Assignment 2

Hand in by April 28 to Søren (soren.wengel_mogensen@control.lth.se).

You need to solve Problem A or Problem B. You do not need to solve both. Hints are provided at the end of this document.

Problem A Prove Rules 1 and 2 of the do-calculus. Rule 2 might be difficult so feel free to use the hints at the end of this document.

Problem B In a linear Gaussian structural causal model, for each $j = 1, \dots, n$

$$X_j = \sum_{X_i \in \boldsymbol{PA}_j} a_{ji} X_i + N_j$$

where N_j is Gaussian. We identify two (distinct) nodes among $X_1, \dots, X_n$ and denote them by X and Y. We say that $\frac{\partial}{\partial x} E^{do(x)}(Y)$ is the *total causal effect of X on Y*. If $Z \subseteq \{X_1, \dots, X_n\} \setminus \{X, Y\}$ is a valid adjustment set for the causal effect from X to Y, then we can estimate the total causal effect from X to Y by linear regression of Y on X and Z. When several valid adjustment sets for the causal effect from X to Y are available they therefore each give rise to an estimator of the total causal effect from X to Y.

We will use an empirical approach to generate hypotheses concerning *optimal* valid adjustment sets in the sense that the resulting asymptotic standard error of the corresponding causal estimator is as small as possible. For this purpose, we consider the following steps.

- Generate a DAG of n nodes, $X_1, X_2, \dots, X_n$, and let $X = X_1$, $Y = X_2$. For simplicity, you may ensure that $X \to Y$ is the only directed path from X to Y.
- Generate data from a linear Gaussian structural equation model corresponding to the DAG.
- Find a valid adjustment set, Z, for the causal effect from X to Y.
- Find another valid adjustment set, $\tilde{Z}$, for the causal effect from X to Y, for instance, by adding a parent of X, or of Y, to Z. Try out different possibilities. Remember to ensure that the new set is also valid.
- Generate data from the model, estimate the total causal effect of X on Y, using linear regression of Y on X and Z and of Y on X and $\tilde{Z}$. Compare their empirical performance.

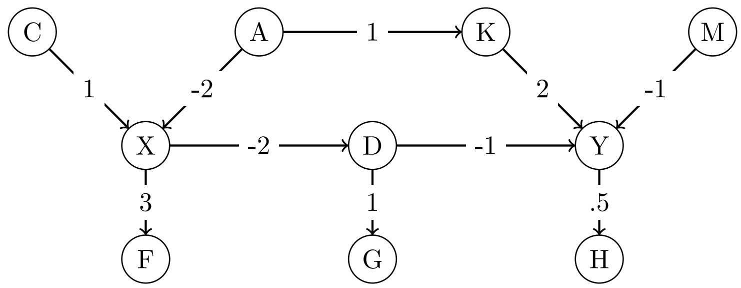

Figure 1: In the above graph, $\{K\}$ is a valid adjustment set for the causal effect from X to Y. We can use the back-door criterion to see this. Sets $\{K, A\}$, $\{K, C\}$, and $\{K, A, C\}$ also satisfy the back-door criterion. For each valid adjustment set, a linear regression allows us to estimate the total causal effect of X on Y. We can generate data from a linear Gaussian structural causal model and check empirically which set leads to the most efficient estimator of the causal effect of X on Y. Similarly, you may try to include M or F in an adjustment set in this example. Construct several example systems to gain more intuition.

There is an example of the above procedure in Figure 1. Use this procedure to generate some hypotheses about how an optimal valid adjustment set might look like. An example hypothesis could be 'adding a parent of X to a valid adjustment set is bad'. Report your experiments and hypotheses.

You may look for inspiration in Henckel, Leonard, Emilija Perković, and Marloes H. Maathuis. "Graphical criteria for efficient total effect estimation via adjustment in causal linear models." Journal of the Royal Statistical Society Series B: Statistical Methodology 84.2 (2022): 579-599.

Hints

Problem A: Rule 1: Use that d-separation also holds in the interventional model. Rule 2: Argue first that the assumption in Rule 2 implies that all back-door paths from Z to Y are blocked by $\{X, W\}$ in the graph obtained from $\mathcal{G}$ by removing all edges pointing into X. Argue that this implies the result. Feel free to look at the proof in Pearl, Judea. "Causal diagrams for empirical research." Biometrika 82.4 (1995): 669-688.