Università degli studi di Milano–Bicocca
Scuola di Economia e Statistica

Corso di Laurea in
Scienze Statistiche ed Economiche

Bayesian Causality

Relatore: Dott. Stefano Peluso

Tesi di laurea di:
Simone Maria Gervasoni
Matricola N. 880068

Anno Accademico 2024/2025

Contents

Ringraziamenti

Inserire qui gli eventuali ringraziamenti, altrimenti eliminare

Chapter 1

Introduzione

The problem of causality is of fundamental interest and it has been for as long as humans have been around. We always have asked why something happens and we tried many possible answers. The question of causality has been explored by philosopher and scientist alike, but its precise statistical definition has remained elusive until the seventies, when Rubin expanded and perfected a theory proposed by Neyman in 1932. This model diverges in a few key points from the definition of some philosopher; John Stuart Mill defines causality as "the antecedent, or the concurrence of antecedents, on which [a given phenomenon] is invariably and unconditionally consequent". Mill thinks that A causes B if and only if each time A happens than B happens. This kind of intuition not only is wrong but can be outright dangerous. Instead in this thesis we rely on the potential outcome model that we believe gives more satisfactory answers to the aleatory nature of reality.

Chapter 2

DAGs

Let's start by introducing the DAGs or Directed Acyclic Graph, we will need this powerful tool to model causal relationships that wont be properly defined until chapter 3. A DAG is a graph that represents the set of random variables $V = (V_1, ..., V_m)$ as nodes and the casual connection between them as edges. It's called "Directed" because the graph can't form loops so in its simplest form it isn't suited to describe reciprocal causal relationships or feedback loops, such as: $A \leftrightarrow B$. We will explore some extension of the model that allow for this kind of interaction (see chapter est-temp), but for now we will stick to easier examples.

[show examples of valid and invalid dags?]

2.1 Nodes and Edges

In this section we will show examples of a DAGs where Y is the variable of interest or *outcome* variable for a statistical analysis (e.g. number of years lived). D (e.g. surgery status) is the exposure variable or the variable for which we want to identify the causal effect on the *outcome* variable . We do this to introduce terminology necessary to describe properties of this model.

2.1.1 Nodes

Descendants Ancestors Parents and Child

We call parents (denoted as PA_m) the set of causal variables that have a directed arrow in to V_m. We call ancestor any variable witch through a sequence of nodes connected by directed edges. The opposite definition goes for kids and ancestors.

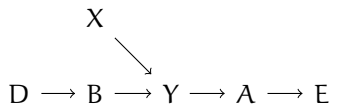


Figure 2.1: DAG: Common effect

In the DAG 2.1 for Y :

- D,B,X are its ancestors
- B,X are its parent
- A, E are its descendants
- A is its child

Note that D,B,X are Y's ancestors but not its descendants because the directed path goes from D,B,X to Y and not the other way around. We can say from this graph that D caused B that caused Y and so on, so by the graph we understand that the treatment had an effect on the health of at least one individual that was *mediated* through B.

Mediators , Common effects , Common causes

Let's focus on B and all the possible connection between it, D and Y:

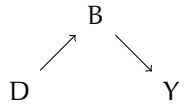


Figure 2.2: DAG: Mediator

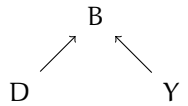


Figure 2.3: DAG: Common effect

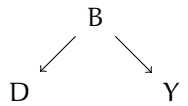


Figure 2.4: DAG: Common cause

In the graph 2.3 B is a common effect of the surgery and the health of the patient; these kind of nodes are called *colliders*. It has to be noted though that

in the graph we don't have any information about the nature of the interaction of the two causes or the strength of the causal relationship. In the graph 2.2 B is the mediator, for witch the causal effect passes through, in this example let B be the presence of tumour and D the operation that removes it, the causal effect of the operation is mediated by the presence of the tumour itself. Meanwhile in the graph 2.3 B is the common cause of both the *exposure* and the *outcome*, for example B could be inflammation and this both caused people to opt in for the surgery and is the cause of a faster death. We can see that this kind of relationship is the most problematic because any association between better health and the operation status is systematically *spurious*; this kinds of nodes are called *confounders*.

2.1.2 Edges

Let's start by defining an important term that will be frequently mentioned: [should i define path and directed path too ?]

Definition 2.1 (Backdoor Path)**.** A backdoor path from variable X to Y is any path from X to Y that starts with an arrow pointing to X

To make any further progress with causal inference it is necessary to outline some assumptions, and especially what makes a DAG *causal* is the following (Hernán & Robins, 2020):

Assumption 2.1 (Markov assumption)**.** *We define the distribution of V to be Markov with respect to a DAG G (equivalently, the distribution factors according to a DAG G) if, for each j,* V_j *is independent of its non-descendants conditional on its parent. Or equivalently, given that V is the set of random variables:*

$$f(v) = \prod_{j=i}^{M} f(v_j|PA_j) \tag{2.1}$$

This may seem like an innocuous assumption but it is a very strong one, because it means that for any two variable V_j V_m for witch there isn't a directed edge between the two, then $V_j \perp\!\!\!\perp V_m|(PA_j, PA_m)$. This leads us to state 3 properties of a causal DAG (Hernán & Robins, 2020):

- The lack of a directed arrow from node V_j to V_m can be interpreted as a lack of causal relationship
- Every common cause has to be present on the graph

- Any variable is cause of its descendants

It's the analyst burden to assert, on a case to case basis, that this condition is indeed credible. If it isn't credible, the problem at hand is often intractable, and a new DAG or new measurements have to be made. The last assumption is similar to the previous one but the direction of the implication is the opposite (Ramsey et al., 2012):

Assumption 2.2 (Faithfulness)**.** *Given a set of variables whose causal structure can be represented by a DAG, no conditional independence holds unless entailed by Markov Assumption.*

This simply means that any independence between variables is due to arise from structure rather than coincidence. To have a clearer picture of this last assumption, we construct an example where faithfulness **doesn't** hold. It's known that $D \perp\!\!\!\perp Y$ but the causal DAG that represents the relationship between the two variable is the following:

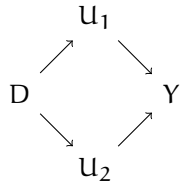
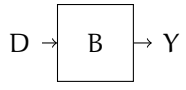


Figure 2.5: DAG:faithfulness

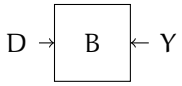
The independence holds if both causal path exactly counteract each other. Needless to say that this kind of occurrence is exceedingly unlikely.

2.2 Conditional independence

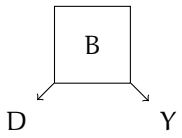
Let's revisit the DAG 2.2, we can see that the surgery is correlated to years lived, but the casual path is mediated by the presence of the tumour. A legitimate question one could ask is: if we already have information on B (status of the tumour) does D give us further information? To answer this kind of question we have to control for B, in this case we can restrict the analysis to either one group (B=0, B=1). We represent this restriction by drawing a box around the variable we wish to condition upon.



We know from assumption 2.1.2 that any variable in the graph G is independent from any other variable conditioned on its parents, so in this graph we have that $Y \perp\!\!\!\perp D|B$. The result is reasonable because in this example life expectancy depends directly only on the presence of the tumour, and weather or not the operation was made is entirely irrelevant. Conditioning on common effects bears different results, let B the mood of the family (B=0 if sad, B=1 if sad), in this example let's say that the operation has no effect on life expectancy. We can represent this examples with this DAG:



Let's say that both D and Y have positive effects on B, now if we restrict the analysis on B=0, we will see that patient that are not treated have a low life expectancy, meanwhile in the B=1 we will have patient that are treated and have a higher life expectancy, this is indeed problematic because if we had not controlled for B, D and Y would have been independent. Even though this is just an example, it is proven (citation needed?) that controlling for a collider, or one of it's descendants introduces spurious systematic correlation. Epidemiologist refer to this spurious correlation as *selection bias under the null*. When we condition on a common effect we open the flow of association through the graph. As we have seen earlier common causes, like DAG 2.4, produce systematic bias that we will refer to as *confounding bias*. Now let B be inflammation status that both causes lower life expectancy (Y) and a higher chance of being operated (D).



Let's imagine restricting the analysis to B = 1 (inflammation present), we can clearly see that a lower life expectancy doesn't tell us any new information on the chance of being operated. We can also deduce this by applying the Markov assumption that tells us that $D \perp\!\!\!\perp Y|B$, since B is the parent of both Y and D. When we condition on a common cause we close the flow of association through the graph.

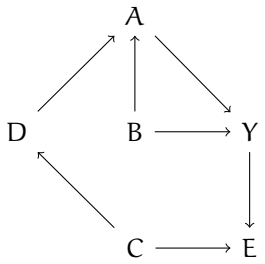
2.3 Back door criterion

If we want to be able to tell weather a casual effect is identifiable we have to satisfy the *backdoor criterium*. We satisfy this criterium if and only if all back-door

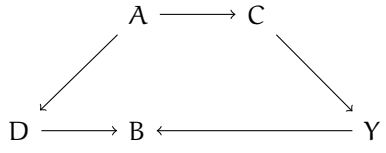
paths are blocked i.e. if all non direct flow of association are blocked and we don't condition on a descendant of the exposure [citation needed for the last part]. We can define a path to be blocked or open by these graphical rules (Hernán & Robins, 2020):

1. if there are no variables being conditioned on, a path is blocked if and only if there is a collider
2. Any path that contains a non-collider that has been conditioned on is blocked.
3. A collider that has been conditioned on does not block a path.
4. A collider that has a descendant that has been conditioned on does not block a path.

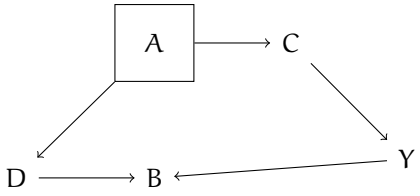
Let's make this rules clearer by applying them to a few examples: Ex1. [should I expand every example with a story]



Let's list all paths from our intervention D to Y : we have $D \to A \to Y$ and $D \to A \leftarrow B \to Y$ these two are not back-door paths, so we don't have to control for anything to avoid modifying the main causal path. The only back-door path we need to worry about is $D \leftarrow C \to E \leftarrow Y$, we see from graphical rule number 1, that a path is closed if there is a collider that in this case is E. We can see than that we don't have to control for anything; the causal effect is identifiable. Example n.2



Here we have two paths $D \to B \leftarrow Y$ and a backdoor path witch is $D \leftarrow A \to C \to Y$. We ought not to condition on B because it's a common cause and would introduce spurious correlation; meanwhile we should condition on A since is a common cause of D and G. The resulting DAG would be:



Once we followed all the rules and no paths the causal effect will be identifiable and any systematic correlation between exposure and outcome can be interpreted as the causal effect D_i has on Y_i.

Chapter 3

Potential outcome model

The potential outcome model defines the causal effect of an event A as the difference between the two possible states of the world, namely the world where A occurs and the one where A does not occur. For example, we would like to understand if a medicine can truly improve headaches. Let's formalize the problem by setting X as the set of patient covariates (Age , Gender,...), D as the treatment regimen (set equal to 1 if the patient takes the medicine and 0 if he takes the placebo), and Y as the number of minutes the headache persists. Under this model the causal effect of the medicine would be calculated by comparing Y_i^1 (the world where is administered the medicine), and Y_i^0 (the world where the patient is given a placebo). These are the so called "potential outcome" because they are both *a priori* observable, but only one can be observed *a posteriori*. Let's introduce a numerical example, where we somehow know both potential outcomes:

Age	Sex	Y_i^0	Y_i^1	δ_i
20	M	20	21	-1
20	F	15	3	12
20	M	8	10	-2
20	F	16	15	1
30	M	12	13	-1
30	F	8	5	3
30	M	2	11	-9
30	F	15	26	11

Table 3.1: Random trial knowing both potential outcomes

We can observe that δ isn't always positive, so the medicine didn't universally improve the situation, but it is seemingly clear that it had positive effects on the whole. This isn't nearly precise enough so we need to introduce some mathematical definitions :

Parameter definition

- CATE or Conditional Average Treatment Effect is defined as $E[Y_i^1 - Y_i^0|X] = E[\delta_i|X]$, so $CATE_{(M,20)} = E[\delta_i|X = (M, 20)] \approx \frac{18-2}{2} = 8$, we can say that on average the medicine caused a reduction of 8 minutes in the length of the headache between 20 year old males.
- ATE or Average Treatment Effect is defined as $E[Y_i^1 - Y_i^0] = E[\delta_i]$, so $ATE = E[\delta_i] \approx \frac{18+12-2+18+3-9+11}{8}$.
- ATT o Average Treatment on the Treated is defined as $E[Y_i^1 - Y_i^0|D = 1] = E[\delta_i|D = 1]$
- ATU o Average Treatment on the Untreated is defined as $E[Y_i^1 - Y_i^0|D = 1] = E[\delta_i|D = 1]$

In this example we somehow know both potential outcomes so we can't calculate the last two quantities. It is useful to make a distinction between *factual* and *counter factual* values, the former refers to the state of the world we are currently in and the latter refers to *what would have been* if the treatment were different. We can better understand the distinction between the two through the switching equation:

$$Y_i^{obs} = D_i \cdot Y_i^1 + (1 - D_i) \cdot Y_i^0 \tag{3.1}$$

We can see that Y_i^1 is either a *factual* value when $D_i = 1$ or *counter factual* when $D_i = 0$. It's obvious that in a real setting, we wont ever be able to fill in all the data like in table 3.1, at least half of the value will be missing. An example of such table could be the following:

Age	Sex	Y_i^0	Y_i^1
20	M	20	?
20	F	15	?
20	M	?	10
20	F	?	15
30	M	12	?
30	F	8	?
30	M	?	11
30	F	?	26

Table 3.2: Table random trial

The potential outcome model manages to transform the intractable problem of causation in to a much more manageable problem of *missing data*.

3.1 Assumptions and exclusion restriction

So far we hid a few of the assumption to avoid making this introduction unnecessary cumbersome however it is now necessary to point them out to avoid confusion. These are the so called *exclusion restriction* : assumption made with substantive knowledge of the subject matter. In fact any of these assumptions can be loosened if the specific case requires it, but they are considered to be the most common assumption when tackling a problem of causal inference (Imbens & Rubin, 2015). This *exclusion restriction* first proposed in (Rubin, 1980), states that :

Assumption 3.1. *SUTVA The potential outcome for any unit do not vary with the treatments assigned to other units and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

(Imbens & Rubin, 2015) This assumption can be split in two parts, the no interference component and the no hidden variation of treatment components. The former implies that any potential outcomes for a given observation are independent from treatment of other units, it can be expressed by :

$$D_j \perp\!\!\!\perp (Y_i^0, Y_i^1) \; \forall i \neq j$$

[to check if the mathematical expression is true] For example in the headache situation we excluded *a priori* that each unit treatment wouldn't effect other patients, of course this kind of assumption ought to be carefully considered by an expert on a case by case basis. However this assumption becomes shaky when we consider time series or pandemics where a unit treatments is very likely to impact the other units. The problem can be solved in two ways either by having a more suiting *exclusion restriction* or by defining a unit in a broader term, in the vaccination example the unit could be the nation (given no spillover between nations), instead of the individual. The second component instead implies that the treatment is homogeneous in the population, so back to the headache example we cannot have a medicine that has different potency or is administered in different ways. This assumption referred to a DAG states the arrow from treatment D to outcome Y corresponds to unambiguous and uniform intervention. The assumption of strong ignorability was first formalized in (Rosenbaum & Rubin, 1983) , and aims to restrict the possible assignment mechanism (should i define a assignment mechanism ?) and has three main components :

Assumption 3.2 (Individualistic assignment)**.** *An assignment mechanism is individualistic if* $Pr(D_i|X_i, Y_i^0, Y_i^1) = q(X_i, Y_i^0, Y_i^1)$ *for some function* $q(.)$

Strong ignorability This means that weather a unit gets a treatment its a function only of its own covariates.

Assumption 3.3 (Probabilistic assignment)**.** *An assignment mechanism is probabilistic if* $Pr(D_i|X_i, Y_i^0, Y_i^1)$ *is bounded strictly between 0 and 1* $\forall X, Y^1, Y^0, i$

This requirement is important because we need to see the counter factual in the limit of all possible subpopulations. The estimation if this condition is not met, is practically impossible, because it means that we don't have any data about how a subset of the population would react to treatment, thereby any causal estimation would rely only on extrapolation. In the causal DAGs drawn so far, probabilistic assignment is implicit, and SUTVA is embedded in the notation because we only consider treatment nodes with binary and well defined alternatives. Probabilistic assignment is concerned with arrows going into the treatment nodes, and SUTVA is only concerned with arrows leaving the treatment nodes. Thus, the treatment nodes are implicitly given a different status compared with all other nodes.(Hernán & Robins, 2020)

Assumption 3.4 (Unconfounded assignment)**.** *An assignment mechanism* $Pr(D_i|X_i, Y_i^0, Y_i^1)$ *is unfounded if it does not depend on the potential outcomes* $Pr(D_i|X_i, Y_i^0, Y_i^1) = Pr(D_i|X_i)$ *or* $D_i \perp\!\!\!\perp (Y^0, Y^1)|X$

This means that once conditioned on the set of variables X , having information on Y doesn't give us further information about the treatment status. The independence above might remind the reader of the consequence of having satisfied the backdoor criterium, in fact once we have a conditioning set X that satisfies it, this implies that the unconfounded assignment holds (Cunningham, 2021). The contrary is also true, unconfounded assignment and faithfulness imply that the backdoor criterium is satisfied. (Hernán & Robins, 2020)
this assumption is not testable (pag 261 IR) [should i add something about this]

3.2 Randomized studies

We often hear about double-blind randomized trials, that is where some of the patients are given the medicine and the other are given a placebo with neither the doctors nor the patients being aware of who received what. Such trial can be expressed through a DAG in this way:

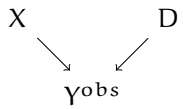


Figure 3.1: Dag for randomized trial

Why is this kind of trial become the golden standard for causal inference? What can such a trial truly tell us? Under the potential outcome model this kind of trial imply this independence:

$$D \perp\!\!\!\perp (Y^0, Y^1) \tag{3.2}$$

Of course this independence is stronger than the assumption 3.1, because we don't have to condition upon X; this is a natural consequence of the assignment mechanism being blind to covariates

$$Pr(D_i|X_i, Y_i^0, Y_i^1) = c$$

This also means that the potential outcomes didn't play a role in determining the treatment regime. However it **doesn't** imply that $D \perp\!\!\!\perp Y^{obs}$, in fact is patently false each time a medicine has an effect on a patient. By virtue of equation 3.2, we can state that $E[Y_i^1] = E[Y_i^1|D_i = 1]$ and the same for $E[Y_i^0] = E[Y_i^0|D_i = 0]$. We can assert that:

$$\begin{aligned} ATE &= E[Y_i^1 - Y_i^0] && (3.3)\\ &= E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 0] && (3.4)\\ &= E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 1] && (3.5)\\ &= E[Y_i^1|D_i = 0] - E[Y_i^0|D_i = 0]. && (3.6) \end{aligned}$$

In the equation 3.4 both quantities are *factual*, through the law of large number we can estimate both as means. We will define as SDO the simple difference in group means in the finite sample:

$$SDO := \frac{1}{N_1}\sum_{i:d_i=1} y_i - \frac{1}{N_2}\sum_{i:d_i=0} y_i \overset{(N_1,N_2)\to\infty}{=} ATE.$$

We can also conclude from equation 3.5 and 3.6 that in a randomized trial $ATE = ATT = ATU$.

3.2.1 Parte empirica

3.3 Observational studies

Observational studies are much different; researchers gather data that isn't intended to be randomized trials, were the researchers cannot intervene and impose the randomization, for two main reason :

1. Ethical or practicability concerns (e.g.: if we wanted to know whether smoking causes cancer)
2. The events might have already taken place (e.g.: if we wanted to study the effect of a certain policy)

The main difference is that $D \not\perp\!\!\!\perp (Y^0, Y^1)$, we can represent this too with a DAG:

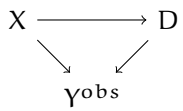


Figure 3.2: Dag per studio osservazionale

if exchengiabilty holds prop score IPW
matching
this part has to be flushed out more fully.

Bibliografia

Baldi, P. & Shahbaba, B. (2020). Bayesian causality. *The American Statistician* **74**, 249–257.

Cunningham, S. (2021). *Causal inference: The mixtape*. Yale university press.

Hernán, M. & Robins, J. (2020). *Causal Inference: What if. Boca Rat Chapman Hill/CRC; 2020*.

Imbens, G. W. & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Ramsey, J., Zhang, J. & Spirtes, P. L. (2012). Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843* .

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66**, 688.

Rubin, D. B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American statistical association* **75**, 591–593.