Ch 8: Causal Models – Assumptions and Challenges

Prashant Shekhar, PhD
Assistant Professor of Data Science
Department of Mathematics
Embry-Riddle Aeronautical University, FL, USA
Email: shekharp@erau.edu

Outline

Things covered in this lecture

The goal of this lecture is to deepen our understanding of when and how to use causal inference methods. We will discuss some more assumptions in order to get a clearer picture of the challenges and limitations that we might face when working with causal models.

1. The challenges of causal inference methods.
2. Identifiability
3. Positivity assumption
4. Exchangeability/ignorability assumption
5. Modularity (independent mechanisms)
6. Consistency
7. SUTVA
8. Selection bias

Identifiability

Definition

A causal effect (or any other causal quantity) is **identifiable** when it can be computed unambiguously from a set of (passive) observations summarized by a distribution $P(V)$ and a causal graph G (Pearl, 2009).

In other words, a causal effect is identifiable if:

1. Enough information to control for non-causal information flow in the graph:
   - This is achievable by blocking all the paths that are leaking non-causal information using the rules of do-calculus and the logic of d-separation.
   - Sometimes its non-achievable.
2. Enough data to estimate the effect of interest
   - Any estimator needs to have a large enough sample size to return meaningful estimates.
   - Probability of every possible value of treatment in our dataset (possibly conditioned on all important covariates) is greater than 0 (**positivity assumption**).

Challenges faced by data scientists

Challenges faced by data scientists

Lack of causal graphs: The lack of a causal model can be a major challenge in implementing causal inference techniques in practice. Usually people:

- **Use domain expertise**: Best but rarely available
- **Use causal discovery techniques**: Doable but difficult to verify
- **Use combination of both**: Is usually best

Combination of both approaches

Some causal discovery methods allow us to easily incorporate our domain knowledge and then they perform the search over the remaining parts of graph search space for us (more on this topic in Part 3, Causal Discovery). This can lead to truly amazing results, yet there are no guarantees.

Challenges faced by data scientists

Not enough data:

- An insufficiently large sample size is not a uniquely causal problem. Any statistical parameter estimate becomes biased under an insufficiently large sample size.
- Depending on the method we choose, some causal methods might require more data than others. For example if we use the **double machine learning (DML)** technique with neural network estimators, we need larger data sizes for training.

Challenges faced by data scientists

Unverifiable assumptions

- While using the back-door criterion sometimes we will be able to rule out the possibility that there are other unobserved variables introducing confounding between treatment and outcome, but in many cases, it might be very difficult.
- In particular, when the research involves human interactions, financial markets, or other complex phenomena, making sure that the effects of interest are identifiable can be difficult if not impossible.

What can we do ?

1. **Refutation tests**: Allows to test the correctness of our causal model. One challenge with these tests is that they check for the overall correctness of the model structure, but they do not say much about how good the obtained estimate is.
2. **Comparison with historical data**: You can compare your observational model with the experimental results and try to adjust your model accordingly.
3. **Evaluation on simulated data with known outcomes**: We can possibly learn a reliable simulator using generative neural networks. One such approach – RealCause – has been proposed by Brady Neal and colleagues (Neal et al., 2020). Such a simulator can help us in testing the model behavior under violation of crucial model assumptions such as **positivity**. Although, synthetic data can lead us astray when assessing the performance of causal models (Reisach et al., 2021; Curth et al., 2021).
4. **Sensitivity analysis**: In certain cases, we might have some idea about the magnitude of possible hidden confounding. In these cases, we can bound the error and check whether the estimated causal effect still holds if the confounders' influence reaches the maximum level that we think is reasonable to assume.

Positivity

Positivity

- **Positivity assumption** is sometimes also called **overlap** or **common support**.
- The assumptions states that the probability of your treatment given all relevant control variables (the variables that are necessary to identify the effect – let's call them Z) has to be strictly positive. Formally:

$$P(T = t|Z = z) > 0$$

The preceding formula must hold for all values of Z that are present in the population of interest (Hernan & Robins, 2020) and for all values of treatment T.

Positivity

At a hospital:

- Suppose we have multiple subjects (patients) described by a continuous variable Z.
- Each subject either received or did not receive a binary treatment T.
- Each subject has some continuous outcome Y.

Additionally, let's assume that in order to identify the causal effect, we need to control for Z, which confounds the relationship between T and Y. We can estimate the causal effect by computing:

$$\mathbf{E}[Y|do(T=1)] - \mathbf{E}[Y|do(T=0)] \tag{1}$$

However, while computing these quantities we can face the following situation...

Positivity

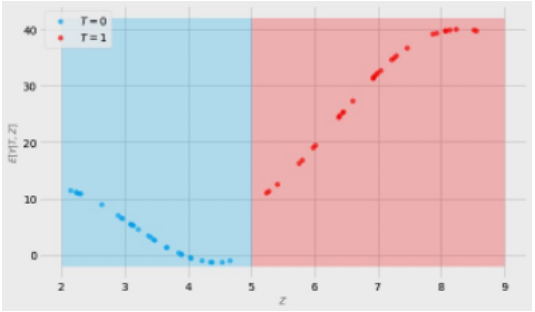


Figure: Here the support of Z (in other words, values of Z) does not overlap at all with the treatment values. In other words for $Z \in [2, 5]$ we only have data for $T = 0$ (only blue curve and no red curve), and for $Z \in (5, 9]$ we only have data for $T = 1$ (only red curve and no blue curve). **Hence we cant compute (1) for either of the regions** $[2, 5]$ **or** $(5, 9]$

Positivity

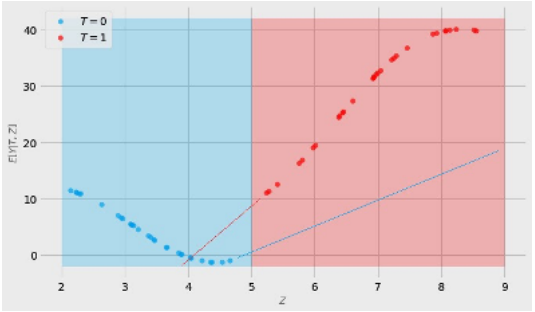


Figure: Hence, we have to extrapolate T $= 0$ curve in the region $Z \in (5, 9]$ and T $= 1$ curve in $Z \in [2, 5]$. With this extrapolation, now computing (1) is possible. **However, this is not reliable as extrapolation is hard.**

Positivity

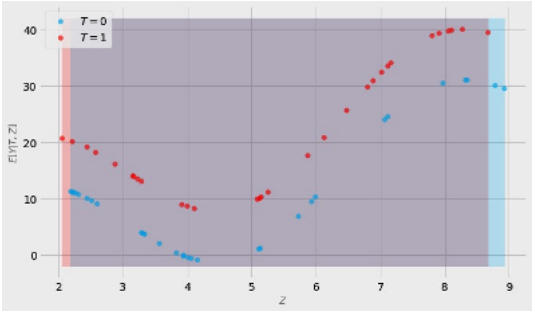


Figure: An example case where the positivity condition is met. Here we just need to interpolate at different Z's. **However, both interpolation and extrapolation in higher dimension can be very challenging.**

Exchangeability

Exchangeability

Here we introduce the **exchangeability** assumption (also known as the **ignorability assumption**) and discuss its relation to confounding.

Definition

The treated subjects, had they been untreated, would have experienced the same average outcome as the untreated did (being actually untreated) and vice versa (Hernán & Robins, 2020). Formally:

$$\{Y^0, Y^1\} \perp\!\!\!\perp T|Z \tag{2}$$

Here Y^0 and Y^1 are counterfactual outcomes under $T = 0$ and $T = 1$ respectively and Z is a vector of control variables.

The idea of exchangeability comes from the **Potential outcomes** framework. In fact, the potential outcomes framework aims to achieve the same goals as SCM/do-calculus-based causal inference, just using different means (see Pearl, 2009, pp. 98-102, 243-245).

Modularity

Modularity assumption is also known as **independent mechanism assumption**.

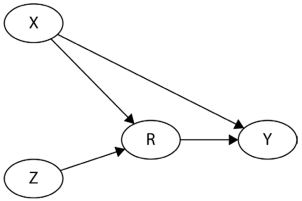
Definition

If we perform an intervention on a single variable X, the structural equation for this variable will be changed (for example, set to a constant), yet all other structural equations in our system of interest will remain untouched.
In other words, changes caused by interventions are **local**: only the mechanism for the intervened variables change (we remove the incoming edges), but other mechanisms in the system remain unchanged.

Modularity example

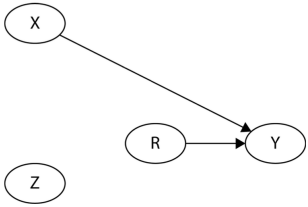
Original SCM with structural equations:

- $X, Z := N(0, 1)$
- $R := X + Z$
- $Y = R + X$



SCM with intervention $R := r$:

- $X, Z := N(0, 1)$
- $R := r$
- $Y = R + X$



- Because on intervention on node R, due to modularity assumption, all the incoming nodes to R got removed (**perfect intervention**). Rest of the SCM remains the same.
- Removing edges from graph under intervention is called **graph mutilation**.

SUTVA

Stable Unit Treatment Value Assumption (SUTVA) also comes from the potential outcome framework.

Definition

One unit (individual, subject, or object) receiving treatment does not influence any other units.

This assumption might often be challenged in the context of interacting subjects.

- For example, if you encourage some users of a social network to send more messages, the recipients of these messages are also likely to start sending more messages (not because they were encouraged by you, but rather because they want to respond to the senders).
- At the experimental level, researchers are trying to overcome these challenges using various techniques such as cluster-level (rather than individual-level) randomization or so-called ego-randomization (Gui et al., 2015; Saint-Jacques et al., 2018).

Consistency

Consistency or **no multiple versions of treatment** also comes from the potential outcome framework.

Definition

Consistency basically encapsulates:

- Treatments should be well defined
- There should not be hidden versions of treatments

Imagine an experiment where people in the treatment group win a brand-new electric BMW while others get a rusty, 20-year-old Mazda without wheels.

- If our outcome variable (Y) is the level of excitement, we'd expect that on average the same person's level of excitement would differ between the two versions of treatment.
- That would be an example of a violation of consistency as we essentially encode two variants of treatment as one.

Selection Bias

Introductory example

- It's 1943. American military planes are engaged in numerous missions and many of them are not coming back home. The ones that are coming back are often damaged.
- An interesting fact is that the damage does not look random. It seems that there's a clear pattern to it. Bullet holes are concentrated around the fuel system and fuselage, but not so much around the engines.
- The military engineers decide to consult a renowned group of statisticians called the Statistical Research Group (SRG) to help them figure out how to optimally distribute the armor so that the places that are at the highest risk of damage are protected.

Introductory example

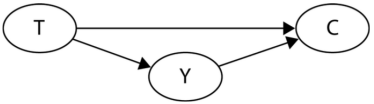
- Instead of answering the questions, Abraham Wald, a Hungarian-born Jewish mathematician, who worked at Columbia University and was a part of the SRG at the time, asked them another question:

Where are the missing holes ?

- What Wald meant were the holes that we've never observed. The ones that were in the planes that never came back.
- So there is selection bias (particularly **survivorship bias**) in our sample of holes.
- So whatever we learn from these holes might not generalize well to the entire population of holes.

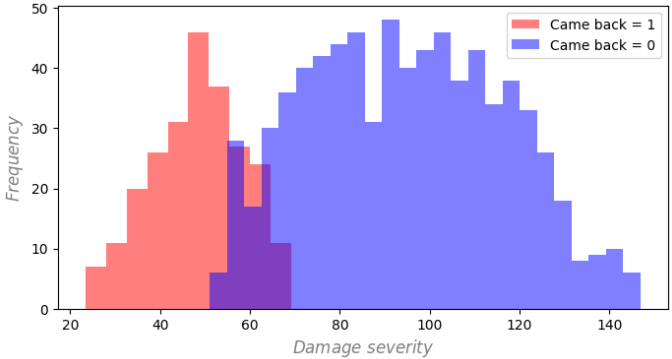
Making it numerical

- Making a SCM out of it:



  - T (Treatment): Number of enemy bullets shot at the engines.
  - Y: Severity of plane damaged
  - C: Binary variable encoding if the plane came back home (1) or not (0).
- Note that what we do by only looking at planes that came back home is **implicitly conditioning on C**.
- However, conditioning on C (a collider) opens a spurious path: $T \rightarrow C \leftarrow Y$
- Hence the spatial distribution of damages we observe on the planes (after conditioning on C) might not give a good overview of the damage distribution

Sample distribution from the SCM

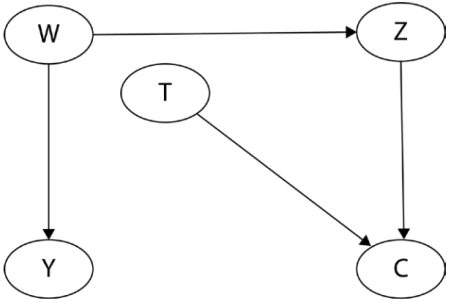


CODE

Refer to the [Python notebook] for implementation of:

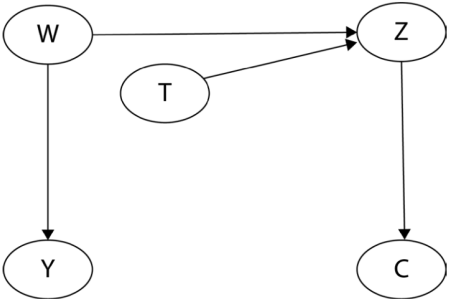
- Data generation from the SCM
- Visualizing samples of Y with C = 0 (blue) and then C = 1 (red).

More Examples



- For this SCM, if we wanted to study T by observing Y, there is no association (C is a collider).
- However, if we condition on C, then there is association. However, it might create bias like before. In such cases its called **selection bias under the null**.

More Examples



- For this SCM, T and C are not even connected, and again there is no association between T and Y.
- However, here again, since C is a child (descendant) of Z, conditioning on C, will also open a path **partially**: $T \rightarrow Z \leftarrow W \rightarrow Y$