Chapter 2

Causal models

Causal inference aims to infer the outcome of an intervention in some system from data obtained by observing (but not intervening in) it. To do this we need terminology to describe actions and how we anticipate the system should respond to them. Three key approaches have emerged; counterfactuals, structural equation models and causal Bayesian networks. In this chapter we will examine the problems these approaches allow us to solve, the assumptions they rely on and how they differ. We will also use them to describe the following simplified examples. The aim is to demonstrate the notations and formalisms needed to tackle more interesting problems later on.

Example 1. Suppose a pharmaceutical company wants to assess the effectiveness of a new drug on recovery from a given illness. This is typically tested by taking a large group of representative patients and randomly assigning half of them to a treatment group (who recieve the drug) and the other half to a control group (who recive a placebo). The goal is to determine the clinical impacts of the drug by comparing the differences between the outcomes for the two groups (in this case, simplified to only two outcomes - recovery or non-recovery). We will use the variable X ($1 =$ drug, $0 =$ placebo) to represent the treatment each person receives and Y ($1 =$ recover, $0 =$ not recover) to describe the outcome.

Example 2. Suppose we want to estimate the impact on high school graduation rates of compusory preschool for all 4 year olds. We have a large cross-sectional dataset on a group of 20 year olds that records if they attended pre-school, if they graduated high school and their parents socio-economic status (SES). We will let $X \in \{0, 1\}$ indicate if an individual attended pre-school, $Y \in \{0, 1\}$ indicate if they graduated high school and $Z \in \{0, 1\}$ represent if they are from a low or high SES background respectively.[1]

2.1 Causal Bayesian networks

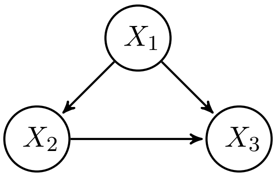
Causal Bayesian networks are an extension of Bayesian networks. A Bayesian network is a graphical way of representing how a distribution factorises. Any joint probability distribution can be factorised into a product of conditional probabilities. There are multiple valid factorisations, corresponding to permutations of variable ordering.

$$P(X_1, X_2, X_3, ...) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)... \tag{2.1}$$

[1] There has been substantial empirical work on the effectevness of early childhood education including a landmark randomised trial, the Perry Preschool project, which ran from 1962-1967 [62].

We can represent this graphically by drawing a network with a node for each variable and adding links from the variables on the right hand side to the variable on the left for each conditional probability distribution, see figure 2.1. If the factorisation simplifies due to conditional independencies between variables, this is reflected by missing edges in the corresponding network. There are multiple valid Bayesian network representations for any probability distribution over more than one variable, see figure 2.2 for an example.

Figure 2.1: A general Bayesian network for the joint distribution over three variables. This network does not encode any conditional independencies between its variables and can thus represent any distribution over three variables.



The statement that a given graph G is a Bayesian network for a distribution P tells us that the distribution can be factorised over the nodes and edges in the graph. There can be no missing edges in G that do not correspond to conditional independencies in P (the converse is not true G can have extra edges). If we let $parents_{X_i}$ represent the set of variables that are parents of the variable X_i in G then we can write the joint distribution as;

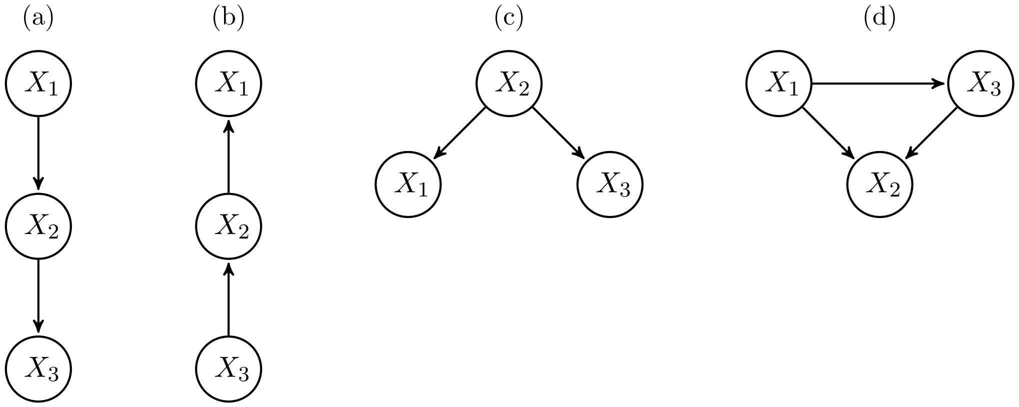
$$P(X_1...X_N) = \prod_{i=1...N} P(X_i|parents_{X_i}) \tag{2.2}$$

A causal Bayesian network is a Bayesian network in which it a link $X_i \rightarrow X_j$, by definition, implies X_i causes X_j. This means that if we intervene and change the value of X_i, we expect X_j to change, but if we intervene to change X_j, X_i will not change. We need some notation to describe interventions and represent distributions over variables in the network after an intervention. In this thesis I use the do operator introduced by Pearl [38].

Definition 3. The do-notation

- do(X=x) denotes an intervention that sets the random variable(s) X to x.
- $\mathrm{P}\left\{Y|do(X)\right\}$ is the distribution of Y conditional on an *intervention* that sets X. This notation is somewhat overloaded. It may be used represent a probability, a probability

Figure 2.2: Some valid Bayesian networks for a distribution that an be factorised as $\mathrm{P}\left\{X_1, X_2, X_3\right\} = \mathrm{P}\left\{X_1\right\}\mathrm{P}\left\{X_2\right\}\mathrm{P}\left\{X_3|X_2\right\}$ (which implies $X_3 \perp\!\!\!\perp X_1|X_2$)



distribution/mass function or a family of distribution functions depending on if the variables are discrete or continuous and whether or not we are treating them as fixed. For example it could represent

- The probability $\mathrm{P}\{Y=1|do(X=x)\}$ as a function of x
- The probability mass function for a discrete Y : $\mathrm{P}\{Y|do(X=x)\}$
- The probability density function for a continuous Y : $f_Y(y|do(X=x))$
- a familiarly of density/mass function for Y paramaterised by x.

Where the distinction is important and not clear from context we will use one of the more specific forms above.

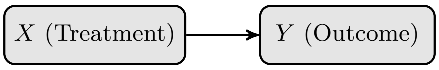
Theorem 4 (Truncated product formula)**.** *If G is a causal network for a distribution P defined over variables $X_1...X_N$, then we can calculate the distribution after an intervention where we set $Z \subset X$ to z, denoted $do(Z=z)$ by dropping the terms for each of the variables in Z from the factorisation given by the network [38].*

$$\mathrm{P}\{X_1...X_N|do(Z=z)\} = \begin{cases} \prod_{i \notin Z} \mathrm{P}\{X_i|parents_{X_i}\} & \textit{if } (X_1...X_N) \textit{ consistent with } Z=z \\ 0 & \textit{otherwise} \end{cases} \tag{2.3}$$

Theorem 4 does not hold for standard Bayesian networks because there are multiple valid networks for the same distribution. The truncated product formula will give different results depending on the selected network. The result is possible with causal Bayesian networks because it follows directly from the assumption that the direction of the link indicates causality. In fact, from the interventionist viewpoint of causality, the truncation product formula defines what it means for a link to be causal.

Returning to example 1, and phrasing our query in terms of interventions; what would the distribution of outcomes look like if everyone was treated $\mathrm{P}\{Y|do(X=1)\}$, relative to if no one was treated $\mathrm{P}\{Y|do(X=0)\}$? The treatment X is a potential cause of Y, along with other unobserved variables, such as the age, gender and the disease sub type of the patient. Since X is assigned via deliberate randomisation we know that it is not effected by any latent variables. The causal Bayesian network for this scenario is shown in figure 2.3 .This network represents the (causal) factorisation $\mathrm{P}\{X,Y\} = \mathrm{P}\{X\}\mathrm{P}\{Y|X\}$, so from equation (2.3), $\mathrm{P}\{Y|do(X)\} = \mathrm{P}\{Y|X\}$. In this example, the interventional distribution is equivalent to the observational one.

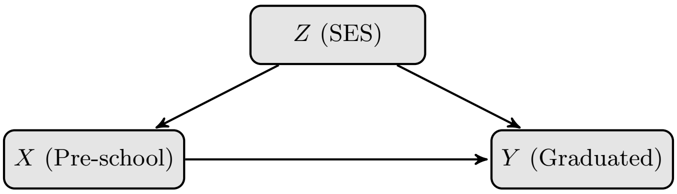
Figure 2.3: Causal Bayesian network for example 1



In example 2 we are interested $\mathrm{P}\{Y|do(X=1)\}$, the expected high-school graduation rate if we introduce universal preschool. We could compare it to outlawing preschool $\mathrm{P}\{Y|do(X=0)\}$ or the current status quo $\mathrm{P}\{Y\}$. It seems reasonable to assume that preschool attendence affects the likelihood of high school graduation [2] and that parental socio-economic status would affect *both* the likelihood of preschool attendance and high school graduation. If we assume that socio-economic status is the only such variable (nothing else effects both attendance *and* graduation), we can represent this problem with the causal Bayesian network in figure 2. In this case, the interventional distribution is not equivalent to the observational one. If parents

[2] The effect does not have to be homogenius, it may depend non-linearly on characteristics of the child, familiy and school.

Figure 2.4: Causal bayesian network for example 2



with high socio-economic status are more likely to send their children to preschool and these children are more likely to graduate high school regardless, comparing the graduation rates of those who attended preschool with those who did not will overstate the benefit of preschool. To obtain the interventional distribution we have to estimate the impact of preschool on high school graduation for each socio-economic level separately and then weight the results by the proportion of the population in that group,

$$\mathrm{P}\left\{Y|do(X=1)\right\} = \sum_{z \in Z} \mathrm{P}\left\{Y|X=1,Z\right\} \mathrm{P}\left\{Z\right\} \tag{2.4}$$

We have seen from these two examples that the expression to estimate the causal effect of an intervention depends on the structure of the causal graph. There is a very powerful and general set of rules that specify how we can transform observational distributions into intervetional ones for a given graph structure. These rules are referred to as the Do-calculus [38]. We discuss them further in section **??**.

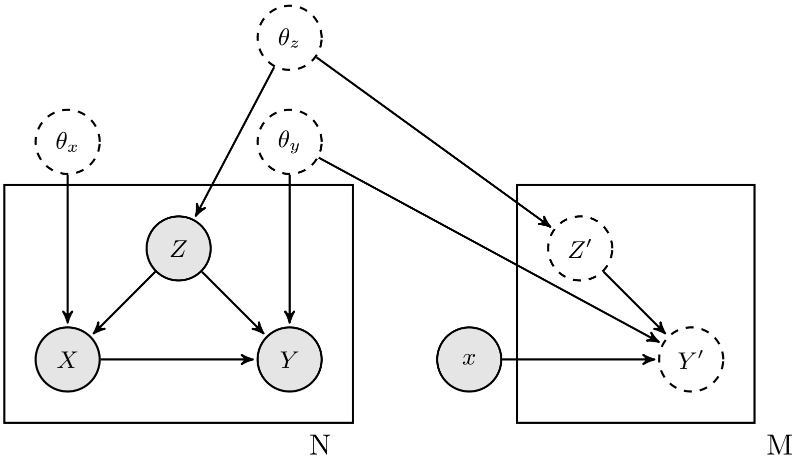
Formalising the definition of an intervention within the framework of causal graphical models provides us with an explicit mechanism to map information from one data generating process, the system pre-intervention, to another, the system post-intervention. The power of defining an intervention in this way stems from the number of things that are invariant between the two processes. All the (conditional) distributions for variables in the graph that were not directly set by the intervention are assumed not be changed by it.

We could represent problems of the type where we try to infer properties of the post-interventional system based on data generated by the pre-interventional distribution by explicitly representing both systems and what they have in common, see figure 2.5. This does not require any special framework or notation. The graphs in figure 2.5 are ordinary Bayesian networks. However, without a causal framework, we have to make assumptions about what will be invarient to the intervention specifically for each such problem we encounter. For complex problems, it is very difficult to conceptualise the assumptions we expect to hold without the benefit of a causal framework.

A causal bayesian network represents much more information than a bayesian network with identical structure. A causal network encodes all possible interventions that could be specified with the do-notation. For example, if the network in figure 2.4 were an ordinary bayesian network and all the variables were binary, the associated distribution could be described by 7 parameters. The equivalent causal bayesian network additionally represents the post-interventional distributions for six possible single variable interventions and twelve possible two variable interventions. Encoding all this information without the assumptions implicit in the causal bayesian network would require an additional 30 parameters. [3]

[3] After each single variable intervention we have a distribution over two variables, which can be represented by three parameters. After each two variable intervention, we have a distribution over one variables which requires one parameter. This takes us to a total of $6*3+12*1=30$ additional parameters.

Figure 2.5: Causal inference with ordinary bayesian networks. The plate on the left represents the observed data generated prior to the intervention and the plate on the right the data we anticipate obtaining after an intervention that the pre-interventional variable X to x. The assumptions characterised by this plate model correspond to those implied by the causal bayesian network in figure 2.4 for the intervention $do(X = x)$. As the networks in this figure are ordinary Bayesian networks, we could have represented the same information with a different ordering of the links within each plate. However, we would then have a complex transformation relating the parameters between the two plates rather than a simple invariance.

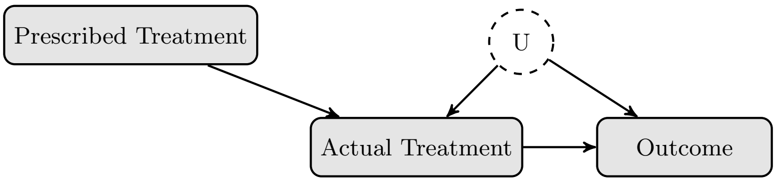


Causal bayesian networks are bayesian networks, so results that apply to bayesian networks carry directly across; the local Markov property states that variables are independent of their non-effects given their direct causes. Similarly the global Markov property and d-separation also hold in causal networks.

Limitations of causal bayesian networks

A number of criticisms have been levelled at this approach to modelling causality. One is that the definition of an intervention only in terms of setting the value of one or more variables is too precise and that any real world intervention will effect many variables in complex and non-deterministic ways [**?** 8]. However, by augmenting the causal graph with additional variables that model how interventions may take effect, the deterministic do operator can model more complex interventions. For example, in the drug treatment case, we assumed that all subjects complied, taking the treatment or placebo as assigned by the experimenter. But what if some people failed to take the prescribed treatment. We can model this within the framework of deterministic interventions by adding a node representing what they were prescribed (the intervention) which probabilistically influences the treatment they actually receive ,see figure 2.6. Note that the fact that we no longer directly assign the treatment opens the possibility that an unobserved latent variable could affect both the actual treatment taken and the outcome.

Figure 2.6: Randomised treatment with imperfect compliance



Another key issue with causal Bayesian networks is that they cannot handle cyclic dependencies between variables. Such feedback loops are common in real-life systems, for example the relationship between supply and demand in economics or predator and prey in ecology. We might regard the underlying causal mechanisms in these examples to be acyclic; the number of predators at one time influences the number of prey in the next period and so on. However, if our measurements of these variables must be aggregated over timeframes that are longer than the scale at which these interactions occur the result is a cyclical dependency. Even were we able to measure on shorter timescales, we might then not have sufficient data on each variable for inference. Such problems have mostly been studied within the dynamic systems literature, typically focusing on understanding the stationary or equilibrium state of the system and making very specific assumptions about functional form in order to make problems tractable. [41] compare the equilibrium approach to reasoning about cyclic problems with strucural equation models, which we discuss in section 2.3 and which can be seen as bayesian causal networks with additional functional assumptions.

2.2 Counterfactuals

The Neyman-Rubin model [47, 48, 46, 49, 50] defines causality in terms of potential outcomes, or counterfactuals. Counterfactuals are statements about imagined or alternate realities, are prevalent in everyday language and may play a role in the development of causal reasoning in humans [63]. Causal effects are differences in counterfactual variables; what is the difference between what would happen if we did one thing versus what would happen if we did something else.

In example 1, the causal effect of the drug relative to placebo for person i is the difference between what would happen if they were given the drug, denoted ${y_i}^1$ versus what would happen if they got the placebo, ${y_i}^0$. The fundamental problem of causal inference is that we can only observe one of these two outcomes, since a given person can only be treated or not treated. The problem can be resolved if, instead of people, there are units that can be assumed to be identical or that will revert exactly to their initial state some time after treatment. This type of assumption often holds to a good approximation in the natural sciences and explains why researchers in these fields are less concerned with causal theory.

Putting aside any estimates of individual causal effects, it is possible to learn something about the distributions under treatment or placebo. Let Y^1 be a random variable representing the potential outcome if treated. The distribution of Y^1 is the distribution of Y if everyone was treated. Similarly Y^0 represents the potential outcome for the placebo. The difference between the probability of recovery, across the population, if everyone was treated and the probability of recovery given placebo is $\mathrm{P}\left\{Y^1\right\} - \mathrm{P}\left\{Y^0\right\}$. We can estimate, (from an experimental or observational study);

$$\begin{aligned}
&\mathrm{P}\{Y|X=1\}, \text{the probability that those who took the treatment will recover}\\
&\mathrm{P}\{Y|X=0\}, \text{the probability that those who were not treated will recover}
\end{aligned}$$

Now, for those who took the treatment, the outcome *had* they taken the treatment Y^1 is the same as the observed outcome. For those who did not take the treatment, the observed outcome is the sames as the outcome *had* they not taken the treatment. Equivalently stated:

$$\begin{aligned}
\mathrm{P}\left\{Y^0|X=0\right\} &= \mathrm{P}\{Y|X=0\}\\
\mathrm{P}\left\{Y^1|X=1\right\} &= \mathrm{P}\{Y|X=1\}
\end{aligned}$$

If we assume $X \perp\!\!\!\perp Y^0$ and $X \perp\!\!\!\perp Y^1$:

$$\begin{aligned} \mathrm{P}\left\{Y^1\right\} &= \mathrm{P}\left\{Y^1|X=1\right\} = \mathrm{P}\left\{Y|X=1\right\} \\ \mathrm{P}\left\{Y^0\right\} &= \mathrm{P}\left\{Y^0|X=0\right\} = \mathrm{P}\left\{Y|X=0\right\} \end{aligned}$$

$$\implies \mathrm{P}\left\{Y^1\right\} - \mathrm{P}\left\{Y^0\right\} = \mathrm{P}\left\{Y|X=1\right\} - \mathrm{P}\left\{Y|X=0\right\}$$

The assumptions $X \perp\!\!\!\perp Y^1$ and $X \perp\!\!\!\perp Y^0$ are referred to as ignore-ability assumptions [46]. They state that the treatment a each person receives is independent of whether they would recover if treated and if they would recover if not treated. This is justified in example 1 due to the randomisation of treatment assignment. In general the treatment assignment will not be independent of the potential outcomes. In example 2, the children who attended preschool may be more likely to have graduated highschool had they in fact not attended than the children who actually did not attend, $X \not\perp\!\!\!\perp Y^0$. Similarly, had the poorer children who did not attend pre-school attended they might not have done as well as the children who did in fact attend, $X \not\perp\!\!\!\perp Y^1$. A more general form of the ignore-ability assumption is to identify a set of variables Z such that $X \perp\!\!\!\perp Y^1|Z$ and $X \perp\!\!\!\perp Y^0|Z$.

Theorem 5 (Ignore-ability)**.** *If $X \perp\!\!\!\perp Y^1|Z$ and $X \perp\!\!\!\perp Y^0|Z$,*

$$\mathrm{P}\left\{Y^1\right\} = \sum_{z\in Z} \mathrm{P}\left\{Y|X=1,Z\right\}\mathrm{P}\left\{Z\right\} \tag{2.5}$$

$$\mathrm{P}\left\{Y^0\right\} = \sum_{z\in Z} \mathrm{P}\left\{Y|X=0,Z\right\}\mathrm{P}\left\{Z\right\} \tag{2.6}$$

Assuming that within each socio-economic status level, attendence at pre-school is independent of the likelihood of graduating high-school had a person attended, then the average rate of high-school graduation given a universal pre-school program can be computed from equation 2.5. Note, that this agrees with the weighted adjustment formula in equation 2.4.

Another assumption introduced within the Neyman-Rubin causal framework is the Stable Unit Treatment Value Assumption (SUTVA) [48]. This is the assumption that the potential outcome for one individual (or unit) does not depend on the treatment assigned to another individual. As an example of a SUTVA violation, suppose disadvantaged four year olds were randomly assigned to attend pre-school. The later school results of children in the control group, who did not attend, could be boosted by the improved behaviour of those did and who now share the classroom with them. SUTVA violations would manifest as a form of model misspecification in causal Bayesian networks.

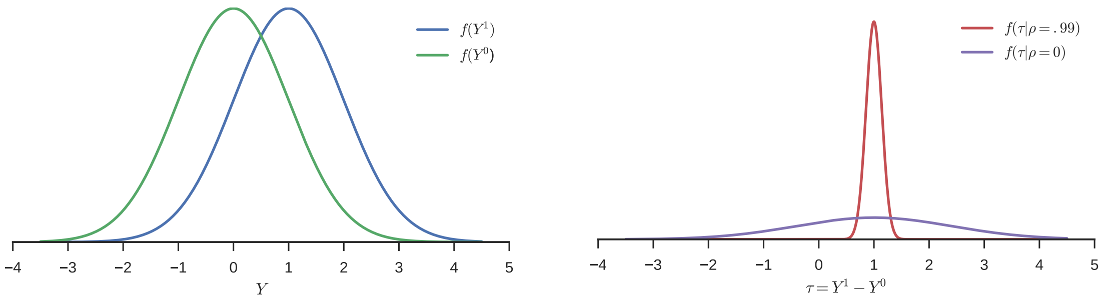
There are complex philosophical objections to counterfactuals arising from the way they describe alternate universes that were never realised. This makes it quite easy to (accidentally) make statements about counterfactuals that cannot be tested with empirical data. Consider the following example based on Dawid [12]. Again we have a drug, where the outcome for an individual if treated is represented by the counterfactual variable Y^1 and the outcome if not treated is Y^0. Suppose these counterfactual variables Y^1 and Y^0 are jointly normal with equal variance (for simplicity).

$$P(Y^1, Y^0) \sim N(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2\end{bmatrix}) \tag{2.7}$$

Their difference is also normal. Let $\tau = Y^1 - Y^0$,

$$P(\tau) = N(\mu_1 - \mu_0, 2\sigma^2(1 - \rho)) \tag{2.8}$$

Figure 2.7: The distribution of individual treatment effects is not identifiable, even from a randomized controlled trial.



(a) Marginal distributions over the potential outcomes Y^1 and Y^0 for $\mu_1 = 1$, $\mu_0 = 0$ and $\sigma = 1$. The blue curve shows the distribution of Y if everyone were to be treated and the blue curve the distribution if no-one was treated.

(b) Two very different distributions of individual causal effects consistent with the potential outcome distributions.

From a (large) randomized controlled trial we can estimate the marginal distributions over the counterfactual variables, see figure 2.7a. These represent the distributions we would expect over the outcome Y if everyone were treated or not treated retrospectively. However, the distribution over the individual causal effects depends on ρ, see figure 2.7b. The key problem is that we can never observe the joint distribution over Y^1 and Y^0. As a result, ρ and thus the variance of τ is not identifiable, even from experimental data. **?**] argues that we should avoid using counterfactuals as they are defined in terms of (metaphysical) individual causal effects. He further points out that the interventional distributions in figure 2.7a, along with a loss function, contain all the information required to decide how to treat a new patient.

This result is unintuitive. It seems on the face of it that the distribution of individual causal effects is relevant to our decision making. If $\rho = 1$ then almost everyone benefits slightly from the treatment whilst if $\rho = 0$, there is a wide range, with some people benefiting a lot and others suffering significant harm. This confusion can be resolved by thinking about personalised rather than individual causal effects. It is entirely possible that potentially observable characteristics (such as gender, age, genetics, etc) affect how people will respond to the treatment. We can partition people into sub-populations on the basis of these characteristics and measure different *personalised* causal effects for each group. The variance of the potential outcome distributions $f(Y^1)$ and $f(Y^0)$ provides bounds on how much can be gained from further personalisation. The metaphysical nature of individual causal effects only arises when we are at the point where the only remaining variation is due to inherent randomness (or variables that we could not even in principle measure).

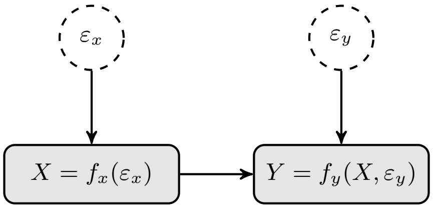
One way of looking at counterfactuals is as a natural language short hand for describing highly specific interventions like those denoted by the do-notation. Rather than taking about the distribution of Y given we intervene to set $X = x$ and hold everything else about the system constant we just say what would the distribution of Y be had X been x. This is certainly convenient, if rather imprecise. However, the ease with which we can make statements with counterfactuals that cannot be tested with empirical data warrants careful attention. We should always be clear what assumptions we are making and have in mind if it is possible (at least in theory) to validate those assumptions.

2.3 Structural Equation models

Structural equation models (SEMs) describe a deterministic world, where some underlying mechanism or function determines the output of any process for a given input. The mechanism (but not the output) is assumed to be independent of what is fed into it. Uncertainties are not inherent but arise from unmeasured variables. Linear structural equation models have a long history for causal estimation [64, 20]. More recently, they have been formalised, generalised to the non-linear setting and connected to developments in graphical models to provide a powerful causal framework [38].

Mathematically, each variable is a deterministic function of its direct causes and a noise term that captures unmeasured variables. The noise terms are required to be mutually independent. If there is the possibility that an unmeasured variable influences more than one variable of interest in a study, it must be modelled explicitly as a latent (unobserved) variable. Structural equation models can be represented visually as a network. Each variable is a node and arrows are drawn from causes to their effects. Figure 2.8 illustrates the SEM for example 1.

Figure 2.8: SEM for example 1



This model encodes the assumption that the outcome y_i for an individual i is caused solely by the treatment x_i they receive and other factors ε_{y_i} that are independent of X. This is justifiable on the grounds that X is random. The outcome of a coin flip for each patient should not be related to any of their characteristics (hidden or otherwise). Note that the causal graph in figure **??** is identical to that the bayesian network for the same problem, figure 2.3. The latent variables ε_x and ε_y are not explicitly drawn in figure 2.3 as are captured by the probabilistic nature of the nodes in a Bayesian network, however adding them would not change the model.

Taking the *action* $X = 1$ corresponds to replacing the equation $X = f_x(\varepsilon_x)$ with $X = 1$. The function f_y and distribution over ε_y does not change. This results in the interventional distribution [4],

$$\mathrm{P}\left\{Y = y | do(X = 1)\right\} = \sum_{\varepsilon_y} \mathrm{P}\left\{\varepsilon_y\right\} \mathbb{1}\{f_y(1, \varepsilon_y) = y\} \tag{2.9}$$

The observational distribution of Y given X is,

[4] We have assumed the variables are discrete only for notational convinience

$$\mathrm{P}\left\{Y=y|X=1\right\}=\sum_{\varepsilon_x}\sum_{\varepsilon_y}\mathrm{P}\left\{\varepsilon_x|X=1\right\}\mathrm{P}\left\{\varepsilon_y|\varepsilon_x\right\}\mathbb{1}\{f_y(1,\varepsilon_y)=y\} \tag{2.10}$$

$$=\sum_{\varepsilon_y}\mathrm{P}\left\{\varepsilon_y\right\}\mathbb{1}\{f_y(1,\varepsilon_y)=y\}\text{, as }\varepsilon_x\perp\!\!\!\perp\varepsilon_y \tag{2.11}$$

The interventional distribution is the same as the observational one. The same argument applies to the intervention $do(X=0)$ and so the causal effect is just the difference in observed outcomes as found via the causal Bayesian network and counterfactual approaches.

The SEM for example 2 is shown in figure 2.9. Intervening to send all children to pre-school replaces the equation $X=f_x(Z,\varepsilon_x)$ with $X=1$, leaving all the other functions and distributions in the model unchanged.

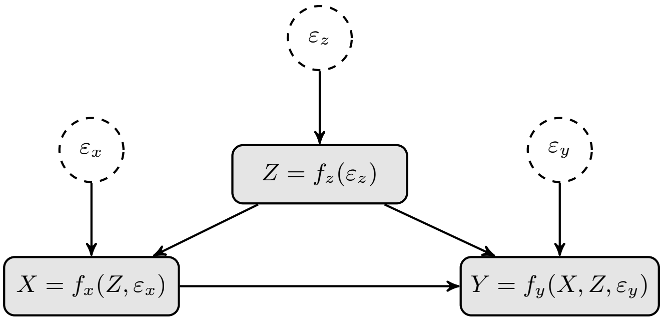
$$\mathrm{P}\left\{Y=y|do(X=1)\right\}=\sum_{z}\sum_{\varepsilon_y}\mathrm{P}\left\{z\right\}\mathrm{P}\left\{\varepsilon_y\right\}\mathbb{1}\{f_y(1,z,\varepsilon_y)=y\} \tag{2.12}$$

$$=\sum_{z}\mathrm{P}\left\{z\right\}\underbrace{\sum_{\varepsilon_y}\mathrm{P}\left\{\varepsilon_y\right\}\mathbb{1}\{f_y(1,z,\varepsilon_y)=y\}}_{\mathrm{P}\{Y=y|X=1,Z=z\}} \tag{2.13}$$

Equation 2.13 corresponds to equations 2.4 and 2.5. It is not equivalent to the observational distribution, given by;

$$\mathrm{P}\left\{Y=y|X=1\right\}=\sum_{z}\sum_{\varepsilon_y}\mathrm{P}\left\{z|X=1\right\}\mathrm{P}\left\{\varepsilon_y\right\}\mathbb{1}\{f_y(1,z,\varepsilon_y)=y\} \tag{2.14}$$

Figure 2.9: SEM for example 2



Structural equation models are generally applied with strong constraints on the functional form of the relationship between the variables and noise is typically assumed to be additive, $X_i=f_i(\cdot)+\varepsilon_i$. A structural equation model with N variables resembels a set of N simultaneous equations, with each variable playing the role of the dependent (left hand side) variable in one equation. However a SEM is, by definition, more than a set of simultaneous equations. By declaring it to be structural we are saying that it represents assumptions about the relationships between variables. When we visualise the model as a network the absence of an arrow between two variables encodes the assumption that one does not cause the other. The similarity between the notation used to describe and analyse structural equation models and simultaneous equations, combined with a reluctance to make explicit statements about causality has led to some confusion in the interpretation of SEMs [23, 38].

2.4 Comparing and unifying the models

Remarkably for models developed relatively independently in fields with very different approaches and problems, the models we have discussed can be nicely unified for interventional queries (those that can be expressed with the do-notation). If the network for a structural equation model is acyclic, that is if starting from any node and following edges in the direction of the arrows you cannot return to the starting point, then it implies a recursive factorisation of the joint distribution over its variables. In other words, the network is a causal Bayesian network. All of the results that apply to causal Bayesian networks also apply to acyclic structural equation models. Taking an action that sets a variable to a specific value equates to replacing the equation for that variable with a constant. This corresponds to dropping a term in the factorisation and the truncated product formula (equation 2.3). Thus, the interventional query $P(Y|do(X))$ is identical in these two frameworks. We can also connect this to counterfactuals via:

$$\begin{aligned} Y^0 &\equiv P(Y|do(X=0)) \\ Y^1 &\equiv P(Y|do(X=1)) \end{aligned} \tag{2.15}$$

The assumption $\varepsilon_X \perp\!\!\!\perp \varepsilon_Y$, stated for our structural equation model, translates to $X \perp\!\!\!\perp (Y^0, Y^1)$ in the language of counterfactuals. When discussing the counterfactual model, we actually made the slightly weaker assumption:

$$X \perp\!\!\!\perp Y^0 \text{ and } X \perp\!\!\!\perp Y^1 \tag{2.16}$$

It is possible to relax the independence of errors assumption for SEMs to correspond exactly with the form of equation (2.16) without losing any of the power provided by d-separation and graphical identification rules [45]. The correspondence between the models for interventional queries (those that can be phrased using the do-notation) makes it straightforward to combine key results and algorithms developed within any of these frameworks. For example, you can draw a causal graphical network to determine if a problem is identifiable and which variables should be adjusted for to obtain an unbiased causal estimate. Then use propensity scores[46] to estimate the effect. If non-parametric assumptions are insufficient for identification or lead to overly large uncertainties, you can specify additional assumptions by phrasing your model in terms of structural equations. The frameworks do differ when it comes to causal queries that involve joint or nested counterfactuals and cannot be expressed with the do-notation. These types of queries arise in the study of mediation [39, 28, 60] and legal discrimination [38].

In practice, differences in focus and approach between the fields in which each model dominates eclipse the actual differences in the frameworks. The work on causal graphical models [38**?**] focuses on asymptotic, non-parametric estimation and rigorous theoretical foundations. The Neyman-Rubin framework builds on the understanding of randomised experiment and generalises to quasi-experimental and observational settings, with a particular focus on non-random assignment to treatment. This research emphasises estimation of average causal effects and provides practical methods for estimation, in particular, propensity scores; a method to control for multiple variables in high dimensional settings with finite data [46]. In economics, inferring causal effects from non-experimental data to support policy decisions is central to the field. Economists are often interested in broader measures of the distribution of causal effects than the mean and make extensive use of structural equation models, generally with strong parametric assumptions [22]. In addition, the parametric structural equation models favoured in economics can be extended to analyse cyclic (otherwise referred to as non-recursive) models.

2.5 What does a causal model give us? Resolving Simpson's paradox

We will now demonstrate our new notation and frameworks for causal inference to resolve a fascinating paradox, noted by Yule [65], demonstrated in real data by Cohen and Nagel [10] and popularised by Simpson [55]. The following example is adapted from Pearl [38]. Suppose a doctor has two treatments, A and B, which she offers to patients to prevent heart disease. She keeps track of which medication her patients choose and whether or not the treatment is successful. She obtains the results in table 2.1.

Table 2.1: Treatment results

Treatment	Success	Fail	Total	Success Rate
A	87	13	100	87%
B	75	25	100	75%

Drug A appears to perform better. However, having read the latest literature on how medications affect men and women differently, she decides to break down her results by gender to see how well the drugs perform for each group and obtains the data in table 2.2.

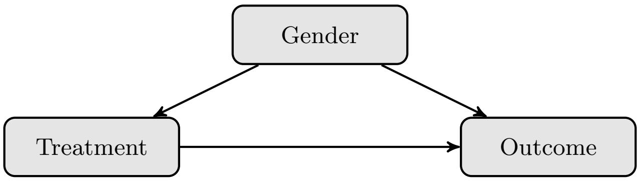
Table 2.2: Treatment results by gender

Gender	Treatment	Success	Fail	Total	Success Rate
M	A	12	8	20	60%
M	B	56	24	80	70%
F	A	75	5	80	94%
F	B	19	1	20	95%

Once the data is broken down by gender, Treatment B looks better for both men *and* women. Suppose the doctor must choose only one drug to prescribe to all her patients in future (perhaps she must recommend which to subsidise under a national health scheme). Should she choose A or B? The ambiguity in this question lies at the heart of Simpson's paradox. How does causal modelling resolve the paradox? The key is that the doctor is trying to choose between *interventions*. She wants to know what the success rate will be if she changes her practice to give all the patients one drug, rather than allowing them to choose as currently occurs.

Let's represent the treatment by the variable T, the gender of the patient by Z and whether or not the treatment was successful by Y. The doctor cares about $\mathrm{P}\{Y|do(T)\}$, not the standard conditional distributions $\mathrm{P}\{Y|T\}$. Unfortunately, the data in tables 2.1 and 2.2 is insufficient to enable estimation of the interventional distribution $\mathrm{P}\{Y|do(T)\}$ or determine if $do(T = A)$ is better or worse than $do(T = B)$. Some assumptions about the causal relationships between the variables are required. In this example, it seems reasonable to conclude that gender may affect the treatment chosen and the outcome. Assuming there are no other such confounding variables (for example income) then we obtain the causal network in figure 2.10.

Figure 2.10: An example of Simpson's Paradox

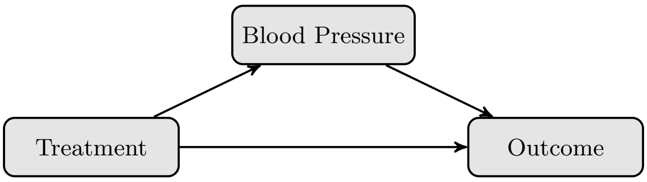


With this model, women are more likely to choose treatment A and are also more likely to recover than men regardless of the treatment they receive. Knowing a patient took drug A indicates they are more likely to be female. When we compare the group of people who took A against those who took B, the effect of the higher share of females in the first group conceals the greater benefit of drug B leading to an apparent reversal in effectiveness. However, when the doctor intervenes to set the treatment each person recieves there will no longer be a link from gender to treatment. So in this case she should choose which drug to prescribe from the gender specific table (and weight by the proportion of the population that belongs to each gender). Drug B is the better choice.

$$\mathrm{P}\{Y|do(T)\} = \mathrm{P}\{Y|T, female\}\,\mathrm{P}\{female\} + \mathrm{P}\{Y|T, male\}\,\mathrm{P}\{male\} \tag{2.17}$$

Is the solution to Simson's paradox to always to break down the data by as many variables as possible? No. Suppose we have the identical data as in 2.1 and 2.2 but replace the column name 'gender' with 'blood preassure', 'M' with 'high' and 'F' with 'normal'. This is a drug designed to prevent heart disease. One pathway to doing so might well be to lower blood pressure. Figure 2.11 shows a plausible causal graph for this setting. It differs from the graph in figure 2.10 only in the direction of a single link. Now, however table 2.2 tells us that people who took treatment A had better blood pressure control and better overall outcomes. In this setting $\mathrm{P}\{Y|do(T)\} = \mathrm{P}\{Y|T\}$. Drug A is the better choice.

Figure 2.11: An example of Simpson's Paradox

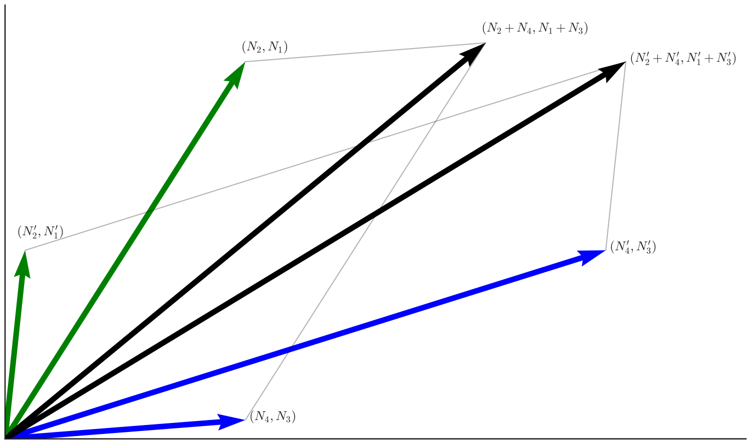


Note that we have not changed the data itself, only the description of the variables that it is associated with. This illustrates that the resolution to Simpson's paradox lies fundamentally not in the data, but in the assumptions we are willing to make. From a purely statistical viewpoint there is no paradox. The reversal just stems from the mathematical property of ratios expressed in equation 2.18 and represented graphically in figure 2.12. The paradox only arises when we attempt to use the data to select an intervention and is resolved when we apply a causal approach to do so.

$$\exists\left\{N_1, ...N_4, N_1'...N_4'\right\} \in \mathbb{N} : \frac{N_1}{N_2} < \frac{N_1'}{N_2'},\ \frac{N_3}{N_4} < \frac{N_3'}{N_4'} \text{ and } \frac{N_1+N_3}{N_2+N_4} > \frac{N_1'+N_3'}{N_2'+N_4'} \tag{2.18}$$

There are many other plausible causal graphs for both scenarios above. Perhaps income affects drug choice as well as gender, or gender might affect treatment choice and blood preassure control given treatment, etc. Causal modeling provides a powerful tool to specify such assumptions and to determine how to estimate causal effects for a given model see section 3.1.

Figure 2.12: Simpson's reversal visualized. The ratios involving N_i' are steeper than than those involving N_i for both the blue and green vectors. However, when we sum them, the ratio is steeper for the un-primed variables.



Chapter 3

Two key questions

We can roughly categorise the problems studied within causal inference into two groups, causal effect estimation and causal discovery. In causal effect estimation we assume (at least implicitly) that key aspects of the causal graph are known. The goal is then to estimate the effect of an intervention or range of interventions in the system. Causal effect estimation is implicit in countless studies in economics, social science and epidemiology into everything from the effect of education on earnings [7], diet on cancer [4] and breastfeeding on intelligence [30] to the effect of pet ownership on survival after a heart attack [15]. Almost every time someone runs a regression model the key quantity of interest is a causal effect. Given how it underlies so much of our scientific progress, there is a enormous potential in properly understanding when we can draw causal conclusions, exactly what assumptions are required to do so and how we can best leverage those assumptions to infer as much information as we can from our data.

Causal discovery aims to leverage much broader assumptions to learn the structure of causal graph from data. This is critical in fields where we are generating a lot of data but have limited theoretical knowledge from which to draw on to determine how variables are related to one another. Causal discovery algorithms are being applied in bioinformatics[3, 51, 42, 1, 57, 16, 56, 58], medical imaging [43] and climate science [59]. An effective and generalisable approach for causal discovery would amount to a major step towards the automation of the scientific endeavour.

3.1 Causal effect estimation

Estimating causal effects from observational data comes down to determining if and how we can write expressions for the interventional distributions of interest in terms of observational quantities, which can be measured. We did this ad-hoc basis to resolve the examples discussed in chapter 2. In this chapter we describe a principled approach to mapping observational quantities to interventional ones and discuss some of the key issues involved in estimating such expressions from finite sample data. We assume the key structure of the graph is known. That is, we assume that we can draw a network containing (at a minimum);

- the target/outcome variable we care about
- the focus/treatment variables on which we are considering interventions
- any variables which act to confound two or more of the other variables we have included.
- any links between variables we have included.

Some of these variables may be latent in that the available data does not record their value, however their position in the network is assumed to be known. For example, consider estimating the impact of schooling on wages. Some measure of inherent ability could influence both the number of years of schooling people choose to pursue and the wages they receive. Even if we have no data to directly assess peoples inherent ability we must include it in the graph because it influences two of the variables we are modelling.

How can the structure of the causal graph be leveraged to compute interventional distributions from observational ones? Given the graph corresponding to observational distribution, the graph after any intervention can be obtained by removing any links into variables directly set by the intervention. The joint interventional distribution is the product of the factors associated with the interventional graph, as given by the truncated product formula 2.3. If there are no latent variables the interventional distribution of interest can be obtained by marginalising over the joint (interventional) distribution. However, if there are latent variables the joint interventional distribution will contain terms that cannot be estimated from the observed data.

The key to estimating causal effects in the presence of latent variables lies combining the assumption of how an intervention changes the graph, encoded by the truncated product formula, with information the graph structure provides about conditional independenties between variables. By leveraging conditional independencies we can effectively localise the effect of an intervention to a specific part of a larger graph. This gives rise to the do-calculus [38]. The do calculus consists of three rules. They are derived from the causal information encoded in a causal network and the properties of d-separation and do not require any addition assumptions other than that of specifying the causal network.

3.1.1 Independence in Bayesian networks: D-separation

Many causal algorithms are based on leveraging the independence properties encoded in Bayesian networks. Therefor, in this section, we briefly review the key results. A more thorough introduction (including proofs) can be found in [**?**]. Recall that a Bayesian network is a way of representing the joint distribution over its variables in a way which highlights conditional independencies between them.

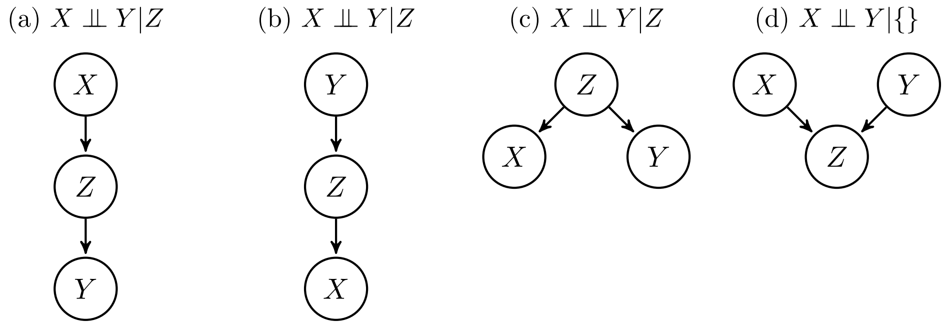
Theorem 6. ***(Local Markov condition)*** *Given a Bayesian network G with nodes $X_1...X_N$, each variable X_i is independent of its non-decedents given its parents in G for all distributions $P(X_1...X_N)$ that are compatible with G.*

The set of conditional independence relations given by the local Markov condition can enforce additional independencies that also hold in all distributions that are compatible with G. D-separation is an algorithm that extends the local Markov property to find these additional independencies. It provides us with a simple way of reading from a network if a given conditional independence statement is true in all distributions compatible with that network.

The statement that X is conditionally independent of Y given Z implies that if we know Z learning the value of Y gives us no additional information about X. From a graphical perspective you can think of this as Z blocks the flow of information from X to Y in the network. Figure 3.1 shows all possible network paths from a variable X to Y via Z. In figures (a) to (c) the path is blocked if we condition on Z and unblocked otherwise. In figure (d) the path is unblocked if we condition on Z and blocked otherwise.

The structure in figure 3.1d is referred to as a collider or v-structure. The somewhat counter-intuitive result than conditioning on Z introduces dependence between X and Y is called the *explaining away phenomena*. As an example, consider a scholarship available to female or disadvantaged students. Let X be gender, Y be family background and Z receipt of the scholarship. There are roughly equal number of boys and girls in both poor and wealthy families so X and Y are independent. However, if we know a student is receiving a scholarship then learning that they are male increases the probability that they are disadvantaged.

Figure 3.1: All possible two edge paths from X to Y via Z



(a) $X \perp\!\!\!\perp Y|Z$ (b) $X \perp\!\!\!\perp Y|Z$ (c) $X \perp\!\!\!\perp Y|Z$ (d) $X \perp\!\!\!\perp Y|\{\}$

Definition 7 (unblocked path)**.** A path from X to Y is a sequence of edges linking adjacent nodes starting at X and finishing at Y, $(X, V_1, V_2...V_k, Y)$. It is unblocked if every triple, $X - V_1 - V_2, V_1 - V_2 - V3, ..., V_{k-1} - V_k - Y$ in the path is unblocked (each triple will belong to one of the cases in figure 3.1)

Definition 8 (d-separation)**.** The variables $\boldsymbol{X}$ are d-separated from $\boldsymbol{Y}$ given $\boldsymbol{Z}$ in the network G if, there are no unblocked paths from any $X \in \boldsymbol{X}$ to any $Y \in \boldsymbol{Y}$ after conditioning on $\boldsymbol{Z}$.

Theorem 9 (d-separation and conditional independence)**.** *If a set of variables $\boldsymbol{Z}$ d-separates $\boldsymbol{X}$ and $\boldsymbol{Y}$ in a Bayesian network G then $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})$ in all distributions P compatible with G. Conversely, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-connected (not d-separated) given $\boldsymbol{Z}$ then it is possible to construct a distribution P' that factorises over G in which they are dependent.*

Theorem 9 says that independencies implied by d-separation on a graph hold in every distribution that can be factored over that graph and that if $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})$ in *all* distributions that can be factored over G then they are d-separated in G. If we denote the independencies implied by d-separation in a graph by $\mathcal{I}(G)$ and the set of independencies in a distribution by $\mathcal{I}(P)$ then $\mathcal{I}(G) \subseteq \mathcal{I}(P)$.

If $\mathcal{I}(G) = \mathcal{I}(P)$ then G is called a perfect map for P. However, it is possible to construct distributions that do not have a perfect-map, that is they contain conditional independencies that cannot be represented by d-separation. A particular case in which this occurs is when there are deterministic relationships between variables. If we have a Bayesian network G in which we specify that some nodes are deterministic we cannot conclude that if X and Y are d-connected then there exists a distribution P' *consistent* with G in which they are dependent. This does not conflict with theorem 9 as *consistent* in this setting requires that P' both factorises over G and satisfies the specified the deterministic relations between variables. This subtlety led to confusion in assessing what independencies hold between counterfactuals via twin networks [38, 45] and demonstrates the caution required in using d-connecteness to assert lack of independence. D-separation can be extended to compute the additional independencies implied by a graph in which certain nodes are known to be deterministic [**?**].

3.2 The Do Calculus

The do-calculus is a set of three rules [37] that can be applied to simplify the expression for an interventional distribution. If by repeated application of the do-calculus, along with standard probability transformations, we can obtain an expression containing only observational quantities then we can use it to estimate the interventional distribution from observational data. Let $\boldsymbol{X}$,$\boldsymbol{Y}$,$\boldsymbol{Z}$ and $\boldsymbol{W}$ be disjoint sets of variables in a causal graph G. We denote the graph G after the an intervention $do(\boldsymbol{X})$, which has the effect of removing all edges into variables in $\boldsymbol{X}$, as $G_{\overline{\boldsymbol{X}}}$

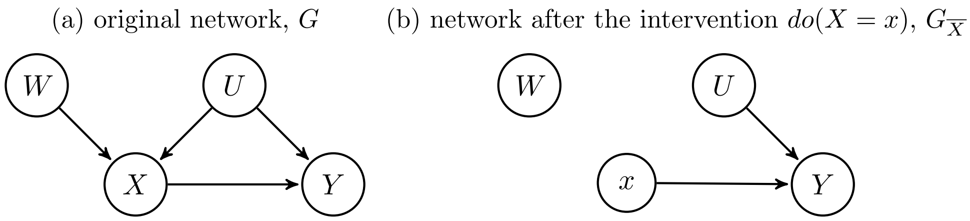
Rule 1: (adding or removing evidence)

Rule 1 allows us to remove (or insert) observational evidence from the right hand side of a conditional interventional distribution. It follows directly from the fact that the relationship between d-separation in a network and independence in the corresponding probability distribution still applies after an intervention.

If $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}|\boldsymbol{Z}, \boldsymbol{X})$ in $G_{\overline{\boldsymbol{X}}}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}), \boldsymbol{Z}=\boldsymbol{z}, \boldsymbol{W}=\boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}), \boldsymbol{Z}=\boldsymbol{z}) \qquad (3.1)$$

Figure 3.2: Rule 1 example. $(Y \perp\!\!\!\perp W|X)$ in $G_{\overline{\boldsymbol{X}}} \implies \mathrm{P}\{Y|do(X), W\} = \mathrm{P}\{Y|do(X)\}$



(a) original network, G

(b) network after the intervention $do(X = x)$, $G_{\overline{X}}$

Rule 2: (exchanging actions with observations)

Rule 2 captures when conditioning on $\boldsymbol{X} = \boldsymbol{x}$ and intervening $do(\boldsymbol{X} = \boldsymbol{x})$ have the same effect on the distribution of the outcome $\boldsymbol{Y}$. Let $G_{\underline{\boldsymbol{X}}}$ denote the causal graph G with all edges *leaving* X removed.

If $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{W})$ in $G_{\underline{\boldsymbol{X}}}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}), \boldsymbol{W}) = P(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}, \boldsymbol{W}) \qquad (3.2)$$

The intuition behind this is that interventional distributions differ from observational ones due to the presence of indirect paths between X and Y. Observing a variable X provides information about Y both directly and indirectly, by changing our belief about the distribution of the parents of X. However setting X tells us nothing about its parents so affects Y only via direct paths out of X. Removing edges *leaving* X removes all the direct paths out of X. If X is then independent of Y (conditional on W) that indicates there are no indirect paths which implies conditioning on X is equivalent to setting X (given W).

Equation 3.2 does not cover cases where acting on one set of variables allows us to replace acting on another set with conditioning (see figure 3.4). The general form of rule 2 is given in equation 3.3.

If $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{W},\boldsymbol{Z})$ in $G_{\underline{\boldsymbol{X}}\overline{\boldsymbol{Z}}}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),do(\boldsymbol{Z}=\boldsymbol{z}),\boldsymbol{W}) = P(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},do(\boldsymbol{Z}=\boldsymbol{z}),\boldsymbol{W}) \tag{3.3}$$

Figure 3.3: An example of rule 2 with a single intervention $(Y \perp\!\!\!\perp X|W)$ in $G_{\underline{\boldsymbol{X}}} \implies \mathrm{P}\{Y|do(X),W\} = \mathrm{P}\{Y|X,W\}$. In this example, observing X provides information about Y both directly and indirectly, because knowing X tells us something about W which also influences Y. If we condition on W, we block this indirect path.

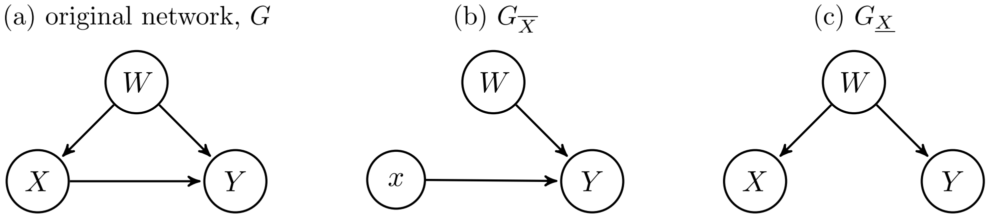
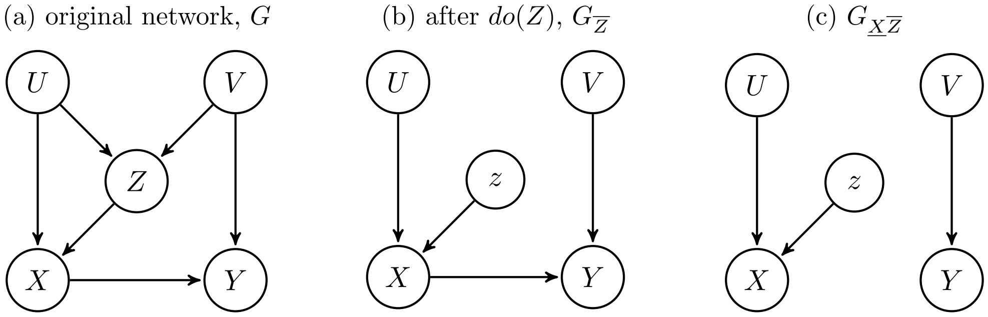

(a) original network, G (b) $G_{\overline{X}}$ (c) $G_{\underline{X}}$

Figure 3.4: An example of applying equation 3.3. In this case $(Y \perp\!\!\!\perp X|Z)$ in $G_{\underline{X}\overline{Z}} \implies \mathrm{P}\{Y|do(X=x),do(Z=z)\} = \mathrm{P}\{Y|X=x,do(Z=z)\}$. Observing, rather than intervening, on Z would not have allowed us to exchange $do(X=x)$ for $X=x$. Conditioning on Z does block the indirect path $X-Z-V-Y$ but opens $X-U-Z-V-Y$.


(a) original network, G (b) after $do(Z)$, $G_{\overline{Z}}$ (c) $G_{\underline{X}\overline{Z}}$

Rule 3: (adding or removing actions)

This rule describes cases where the intervention $do(\boldsymbol{X}=\boldsymbol{x})$ has no effect on the distribution of the outcome $\boldsymbol{Y}$. A simple case of rule 3 is given in equation 3.4. If $\boldsymbol{Y}$ is independent of $\boldsymbol{X}$ in G after removing links entering $\boldsymbol{X}$ then can be no direct path from $\boldsymbol{X}$ to $\boldsymbol{Y}$ and any intervention on $\boldsymbol{X}$ should not affect $\boldsymbol{Y}$.

if $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X})$ in $G_{\overline{\boldsymbol{X}}}$:

$$\mathrm{P}\{\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})\} = \mathrm{P}\{\boldsymbol{Y}\} \tag{3.4}$$

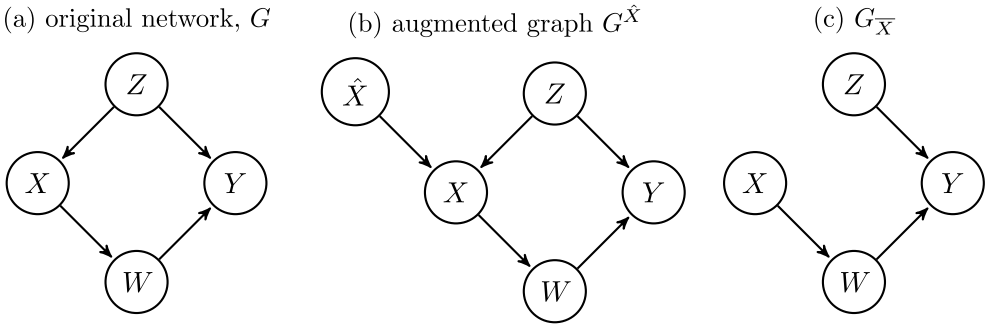
The general case of rule 3 is easier to state by explicitly representing the intervention in the graphical model . Let $G^{\hat{\boldsymbol{X}}}$ denote the graph G after adding a variable $\hat{X}_i$ as a parent of each variable $X_i \in \boldsymbol{X}$ (see figure 3.5b). The variable $\hat{X}_i$ can be thought of as representing the mechanism by which X_i takes its value, either by being set via intervention or as a stochastic function of its other parents [**?**].

if $(\boldsymbol{Y} \perp\!\!\!\perp \hat{\boldsymbol{X}}|\boldsymbol{Z},\boldsymbol{W})$ in $G^{\hat{\boldsymbol{X}}}_{\overline{\boldsymbol{Z}}}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{Z}=\boldsymbol{z}),do(\boldsymbol{X}=\boldsymbol{x}),\boldsymbol{W}=\boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{Z}=\boldsymbol{z}),\boldsymbol{W}=\boldsymbol{w}) \tag{3.5}$$

The statement that $\boldsymbol{Y} \perp\!\!\!\perp \hat{\boldsymbol{X}}$ (without conditioning on $\boldsymbol{X}$) implies that there is no unblocked path from $\boldsymbol{X}$ to $\boldsymbol{Y}$ in G which *includes* an arrow leaving $\boldsymbol{X}$. These are the only paths by which intervening in $\boldsymbol{X}$ can effect $\boldsymbol{Y}$.

Figure 3.5: Example application of equation 3.5. $(Y \perp\!\!\!\perp \hat{X}|W, Z) \implies \mathrm{P}\{Y|do(X), W, Z\} = \mathrm{P}\{Y|W, Z\}$. We have to condition on Z because conditioning on W blocks the path $\hat{X} - X - W - Y$ but opens $\hat{X} - X - Z - Y$.



3.2.1 Identifiability

A natural question to ask is, given a set of assumptions about the causal graph, is it possible to estimate a given interventional distribution from observational data? This is the identifiablity problem. It asks if we can obtain an unbiased point estimate for the causal query of interest in the infinite data limit. A query is non-parametrically identifiable if it is identifiable without assumptions on the functional form of the dependencies between variables in the graph.

Definition 10 (Non-parametric identifiablity)**.** Let G be a causal graph containing observed variables $\boldsymbol{V}$ and latent variables $\boldsymbol{U}$ and let $\mathrm{P}\{\cdot\}$ be any positive distribution over $\boldsymbol{V}$. A causal query of the form $\mathrm{P}\{\boldsymbol{Y}|do(\boldsymbol{X}), \boldsymbol{W}\}$, where $\boldsymbol{Y}$,$\boldsymbol{X}$ and $\boldsymbol{W}$ are disjoint subsets of $\boldsymbol{V}$, is non-parametrically identifiable if it is uniquely determined by $\mathrm{P}\{\cdot\}$ and G.

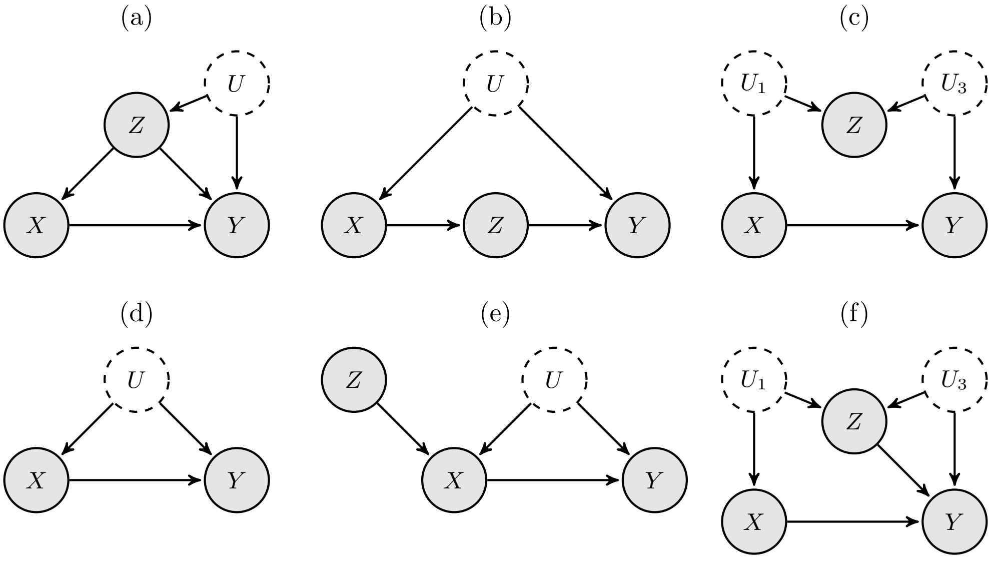
The question of non-parametric identifiability is solved. The do calculus is complete [**?** 26]. A problem is identifiable if and only if the interventional distribution of interest can be transformed into term containing only observational quantities via repeated application of the do calculus. There is a polynomial time algorithm [**?**] based on these properties that, for a given network and interventional (do-type) query, can:

1. determine if the query can be translated into an expression involving only distributions over observed variables. In other words, determine if the query is identifiable given the assumptions encoded by the network.
2. if it is identifiable, return the required expression.

Figure 3.6 shows some examples of identifiable and non identifiable queries. I have created a javascript implementation of the identifiability algorithm [**?**] on which you can test your own queries http://finnhacks42.github.io.

Many interesting questions relating to identifiability remain open. What is the minimal (by some metric) additional information that would be required to make a non-identifiable query identifiable? What if we assume various restrictions on the functional form of the relationships between the variables? Some queries which are not non-parametrically identifiable can be identified by additional assumptions such as linearity. A complete algorithm for the problem of linear identifiability is yet to be found, despite a rich body of work [**? ? ?**].

Figure 3.6: Examples of identifiable and non-identifiable queries. In subfigures (a), (b) and (c) the causal query $\mathrm{P}\left\{Y|do(X)\right\}$ is identifiable. In subfigures (d), (e) and (f) it is not.



Although identifiability is a natural and important question to ask, it does not partition causal questions into solvable and unsolvable. Estimators for identifiable queries can be slow to converge and we may be able to obtain useful bounds on causal effects in cases where point estimates are not identified.

3.3 Estimation

3.3.1 Defining causal effects

So far we have described causal effect estimation in term of identifying the interventional distribution $\mathrm{P}\left\{Y|do(X)\right\}$ from observational data. This interventional distribution is in fact a family of distributions parametrised by the value, x, to which the treatment variable X is set. From a decision theoretic viewpoint, we can select an optimal action x by specifying a utility function $\mathcal{U} : y \in \mathcal{Y} \to \mathbb{R}$ that assigns a value to each outcome y and then selecting the action that maximises the expected utility.

$$x* = \arg\max_{x} \mathbb{E}_{y \sim \mathrm{P}\{Y|do(X=x)\}}\left[\mathcal{U}(y)\right] \tag{3.6}$$

Frequently however, studies wish to define and estimate a causal effect without reference to a specific utility function. There are a variety of ways of defining causal effects that can be viewed as different ways of summarising the family of interventional distributions. For a binary treatment variable X, the average causal effect, ACE [1] is defined as;

$$ACE = \mathbb{E}\left[Y|do(X=1)\right] - \mathbb{E}\left[Y|do(X=0)\right] \tag{3.7}$$

Assuming the expectations in equation 3.7 are well defined, the ACE captures the shift in the mean outcome that arises from varying X. It does not capture changes in variance or higher

[1] also referred to as the average treatment effect (ATE)

moments of the distribution. The ACE can be generalised to non-discrete interventions by considering the effect on the expectation of Y of an infinitesimal change in x. If X is linearly related to Y then the ACE is constant and equivalent to the corresponding coefficient in the linear structural equation model.

$$ACE(x) = \frac{d}{dx}\mathbb{E}\left[Y|do(X = x)\right] \tag{3.8}$$

The average causal effect is often introduced as the average over individual causal effects as discussed in section 2.2. Individual causal effects are deterministic and cannot be expressed as properties of the iterventional distribution. However we can personalise the average causal effect by stating it with respect to some observed context. I will refer to this as the personalised causal effect (PCE) [2].

$$PCE(z) = \mathbb{E}\left[Y|do(X = 1), z\right] - \mathbb{E}\left[Y|do(X = 0), z\right] \tag{3.9}$$

In some cases we may be interested in the average causal effect for some sub-group of the population. A particularly common example of this is the average treatment effect of the treatment of the treated (ATT). This would be the key quantify of interest is we had to decide whether or not to continue providing a program or treatment for which we could not control the treatment assignment process.

$$ATT = \mathbb{E}_{z\sim \mathrm{P}\{Z|x=1\}}\left[Y|do(X = 1)\right] - \mathbb{E}_{z\sim \mathrm{P}\{Z|x=1\}}\left[Y|do(X = 0)\right] \tag{3.10}$$

We can also summarise causal effects with counterfactuals. The ACE is $\mathbb{E}\left[Y^1 - Y^0\right]$. We could also estimate the ratio of expectations $\frac{\mathbb{E}[Y^1]}{\mathbb{E}[Y^0]}$. However, the quantity $\mathbb{E}\left[\frac{Y^1}{Y^0}\right]$ depends on the joint distribution over the counterfactual variables (Y^1, Y^0) and thus cannot be computed from the interventional distribution.

Another way of conceptualising causal effects is as a property indicating the strength of the causal link between two variables. This notion is complex to formalise when the relationship between variables is non-linear. Suppose $Y = X \oplus Z$ with $P(Z = 1) = \frac{1}{2}$, the interventional distributions over X are identical after marginalising out Z. Janzing et al. [31] propose a number of postulates that a notion of casual strength could satisfy, demonstrate why previous measures fail these postulates and propose an alternative based on information flow.

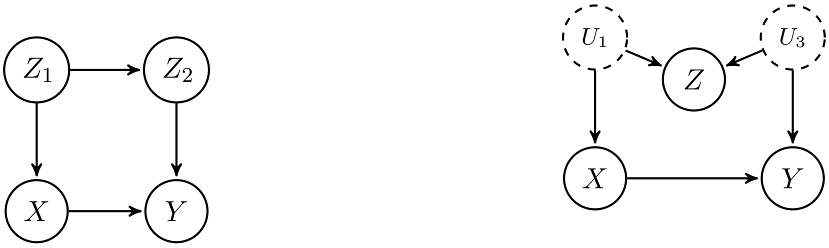
3.3.2 Estimating causal effects by adjusting for confounding variables

Probably the two most frequently applied approaches to estimating causal effects from observational data are instrumental variables and adjusting for confounding factors. Instrumental variables correspond to the graph in figure 3.6e, which is not identifiable without parametric assumptions, however they can provide tight bounds. Adjusting for confounding equates to identifying a set of variables $\boldsymbol{Z}$ such that the ignorability assumption discussed in section 2.2 holds. This corresponds to a simple graphical test known as the backdoor criterion [38]. The setting is also referred to as unconfounded.

[2] This quantity is sometimes called the conditional average treatment effect CATE, however that term is also used for the sample rather than population effect

Figure 3.7

(a) There can be multiple valid adjustment sets. (b) Conditioning on Z opens the backdoor path
Z_1 or Z_2 or $\{Z_1, Z_2\}$ all block the backdoor path $X - U_1 - Z - U_2 - Y$
from X to Y.



Theorem 11 (The backdoor criterion)**.** *[38] Let $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{Y}$ be disjoint sets of vertices in a causal graph G. If $\boldsymbol{Z}$ blocks (see Definition 7) for every path from X_i to Y_j that contains a link into X_i, for every pair $(X_i \in \boldsymbol{X}, Y_j \in \boldsymbol{Y})$, and no node in $\boldsymbol{Z}$ is a decedent of a node in $\boldsymbol{X}$ then the backdoor criterion is satisfied and;*

$$\mathrm{P}\left\{\boldsymbol{y}|do(\boldsymbol{x})\right\} = \sum_{\boldsymbol{z}} \mathrm{P}\left\{\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}\right\} \mathrm{P}\left\{\boldsymbol{z}\right\} \tag{3.11}$$

The backdoor criterion derives from rule 2 of the do-calculus. Selecting which covariates should be adjusted for to estimate a causal effect reduces to identifying a set which satisfies the backdoor criterion. There may be more than one valid adjustment set, see figure 3.7a. The seemingly simple problem of determining if a variable should be adjusted for when estimating causal effects has been the subject of substantial debate and controversy []. Adjusting for the wrong variables (even pre-treatment variables) can introduce or magnify bias, see figure 3.7b. Causal graphs and the back door criterion provide a clear mechanism by which to decide which variables should be adjusted for. For a practical example, see the discussion in **?**] on whether birth weight should be adjusted for to estimate the causal effect of smoking on neonatal mortality.

Given a set of variables $\boldsymbol{Z}$ satisfies the backdoor criterion (or equivalently the conditional ignorability assumption), the interventional distribution is asymptotically identifiable and can be estimated from equation 3.11 and the expected value of Y after the intervention $do(X = x)$ is given by equation 3.12 and the average causal effect for a binary intervention $x \in \{0, 1\}$ is given by equation 3.13.

$$\mathbb{E}\left[Y|do(X = x)\right] = \mathbb{E}_{z \sim \mathrm{P}\{\boldsymbol{Z}\}}\left[\mathbb{E}\left[Y|x, \boldsymbol{z}\right]\right] \tag{3.12}$$

$$ACE = \mathbb{E}_{z \sim \mathrm{P}\{\boldsymbol{Z}\}}\left[\mathbb{E}\left[Y|1, \boldsymbol{z}\right] - \mathbb{E}\left[Y|0, \boldsymbol{z}\right]\right] \tag{3.13}$$

Assuming x and $\boldsymbol{z}$ are discrete, equation 3.12, and thus the ACE, can be estimated by selecting the data for which $X = x$, stratifying by $\boldsymbol{Z}$, then computing the mean outcome within each strata and finally weighting the results by the number of samples in each strata. However this approach is not workable for most real problems with finite samples as the number of strata grows exponentially with the dimension of $\boldsymbol{Z}$ and it cannot handle continuous covariates. There is a substantial work within in the statistics and econometrics literature on estimating average causal effects assuming conditional ignorability, see Imbens [29] for a comprehensive review. The key approaches are based on matching on covariates, propensity score methods and regression. We now examine these approaches from a machine learning perceptive.

In standard supervised learning, we have a training set $(\boldsymbol{x_1}, y_1), ..., (\boldsymbol{x_n}, y_n)$ assumed to be sampled i.i.d from an unknown distribution $\mathrm{P}\left\{\boldsymbol{x}, y\right\} = \mathrm{P}\left\{\boldsymbol{x}\right\}\mathrm{P}\left\{y|\boldsymbol{x}\right\}$. The goal is to select a hypothesis $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$ such that, on unseen data $\sim \mathrm{P}\left\{\boldsymbol{x}, y\right\}$, $h(\boldsymbol{x})$ is close (by some metric) in expectation to y. In other words we wish to minimise the generalisation error $E_{out}(h)$,

$$E_{out}(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathrm{P}\{\boldsymbol{x}\}\,\mathrm{P}\{y|\boldsymbol{x}\}}\left[L(h(\boldsymbol{x}), y)\right] \tag{3.14}$$

We cannot directly compute the generalisation error as $\mathrm{P}\left\{\boldsymbol{x}, y\right\}$ is unknown, we only have access to a sample. We could search over $\mathcal{H}$ and select a hypothesis $h^*(\boldsymbol{x})$ that minimises some loss function on the sample data.

$$E_{in}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(\boldsymbol{x}_i), y_i) \tag{3.15}$$

The VC-dimension of the hypothesis space provides (typically loose) bounds on the probability that $E_{out} >> E_{in}$. However, in practice, the generalisation error is usually estimated empirically from a hold out set of the sample that was not used to train the model, or via cross-validation.

In the causal effect estimation under ignoreability, we have training data $(\boldsymbol{x}_1, \boldsymbol{z}_1, y_1), ..., (\boldsymbol{x}_n, \boldsymbol{z}_n, y_n)$ sampled i.i.d from $\mathrm{P}\left\{\boldsymbol{z}\right\}\mathrm{P}\left\{\boldsymbol{x}|\boldsymbol{z}\right\}\mathrm{P}\left\{y|\boldsymbol{x}, \boldsymbol{z}\right\}$. Estimating $\mathbb{E}\left[Y|do(\boldsymbol{X} = \boldsymbol{x})\right]$ corresponds to selecting a hypothesis $h \in \mathcal{H} : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ that minimises;

$$\begin{aligned} E_{out} &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{z},y)\sim\delta(\boldsymbol{x}-\boldsymbol{x'})\,\mathrm{P}\{\boldsymbol{z}\}\,\mathrm{P}\{y|\boldsymbol{x},\boldsymbol{z}\}}\left[L_2(h(\boldsymbol{x}, \boldsymbol{z}), y)\right], & (3.16) \\ &= \mathbb{E}_{(\boldsymbol{z},y)\sim\mathrm{P}\{\boldsymbol{z}\}\,\mathrm{P}\{y|\boldsymbol{x},\boldsymbol{z}\}}\left[L_2(h(\boldsymbol{x}, \boldsymbol{z}), y)\right], & (3.17) \end{aligned}$$

?] identified that this is equivalent to the covariate shift problem. If we let $\boldsymbol{v} = (\boldsymbol{x}, \boldsymbol{z})$ then we have training data sampled from $\mathrm{P}_{train}\left\{\boldsymbol{v}\right\}\mathrm{P}\left\{y|\boldsymbol{v}\right\}$ where $\mathrm{P}_{train}\left\{\boldsymbol{v}\right\} = \mathrm{P}\left\{\boldsymbol{z}\right\}\mathrm{P}\left\{\boldsymbol{x}|\boldsymbol{z}\right\}$ but at test time the data will be sampled from $\mathrm{P}_{test}\left\{\boldsymbol{v}\right\}\mathrm{P}\left\{y|\boldsymbol{v}\right\}$, where $\mathrm{P}_{test}\left\{\boldsymbol{v}\right\} = \delta(\boldsymbol{x} - \boldsymbol{x'})\mathrm{P}\left\{\boldsymbol{z}\right\}$. [3] With this connection to covariate shift in mine, let us return to regression, matching and propensity scores.

Regression

The regression approach is to learn a function that is a good approximation to the output surface $\mathbb{E}\left[Y|X, Z\right]$. Let $f_1(z) = \mathbb{E}\left[Y|X = 1, Z = z\right]$. The expectation of Y after the intervention $X = 1$ is then obtained by taking the expectation with respect to Z, $\mathbb{E}\left[Y|do(X = 1)\right] = \mathbb{E}_{z\sim\mathrm{P}\{Z\}}\left[\mathbb{E}\left[Y|X = 1, z\right]\right]$. We can learn a parametric regression model $\hat{f}_1(z)$ via empirical risk minimisation.

[3] It is not obvious that the question of estimating causal effects under ignorability entirely reduces to covariate shift. Take the case where we have a binary intervention $x \in \{0, 1\}$. Suppose we learn $h(1, \boldsymbol{z}) = \mathbb{E}\left[Y|x = 1, \boldsymbol{z}\right] + g(\boldsymbol{z})$ and $h(0, \boldsymbol{z}) = \mathbb{E}\left[Y|x = 0, \boldsymbol{z}\right] + g(\boldsymbol{z})$, then the estimated average causal effect equals the true average causal effect for any function g, $\mathbb{E}\left[h(1, \boldsymbol{z}) - h(0, \boldsymbol{z})\right] = \mathbb{E}\left[Y|x = 1, \boldsymbol{z}\right] - \mathbb{E}\left[Y|x = 0, \boldsymbol{z}\right]$. More generally, if the goal is to select an optimal action x^* from a continuous space of possible interventions we need algorithms capable of leveraging any structure in the relationship between x and y as well as a means of focusing the loss on regions of the sample likely to affect x^*.

$$\hat{f}_1(z) = h_1(z; \hat{\theta}_{obs}), \text{ where } \hat{\theta}_{obs} = \arg\min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = 1\} \, L\left(h_1(z_i; \theta), y_i\right) \right] \tag{3.18}$$

This estimator is consistent with respect to the observational distribution. As the sample size tends to infinity, $\hat{\theta}_{obs}$ approaches the parameter within the hypothesis space that minimises the expected loss given data sampled from the observational distribution.

$$\lim_{n \to \infty} \hat{\theta}_{obs} = \arg\min_{\theta \in \Theta} \mathbb{E}_{(z,y) \sim \mathrm{P}\{z|x=1\}\,\mathrm{P}\{y|x=1,z\}} \left[L\left(h_1(z; \theta), y\right) \right] \tag{3.19}$$

If the model is correctly specified such that $f_1(z) = h_1(z; \theta^*)$ for some $\theta^* \in \Theta$ then the empirical risk minimisation estimate is consistent with respect to the loss over any distribution of Z [**?**], including the interventional one.

$$\lim_{n \to \infty} \hat{\theta}_{obs} = \theta^* = \arg\min_{\theta \in \Theta} \mathbb{E}_{(z,y) \sim \mathrm{P}\{z\}\,\mathrm{P}\{y|x=1,z\}} \left[L\left(h_1(z; \theta), y\right) \right] \tag{3.20}$$

The average causal effect can then be estimated by,

$$\hat{\tau}_{reg} = \sum_{i=1}^{n} \left(\hat{f}_1(z_i) - \hat{f}_0(z_i) \right) \tag{3.21}$$

Regression thus has a direct causal interpretation if the parametric model is correctly specified and the covariates included form a valid backdoor adjustment set for the treatment variable of interest in the corresponding structural equation model.

Propensity scores

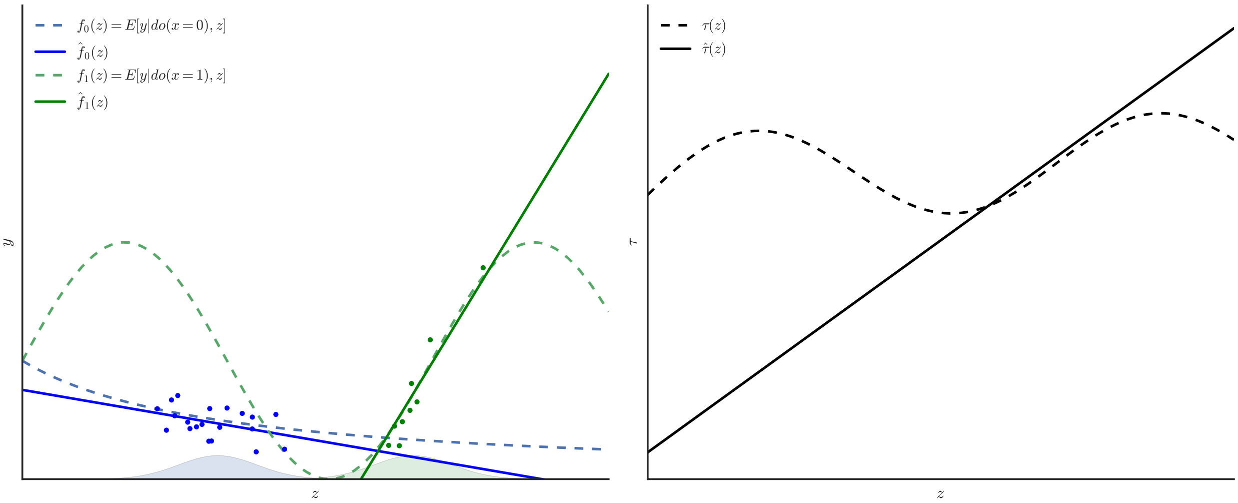
If the parametric model is miss-specified then the parameter that minimises the loss depends on the distribution from which the covariates z are sampled. The model learned by ERM could perform very well in a validation set (which estimates the generalisation error over the observational distribution of (x, z)) but yield very poor estimates of the causal effect see figure 3.8.

A general approach to estimating the expectation of some function $f(\cdot)$ with respect to data from some distribution $\mathrm{P}\{\cdot\}$, when we have data sampled from a different distribution $\mathrm{Q}\{\cdot\}$ is importance sampling [**? ?**].

$$\mathbb{E}_{\boldsymbol{v} \sim \mathrm{P}\{\boldsymbol{v}\}} \left[f(v) \right] = \mathbb{E}_{\boldsymbol{v} \sim \mathrm{Q}\{\boldsymbol{v}\}} \left[f(v) \frac{\mathrm{P}\{\boldsymbol{v}\}}{\mathrm{Q}\{\boldsymbol{v}\}} \right] \tag{3.22}$$

This importance weighting approach can be applied to the covariate shift/average causal effect problem by weighting the terms in the empirical risk minimisation estimator [**?**].

Figure 3.8: Estimating causal effects with parametric regression can do badly when the model is miss-specified even if the regression models fit well. In this example, $\mathrm{P}\{Z|X=0\} \sim N(\mu_0, \sigma_0^2)$ and $\mathrm{P}\{Z|X=1\} \sim N(\mu_1, \sigma_1^2)$ with little overlap in the densities. If $X=0$ then $Y \sim N(f_1(x) = sin(x), \sigma_y^2)$ and if $X=1$ then $Y \sim N(f_0(x) = \frac{1}{x+1}, \sigma_y^2)$. We estimate $f_1(z)$ from the sample in which $X=1$ (green points) and $f_0(z)$ from the sample for which $X=0$ (blue points). In both cases the linear model is a good fit to the data. However, the resulting estimate of the causal effect is very poor for the lower values of z.



$$\hat{\theta}_{iw} = \operatorname*{arg\,min}_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = 1\} \, L\left(h_1(z_i;\theta), y_i\right) \frac{\mathrm{P}\{z_i\}\,\delta(x_i - 1)}{\mathrm{P}\{z_i\}\,\mathrm{P}\{x_i = 1|z_i\}} \right] \tag{3.23}$$

$$= \operatorname*{arg\,min}_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = 1\} \, L\left(h_1(z_i;\theta), y_i\right) \frac{1}{e(z_i)} \right], \tag{3.24}$$

where $e(\boldsymbol{z})$ is the propensity score, defined by [46];

$$e(\boldsymbol{z}) \equiv \mathrm{P}\{x = 1|\boldsymbol{z}\} \tag{3.25}$$

The estimator in equation 3.23 is an example of a doubly robust estimator [**?** **?**]. Doubly robust methods are asymptotically unbiased as long as either the regression model h or propensity score e are correctly specified [**?**].

The propensity score can be used to estimate the average causal effect without specifying a regression model for $\mathbb{E}[Y|X, \boldsymbol{Z}]$. Rosenbaum and Rubin [46] demonstrated that if the ignoreability assumption is satisfied by conditioning on $\boldsymbol{Z}$, then it is also satisfied by conditioning on $e(\boldsymbol{z})$. This allow for estimators based on stratifying, matching or regression on the propensity score rather than the covariates $\boldsymbol{Z}$. Inverse propensity weighting can also be combined with empirical estimation of $\mathbb{E}[Y|X, Z]$ yielding the simple, albeit inefficient, estimator in equation 3.27 [29]. In some settings, such as stratified randomised trials [] or bandit problems [] the propensity score may be known. However in general, it must be estimated from data. Frequently this is done with parametric model such as logistic or probit regression []. **?**] review the theoretical properties of key propensity score based estimators, including stratification and inverse propensity weighting.

$$\mathbb{E}\left[Y|do(X=x)\right] = \mathbb{E}_{z\sim \mathrm{P}\{\boldsymbol{Z}\}}\left[\mathbb{E}\left[Y|x,\boldsymbol{z}\right]\right] = \mathbb{E}_{z\sim \mathrm{P}\{\boldsymbol{Z}|\boldsymbol{x}\}}\left[\mathbb{E}\left[Y|x,\boldsymbol{z}\right]\frac{1}{e(\boldsymbol{z})}\right] \tag{3.26}$$

$$\hat{\tau}_{ip} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}\{x_i=1\}\,y_i}{e(\boldsymbol{z}_i)} - \frac{\mathbb{1}\{x_i=0\}\,y_i}{1-e(\boldsymbol{z}_i)}\right) \tag{3.27}$$

Matching

There is a straightforward connection between matching and regression for causal effect estimation. If $h \in \mathcal{H} \implies h + a \in \mathcal{H}$ for any constant a and $\hat{f}$ is selected by minimising empirical risk with an L_2 loss then $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i=1\}\,\hat{f}_1(z_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i=1\}\,y_i$ [4], and equation 3.21 can be re-written as,

$$\hat{\tau}_{reg} = \sum_{i=1}^{n}\left[\mathbb{1}\{x_i=1\}\left(y_i - \hat{f}_0(z_i)\right) + \mathbb{1}\{x_i=0\}\left(\hat{f}_1(z_i) - y_i\right)\right] \tag{3.28}$$

This formulation of the regression estimator highlights the missing data aspect of casual effect estimation. For each instance, the regression models are used to estimate the counterfactual outcome had the instance received the alternate treatment. Matching estimates the counterfactual outcome for an instance from the outcome of *similar* instances that received a different treatment.] analysis an estimator where both target and control instances are matched and the matching is done with replacement, let $j \in J_k(i)$ be the indices of the k instances closest to i by some metric $d(z_i, z_j)$ such that $x_i \neq x_j$.

$$\hat{\tau}_{match} = \sum_{i=1}^{n}\left[\mathbb{1}\{x_i=1\}\left(y_i - \frac{1}{k}\sum_{j\in J_k(i)} y_j\right) + \mathbb{1}\{x_i=0\}\left(\frac{1}{k}\sum_{j\in J_k(i)} y_j - y_i\right)\right] \tag{3.29}$$

This estimator is equivalent to equation 3.28 with k nearest neighbour regression. There are many variants of matching estimators utilising different distance metrics, matching with or without replacement (and in the latter case, greedy or optimal matching) and with or without discarding matches beyond some threshold [**?** **?**]. Although intuitive, matching estimators in general have poor large sample properties [**?**]. An exception is where the goal is to estimate the average treatment effect of treatment on the treated in settings where there is a large set of control instances (compared to treatment instances) [29].

The practical performance of the estimation approaches discussed in this section will depend on the sample size, dimensionality of the covariates, the complexity of the treatment assignment mechanism and output function and the degree of prior knowledge available about these functions. A key difference between causal effect estimation and standard machine learning problems is that we cannot directly apply cross-validation or a hold out set for model selection because we lack samples from the counterfactual.

The significance to this should not be underestimated. Cross-validation has allowed applied machine learning to succeed with a very a-theoretical approach on the basis that we can identify

[4] [] state this holds for most implementations

when a model is successful. With causal effect estimation there is no guarantee that a model which performs well at prediction (even out of sample) will accurately estimate the outcome of an intervention.] propose inverse propensity weighted cross validation for the covariate shift problem. There is relatively little theory on model selection for estimating the propensity score. To achieve asymptotically unbiased estimates, the covariates should be satisfy the backdoor criterion. It is also known that including variables that influences X but not Y increases variance without any reduction in bias [] and can only increase bias if there are unmeasured confounding variables []. With doubly robust estimators, one could apply an iterative approach, fitting a propensity score model, using the results for inverse propensity weighted cross-validation of the regression model and then selecting covariates for the propensity model on the assumption the estimated regression function was correct. I have found no examples of such approaches.

There have been many studies attempting to quantify the success of different methods for causal estimation on simulated data []. Another approach is to compare estimates from observational studies with the results from corresponding experiments [2]. Unfortunately there are a relatively small number of examples where comparable observational and experimental data are available. The results have been mixed with later studies finding generally better alignment of results but it is hard to ascertain if this is due to improved methodologies or over-fitting to the available data-sets.