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 of everything from the effect of education on earnings [26], diet on cancer [17] and breastfeeding on intelligence [66] to the effect of pet ownership on survival after a heart attack [43]. 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 graphs 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 [15, 107, 94, 4, 121, 47, 117, 123], medical imaging [95] and climate science [125]. An effective and generalisable approach for causal discovery would be 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 on an 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 basic 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, and
- 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 people's 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 the 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 in combining the assumption of how an intervention changes the graph, encoded by the truncated product formula, with information the graph structure provides about conditional independencies 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 [87]. 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 additional 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 properties of Bayesian networks. A more thorough introduction (including proofs) can be found in [74]. 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.*

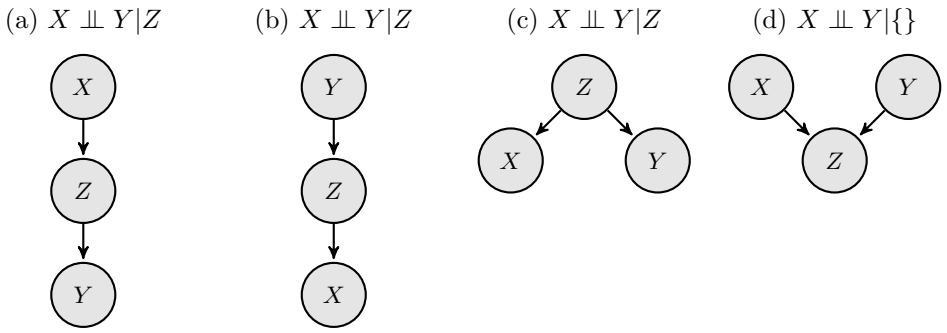
The set of conditional independence relations given by the local Markov condition can enforce additional independencies that also hold in all distributions that are compatible with G. D-separation is an algorithm that extends the local Markov property to find these additional independencies. It provides 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 that 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 disad-

vantaged students. Let X be gender, Y be family background and Z receipt of the scholarship. There are roughly equal numbers 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



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 [87, 97] 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 [48].

3.2 The Do Calculus

The do-calculus is a set of three rules [86] 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)

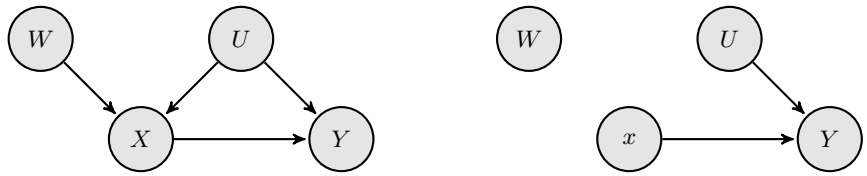
Rule 1 allows us to remove (or insert) observational evidence from the right hand side of a conditional interventional distribution. It follows directly from the fact that the relationship between d-separation in a network and independence in the corresponding probability distribution still applies after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

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

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

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 describes when conditioning on $\boldsymbol{X} = \boldsymbol{x}$ and intervening, $do(\boldsymbol{X} = \boldsymbol{x})$, have the same effect on the distribution over $\boldsymbol{Y}$. 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 $G_{\underline{\boldsymbol{X}}}$:

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

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 and therefor 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. This 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.

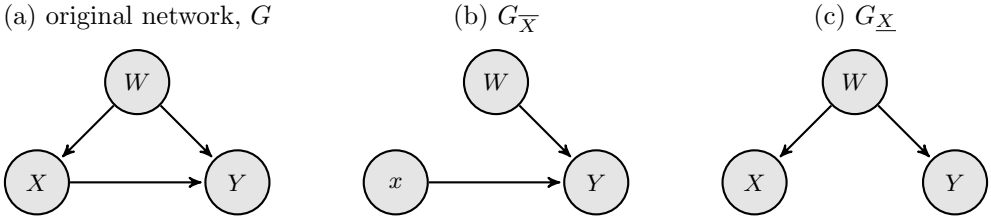
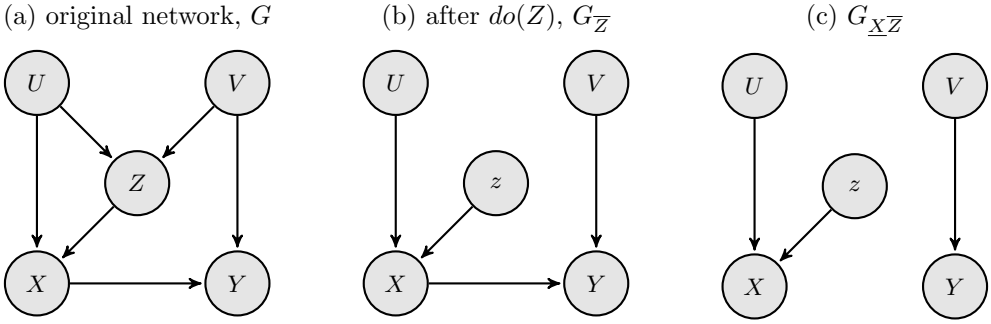


Figure 3.4: An example of applying equation 3.3. In this case $(Y \perp\!\!\!\perp X|Z)$ in $G_{\underline{X}\overline{Z}} \implies \mathrm{P}(Y|do(X = x), do(Z = z)) = \mathrm{P}(Y|X = x, do(Z = z))$. Observing, rather than intervening, on Z would not have allowed us to exchange $do(X = x)$ for $X =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



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 $\boldsymbol{X}$ will not affect $\boldsymbol{Y}$.

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

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

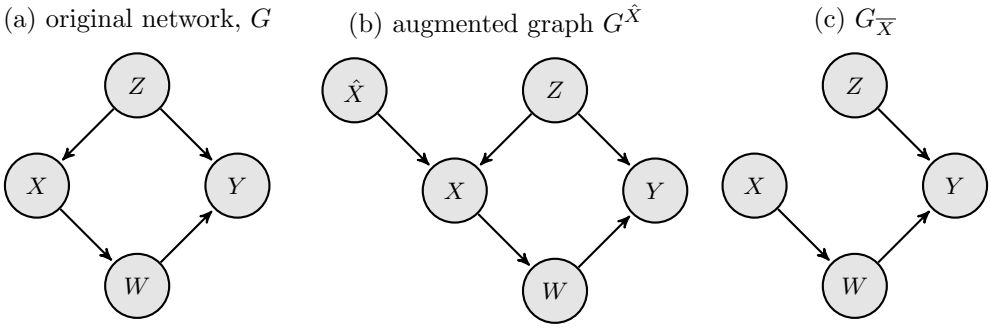
The general case of rule 3 is easier to state by explicitly representing the intervention in the graphical model. Let $G^{\hat{\boldsymbol{X}}}$ denote the graph G after adding a variable $\hat{X}_i$ as a parent of each variable $X_i \in \boldsymbol{X}$ (see figure 3.5b). The variable $\hat{X}_i$ can 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 [74].

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. $(Y \perp\!\!\!\perp \hat{X}|W,Z) \implies \mathrm{P}\left(Y|do(X),W,Z\right) = \mathrm{P}\left(Y|W,Z\right)$. 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 about 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 $\mathrm{P}\left(\cdot\right)$ be any positive distribution over $\boldsymbol{V}$. A causal query of the form $\mathrm{P}\left(\boldsymbol{Y}|do(\boldsymbol{X}),\boldsymbol{W}\right)$, where $\boldsymbol{Y}$,$\boldsymbol{X}$ and $\boldsymbol{W}$ are disjoint subsets of $\boldsymbol{V}$, is non-parametrically identifiable if it is uniquely determined by $\mathrm{P}\left(\cdot\right)$ and G.

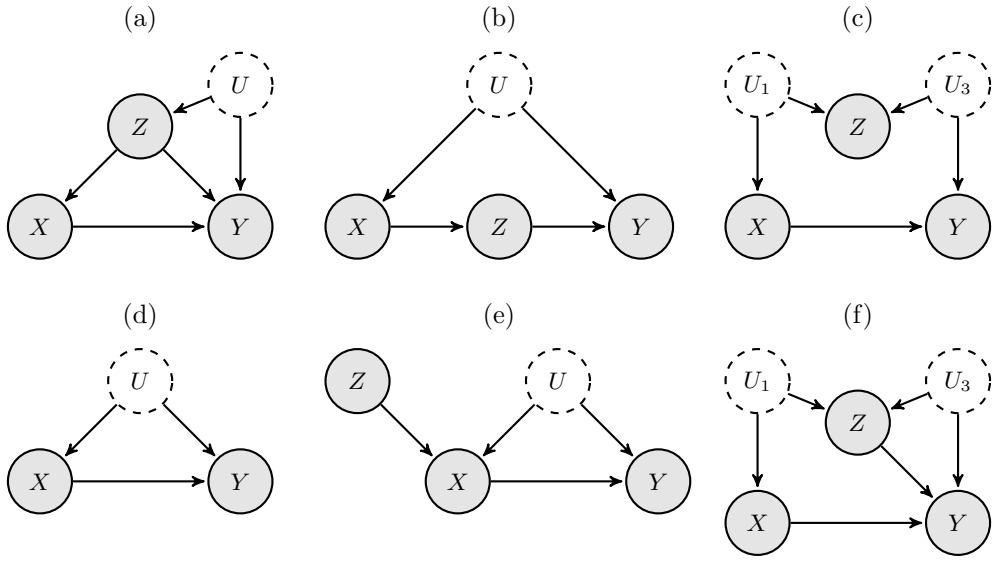
The question of non-parametric identifiability is solved. The do calculus is complete [113, 62]. A problem is identifiable if and only if the interventional distribution of interest can be transformed into term containing only observational quantities via repeated application of the do-calculus. There is a polynomial time algorithm [112] 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, and
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 [112] 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 identifiable non-parametrically 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 [28, 124, 37].

Figure 3.6: Examples of identifiable and non-identifiable queries. In subfigures (a), (b) and (c) the causal query $\mathrm{P}(Y|do(X))$ 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 $\mathrm{P}(Y|do(X))$ from observational data. This interventional distribution is in fact a family of distributions parameterised 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 $\mathcal{U} : y \in \mathcal{Y} \rightarrow \mathbb{R}$ 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 interventional 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 quantity of interest if we had to decide whether or not to continue providing a program or treatment for which we could not control the treatment assignment process.

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

Causal effects can also be written in terms of counterfactuals. The ACE is $\mathbb{E}\left[Y^1 - Y^0\right]$. We could estimate the ratio of expectations $\frac{\mathbb{E}\left[Y^1\right]}{\mathbb{E}[Y^0]}$ instead of the difference. 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. [67] propose a number of postulates that a notion of causal 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 [87]. 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)**.** *[87] 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 [88]. 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 for deciding which variables should be adjusted for. For a practical example, see the discussion in Schisterman et al. [109] on whether birth weight should be adjusted for to estimate the causal effect of smoking on neonatal mortality.

Given that 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. 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 stratum and finally weighting the results by the number of samples in each strata. However this approach is not workable for most real world 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 body of work within in the statistics and econometrics literature on estimating

average causal effects assuming conditional ignorability, see Imbens [64] 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 $\mathrm{P}(\boldsymbol{x}, y) = \mathrm{P}(\boldsymbol{x})\,\mathrm{P}(y|\boldsymbol{x})$. The goal is to select a hypothesis $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$ such that, on unseen data $\sim \mathrm{P}(\boldsymbol{x}, y)$, $h(\boldsymbol{x})$ 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 $\mathrm{P}(\boldsymbol{x}, y)$ 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 ignorability, 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 $\mathrm{P}(\boldsymbol{z})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{z})\,\mathrm{P}(y|\boldsymbol{x}, \boldsymbol{z})$. 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;

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

$$= \mathbb{E}_{(\boldsymbol{z},y)\sim \mathrm{P}(\boldsymbol{z})\,\mathrm{P}(y|\boldsymbol{x},\boldsymbol{z})}\left[L_2(h(\boldsymbol{x}, \boldsymbol{z}), y)\right], \tag{3.17}$$

Johansson et al. [70] 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 $\mathrm{P}_{train}\{\boldsymbol{v}\}\,\mathrm{P}(y|\boldsymbol{v})$ where $\mathrm{P}_{train}\{\boldsymbol{v}\} = \mathrm{P}(\boldsymbol{z})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{z})$ but at test time the data will be sampled from $\mathrm{P}_{test}\{\boldsymbol{v}\}\,\mathrm{P}(y|\boldsymbol{v})$, where $\mathrm{P}_{test}\{\boldsymbol{v}\} = \delta(\boldsymbol{x} - \boldsymbol{x'})\,\mathrm{P}(\boldsymbol{z})$. [3] With this connection to covariate shift in mind, 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, $\mathbb{E}\left[Y|do(X = 1)\right] =$

[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^*.

$\mathbb{E}_{z\sim \mathrm{P}(Z)}\left[\mathbb{E}\left[Y|X=1,z\right]\right]$. We can learn a parametric regression model $\hat{f}_1(z)$ via empirical risk minimisation.

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

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.

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

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 [122], including the interventional one.

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

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

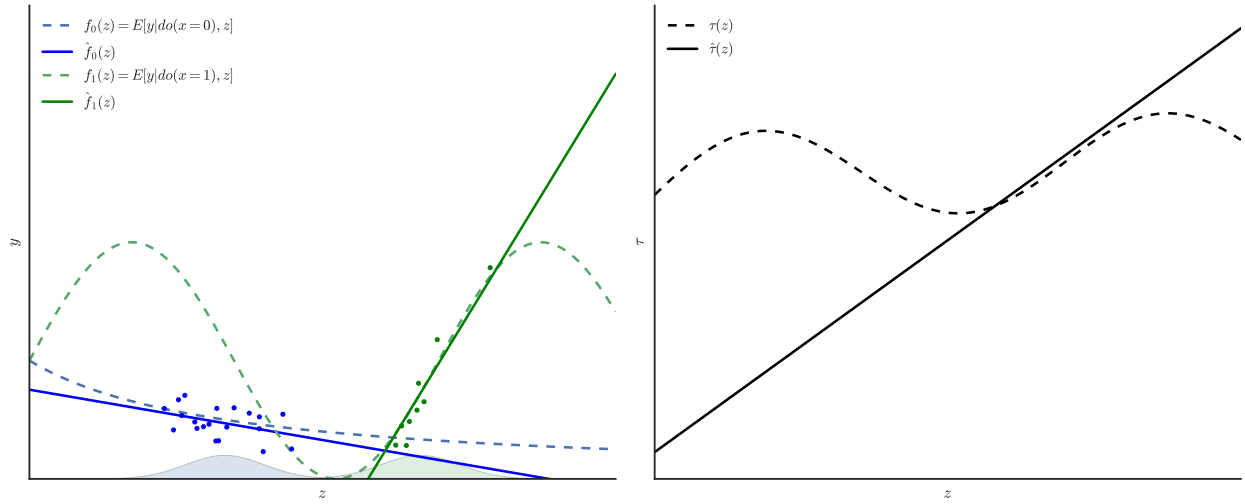
If the parametric model is miss-specified then the parameter that minimises the loss depends on the distribution from which the covariates z are sampled. The model learned by ERM could perform very well in a validation set (which estimates the generalisation error over the observational distribution 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 [59, 74].

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

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 [122].

Figure 3.8: Parametric regression may yeild poor estimates of causal effects if the model is miss-specified, even if the model fits well over the domain of the training data. In this example, $\mathrm{P}\left(Z|X=0\right) \sim N(\mu_0, \sigma_0^2)$ and $\mathrm{P}\left(Z|X=1\right) \sim N(\mu_1, \sigma_1^2)$ 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.



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

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

where $e(\boldsymbol{z})$ is the propensity score, defined by [101];

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

The estimator in equation 3.23 is an example of a doubly robust estimator [108**?**]. Doubly robust methods are asymptotically unbiased as long as either the regression model h or propensity score e are correctly specified [100].

The propensity score can be used to estimate the average causal effect without specifying a regression model for $\mathbb{E}\left[Y|X, \boldsymbol{Z}\right]$. Rosenbaum and Rubin [101] demonstrated that if the ignorability assumption is satisfied by conditioning on $\boldsymbol{Z}$, then it is also satisfied by conditioning on $e(\boldsymbol{z})$. This allows 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 $\mathbb{E}\left[Y|X, Z\right]$ yielding the simple, albeit inefficient, estimator in equation 3.27 [64]. In some settings, such as stratified randomised trials [65] or learning from logged bandit feedback [18] the propensity score may be known. However in general, it must be estimated from data. Frequently this is done with a simple parametric model such as logistic regression, but a wide range of standard machine learning algorithms including bagging and boosting, random forests and nueral networks can also be applied [12]. Lunceford et al. [79] 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. Abadie and Imbens [1] analyse 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 [30, 102]. Although intuitive, matching estimators in general have poor large sample properties [2]. 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) [64].

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 standard machine learning problems and causal effect estimation is that when estimating causal effects we cannot directly apply cross-validation or a hold-out set for model selection because we lack samples from the counterfactual.

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

[4] [64] state this holds for most implementations

when a model is successful. With causal effect estimation there is no guarantee that a model that performs well at prediction (even out of sample) will accurately estimate the outcome of an intervention. Sugiyama et al. [122] 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 satisfy the backdoor criterion. It is also known that conditioning on instrumental variables, which directly influence X but not Y, increases variance without any reduction in bias and can increase bias if there are unmeasured confounding variables [130, 16, 89, 83]. 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 an approach.

The performance of methods for causal effect estimation can be tested on simulated data [44, 137, 58, 36] or by comparing estimates from observational studies with the results from corresponding exeriments [75, 42, 57, 56, 35, 116, 7]. Unfortunately there are a relatively small number of examples where comparable observational and experimental data are available. The results are mixed with later studies finding generally better alignment of results but it is hard to ascertain if this is due to improved methodological approaches or over-fitting to the available data-sets.

3.4 Causal Discovery

We now move to the much more general problem of learning a causal graph from observational data. In this setting we make much broader assumptions about the structure of the graph. For example, that it is acyclic or that we have no unmeasured confounding variables. We do not assume the existence or directions of any links between the variables. Amazingly, it is possible to infer some aspects of causal structure with such general assumptions. The set of conditional independence in a non-experimental data set indicates some causal structures are more likely than others. In addition, there can be subtle asymmetries in the relationship between the joint distribution of cause and effect and the distributions of cause given effect and effect given cause. These clues are the key to causal discovery algorithms.

Causal discovery is a much grander goal than causal effect estimation given a known causal network. Arguably, if achieved, it would equate to the automation of scientific discovery. We need simply supply our algorithm with a vast collection of variables (regardless of their relevance to the problem) and it would learn the causal structure and from that allow us to estimate the effects of any intervention we cared to make. Unfortunately, causal discovery is very hard. Even with the assumption that the causal graph is acyclic and there are no latent variables, the number of possible graphs grows exponentially with the square of the number of variables .

In the next sections we briefly survey the key approaches to causal discovery. We roughly divide the methods into those based on those that exploit the connection between the conditional independencies in a joint distribution and the structure of a causal model and those that leverage assumptions about the functional form of the relationships between cause and effect.

3.4.1 Conditional independence based methods

One general approach is to look for clues about the structure of the network in the conditional independence relations in the distribution. For any Bayesian network, G, (causal or not) we can read of conditional independencies in the joint distribution from the structure of the network. If

a set of variables Z d-separates X and Y in G then $(X \perp\!\!\!\perp Y|Z)$ in the distribution P. However, we want to work in the other direction, from conditional independence in the distribution to the structure of the network. This requires that we assume the reverse condition: $(X \perp\!\!\!\perp Y|Z)$ in P must imply Z d-separates X and Y in G. This assumption, commonly referred to as ***faithfulness* ??**, says there are no additional independence relations that are satisfied in P but not in all distributions $\boldsymbol{P'}$ that are compatible with G. Stating that P is faithful to G is equivalent to G is a ***perfect map*** [85] for P.

Faithfulness is an assumption. It does not always hold and we cannot verify it from the observational data we wish to use for causal inference. However, most distributions generated by a causal Bayesian network will be faithful to that network. For faithfulness to be violated, different causal effects must exactly balance one-another out. For example, consider a simple binary variable model of chocolate consumption, income and obesity (figure). If the coefficients in the conditional probability tables are just right then the direct effect of chocolate on obesity will exactly balance the indirect effect through income and obesity will appear independent of chocolate consumption. However, this independence is not stable. It would disappear under a small perturbation to any of the parameters. In discrete systems, violations correspond to the solutions to polynomial equations over values in the CPD tables and thus are a space of measure zero with respect to all possible distributions associated with the graph [74].

Given the faithfulness assumption, our causal discovery problem reduces to finding the set of Bayesian networks that have exactly the dependency structure as we observe in P. This set can also be referred to as the Markov equivalence class compatible with P.

Without hidden common causes

The strong assumption that there are no hidden variables that cause two or more variables in $\boldsymbol{V}$ significantly reduces the 'search space' of Bayesian networks we must consider. This assumption is referred to as causal sufficiency [118].

We will begin with a brute force algorithm (described as the SGS algorithm in Spirtes et al. [118] and IC algorithm in Pearl [87]). While it is impractical for all but the smallest of networks, it demonstrates key concepts that also underlie the more useful and complex algorithms we will discuss later.

The SGS (or IC) Algorithm

Input: A distribution P, over variables $\boldsymbol{V}$, that was generated by and is faithful to an (unknown) Bayesian network G

Output: A partially directed network that represents the Markov equivalence class of G

1. Join all pairs of vertices $(a, b) \in \boldsymbol{V}$ with an un-directed link to form a complete graph.
2. For each link $a - b$ search for a set $\boldsymbol{S}_{a,b} \subseteq V \setminus \{a, b\}$ that renders a and b conditionally independent. If such a set (including the empty set) exists then a and b cannot be directly connected in G so delete the link.
3. For all pairs of non-linked variables (α, β) with a common neighbour, c, if $c \notin \boldsymbol{S}_{\alpha,\beta}$, then c must be a collider in the path α, c, β so add arrows to direct the links $\alpha - c$ and $\beta - c$ towards c.
4. Recursively try to orient any edges that remain un-directed to avoid creating cycles (because they are not there by assumption) and additional colliders (because any colliders were found in step 3).

The SGS algorithm utilises the fact that a collider structure (figure 3.1d) induces a distinct conditional independence relation. Assuming you have a consistent conditional independence test, it converges to return a partially directed network that represents the Markov equivalence class for the generating causal model. Unfortunately the number of conditional independence tests required for step 2 grows exponentially (in the worst case) with the number of variables. Not only that, but for each edge that is in the true network, the algorithm will always tests all other possible subsets of variables. If the assumption that there are no hidden common causes or that the distribution is faithful are violated, step 3 of the SGS algorithm can produce double headed arrows.

The PC algorithm Spirtes et al. [118] modifies step 2 of the SGS algorithm to utilise the fact that if two variables (a, b) are conditionally independent given some set, they will also be conditionally independent given a set that contains only variables adjacent to a or b. It also checks for low order conditional independence relations before higher order ones. This allows it to exploit any sparsity in the true network, leading to much better average case performance [118] (although the worst case, where the true network is complete, is still exponential). With finite data, the order in which the links are considered can change the output (unlike for SGS). The effect of wrongly removing a link early on flows through to later conditional independence tests by changing which nodes are considered adjacent.

The PC Algorithm

Input: A distribution P, over variables $\boldsymbol{V} = \{V_1 ... V_k\}$, that was generated by and is faithful to an (unknown) Bayesian network G

Output: A partially directed network that represents the Markov equivalence class of G

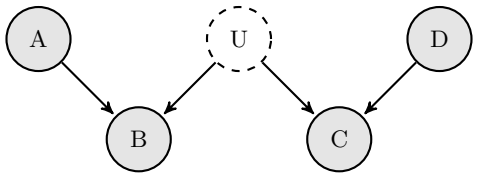
1. As for SGS
2. **for** each link $a - b$:
   $n = 0$
   $\boldsymbol{A}_{a,b} = \{A_1 ... A_j\}$ be the set of nodes adjacent to a and/or b
   while a and b are connected and $n < j$:
   if any subset of size n of $\boldsymbol{A}$ makes a and b conditionally independent:
   delete the link
   $n = n + 1$
3. as for SGS
4. as for SGS

The PC algorithm also returns a set of Markov equivalent networks consistent with the distribution. Since we have assumed there are no hidden variables, for any single graph in this set we can calculate causal effects from the truncated product formula 2.3. We can then bound the true causal effect by combining the results for the all the networks. This procedure is the IDA algorithm [81] and has been found to outperform standard regularisation techniques at finding causal effects in a high-dimensional yeast gene expression data set [80]. An implementation is available in the R package [71]

With hidden variables

There are an number of difficulties in extending the approach of the last section to deal with the case where there are latent variables. With an unknown number of hidden variables there

Figure 3.9: A distribution faithful to this DAG is not faithful to any DAG over the variables $\{A, B, C, D\}$ after marginalising over U.



are infinity many possible structures to search over. In addition, the space of causal networks is not closed under marginalisation. If we have a distribution that $P'(\boldsymbol{O}, \boldsymbol{U})$ generated by and is faithful to a network G the distribution $P(\boldsymbol{O})$, that results from marginalising over $\boldsymbol{U}$, may not be faithful to any Bayesian network (see figure 3.9). The key to constraining the space of possible models is that many latent structures are equivalent (under transforms of the hidden variables).

Theorem 12. *[127] For every latent structure there is a dependency equivalent structure such that every latent (unobserved) variable is a root node with exactly two children .*

Since we only care about the causal relationships between observed variables, it is sufficient to search over networks where any hidden variables have no parents and directly cause two of the observed variables. Instead of representing hidden variables explicitly we can capture the necessary independence relations with a more general graphical model that supports bi-directed edges that play the role of a hidden confounding variable. These models, referred to as maximal ancestral graphs (MAGs) are closed under marginalisation and conditioning.

For any DAG with latent (and selection) variables there is a unique MAG [96]. This makes it possible to extend the PC algorithm to latent structures, resulting in the FCI algorithm [**?**]. The logic behind the algorithm is very similar. Certain structures are ruled out as a consequence of being inconstant with the observed conditional independence relations. The output is an equivalence class of MAGs, which can be represented graphically as a partial ancestral graph PAG [119]. Assuming there are no selection variables, the PAG can contain four types of link:

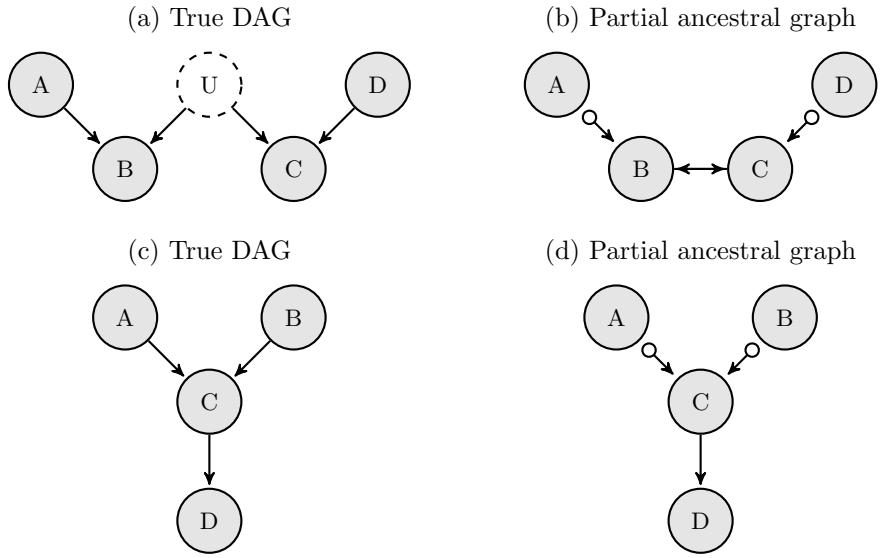
1. $X \to Y$, meaning X causes Y
2. $X \leftrightarrow Y$, meaning there is a latent variable that causes X and Y.
3. $X \circ\!\!\to Y$, either X causes Y or a latent variable causes both.
4. $X \circ\!\!-\!\!\circ Y$, either X causes Y or Y causes X or a latent variable causes both.

The circles indicate where it is ambiguous if there should be an arrowhead (IE where there is one in some MAGs and not in others in the equivalence class). Counter-intuitively it is sometimes possible to rule out or confirm the existence of a confounding variable and fully determine the causal type of a link (see examples in figure 3.10).

The FCI algorithm can be made complete such that it discovers all aspects of the true causal structure that are identifiable from the conditional independence relations of a distribution over observed variables and the faithfulness assumption [134]. More recently [32] have proposed the RFCI algorithm, which in some cases returns more ambiguous links than FCI but is substantially faster. [29] point out that the problem of learning sparse causal networks from data is not NP-hard and propose the FCI+ algorithm, that requires $O(N^{2(k+2)})$ conditional independence tests, where k is the maximum node degree over the observed variables.

Latent variables can create constraints on the marginal distribution over the observed variables that cannot be expressed in term of conditional independencies. These generalised constraints can be expressed and leveraged within nested Markov models [114, 111]

Figure 3.10: FCI examples: true graph and FCI output



All the algorithms discussed in this section rely on being able to perform conditional independence tests. This is non-trivial with high dimensional data. If the functional relationship between the variables is linear with Gaussian noise then the network represents a multivariate normal distribution and a pair of variables are conditionally independent if and only if the corresponding entry in the inverse correlation matrix is non-zero [74]. Where the functions are non-linear one can apply kernelised independence tests [51, 136]

3.4.2 Discovery with functional models

The algorithms we have considered so far return a Markov equivalence class. They cannot distinguish between two models that result in the same set of conditional independence relations. Consider the very simple case where there are only two variables and the possible causal structures are $X \to Y$ or $Y \to X$. These models have the same dependency structure but in one case $P(Y|do(X)) = P(Y|X)$ and in the other $P(Y|do(X)) = P(Y)$. No algorithm relying purely on conditional independence relations can separate these two cases.

Let us focus only on the two variable case $X \to Y$ or $Y \to X$. What possible clues could there be in the distribution $P(X, Y)$ that could indicate which causal model it was generated from? Recall the functional definition of causality (section 2.3). There are a number of assumptions about the form of the functions that can allow us to identify the causal direction: non-invertable functions, additive noise [61], post-non-linear additive noise [135] or linear models with non-Gaussian noise [60].

The causal direction can also be identified via a connection between casual discovery and semi-supervised learning [69]. Suppose we are trying to learn $\mathrm{P}(Y|X)$. The goal of semi-supervised learning is to improve our estimate of $\mathrm{P}(Y|X)$ by leveraging additional data sampled from $\mathrm{P}(X)$. However, if the true causal model is $X \to Y$ then their is some function mapping values of X to Y which should be invariant to any changes in the input distribution $\mathrm{P}(X)$. Therefor $\mathrm{P}(X)$ should be independent of $\mathrm{P}(Y|X)$ and semi-supervised learning should not perform any better than standard supervised learning. However if the true causal model is $Y \to X$ then variations in the $\mathrm{P}(X)$ can result from both the input distribution over Y and the mapping from X to Y and semi-supervised learning could help. This assumption of independence of mechanism and input can also allow the identification of the causal direction in SEMs even where

the functions are deterministic and invertable [33]. Janzing et al. [68] leverage an information geometric viewpoint of the independence of mechanism and input to infer the causal direction between a pair of associated variables.

Instead of positing a functional restriction on the relationship between variables and then developing theory to exploit that assumption, [78] propose learning what the causal relationship looks like from data. They assume there will be a difference between the relationship of $\mathrm{P}(X)$, $\mathrm{P}(Y)$ and $\mathrm{P}(X|Y)$ between $X \to Y$ versus $Y \to X$. Their algorithm requires a data set in which each row is itself a data set consisting of pairs of variables (x_i, y_i) with a label indicating the direction of causality between X and Y. They use a kernel mean embedding to represent the distributions $\mathrm{P}(X)$, $\mathrm{P}(Y)$ and $\mathrm{P}(X|Y)$ as features for each individual sub-data set and train an algorithm to learn the direction of causality. Unfortunately we do not have a large collection of data sets where the causal direction is known to train such a model. Lopez-Paz et al. [78] instead use a simulated data set so their model will necessarily be based on the assumptions they make when generating the data. Nonetheless this approach makes it possible to rapidly construct a model from a wide range of possible assumptions, without doing a lot of theory to design a specific algorithm optimised to that setting.

[91] have extended results from the bi-variate case to the multivariate setting. They show that if we can come up with a condition that guarantees identifiability for the bi-variate case, we can extend that result to get the conditions under which the multivariate case is identifiable. They build on this to develop an algorithm that allows the construction of causal graphs based on the additive noise assumption.