Causality
10716, Advanced ML
Pradeep Ravikumar (with some notes from Larry Wasserman)

Prediction and causation are very different. Typical questions are:

Prediction:	Predict Y after **observing** $X = x$
Causation:	Predict Y after **setting** $X = x$.

Causation involves predicting the effect of an intervention. For example:

Prediction:	Predict health given that a person takes vitamin C
Causation:	Predict health if I give a person vitamin C

The difference between passively observing $X = x$ and actively intervening and setting $X = x$ is significant and requires different techniques and, typically, much stronger assumptions. This is the area known as *causal inference*.

For years, causal inference was studied by statisticians, epidemiologists and economists. The machine learning community was largely uninterested. This has changed. The ML community now has an active research program in causation. This is because it is now recognized that many problems that were once treated as prediction problems are actually causal questions. Questions like: "If I place this ad on a web page, will people click on it?" and "If I recommend a product will people buy it?" are causal questions, not predictive questions.

1 Preliminaries

Before we jump into the details, there are a few general concepts to discuss.

1.1 Two Types of Causal Questions

There are two types of causal questions. The first deals with questions like this: do cell phones cause brain cancer? In this case, there are variables X and Y and we want to know the causal effect of X on Y. The challenges are: formalize the causal influence of X on Y via some parameter θ and find a way to estimate θ. This is usually what we mean when we refer to *causal inference*.

The second question is: given a set of variables, determine the causal relationship between the variables. This is called *causal discovery*.

1.2 Two Types of Data

Data can be from a controlled, randomized experiment (or more generally from *interventions*, which we will define shortly) or from an observational study. In the former, X is randomly set for the various subjects. In the latter, it is not randomly set. In randomized experiments, causal inference is straightforward. In observational (non-randomized) studies, the problem is much harder and requires stronger assumptions and also requires subject matter knowledge. Statistics and Machine Learning cannot solve causal problems without background knowledge.

1.3 Two Languages for Causation

There are two different mathematical languages for studying causation. The first is based on *potential outcomes*. The second is based on *structural causal models*. It will not seem obvious at first, but the two are mathematically equivalent (apart from some small details).

1.4 Example

Consider this story. A mother notices that tall kids have a higher reading level than short kids. The mother puts her small child on a device and stretches the child until he is tall. She is dismayed to find out that his reading level has not changed.

The mother is correct that height and reading skill are **associated**. Put another way, you can use height to predict reading skill. But that does not imply that height *causes* reading skill. This is what statisticians mean when they say:

> **correlation is not causation.**

On the other hand, consider smoking and lung cancer. We know that smoking and lung cancer are associated. But we also believe that smoking causes lung cancer. In this case, we recognize that intervening and forcing someone to smoke does change his probability of getting lung cancer.

1.5 Prediction Versus Causation

The difference between prediction (association/correlation) and causation is this: in prediction we are interested in

$$\mathbb{P}(Y \in A | X = x)$$

which means: the probability that $Y \in A$ given that we **observe** that X is equal to x. For causation we are interested in

$$\mathbb{P}(Y \in A | \mathsf{set}\ X = x)$$

which means: the probability that $Y \in A$ given that we **set** X equal to x. Prediction is about passive observation. Causation is about active intervention. The phrase **correlation is not causation** can be written mathematically as

$$\mathbb{P}(Y \in A | X = x) \neq \mathbb{P}(Y \in A | \mathsf{set}\ X = x).$$

Despite the fact that causation and association are different, people confuse them up all the time, even people trained in statistics and machine learning. On TV recently there was a report that good health is associated with getting seven hours of sleep. So far so good. Then the reporter goes on to say that, therefore, everyone should strive to sleep exactly seven hours so they will be healthy. Wrong. That's confusing causation and association. Another TV report pointed out a correlation between people who brush their teeth regularly and low rates of heart disease. An interesting correlation. Then the reporter (a doctor in this case) went on to urge people to brush their teeth to save their hearts. Wrong!

To avoid this confusion we need a way to discuss causation mathematically. That is, we need someway to make $\mathbb{P}(Y \in A | \mathsf{set}\ X = x)$ formal. We will be looking at two ways to do this: **potential outcomes**, and **structural causal models**. There are two different languages for saying the same thing.

Causal inference is tricky and should be used with great caution. The main messages are:

1. Causal effects can be estimated consistently from randomized experiments.
2. It is difficult to estimate causal effects from observational (non-randomized) experiments.
3. All causal conclusions from observational studies should be regarded as very tentative.

2 Introduction

So far we have studied statistical models. Given such a statistical model, one could use *probabilistic reasoning* to deduce likely observations and outcomes. And conversely, given a set of observations and outcomes, one could then use *statistical learning* to infer the likely underlying statistical model. But in the real world, we do not just have passively observed data, but also outcomes of active interventions. The counterpart of statistical models here is a causal model. We could use *causal reasoning* to infer the likely outcomes of interventions and changes to the environment, and conversely, use *causal learning* to learn such causal models given data comprising passive observations as well as actively intervened outcomes.

2.1 Interventions and Counterfactuals

The key distinction between statistical and causal setting arises from outcomes of *interventions*. An intervention comprises simply of setting a variable to a particular value. We often assume this is an *ideal intervention*: where we set the value of a specific variable without "setting" values of other variables immediately. There could still be *causal* effects consequently, which we are actually interested in measuring. Interventions might not be possible per se. For instance, we cannot simply instantaneously intervene on age of a person.

As we will see *counterfactuals* are a slightly more subtle notion: here, we do observe the value of the specific variable, but want to reason about the outcome if we set the variable to some other value. This results in some slightly different conclusions, as we will see in the sequel. But what is the connection between a causal model and a statistical model? The following important principle could be viewed as providing one such link.

2.2 Hierarchy of Models

We thus have the following hierarchy of mathematical models for reasoning in complex environments. A statistical model can answer observational queries, but not interventional, or counterfactual queries.
A causal graphical model can answer both observational and interventional queries, but not counterfactual ones.
A structural causal model can answer observational, interventional, as well as counterfactual queries.
Mechanistic or Physical Models can not only all of above queries, but moreover have components that can be mapped to the real world, so that they additionally provide "scientific/physical insight".

2.3 Causal vs Statistical Models

Thus, a causal model could be viewed as subsuming a statistical model, just as the corresponding set of outcomes, both passive and active, subsumes the set of passively observed outcomes.

Definition 1 (Reichenbach's Common Cause Principle) *If two random variables X and Y are statistically dependent, then there exists a third random variable Z that is a common cause (possibly coinciding with either X or Y), such that $X \perp\!\!\!\perp Y \,|\, Z$.*

Remark: Spurious Dependences. Note that in some cases where we only observe the random variables via a finite set of samples from their corresponding distributions, it is possible for there to be a spurious observed dependence: the principle above then does not apply. When we do not have iid data, and instead see samples from two time-varying stochastic processes, then it might seem that the two variables are dependent. This is for instance what leads people to claim analogues of the increasing number of Shrek movies being associated with global warming. In such a case, we could view time itself as a common cause. In certain cases, the samples we observe of X and Y are implicitly conditioned on a specific value of some other variable Z. When this conditioning variable is a downstream "effect" of X and Y, then we know that this results in observed conditional dependence of X and Y: this is also called selection bias. In this case again the Reichenbach's principle does not apply, since we do not observe marginal dependence of X and Y.

Time and Causality. In certain accounts, causality is intrinsically associated with time in the sense that the effect variable is observed after the cause variable. However, it is not necessary we incorporate time in our causal modeling machinery. It could be that the observations cannot be mapped to specific time, or might even be equilibrium observations. For instance, consider motivation as cause, and grades as the effect: the former cannot be mapped to a specfic time instance. Mapping to time is most common in hard sciences such as physics and chemistry.

3 Independent Mechanisms

Consider the two variables: altitude, denoted by A, and temperature denoted by T. By the chain rule, we can express $P(A, T) = P(A)P(T|A)$ or as $P(A, T) = P(T)P(A|T)$. Suppose we are interested in determining which is the cause and which is the effect among these two variables. One way to do so is to intervene on either A or T, and observe the effects. If A were the cause, intervening on A would mean we observe changes in the effect T. On the other hand, if A were the effect, intervening on A would mean we do not observe changes to the cause T. To formalize this, we need a notion of "independence of mechanisms".

One way to restate above is that if A were the cause, we would expect that $P(A)$ and $P(T|A)$ are "independent mechanisms". We will formalize this in the sequel, but loosely, even if we were to change the altitude (or the distribution of altitudes) of a city/country, we would expect that the mechanism $P(T|A)$, which specifies how temperatures are affected by altitudes, to **not change**. On the other hand, if T were the cause, we would expect that $P(T)$ and $P(A|T)$ are independent mechanisms. Thus, if we were to somehow change the temperature of the city, the mechanism specifying its altitude given the temperature should not be affected. But the mechanism specifying the temperature given the altitude could be affected.

In other words, if A is the cause, we can perform a localized intervention on A, i.e. change $P(A)$ without affecting $P(T|A)$: $P(A)$ and $P(T|A)$ are modular, autonomous, invariant mechanisms. Whereas we would get different autonomous mechanisms if T were the cause.

Suppose we have data from multple countries, each with different distributions of altitudes and temperatures. Then just from observational data, we could check if $P(T|A)$ or $P(A|T)$ is an invariant mechanism (i.e. same for different $P(A)$ or $P(T)$), which in turn could help us find the causal direction. We will return to this when we study learning of causal models from observational data.

Definition 2 (Principle of Independent Mechanisms) *A Causal Generative Process consists of autonomous models that do not inform or influence each other. When the causal generative process specifies a joint probability distribution, the conditional distributions of effects given immediate causes do not inform or influence other conditional distributions.*

In the two variable case, the principle above is also referred to as independence of cause and mechanism. This principle has three facets.

- We should be able to change one mechanism (*intervene*) without affecting others. In other words, there is no pathway connecting different mechanisms via some "meta mechanisms". Thus, mechanisms are invariant to *changes in other mechanisms*. Such autonomy is critical for transfer of knowledge from one domain to another. In a new domain, most if not all of the mechanisms are the same, and hence can directly transfer.

- Each mechanism should not provide "information" about other mechanisms. One aspects of this is with respect to changes: change in one mechanism should not provide information on how other mechanisms have changed. But there should be no information flow even in the absence of changes.

- Suppose the conditional distributions can be specified deterministically given the set of observed random variables, and additional noise variables. In the two variable, cause-effect model, case, suppose that C, E are two variables, with conditional distributions specified as:

  $$\begin{aligned} C &= N_C \\ E &= f_E(C, N_E), \end{aligned}$$

  where N_C, N_E are additional noise variables, and f_E is a deterministic function. Then N_C, N_E are statistically independent.

  It can be seen that such statistical independence is necessary for independence of mechanisms.

> To see this, note that the noise variable N acts as a gate, whose value specifies one of many deterministic mechanisms, which we can rewrite as $E = f_N(C)$. So if N is dependent on some other noise variable M for some other mechanism say $E' = g_M(C)$, then there is information leakage between different mechanisms.

The third facet is something we have a good understanding of, namely statistical independence of noise variables. But how do we formalize the first two facets which need some way of specifying independence of mechanisms (rather than random variables). We will see that in the sequel.

4 Cause Effect Models

It is instructive to first study causal models in the simplest possible setting with just two variables: a cause variable C and an effect variable E. This would allow us to isolate the additional facets of a causal model beyond that of a statistical model.

A common approach of specifying a causal model is via a so-called Structural Causal Model (SCM):

$$\begin{aligned} C &= N_C \\ E &= f_E(C, N_E), \end{aligned}$$

where N_C, N_E are independent noise variables, and f_E is a deterministic function. This is associated with a "causal graph" with nodes $\{C, E\}$ and a directed edge $C \to E$.

4.1 Interventions

When we intervene on a variable, say E, we simply substitute its existing causal mechanism ($E = f_E(C, N_E)$),for a substitute intervened one.

The simplest such intervention is where the substitute mechanism is simply setting E to a constant. This is referred to as a **hard intervention**, and we will denote the resulting causal model as $P^{do(E=e)}$. This is to be contrasted with the more general **soft interventions**, where the substitute mechanism could be a more general e.g. $E = g(C, \widetilde{N}_E)$. We will denote the resulting intervened causal model via $P^{do(E=g(C,\widetilde{N}_E))}$.

If the causal graph is $C \to E$, when we intervene on the effect variable, we do not expect the cause mechanism to change. Accordingly,

$$P_C^{do(E=e)} = P_C \neq P_{C|E=e}.$$

This clearly shows that intervening is not the same as conditioning. But on the other hand when we intervene on the cause variable:

$$P_E^{do(C=c)} = P_{E|C=c} \neq P_E.$$

Moreover, in the effect-intervened causal model $P_{C,E}^{do(E=\tilde{N}_E)}$, we have that $C \perp\!\!\!\perp E$, but which does not hold in the cause-intervened causal model: $P_{C,E}^{do(C=\tilde{N}_C)}$.

4.2 Counterfactuals

Suppose we have a disease that could result in blindness, and a candidate treatment for this disease. Let T denote the binary variable on whether or not to administer the treatment and let B denote the binary variable on whether or not the disease causes blindness. Now suppose that for 99% of patients, the treatment leads to a cure, and if not treated, they get blind. While for the remaining 1% of patients, the treatment leads to blindness, and if not treated, they get cured. So any patient belongs to one of these two categories, which in turn depends on a condition $N_B \in \{0,1\}$ which is unknown to the doctor. We thus have the following causal model:

$$\begin{aligned} T &= N_T \\ B &= T\,N_B + (1-T)(1-N_B), \end{aligned}$$

where $N_B \sim \text{Ber}(0.01)$.

Suppose somebody gets the treatment, but are blinded. A natural question would then be: what would have happened if we had not given that person the treatment?

This is a *counterfactual* question. Note that unlike the intervention case, here we do observe the values of the variables: $T = 1, B = 1$. Plugging these into the SCM above, we get that $N_T = 1, N_B = 1$. Suppose we plug these values of the noise variables back into the SCM: this then yields the **counterfactual SCM**:

$$\begin{aligned} T &= 1 \\ B &= T + (1-T)0 = T. \end{aligned}$$

Let us denote this counterfactual SCM by $\mathcal{C}'$. The counterfactual question asked above is then reasoning with an intervention on this counterfactual SCM: $\mathcal{C}'^{do(T=0)}$. This can be seen to be the SCM: is then the SCM:

$$\begin{aligned} T &= 0 \\ B &= T, \end{aligned}$$

which places all mass on $(0, 0)$. So the answer to the counterfactual query is that the patient would have been cured if not treated!

Does that mean the doctor commited malpractice? Note however that the condition (affecting the noise variable N_B) is unknown apriori. All the doctor could thus use apriori are the *interventional probabilities*:

$$\begin{aligned} P^{do(T=1)}(B = 0) &= P(N_B = 0) = 0.99 \\ P^{do(T=0)}(B = 0) &= P(N_B = 1) = 0.01, \end{aligned}$$

which does warrant the doctor treating the patient.

So for answering counterfactual questions, we first set up a counterfactual SCM by inferring noise variables given the observed values of the observed variables, and then use intervention based causal reasoning with this SCM. A consequence of this is that two SCMs with *the same interventional probabilities* can lead to different counterfactual probabilities.

Consider the following SCM:

$$\begin{aligned} C &= N_C \\ E &= n_E(C), \end{aligned}$$

where without loss of generality we have combined the functional and noise elements of the effect mechanism in the earlier SCM into a random function n_E . Suppose C takes values in the finite set $\mathcal{C} = [k]$, and E in some other finite set $\mathbb{E}$. Then any deterministic function g from $\mathcal{C}$ to $\mathbb{E}$ can be associated with the vector $(g(1), \ldots, g(k)) \in \mathbb{E}^k$. Thus the set of random functions g is associated with a *random vector* $(g(1), \ldots, g(k))$. Since C is the cause, and E is the effect, $P_E^{do(C=j)} = P_{E|C=j} = P_{g(j)}$, so that the observational and interventional distributions (when we intervene on the cause) coincide, and moreover only depend on the *marginal distributions* $P_{g(j)}$ of the random noise function g.

Suppose we observe some values (j, e) of the cause and effect variables. And suppose as before we identify n_E with a random vector $(g(1), \ldots, g(k))$. Then, in the counterfactual SCM above:

$$\begin{aligned} C &= j \\ E &\sim g(C) \,|\, g(j) = e \end{aligned}$$

Now consider the counterfactual query of what would have happened were $C = j'$ instead. This could be answered by $P_E^{do(C=j')} = P_{E|C=j'} = P_{g(j')|g(j)=e}$, which explicitly involves the joint distribution of the noise distribution g.

5 Causality: Connections to Unsupervised Learning

In the Unsupervised Learning task of clustering, we are given (samples based access to) the distribution P_X of some variables X, and are asked to infer $P_{Y|X}$ of some "cluster" label Y. Note that we do not have any access, sampling based or otherwise, to P_Y or $P_{X,Y}$.

Suppose X is the cause, and Y is the effect. Then by independence of causal mechanisms, P_X has no information about $P_{Y|X}$. So unsupervised learning "in a causal direction" is not possible.

On the other hand, suppose X is the effect, and Y is the cause. then P_X may have info about $P_{Y|X}$. As an example, suppose $Y \in \{-1, 1\}$, and $X = Y\,\mu + N_X$ where $N_X \sim \mathcal{N}(0, 1)$, then, P_X is a Gaussian mixture model, from which can recover the Gaussian components $P_{X|Y}$ and hence via Bayes rule $P_{Y|X}$.

Remark. Note that if we aim to obtain an optimal decision function $f^* \in \mathcal{F}$, and optimize out $P_{Y|X}$ over a set $\mathcal{P}$, wrt some optimality criterion, then this optimal f^* by definition will depend on P_X.

As an example, suppose we use a Bayesian optimality criertion, with respect to some prior π over distributions $P_{Y|X}$. Then solving for:

$$\arg \inf_{f in \mathcal{F}} \mathbb{E}_{P_{Y|X} \sim \pi} \textsc{loss}(f, P_{Y|X}, P_X),$$

clearly only depends on P_X.

This is also the case with a minimax optimality criterion:

$$\arg \inf_{f \in \mathcal{F}} \sup_{P_{Y|X} \in \mathcal{P}} \textsc{loss}(f, P_{Y|X}, P_X),$$

clearly only depends on P_X.

Thus the key point is not whether the decision function of interest depends only on P_X, but rather is an informational statement about P_X itself: whether $P_{Y|X}$ is idenfitiable given P_X.

6 Causality: Connections to Domain Adaptation

Suppose X is the cause, and Y the effect. Then even if we change P_X (covariate shift), the "autonomous mechanism" $P_{Y|X}$ need not change, and hence could be used even in the covariate shifted domain. Of course, it is still possible that $P_{Y|X}$ also changes (independent of the changes in P_X), but by the independence principle, the new P'_X has no info about $P'_{Y|X}$ so we might as well use the old $P_{Y|X}$ in the absence of further information.

The above is only justified in the causal direction. Whereas, in the anti-causal direction, any change in P_X could also entail changes in $P_{Y|X}$, and indeed we could even aim to estimate this change. In one extreme, P'_X could make $P'_{Y|X}$ identifiable, so we could use unsupervised learning to infer this.

Causality seems to have a lot to say about domain adaptation and transfer problems, but much is still open.

7 General Structural Causal Models (SCMs)

An SCM $\mathcal{C}$ consists of a DAG G, and a collection of "structural assignments":

$$X_j = f_j(\mathrm{PA}_j, N_j),$$

where PA_j are parents of X_j in G, and $(N_1, ..., N_p)$ are independent noise terms. Each such assignment is simply an alternative characterization of the conditional distribution of X_j given PA_j. The variables PA_j are called direct causes of X_j, and X_j is a direct effect of its causes. SCMs are also called Structural Equation Models (SEMs). When $f_j(\cdot)$ are non-linear, these are called non-linear SCMs/SEMs.

Proposition 3 *An SCM entails a unique joint distribution over* $(X_1, ..., X_p)$.

This just follows from how in the case of DGMs, the conditional distributions of variables given their parents specifies a unique joint distribution that is simply the product of these node conditional distributions.

Remark. SCMs are still not "mechanistic" enough for "scientific theories", many of which for instance take the form of PDEs, with components corresponding to physical quantities.

But one can analyze the behavior of these PDEs in their equilibrium state, which can then yield an SCM (Dash 2005, Hansen & Sokol 2014).

7.1 Interventions

Given an SCM $\mathcal{C}$ with DAG G, suppose we replace one or more structural assignments to obtain a new SCM $\widetilde{\mathcal{C}}$:

$$X_k = \widetilde{f}(\widetilde{PA}_k, \widetilde{N}_k).$$

We then call this an intervention SCM, and the resulting distribution $P^{\widetilde{\mathcal{C}}}$ as the intervention distribution. Variables whose structural assignments are replaced are said to be intervened on.

We denote this as:

$$P^{\widetilde{\mathcal{C}}} = P^{do(X_k:=\widetilde{f}(\widetilde{\mathrm{PA}}_k,\widetilde{N}_k))}.$$

When $\widetilde{f}(\widetilde{PA}_k, \widetilde{N}_k)$ places a point mass on a constant a, we simply write:

$$P^{\widetilde{\mathcal{C}}} = P^{do(X_k:=a}.$$

Example: Consider the DAG $X_1 \to Y \to X_2$, and the structural assignments:

$$\begin{aligned} X_1 &= N_{X_1} \\ Y &= X_1 + N_Y \\ X_2 &= Y + N_{X_2}, \end{aligned}$$

where $N_{X_1}, N_Y \sim N(0,1)$, and $N_{X_2} \sim N(0, 0.1)$ are all independent noise terms. Suppose we are interested in predicting Y given (X_1, X_2). Clearly X_2 is a better predictor of Y: it has lower variance, and hence a linear model with X_2 has much lower MSE than that with X_1. However X_2 is useless wrt interventions:

$$P_Y^{\mathcal{C};do(X_2=\widetilde{N})} = P_Y^{\mathcal{C}}.$$

Note in particular that intervention is distinct from conditioning:

$$P_Y^{\mathcal{C};do(X_2=x)} = P_Y^{\mathcal{C}} \neq P_{Y|X_2=x}^{\mathcal{C}}.$$

On the other hand:

$$P_Y^{\mathcal{C};do(X_1=\widetilde{N})} \neq P_Y^{\mathcal{C}}.$$

7.2 Total Causal Effect

Given an SCM $\mathcal{C}$, there is a total causal effect from X to Y iff

$$X \not\perp\!\!\!\perp Y \text{ in } P_X^{do(X=\widetilde{N}_X)},$$

for some RV $\widetilde{N}_X$.

Proposition 4 *Given an SCM $\mathcal{C}$, the following statements are equivalent:*

- *There is a total causal effect from X to Y.*
- $X \not\perp\!\!\!\perp Y$ *in* $P_X^{do(X=\widetilde{N}_X)}$ *for any RV* $\widetilde{N}_X$ *with full support*
- *There exists* $x, x' \in \mathcal{X}$ *s.t.* $P_Y^{do(X=x)} \neq P_Y^{do(X=x')}$
- *There exists* $x \in \mathcal{X}$ *s.t.* $P_Y^{do(X=x)} \neq P_Y^{\mathcal{C}}$.

Remark. If there is no directed path from X to Y, then clearly there is no total causal effect. On the other hand, even if there is a directed path, there could be no causal effect (due to cancellations).

8 Counterfactuals

Consider an SCM $\mathcal{C}$ with DAG G over a random vector X. Given discrete observations x, the counterfactual SCM, denoted by $\mathcal{C}_{X=x}$ is the same set of structural assignments, but which the noise variables having distribution $P_{N|X=x}$, where P_N is the noise distribution in the original SCM $\mathcal{C}$. Note that in this counterfactual SCM, the new set of noise variables need not be independent anymore.

Consider an SCM $\mathcal{C}_1$:

$$\begin{aligned}
&X_1 = N_1\\
&X_2 = N_2\\
&X_3 = (I[N_3 > 0]X_1 + I[N_3 = 0]X_2)I(X_1 \neq X_2) + N_3 I[X_1 = X_2].
\end{aligned}$$

And a slightly different SCM $\mathcal{C}_2$:

$$\begin{aligned}
&X_1 = N_1\\
&X_2 = N_2\\
&X_3 = (I[N_3 > 0]X_1 + I[N_3 = 0]X_2)I(X_1 \neq X_2) + (2 - N_3)I[X_1 = X_2].
\end{aligned}$$

And suppose $N_3 \sim \text{Unif}[0, 2]$. Then it can seen that we also have $2 - N_3 \sim \text{Unif}[0, 2]$. Thus, both SCMs entail the same observation distribution $P(X)$. They also correspond to the same causal graphical model, since the distribution of each node given their parents is the same in both cases. They have the same intervention distributions as well. But they have different counterfactual distributions.

Suppose we observe $(X_1, X_2, X_3) = (1, 0, 0)$. It then follows that $N_1 = 1, N_2 = 0, N_3 = 0$. The corresponding counterfactual SCM $\mathcal{C}_{1;X=(1,0,0)}$ is given as:

$$\begin{aligned} X_1 &= 1 \\ X_2 &= 0 \\ X_3 &= X_2 I(X_1 \neq X_2) \end{aligned}$$

while the counterfactual SCM $\mathcal{C}_{2;X=(1,0,0)}$ is given as:

$$\begin{aligned} X_1 &= 1 \\ X_2 &= 0 \\ X_3 &= X_2 I(X_1 \neq X_2) + 2I[X_1 = X_3]. \end{aligned}$$

It can be seen that the counterfactual SCMs are different. In particular, if we were to ask the counterfactual query of consequences if we were to change $X_1 to 0$, we would get different answers from the two counterfactual SCMs.

8.1 Falsifiability

One could verify observational distributions via observational data. And verify interventional distributions via interventions i.e. randomized experiments. But there is no counterpart of counterfactual distributions in the real world.

It is possible in certain cases for counterfactual distributions to be falsifiable. For instance, if we can observe the specific noise samples as entailed by the counterfactual distribution (e.g. imagine if the noise variables could be obtained via some measurement). One could also falsify the counterfactual SCM by drawing upon scientific theories or domain knowledge. But in general, it is not falsifiable. That said, humans often think in terms of counterfactuals, and indeed, counterfactuals have occured in literature and philosophy throughout human history. What if I was on the plane that crashed, rather than the plane that left an hour later? What if I had picked the winning lottery numbers? And so on.

8.2 Equivalence of Causal Models

Definition 5 *We say two models as probabilistically/interventionally/counterfactually equivalent if they entail the same obs./obs. and interv./obs., interv., and counterfactual distributions.*

It turns out that for a pair of positive distributions, for them to be interventionally equivalent it suffices for them to agree on simple single-node interventions where $X_k = \widetilde{N}_k$ for some independent noise distribution (rather than an entirely different SCM component that could depend on some other subsets of nodes). This is convenient because such interventions are easier to setup in the real world, and are also mathematically easier.

Proposition 6 *Suppose that two SCMs $\mathcal{C}_1, \mathcal{C}_2$ with strictly positive SCM components, satisfy:*

$$P_X^{\mathcal{C}_1;do(X_j=\widetilde{N}_j)} = P_X^{\mathcal{C}_2;do(X_j=\widetilde{N}_j)},$$

for all $j \in [p]$, and all distributions $\widetilde{N}_j$ with full support. Then $\mathcal{C}_1$ and $\mathcal{C}_2$ are fully interventionally equivalent.

9 SCM Calculus

While the truncated factorization in (**??**) specifies a new intervention SCM, and one would need to perform probabilistic reasoning on this new SCM to derive conditional probabilities of interest. But could we directly express the conditional probabilities in the intervention SCM in terms of conditional probabilities in the original SCM?

Definition 7 (Confounding) *Consider an SCM $\mathcal{C}$, with a directed path from X to Y, for some nodes $X, Y \in V$. The causal effect from X to Y is said to be* ***confounded*** *if:*

$$P^{\mathcal{C};do(X:=x)}(y) \neq P^{\mathcal{C})}(y).$$

Otherwise the causal effect is said to ***unconfounded****.*

Thus, there exist some confounding variables that account for the dependence between X and Y. Consider the following simple instance of such confounding.

Example 8 (Kidney Stones) *Suppose somebody has kidney stones for which they seek treatment. Let Z be a RV that denotes the size of the kidney stones, T a binary treatment RV (with $T = 0$ indicating one treatment, and $T = 1$indicating the other), and R a binary recovery RV (with $R = 1$ indicating recovery, and $R = 0$ otherwise.) We are interested in measuring the average causal effect between the treatments:*

$$P^{\mathcal{C};do(T=1)}(R = 1) - P^{\mathcal{C};do(T=2)}(R = 1).$$

Note that this is different from the difference between the conditionals:

$$P^{\mathcal{C}}(R = 1|T = 1) - P^{\mathcal{C}}(R = 1|T = 2).$$

But we can connect some *conditionals in $P^{\mathcal{C};do(T=t)}$ to conditionals in $P^{\mathcal{C}}$. For instance, it can be seen from inspecting the intervention SCM that:*

$$\begin{aligned} P^{\mathcal{C};do(T=1)}(R=1|T=1, Z=z) &= P^{\mathcal{C}}(R=1|T=1, Z=z \\ P^{\mathcal{C};do(T=1)}(Z=z) &= P^{\mathcal{C}}(Z=z). \end{aligned}$$

This thus allows us to perform the following calculation:

$$\begin{aligned} P^{\mathcal{C};do(T=1)}(R=1) &= \sum_z P^{\mathcal{C};do(T=1)}(R=1, T=1, Z=z) \\ &= \sum_z P^{\mathcal{C};do(T=1)}(R=1|T=1, Z=z)P^{\mathcal{C};do(T=1)}(T=1, Z=z) \\ &= \sum_z P^{\mathcal{C};do(T=1)}(R=1|T=1, Z=z)P^{\mathcal{C};do(T=1)}(Z=z) \\ &= \sum_z P^{\mathcal{C}}(R=1|T=1, Z=z)P^{\mathcal{C}}(Z=z), \end{aligned}$$

In the example above, by adjusting for Z, we could compute the effect of the treatment entirely from observational conditional probabilities. This leads us to the following concept.

Definition 9 (Valid Adjustment Set) *Consider an SCM $\mathcal{C}$, and two nodes $X, Y \in V$, where $Y \notin \mathrm{PA}_X$. We call a set $\mathbf{Z}$ a **valid adjustment set** for the ordered pair (X, Y) if:*

$$P^{\mathcal{C};do(X:=x)}(y) = \sum_{\mathbf{z}} P^{\mathcal{C})}(y|x, \mathbf{z})P^{\mathcal{C}}(\mathbf{z}).$$

Let us now perform a calculation similar to the kidney stones example. For any set of nodes $\mathbf{Z} \subseteq V$, we have that:

$$\begin{aligned} P^{\mathcal{C};do(X:=x)}(y) &= \sum_{\mathbf{z}} P^{\mathcal{C};do(X:=x)}(y, \mathbf{z}) \\ &= \sum_{\mathbf{z}} P^{\mathcal{C};do(X:=x)}(y|x, \mathbf{z})P^{\mathcal{C};do(X:=x)}(\mathbf{z}). \end{aligned}$$

So, if the set of nodes $\mathbf{Z}$ satisfies:

$$\begin{aligned} P^{\mathcal{C};do(X:=x)}(y|x, \mathbf{z}) &= P^{\mathcal{C}}(y|x, \mathbf{z}) \\ P^{\mathcal{C};do(X:=x)}(\mathbf{z}) &= P^{\mathcal{C}}(\mathbf{z}), \end{aligned} \tag{1}$$

it then follows that $\mathbf{Z}$ is a valid adjustment set.

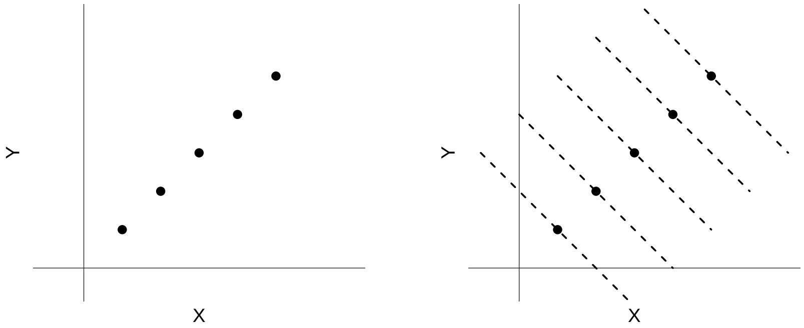

Figure 1: *Left: X and Y have positive association. Right: The lines are the potential outcomes, i.e. what would happen to each person if I changed their X value. Despite the positive association, the causal effect is negative. If we increase X everyone's Y values will decrease.*

10 Potential Outcomes

Suppose that X is a binary variable that represents some exposure. So $X = 1$ means the subject was exposed and $X = 0$ means the subject was not exposed. And Y is some "outcome" variable measuring how well the treatment worked.

We can address the problem of predicting Y from X by estimating $P(Y|X = x)$. But that does not get at causal dependence between X and Y. Let Y_1 denote the response if the subject is exposed. Let Y_0 denote the response if the subject is not exposed. If we expose a subject, we observe Y_1 but we do not observe Y_0. Instead, Y_0 is the value we would have observed if the subject had NOT been exposed.

Thus,

$$Y = \begin{cases} Y_1 & \text{if } X = 1 \\ Y_0 & \text{if } X = 0. \end{cases}$$

More succinctly

$$Y = XY_1 + (1 - X)Y_0. \tag{2}$$

The variables (Y_0, Y_1) are also called *potential outcomes.*

We have replaced the random variables (X, Y) with the more detailed variables (X, Y_0, Y_1, Y) where $Y = XY_1 + (1 - X)Y_0$.

A small dataset might look like this:

X	Y	Y_0	Y_1
1	1	*	1
1	1	*	1
1	0	*	0
1	1	*	1
0	1	1	*
0	0	0	*
0	1	1	*
0	1	1	*

It is important to keep in mind here that each row corresponds to a different subject. Thus, $X = 1$ in any row indicates that subject was given the treatment, and $X = 0$ indicates they were not. The asterisks indicate unobserved variables. So, for those subjects for whom $X = 1$, for *those specific subjects*, we only observe their Y_1 value, and do not observe their Y_0 value.

Causal questions involve the the distribution $P(Y_0, Y_1)$ of the potential outcomes. For instance, the treatment could be said to be effective if $\mathbb{E}[Y_1] - \mathbb{E}[Y_0]$ is large.

The *mean treatment effect* or *mean causal effect* is defined by

$$\theta = \mathbb{E}(Y_1) - \mathbb{E}(Y_0) = \mathbb{E}(Y|\mathsf{set}\ X = 1) - \mathbb{E}(Y|\mathsf{set}\ X = 0).$$

The parameter θ has the following interpretation: θ is the mean response if we exposed everyone minus the mean response if we exposed no-one.

Lemma 10 *In general,*

$$\mathbb{E}[Y_1] \neq \mathbb{E}[Y|X = 1] \quad \text{and} \quad \mathbb{E}[Y_0] \neq \mathbb{E}[Y|X = 0].$$

To see this, note that since $Y = XY_1 + (1 - X)Y_0$, we have that $\mathbb{E}[Y|X = 1] = \mathbb{E}[Y_1|X = 1] \neq \mathbb{E}[Y_1]$ unless for instance Y_1 is independent of X.

Suppose now that we observe a sample $(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)})$. Can we estimate θ? In general the answer is no. We can estimate

$$\alpha = \mathbb{E}(Y|X = 1) - \mathbb{E}(Y|X = 0)$$

but α is not equal to θ. Quantities like $\mathbb{E}(Y|X = 1)$ and $\mathbb{E}(Y|X = 0)$ are predictive parameters. These are things that are commonly estimated in statistics and machine learning.

Let's formalize this. Let $\mathcal{P}$ be the set of distributions for (X, Y_0, Y_1, Y) such that $P(X = 0) > \delta$ and $P(X = 1) > \delta$ for some $\delta > 0$. (We have no hope if we do not have positive probability of observing exposed and unexposed subjects.) Recall that $Y = XY_1 + (1-X)Y_0$.

The observed data are $(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)}) \sim P$. Let $\theta(P) = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$. An estimator is uniformly consistent if, for every $\epsilon > 0$,

$$\sup_{P \in \mathcal{P}} P(|\widehat{\theta}_n - \theta(P)| > \epsilon) \to 0$$

as $n \to \infty$.

Theorem 11 *In general, there does not exist a uniformly consistent estimator of* θ.

Proof. It is easy construct $P(X, Y_0, Y_1)$ and and $Q(X, Y_0, Y_1)$ such that $\theta(P) \neq \theta(Q)$ and yet $P(X, Y) = Q(X, Y)$. □

In the case that X is continuous, the causal quantity (or rather, an example of a causal quantity) is

$$\theta(x) = \mathbb{E}[Y(x)]$$

which, in general, is NOT equal to $m(x) = \mathbb{E}[Y|X = x]$.

Thus, in the potential outcomes view, causal analysis involves reasoning about some random variables — potential outcomes — which by definition are rife with missingness. How do we relate this to causal SCMs, and the causal graph approach we have studied so far?

Given an SCM $\mathcal{C}$, we can interpret Y_1 as $Y^{\mathcal{C};do(X=1)}$ and Y_0 as $Y^{\mathcal{C};do(X=0)}$ We can verify that for any SCM $\mathcal{C}$:

$$Y^{\mathcal{C}} = Y^{\mathcal{C};do(X=1)}\mathbb{I}[X^{\mathcal{C}} = 1] + Y^{\mathcal{C};do(X=0)}\mathbb{I}[X^{\mathcal{C}} = 0].$$

But there are other interpretations as well, since the potential outcomes literature focuses on *individual subjects*. Pearl (2009) thus suggests that any individual subject u can be associated with some noise variables N_u, so that we can consider the resulting counterfactual SCM $\mathcal{C}_{N=N_u}$. The potential outcomes for subject u can then be specified as:

$$\begin{aligned} Y_0(u) &= Y^{\mathcal{C}_{N=N_u};do(X=0)} \\ Y_1(u) &= Y^{\mathcal{C}_{N=N_u};do(X=1)}, \end{aligned}$$

but specializes our earlier specification to an individual for whom treatment outcomes for instance could be deterministic.

But even if you find this specific connection convoluted, it is instructive to look at how potential outcomes are measured given data.

One approach is to use a randomized trial, where we ensure that $X \perp\!\!\!\perp (Y_0, Y_1)$: since who gets treatment is chosen completely independently (for instance by uniformly random assignment) of how they may fare given the treatment. We can then simply write: $P(Y_1) = P(Y_1|X = 1) = P(Y|X = 1)$. And similarly, $P(Y_0) = P(Y|X = 1)$, so that we could estimate the potential outcome probabilities just from conditionals over the observed RVs. This is why randomized control trials are used for measuring causal effects.

But suppose X was not chosen independent of Y_0, Y_1, for instance, we only had access to observational data. But suppose we knew all the "confounding" variables $\mathbf{Z}$ such that:

$$\begin{aligned} X &\perp\!\!\!\perp (Y_0, Y_1) \,|\, \mathbf{Z} \\ (\mathbf{Z}_0, \mathbf{Z}_1) &\sim_d \mathbf{Z} \end{aligned}$$

Thus conditioned on the confounding variables, the treatment outcomes are independent of who gets the treatment. An example of such confounding variables could be age, or socio-economic status. We also assume that intervening on the treatment would not have affected the distribution of the confounding variables (e.g. we would not affect the age by deciding to give somebody the treatment!). Given such a confounding set, we then have:

$$\begin{aligned} P(Y_1) &= \sum_z P(Y_1|\mathbf{Z}_1 = z)P(\mathbf{Z}_1 = z) \\ &= \sum_z P(Y_1|X = 1, \mathbf{Z}_1 = z)P(\mathbf{Z} = z) \\ &= \sum_z P(Y|X = 1, Z = z)P(Z = z), \end{aligned}$$

where the second equality used the independence of X, as well as that $\mathbf{Z}_1 \sim_d \mathbf{Z}$, while the last equality simply used the definition of potential outcomes. This can be seen to be precisely the computation using a valid adjustment set, and the conditions above are precisely those specifying valid adjustment sets. Thus, causal SCMs and potential outcomes result in the same calculations under the same assumptions.

One can even show that theorems that hold in one framework also hold under equivalent assumptions in the other framework. Nonetheless, they might be easier to prove in one framework vs the other. Potential outcomes seem preferable for a smaller set of discrete variables, while SCMs might be preferable for larger scale settings. Potential outcomes also focus on assumptions that relate interventional quantities to observed conditionals, while SCMs allow for more complex causal reasoning.

11 Algorithmic Notion of Causality

So far we have discussed conditional independence properties among random variables. But if we are to ask that causal mechanisms i.e. conditional distributions of nodes given their parents are to be "independent," it might seem we lack the technical tools to define what we mean when we say that two *conditional distributions* (in contrast to random variables) are independent.

Consider objects (not necessarily random variables) from some set Ω. And assume that we have access to an "information" functional:

$$R : 2^{\Omega} \mapsto \mathbb{R},$$

that given a set of objects, can quantify the amount of information in this set. For any two sets $\mathbf{x}, \mathbf{y} \subseteq \Omega$, we would have that: $R(\mathbf{x}, \mathbf{y}) \geq R(\mathbf{y})$, so that R is monotone. We could thus interpret: $R(\mathbf{x} \,|\, \mathbf{y}) = R(\mathbf{x}, \mathbf{y}) - R(\mathbf{y})$ as the conditional information in $\mathbf{x}$ given $\mathbf{y}$. Similarly we can define a counterpart of mutual information:

$$I(\mathbf{x}; \mathbf{y}) = R(\mathbf{x}) + R(\mathbf{y}) - R(\mathbf{x}, \mathbf{y}),$$

as well as conditional mutual information:

$$I(\mathbf{x}; \mathbf{y} \,|\, \mathbf{z}) = R(\mathbf{x}, \mathbf{z}) + R(\mathbf{y}, \mathbf{z}) - R(\mathbf{x}, \mathbf{y}, \mathbf{z}) - R(\mathbf{z}).$$

We can also define generalized SCMs over a set of objects $(x_1, \dots, x_p)$ by requiring that a node $\mathbf{x}_j$ not contain more information than its parents $\textsc{pa}_j$ and an unobserved independent noise object n_j:

$$R(x_j, \textsc{pa}_j, n_j) = R(\textsc{pa}_j, n_j),$$

and further that the noise objects are independent:

$$R(n_1, \dots, n_p) = \sum_{j=1}^{p} R(n_j).$$

This can also be stated simply as:

$$I(n_1, \dots, n_p) = 0.$$

Janzing and Scholkopf (2010) suggest the use of Kolmogorov complexity as the notion of information above, to derive an "algorithmic model of causality". This might be suitable to cases where the data consists of non-stationary perhaps even deterministic time-series, for which such algorithmic notions of dependence might be better suited.

11.1 Algorithmic Independence of Conditionals

Given such an algorithmic notion of dependence, we can now formalize one notion of "independence" of causal mechanisms. We say that an SCM has algorithmically independent conditionals if:

$$I(P_{X_1 \,|\, \textsc{pa}_1}, \dots, P_{X_p \,|\, \textsc{pa}_p}) = 0.$$

References

Two excellent books (which the lecture notes draw from):

Elements of Causal Inference, Foundations and Learning Algorithms, Jonas Peters, Dominik Janzing and Bernhard Scholkopf, available for free: https://www.dropbox.com/s/gkmsow492w3oolt/11283.pdf?dl=1

Causal Inference, Miguel Hernan, also available for free: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/