Causal Discovery
Tutorial Teaching Reflection

Taha Bouhoun

January 15, 2021

1 Preparing the session's materials

The topics that we've selected for the tutorial jumped from Pearlian causality theory straight to causal effect estimation algorithms (EconML, DoWhy). I was then inspired to devote my session for causal discovery as an attempt to cover at least the major algorithms that have been central to many breakthroughs in genetics.

Throughout the tutorial sessions that my classmates have conducted, it was clear to me that focusing on fewer takeaways that are much easier to retain and understand is far better than overloading students with readings and aspiring that they would catch a fraction of it. Dean Kosslyn refers to this as "Desirable Difficulty" which insists on developing materials that are not too easy to be boring but also not too hard that it's unrealistic to retain in one session. Hence my decision to pick a limited yet focused set of readings which contains:

- a 15-pages paper offering a literature review of causal search theories, their assumptions, advantages, and drawbacks.
- A 40-minutes video by Frederick Eberhardt on causal discovery accompanied by a detailed slideshow for all the relevant theoretical details.
- Lecture by Elizabeth Silver on Causality and Causal Discovery along with a well-commented slideshow.

The readings for Minerva classes can be overwhelming at times, and students can be drifted away by the content (especially if the readings are broad). Thus the necessity for a study guide that emphasizes the following points:

- The reason behind the readings' choice and their purpose
- Definitions and Keywords related to the topic of interest
- Bullet-points summary of the expectations before turning up to class

In an attempt to bridge the gap between theory and practice, the pre-class work pushes the students to experiment with Python implementation of causal search algorithms. Given the time constraints, I set up a choice for either working with synthetic data or to pick a dataset of their choice where the SCM is unknown. The advantage of picking a synthetic dataset was to check if the Markov Equivalent DAGs that the students came up with would match the structural equations

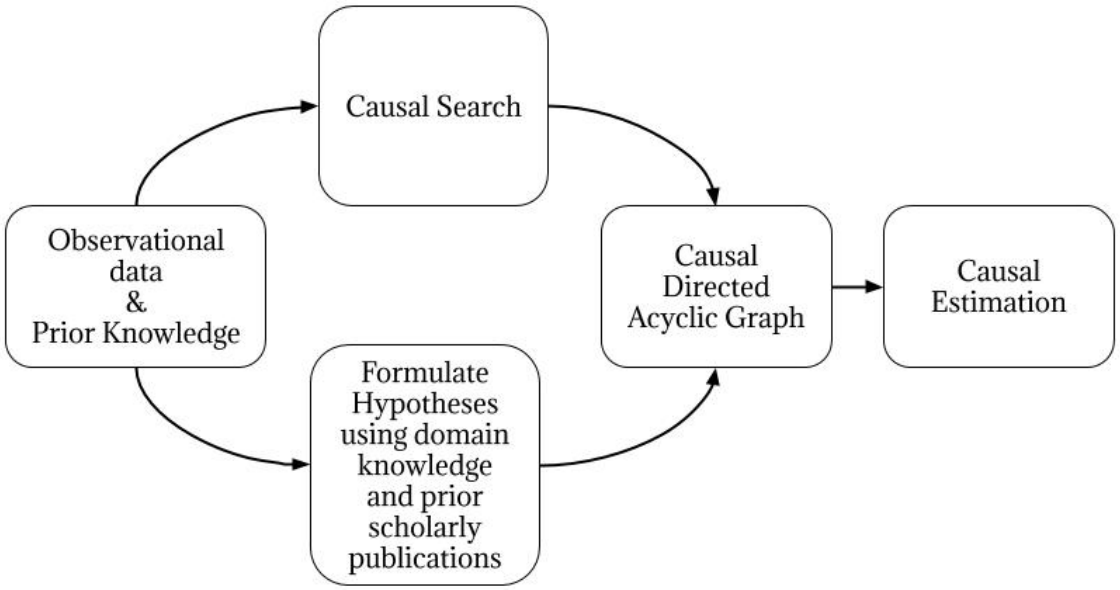
I used to generate the data. Deliberate practice is the closest principle to coding implementation as a means to apply a newly-introduced theory or concept. The pre-class along with a 20-minutes breakout session is an opportunity to experiment with causal search packages. Although it might not be enough for students to become experts (Hambrick et al., 2014), it sheds light on how we can tool-ify and put causal search theories to use.

2 Conducting the session

Being newly introduced to the technicalities of causal discovery, I've done far more readings than what I've proposed for the class which drove me straight to the "valley of despair" on the Dunning-Kruger curve. In other words, the more I dive into the readings the more self-conscious I felt about how much I still ignore. Nonetheless, I had a planned sequence to conduct the class and cover the three main takeaways:

- Assumptions of causal discovery
- Constraint and Score-based algorithm
- LiNGAM

Anchoring the discussion to the above-mentioned themes made it easier to manage the class effectively. Dean Kosslyn argues that the most effective starting point is to build on prior knowledge and associations, hence the following diagram that I have prepared before class to draw a map between causal discovery and causal search.



The science of learning literature suggests using examples to distill abstract ideas. However, the background of our causal inference cohort is diverse with students oriented mostly to CS (Oscar F., Sanny, Oscar E., Asmaa), SS (Johannes, Berfin), and NS (Micheal). As a result, I proposed materials in the readings with applications in Economics and Genetics because, after all, the common factor is that we're mostly dealing with observational data. Perhaps the most important question that a student would ask after being introduced to new concepts is how to employ them in their field of interest. I realized that being mindful of the students' expectations of the class content can serve as a motivation to make an effort and prepare well for the session.

Throughout the session, I either picked on students to answer questions on the readings or redirect a student's question. It turned out that even with planning, I could've missed some crucial details about causal discovery. An example was when Micheal asked about the Markov Equivalence class. Although we visited the concept during Pearl's book sessions, Micheal pointed out that he thought a MEC means an undirected graph (an SCM with no arrows). In fact, a MEC graph can have both directed and undirected edges, and according to Pearl's definition:

A set of DAGs are Markov equivalent if and only if they have the same adjacencies and the same unshielded colliders

The breakout session was the highlight of my session, many peers had positive feedback on seeing LiNGAM at work. The pre-class was about using the PC and GES algorithms which they only output the Markov Equivalent Class. The breakout code files include LiNGAM implementation of the synthetic data which outputs the exact SCM of the model, also known as identifiability.

3 Causal Discovery Content

Elizabeth Silver in her lecture mentions how the search space of causal search is super-exponential. Formally, for all acyclic graphs, the number of potential DAGs is:

$$2^{\frac{1}{2}(n-1)n} \leq S \leq 3^{\frac{1}{2}(n-1)n}$$

I came across this visualization after watching a conference lecture by Jonas Peters, it captures the increase in search space using only numbers retrieved from The Online Encyclopedia of Integer Sequences.

d	Number of DAGs with d nodes
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103
15	237725265553410354992180218286376719253505
16	83756670773733320287699303047996412235223138303
17	62707921196923889899446452602494921906963551482675201
18	99421195322159515895228914592354524516555026878588305014783
19	33277190122710759173617757331126112588358307625842190258354677350 5

Table B.1: The number of DAGs depending on the number d of nodes, taken from `http://oeis.org/A003024` [OEIS Foundation Inc., 2017]. The length of the numbers grows faster than any linear term.

Although constraint and score-based algorithms are useful in narrowing down the potential DAGs, the MEC can be intractable quickly if the number of variables hits double digits. Therefore, the idea behind LiNGAM as an alternative search strategy that guarantees identifiability (unique DAG) by assuming a non-Gaussian error or a non-Linear functional relationship between variables. The following example of two variables illustrates how assuming non-Gaussian error leads to identifying the causal direction: Assuming a true model where X causes Y, the structural equation can be written as:

$$y = \beta \times x + \epsilon_y \qquad \text{where} \qquad x \perp \epsilon_y$$

In case we assumed the opposite backward model, the structural equation would be:

$$x = \alpha \times y + \epsilon_x \qquad \text{where} \qquad y \perp \epsilon_x$$

$$\epsilon_x = x - \alpha \times y$$

We substitute the true equation for the variable Y:

$$\epsilon_x = x - \alpha \times (\beta \times x + \epsilon_y)$$

$$\epsilon_x = (1 - \alpha \times \beta)\, x - \alpha \times \epsilon_y$$

This violates the independence assumption between Y and the error of X according to the Darmois-Skitovich theorem where it states that:

> *If two linear combinations L1 and L2 of random variables X1, X2, .., Xn are independent, then each random variable Xi is normally distributed.*

In this case, if the error is non-Gaussian, then the two linear combinations have to be independent. The same rule applies when the functional relationships between the variables in the DAG are non-linear. It's a very interesting finding by Shimizu et al. since assuming a Gaussian error or linear equations is supposed to simplify the computation. However, it seems that, in the context of causal discovery, Occam's razors doesn't apply

4 Appendix

- Readings, Study Guide, and Pre-class work
- Session plan & side notes

5 References

- Glymour C, Zhang K, and Spirtes P (2019) Review of Causal Discovery Methods Based on Graphical Models.
- Hambrick, Zach & Altmann, Erik & Oswald, Frederick & Meinz, Elizabeth & Gobet, Fernand & Campitelli, Guillermo. (2014). Accounting for expert performance: The devil is in the details. Intelligence. 45. 112-114. 10.1016/j.intell.2014.01.007.
- Kosslyn, S. (2017-10-06). The Science of Learning: Mechanisms and Principles. In Building the Intentional University: Minerva and the Future of Higher Education. : The MIT Press.