = E[y|do(t=1), w] - E[y|do(t=0), w]$
- T-learner: model $T = 0$ and $T = 1$ separately (e.g. regression separetely for treated and untreated):

$$\mu_0(w) = E[Y|do(T=0), W=w]$$
$$\mu_1(w) = E[Y|do(T=1), W=w]$$
$$\tau(w) = \mu_1(w) - \mu_0(w)$$

- S-learner: use T as just another feature (assuming W is a sufficient set)

$$\mu(t, w) = E[Y|T=t, W=w]$$
$$\tau(w) = \mu(1, w) - \mu(0, w)$$

- (many other variants combinations: this is a whole literature)

Intuitive way-pointers:

- where does the complexity come from?

  a. variance in outcome under control: $E[y|do(T = 0), w]$

  b. variance CATE: $\tau(w)$ (in statistics: *interaction* between treatment and covariate)

Where does the variance come from?

$$\begin{tikzcd} & X \arrow[d] \\ T \arrow[r] & Y \end{tikzcd}$$

DAG

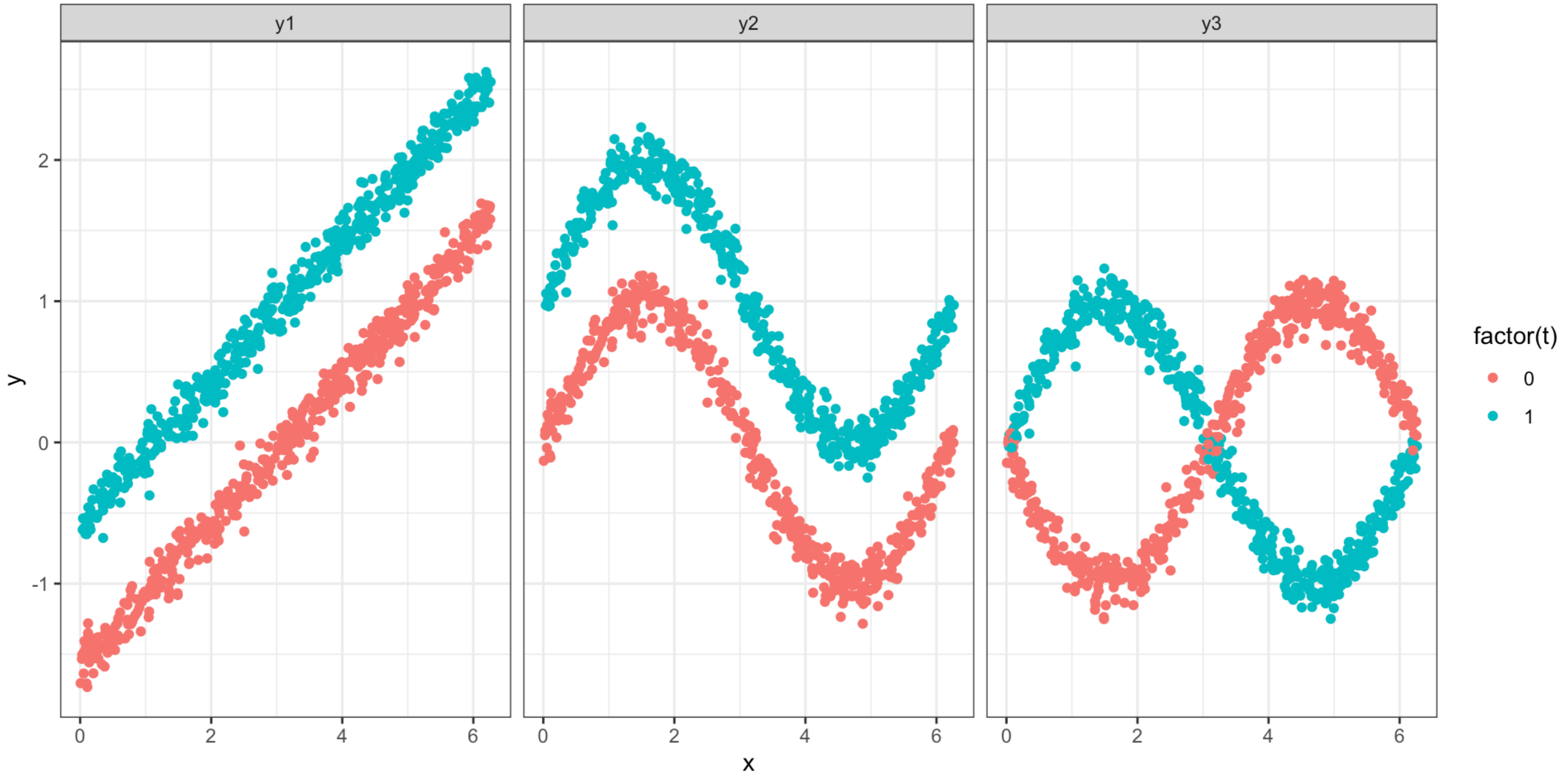


Figure 1: Three datasets with the same DAG

1. $Y = T + 0.5(X - \pi) + \epsilon$ (linear)
2. $Y = T + \sin(X) + \epsilon$ (non-linear additive)
3. $Y = T * \sin(X) - (1 - T)\sin(x) + \epsilon$ (non-linear + interaction)

Limitations of DAGs and SCMs

Making DAGs

- how do you get a DAG? up to now we assumed we had one
- based on prior evidence, expert knowledge
- “no causes in, no causes out”

A003024: The death of DAGs?

The number of possible DAGs grows super-exponentially in the number of nodes

n_nodes	n_dags	time at 1 sec / DAG
1	1	
2	3	
3	25	
4	543	
5	29281	> an hour
6	3781503	> a day
7	1138779265	> a year
8	783702329343	
9	1213442454842881	> human species
10	4175098976430598143	> age of universe

Do we need to consider all DAGs?

- a single sufficient set suffices
- adjusting for all direct causes of the treatment or all direct causes of the outcome are always sufficent sets
- can we judge these without specifying all covariate-covariate relationships?
- potential approach:
  - put all potential confounders in a cluster (e.g Anand et al. 2023)
  - ignore covariate-covariate relationships in that cluster
  - what happens when (partial) missing data?

SCM vs potential outcomes

- definition of causal effect
  - PO: averages of individual potential outcomes
  - SCM: submodel or mutilated DAG
- both require positivity
- d-separation implies conditional independence (exchangeability)

References

Anand, Tara V., Adele H. Ribeiro, Jin Tian, and Elias Bareinboim. 2023. "Causal Effect Identification in Cluster DAGs." *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (10): 12172–79. https://doi.org/10.1609/aaai.v37i10.26435.