1 Why is causality important

(this section should include explanation of why correlation is not causation)

There are two ways in which correlation is not causation. Firstly correlation may not be statistically significant, - ie not even a genuine association. secondly it can be significant but spurious.

2 Models of causality

To understand causality we first need a more formal definition of what we are talking about. In this section we will look at three slightly different definitions of causality and use them to describe the following very simple example. The aim is to demonstrate the notations and formalisms we will need to tackle more interesting problems later on.

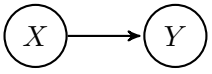
We have developed a new drug for some illness and wish to determine how effective it is. We take a large group of patients and randomly assign half of them to a treatment group and the other half to a control group. The people in the treatment group get the drug, everyone else gets a placebo pill. The question we want to answer is does giving people the active drug improve their changes of recovery relative to giving them the placebo. 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.

2.1 Structural Equation Models

Structural equation models (SEMs) describe a deterministic world, where underlying mechanisms determine 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. Linear structural equation models have a long history for causal estimation [43, 8]. More recently, they have been formalized, generalized to the non-parametric setting and connected to developments in graphical models to provide a powerful causal framework [23].

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. For our example the SEM is:

$$\begin{aligned} X &= \epsilon_x \\ Y &= f(X, \epsilon_y) \end{aligned} \qquad \text{(1)}$$



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).

We want to estimate the causal effect of treatment; what is the probability of recovery if we **take the action** 'treat' versus the **action** 'placebo'? Taking the action 'treat' corresponds to replacing the equation $X = \epsilon_x$ with $X = 1$. The probability distribution over Y given we

set $X = 1$ is then $P(Y = y|do(X = 1)) = P(f(1, \epsilon_y) = y)$, where we have introduced the *do* notation to distinguish setting a variable from observing it [22]. However, for this model, The probability of observing Y given $X = 1$, $P(Y = y|X = 1)$ is also given by $P(f(1, \epsilon_y) = y)$ because $\epsilon_y \perp\!\!\!\perp \epsilon x$. The causal effect is exactly the difference in observed outcomes, as we would intuitively expect for a randomized experiment. In this case, due to the assumption that $X \to Y$ and that there is no hidden common cause, correlation is causation.

For a model with N variables, a structural equation model looks like 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.

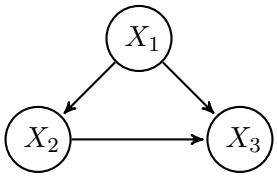
2.2 Causal Bayesian Networks

Causal Bayesian Networks are, unsurprisingly, an extension of Bayesian networks. Any joint probability distribution can be factorized into a product of conditional probabilities. There are multiple valid factorizations, corresponding to permutations of variable ordering.

$$P(X_1, X_2, X_3, ...) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)... \tag{2}$$

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 (figure 1). If there are conditional independencies between variables the factorization simplifies, which is reflected by missing edges in the corresponding network.

Figure 1: General Bayesian network over 3 variables



The statement that a given graph G is a Bayesian network for a distribution P tells us that the distribution can be factorized 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;

$$P(X_1...X_N) = \prod_{i=1...N} P(X_i|parents_{X_i}) \tag{3}$$

A causal Bayesian network is a Bayesian network in which it a link $X_i \to X_j$, 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. More generally, 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 simply dropping the terms for each of the variables in Z from the factorization given by the network. This is referred to as the truncated product formula [23].

$$P_{do(Z=z)}(X_1...X_N) = \begin{cases} \prod_{i \notin Z} P(X_i|parents_{X_i}) & \text{if } (X_1...X_N) \text{ consistant with } Z = z \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Causal Bayesian networks encode information on all interventional distributions over a set of variables. A causal network represents (much) more information than a Bayesian network with identical structure due to the assumption that links can be interpreted causally. 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. There is an algorithm [36] based on these properties that, for a given network and interventional (do-type) query, can:

a) 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

b) if it is identiable, return the required expression

Returning at last to our simple example, and phrasing our query in terms of interventions; what would the distribution of outcomes look like if everyone was treated $P(Y|do(X = 1))$, relative to if no one was treated $P(Y|do(X = 0))$? If we assume that X causes Y and there are no common causes then the causal Bayesian network is just $X \rightarrow Y$, the same structure as the network for the SEM specified by equation (1). This network represents the (causal) factorization $P(X, Y) = P(X)P(Y|X)$, so from equation (4), $P(Y|do(X)) = P(Y|X)$. In this example, the interventional distribution is equivelent to the observational one as we found earlier using the structural equation model.

2.3 Counterfactuals and the Neyman-Rubin framework

The Neyman-Rubin model [30, 31, 29, 32, 33] 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 [41]. 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 our example, 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, you have units you can assume are 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.

Instead of trying to estimate individual effects, lets see if we can 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 we would see of Y if everyone was treated. Similarly Y^0 represents the potential outcome for the placebo. We want to know the difference between the probability of recovery, across the population if everyone was treated,

and the probability of recovery given placebo $P(Y^1) - P(Y^0)$. We can estimate, from an experimental or observational study, the probability that people recover if treated $P(Y|X = 1)$ and the probability that they recover if not treated $P(Y|X = 0)$. Now if $X = 0$ then $Y = Y^0$. Equivalently stated:

$$\begin{aligned} P(Y^0|X = 0) &= P(Y|X = 0) \\ P(Y^1|X = 1) &= P(Y|X = 1) \end{aligned} \tag{5}$$

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

$$\begin{aligned} P(Y^1) = P(Y^1|X = 1) = P(Y|X = 1) \\ P(Y^0) = P(Y^0|X = 0) = P(Y|X = 0) \end{aligned} \tag{6}$$

$$\implies P(Y^1) - P(Y^0) = P(Y|X = 1) - P(Y|X = 0) \tag{7}$$

The assumptions $X \perp\!\!\!\perp Y^1$ and $X \perp\!\!\!\perp Y^0$ are referred to as ignoreability assumptions [29]. 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. Again this is justified in our example due to the randomization of treatment assignment. In general these assumption do not hold. If people were deciding whether or not to buy the treatment, rather than it being randomly assigned, there could be a variable, for example income,D-separation still applies in the augmented model. that influenced both the decision to get treatment and the likelyhood of recovery given treatment or placebo.

2.4 Tying it all together. How do these models relate?

Remarkably for models developed relatively independently in fields with very different approaches and problems, the models we have discussed are definitionally (almost) equivalent.

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 factorization 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 factorization and the truncated product formula (equation 4). Thus, the interventional query $P(Y|do(X))$ is identical in these two frameworks. We can also connect this to counterfactuals via:

$$\begin{aligned} Y^0 &\equiv P(Y|do(X = 0)) \\ Y^1 &\equiv P(Y|do(X = 1)) \end{aligned} \tag{8}$$

The assumption $\epsilon_X \perp\!\!\!\perp \epsilon_Y$, 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:

$$X \perp\!\!\!\perp Y^0 \text{ and } X \perp\!\!\!\perp Y^1 \tag{9}$$

It is possible to relax the independence of errors assumption we made for SEMs slightly to correspond exactly the form of equation (9) without losing any of the power provided by d-separation and graphical identification rules [27]. To determine if and how an interventional query can be non-parametrically identified, it is equivalent to specify assumptions graphically in terms of functional models or bayesian networks or as conditional independence statements involving counterfactual variables (ignorability assumptions). By non-parametrically, I mean that we are not making any assumptions about the form of the relationships between variables. Models that are not non-parametrically identifiable can still be identified given assumptions about the distributions of variables and the functional relationship between them, for example, that the functions are linear or that the noise is additive [25]. This form of assumption fits extremely naturally into the structural equation framework.

We can also pose causal queries that are not interventional and cannot be phrased in terms of the do-notation. Our patients can be broken down into four groups. The first group will recover whether or not they receive treatment, the second group will recover if treated but not on the placebo, the third group will recover on the placebo and not if treated, and the last group will not recover on treatment or placebo. Unfortunately, we don't know which group each person belongs to. Drawing this up as a table:

group	placebo	treatment	probability of group
1	die	die	$\alpha = P(Y^0 = 0, Y^1 = 0)$
2	die	recover	$\beta = P(Y^0 = 0, Y^1 = 1)$
3	recover	die	$\gamma = P(Y^0 = 1, Y^1 = 0)$
4	recover	recover	$\delta = P(Y^0 = 1, Y^1 = 1)$

The queries we have been asking thus far are about $P(Y^0 = 1) = \gamma + \delta$ and $P(Y^1 = 1) = \beta + \delta$, but suppose we asked the question; what is the probability that this patient, who was not treated and died, would have recovered if they had been treated? We know they are in either group 1 or 2 since they died without treatment, so the answer is $\frac{\beta}{\alpha+\beta}$. Can we estimate the $\alpha, \beta, \gamma, \delta$ or in other words, identify the joint distribution over the counterfactuals $P(Y^0, Y^1)$ given the interventional distributions, $P(Y^0)$ and $P(Y^1)$? The answer is no, putting our constraints and unknowns in matrix form:

$$\begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} P(Y^0) \\ 1 - P(Y^0) \\ P(Y^1) \\ 1 - P(Y^1) \\ 1 \end{pmatrix} \implies \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} 1 - P(Y^0) - P(Y^1) + \delta \\ P(Y^1) - \delta \\ P(Y^0) - \delta \\ \delta \end{pmatrix} \tag{10}$$

The value of δ is not determined so the query is not identifiable. However we do get bounds on the terms. Since probabilities cannot be negative, $P(Y1) - P(Y^0) - 1 \leq \delta \leq min(P(Y^1), P(Y^0)$. Note, if we made the additional assumption $\gamma = 0$; that the drug did not cause anyone to die who would otherwise have survived, then we can determine the joint distribution over counterfactuals. Alternatively, if we could assume that after treatment people returned to their initial state after some period of time, (say we were testing a drug for acne) then we could run a crossover study to determine the joint distribution. In a crossover study, the participants are randomly assigned to treatment and placebo, results are measured and then the groups are swapped. The scientific and philosophical validity of counterfactual queries remains under question [4, 5], however they are nonetheless widely posed in the form of attribution of causal effects to different pathways and mediation [24, 12, 39].

There are differences between the models we have considered when it comes to counterfactual

queries. Counterfactuals are not defined in causal Bayesian networks, as they only encode information on the interventional distribution over variables. Counterfactuals can be defined in terms of structural equation models [23] but there are subtle differences depending on the form of assumptions made. Structural equation models with independent errors allows the identification of quantities in mediation studies, which are not identifiable with the weak ignorability assumptions and cannot be tested experimentally [27].

In practice, differences in focus and approach across different fields eclipse these actual differences in the models. The work on causal graphical models [23, 38] focuses on non-parametric estimation in the population limit and rigorous theoretical foundations. The Neyman-Rubin framework builds on our understanding of randomized experiment and generalizes to quasi-exeperimental and observational settings, with a particular focus on non-random assignment to treatment. This research emphasises estimating 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 [29]. In economics, inferring causal effects from non-experimental data so as 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 [9]. In addition, the parametric structural equation models favoured in economics can be extended to analyse cyclic (otherwise referred to as non-recursive) models.

The equivalence between models for interventional queries allows researchers to combine the best aspects of all of these models. You can define your assumptions and determine which variables must be controlled for using a graphical model and apply propensity scores to make the adjustment. [19] provides an excellent and pragmatic introduction to this combined approach. 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.

3 The Do Calculus

If you state your assumptions in the form of an acyclic structural equation model or a causal bayesian network, the do calculus provides the methodology to determine if a given causal effect can be estimated from observational data and, if it can, provides to formula to do so. The calculus relies heavily on the understanding how conditional independence properties are encoded in bayesian networks. Therefore, in the next section we give a short review of independence in bayesian networks.

3.1 Independence in bayesian networks: D-separation

Recall from section 2.2 the local markov condition:

Theorem 3.1 ***(Local markov condition)*** *Given a bayesian network G with nodes $X_1...X_N$, each variable X_i is independent of its non-decendents 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 independences that also hold in all distributions that are compatible with G. D-separation extends the local markov property to find these additional independences and allows us to read from a network if a given conditional independence statement is true in all distributions compatible with that network.

Theorem 3.2 ***(D-separation)*** *If a set of variables* $\boldsymbol{Z}$ *d-separates* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *in* 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 given* $\boldsymbol{Z}$ *then there will be at least one distribution* P' *that is compatible with* G *where they are dependent.*

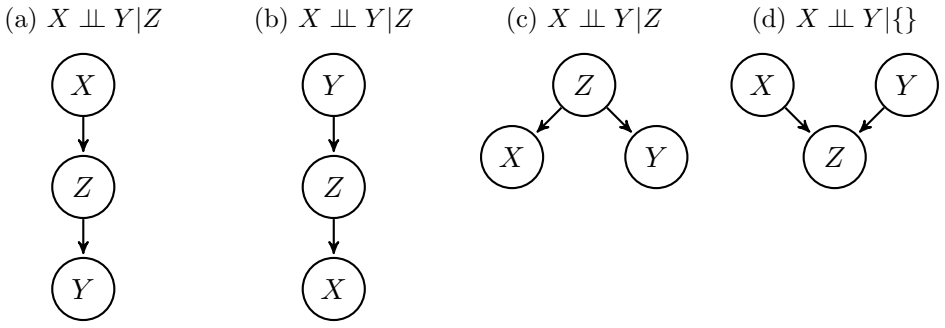
X conditionally independent of Y given Z means, if we know Z, learning the value of Y gives us no additional information about X. You can also think of this as Z blocks the flow of information from X to Y in the network. Let's examine the four possible bayesian networks over the variables X, Y, Z that do not have a direct link $X \to Y$ (figure 2).

In figure (2a) conditioning on Z renders X and Y conditionally independent. For example, X could be number of years a person smoked, Z the amount of tar in their lungs and Y lung cancer. If you don't know Z, knowing how much they smoked changes your estimate of the likelyhood of their having lung cancer, but, if smoking causes cancer purely through the build up of tar in the lungs as we have assumed in this model, then once you measure the amount of tar in the lungs knowing how many years a person smoked for gives you no additional information on the probability that they will develop lung cancer. Figure (2b) is the same as (2a) with X and Y exchanged, which does not change anything since conditional independence is symmetric $(X \perp\!\!\!\perp Y|Z) \Leftrightarrow (Y \perp\!\!\!\perp X|Z)$.

In figure (2c), Z could represent having an illness and X and Y could be two independent tests that can detect (with some error) that illness. Without conditioning on Z you would expect results from the two different tests to be highly correlated, but if you already know someone has the illness then learning they tested positive to the test X doesn't change the likelyhood they test positive in test Y. Again, conditioning on Z blocks the path from X to Y.

The arrangement of variables in figure (2d), often referred to as a collider or v-structure, is the odd one out. In this case X is independent of Y without conditioning on Z but if you do condition on Z then X and Y actually become dependent. Suppose X is socio-economic status, Y is gender and Z is receipt of a scholarship available to female or disadvantaged students. Roughly equal numbers of women and men grow up in poorer families so X is independent of Y, but if you know that a given student received a scholarship, then learning that they are male tells you they must have a low socio-economic background. Conditioning on Z unblocks the path from X to Y.

Figure 2: All possible two edge paths from X to Y via Z



For a general bayesian network a path between two variables may involve more than three variables.

Definition 3.1 ***(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 2)*

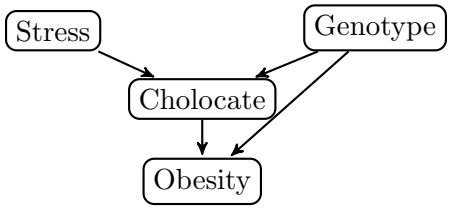
If all paths between two nodes are blocked by conditioning on $\boldsymbol{Z}$ then no information can flow

between them and they are d-separated given $\boldsymbol{Z}$.

Definition 3.2 *The variables $\boldsymbol{X}$ are d-separated from $\boldsymbol{Y}$ given $\boldsymbol{Z}$ in the network G if, after conditioning on $\boldsymbol{Z}$ there are no unblocked paths from $\boldsymbol{X}$ to $\boldsymbol{Y}$.*

Let us conclude with an example. Assuming the model in figure 3: Is obesity independent of stress given chocolate consumption?

Figure 3: D-separation example



There are two paths from stress to obesity in figure 3, $Stress \rightarrow Chocolate \rightarrow Obesity$ and $Stress \rightarrow Chocolate \leftarrow Genotype \rightarrow Obesity$. In the first, the variable chocolate is an intermediary as in figure 2a so this path is blocked when we condition on chocolate. However, in the second path chocolate is a collider as in figure 2d so this path is unblocked when we condition on chocolate. Since there is an unblocked path from stress to obesity after conditioning on chocolate, obesity is not independent of stress given chocolate consumption. This example shows the role of a variable (if it is a collider or not) is relative to the path you are considering. Conditioning on a variable can block some paths and unblock others.

For a more complete introduction to d-separation (with proofs) see [16].

3.2 The three rules

The do calculus consists of three rules. They derive from the causal information encoded in a causal network and the properties of d-separation and do not imply any addition assumptions other than that of specifying the causal network. The key property of these three rules is that they are complete. If an interventional (do-type) query can be transformed to a do-less statement via repeated application of these rules we can get an unbiased point estimate of the causal effect from observational data. Conversely, if the query is not identified via these rules we cannot get an unbiased estimate of the causal effect without additional assumptions, even with an infinite amount of observational data.

3.2.1 Rule 1

This rule describes which variables, on which we have not intervened, effect the distribution of the outcome given some intervention. The intervention $do(\boldsymbol{X} = \boldsymbol{x})$ changes a causal network, G, in a simple way. Variables in $\boldsymbol{X}$ are no longer determined by their parents but instead take on fixed values specified by $\boldsymbol{x}$. This corresponds to deleting the edges with arrows into variables in $\boldsymbol{X}$ (see figure 4). The resulting 'mutilated' network $G_{\overline{\boldsymbol{X}}}$ remains a causal network and d-separation still applies.

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

$$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{W} = \boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{Z} = \boldsymbol{z}) \tag{11}$$

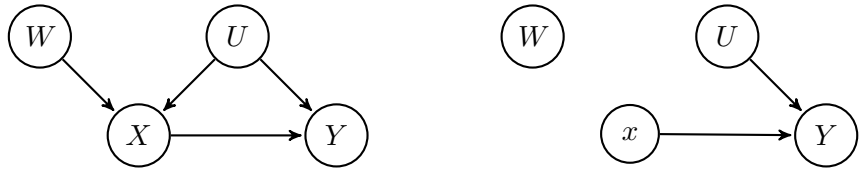
$\implies$ if $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}|\boldsymbol{X})$ in $G_{\overline{\boldsymbol{X}}}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{W} = \boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})) \tag{12}$$

This stems directly from the fact that the relationship between d-separation in a network and independence in the corresponding probability distribution still holds in the mutilated network.

Figure 4: Intervention in a causal bayesian network

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



3.2.2 Rule 2

Rule 2 states when conditioning on $\boldsymbol{X} = \boldsymbol{x}$ and intervening $do(\boldsymbol{X} = \boldsymbol{x})$ have the same effect on the distribution of the outcome $\boldsymbol{Y}$. You can think of this as when correlation is causation. It is easiest to understand by explicitly including the intervention process in the graphical model. We can depict the possibility of intervention by adding a new decision node $\hat{X}$ as a parent of each X in the set of nodes we are intervening on (figure 5). Let ϵ be some arbitrary value not in the set of possible values for X. If $X = \epsilon$ the distribution of X is what it was without intervention. Otherwise, if $X = x$, X deterministically takes the value x and is independent of its previous parents, representing the intervention $do(X = x)$. We use the notation $G^{\dagger}$ to represent G augmented with these decision nodes.

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

$$P(\boldsymbol{Y}|do(\boldsymbol{Z} = \boldsymbol{z}), do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{W} = \boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{Z} = \boldsymbol{z}), \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W} = \boldsymbol{w}) \tag{13}$$

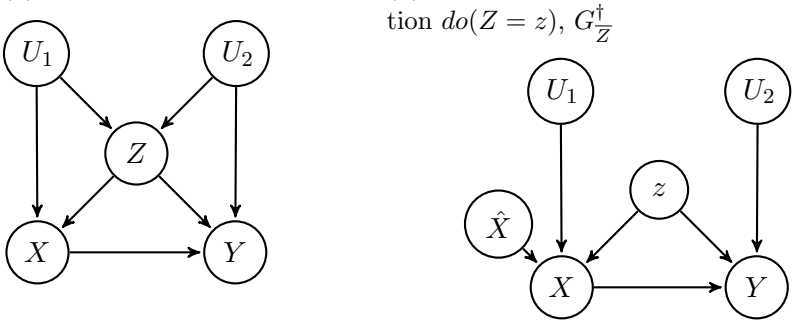
$\implies$ if $(\boldsymbol{Y} \perp\!\!\!\perp \hat{\boldsymbol{X}}|\boldsymbol{X})$ in $G^{\dagger}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})) = P(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}) \tag{14}$$

If the outcome does not depend on how the decision to assign the interventional variables was made, then the interventional distribution equals the observational one.

Figure 5: Augmentated causal network with intervention

(a) original network, G (b) augmented network after the intervention $do(Z = z)$, $G^{\dagger}_{\overline{Z}}$



3.2.3 Rule 3

This rule describes cases where the intervention $do(\boldsymbol{X} = \boldsymbol{x})$ has no effect on the distribution of the outcome $\boldsymbol{Y}$.

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

$$P(\boldsymbol{Y}|do(\boldsymbol{Z} = \boldsymbol{z}), do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{W} = \boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{Z} = \boldsymbol{z}), \boldsymbol{W} = \boldsymbol{w}) \tag{15}$$

$\implies$ if $(\boldsymbol{Y} \perp\!\!\!\perp \hat{\boldsymbol{X}})$ in $G^{\dagger}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})) = P(\boldsymbol{Y}) \tag{16}$$

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 $\boldsymbol{Y}$ which starts on an arrow leaving $\boldsymbol{X}$. This means there is no causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ and the intervention $do(\boldsymbol{X} = \boldsymbol{x})$ does not alter the distribution of $P(\boldsymbol{Y})$.

3.3 Easy graphical tests for identifiability

Determining if a network is identifiable via the two calculus still requires some algebra and it may not always be obvious what sequence of rules are required for a given network. From the do-calculus we can derive some quick graphical tests that are sufficient (but not necessary) for non-parametric identifiability, and align with intuitive ideas about estimating causal effect by adjusting for confounding variables.

3.3.1 If all variables are observed

If all the variables in a causal network can be measured any interventional query on that network are identifiable. You just have to condition on the parents of the variables on which you intervene. For discrete variables, the causal effect is given by:

$$P(\boldsymbol{Y}|do(\boldsymbol{X})) = \sum_{\boldsymbol{Pa_X}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Pa_X}) P(\boldsymbol{Pa_X}) \tag{17}$$

where $\boldsymbol{Pa_X}$ are the parents of $\boldsymbol{X}$.

Conditioning on the parents blocks all information that does not flow causally (along the arrows) from the variables on which we intervened which eliminates spurious correlations and gives the causal effect.

3.3.2 The Back Door criterion

The back door criterion is simplified version of rule 2. It generalizes from the case where all variables can be measured by observing we don't need to condition on all the parents of the intervened on variables: it is sufficient to condition on a set of variables that block all spurious paths from $\boldsymbol{X}$ to $\boldsymbol{Y}$. A spurious, or back door, path is one that leaves $\boldsymbol{X}$ against the direction of the causal arrows. If we have a set of variables $\boldsymbol{Z}$ that block all back door paths from $\boldsymbol{X}$ to $\boldsymbol{Y}$, where no variable in $\boldsymbol{Z}$ is a descendent of $\boldsymbol{X}$ then:

$$P(\boldsymbol{Y}|do(\boldsymbol{X})) = \sum_{Z} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) P(\boldsymbol{Z}) \tag{18}$$

The requirement that $\boldsymbol{Z}$ should not include descendants of $\boldsymbol{X}$ stems from the fact that we don't want to open any new spurious paths. If Z descends from X then it may lie on a path of the form $X \rightarrow ... \rightarrow Z \leftarrow ...Y$, which conditioning on Z would open. The back door criterion corresponds exactly to the strong ignoreability assumption [29], used to justify the validity of a causal estimate in the counterfactual approach . If $\boldsymbol{Z}$ satisfies the backdoor criterion then, for a single binary treatment variable X and outcome variable Y, $\{Y^0, Y^1\} \perp\!\!\!\perp X|\boldsymbol{Z}$.

3.3.3 The Front Door Criterion

It we have an unobservable variable creating a spurious path from $\boldsymbol{X}$ to $\boldsymbol{Y}$ then we cannot apply the back door criterion to directly estimate $P(\boldsymbol{Y}|do(\boldsymbol{X}))$. However, if there is a set of mediating variables $\boldsymbol{M}$ that intercepts all all the direct (non-spurious) paths from $\boldsymbol{X}$ to $\boldsymbol{Y}$ and we can estimate the causal effect of $\boldsymbol{X}$ on $\boldsymbol{M}$ and the causal effect of $\boldsymbol{M}$ on $\boldsymbol{Y}$ then we can combine these terms to get the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$.

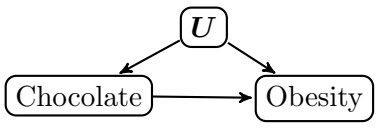
$$\begin{aligned} P(\boldsymbol{Y}|do(\boldsymbol{X})) &= \sum_{\boldsymbol{M}} P(\boldsymbol{M}|do(\boldsymbol{X}))P(\boldsymbol{Y}|do(\boldsymbol{M})) \\ &= \sum_{\boldsymbol{M}} P(\boldsymbol{M}|\boldsymbol{X}) \sum_{\boldsymbol{X}} P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{M})P(\boldsymbol{X}) \end{aligned} \tag{19}$$

3.3.4 Examples

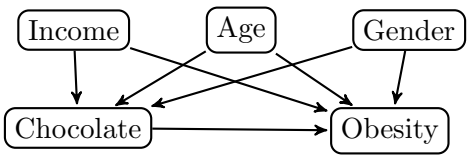
We are now in a position to consider some more interesting examples than the one in section 2. Consider the question; does eating chocolate increase or decrease the risk of obesity? Assume we have data from a cross-sectional study where a large group of people were weighed and asked how much chocolate they ate. It is easy to come up with a whole lot of variables that may causally effect chocolate consumption and obesity, ie. age, income and gender, so we cannot conclude that the causal effect is equivalent to the correlation. If we assume there may be unknown variables (denoted by U in all the following examples) that effect both obesity and chocolate consumption, (figure 6a), the causal effect cannot be identified as there is no way to block the spurious path through the unobserved variables. If we assume the only confounding variables are income, age and gender, (figure 6b), we can calculate the causal effect by conditioning on these variables.

Figure 6: Chocolate and Obesity

(a) Not identifiable



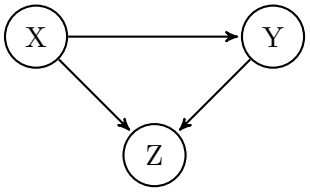
(b) Identifiable



In reality of course there are many other variables involved [6] and it may not be obvious from prior knowledge or theory what role the variables play in the network. This is important as conditioning on the wrong variables can increase the bias in a causal estimate: Conditioning on a collider variable, $\rightarrow Z \leftarrow$, may open a previously blocked, spurious path [37] (figure 7) . Conditioning on a variable that mediates between the cause and effect biases the causal estimate by blocking a true casual path (figure 8). Conditioning on a variable that is a true confounder, that is a common cause of both X and Y, may increase the bias in our causal estimate if it moves the causal estimate in the opposite direction than would correcting for a hidden (and stronger)

Figure 7: Conditioning on a collider variable can introduce bias

(a) post-treatment bias: conditioning on Z opens the spurious path X, Z, Y



(b) M-bias: conditioning on Z opens the path X, U_1, Z, U_2, Y

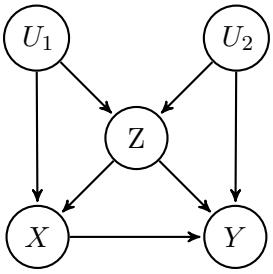
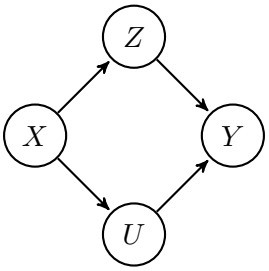


Figure 8: Conditioning on a mediating variable (Z) can introduce bias



confounder U. Finally, (although it is not obvious graphically) conditioning on a variable that is a cause of X but not (or only weakly) a cause of Y can amplify or introduce bias [21, 42].

For some problems, the temporal ordering of events imposes restrictions on the causal structure. Consider the example of a non-randomized trial. A new treatment is developed and made available to a large group of patients some of whom decide to take. Researchers have variables on the health, income, gender, age, etc of the patients before the treatment was offered (pre-treatment variables) and collect further data on the health of all patients after treatment has begun (post-treatment variables). It has been understood for a long time that conditioning on post-treatment variables can increase bias [28]. The realization that that conditioning on pre-treatement can also introduce bias is more recent. Not conditioning on post-treatment variables avoids the biases in figure 7a and figure 8. However, M-bias (figure 7b) and bias amplification (figure 10) are consequences of conditioning on pre-treatment variables.

It is clear we need to be wary of the condition on everything approach and the condition on all pre-treatment variables approach. But what is the alternative if the structure of the causal network is not clear from prior knowledge? One approach would be to construct a number of different plausible structures and evaluate the desired causal effect for each of them. This would give more accurate uncertainty bounds for the causal effect. Alternatively, one could apply a causal discovery algorithm (see section 5).

It may be that there are some general conditions under which certain types of biases are less

Figure 9: Conditioning on a confounder (Z) can increase bias

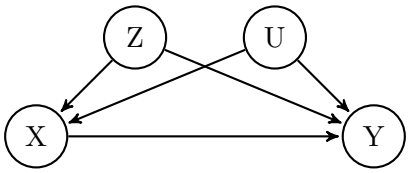
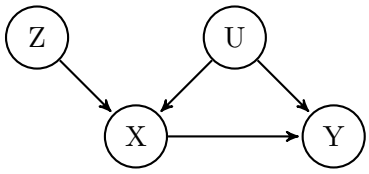
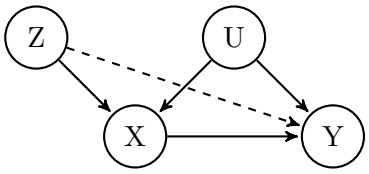


Figure 10: Bias amplification

(a) Conditioning on an instrumental variable Z can increase bias in estimate of effect of X on Y



(b) Conditioning on a confounding variable that is strongly linked to X and weakly linked to Y can increase bias.



likely or weaker than other. For example, in the situation in figure 7b, where Z is both a confounding variable and part of an M-structure. Conditioning on Z blocks the spurious path $X \leftarrow Z \rightarrow Y$ and opens the spurious path $X \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y$. Should we condition on Z or not? Under what general assumptions is one of these paths weaker than the other. There has been relatively little work in this area. [**?**] gives a description of the problem and links it to selecting covariates when the goal is causal estimation. [20, 21] consider if one should adjust for a variable that could be either an instrumental variable or a confounder for the effect of X on Y.

4 It's not (non-parametrically) identifiable! What can I do?

4.1 Bounding causal effects

Causal effects that cannot be identified can often still be bounded. In other words, bounds require weaker assumptions than point identification.

4.2 Additional assumptions

additive noise causal anti-causal [14] learning what indicates causality.

4.3 Instrumental variables

5 Causal structure learning (causal discovery)

In previous sections we discussed when assumptions about the structure of the variables in a specific problem is sufficient to identify a causal effect. This approach relies on having enough prior knowledge or theory about the problem to allow you to, at least partially, specify the causal network. In this section, we consider the much harder problem of causal inference where you need to learn the network. Causal inference might seem impossible without specific assumptions about the structure of the variables involved but, amazingly, some aspects of causal structure can be determined from much more general assumptions.

In this section we consider variants of the following problem: Assume there is some acyclic causal network G that generated the distribution $P(\boldsymbol{V})$ from which our data has been sampled. Our goal is to recover the network from this data.

What can we look for in the distribution that could give us clues as to the structure of the network?

5.1 Causal discovery with conditional independence tests

One general approach is to look for clues about the structure of the network in the conditional independence relations in the distribution. We know, (section 3.1), if Z d-separates X and Y in G then $(X \perp\!\!\!\perp Y|Z)$ in 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*** 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.

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.

5.1.1 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.

We will begin with a brute force algorithm (described as the SGS algorithm in [38] and IC algorithm in [23]). 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 equivelence class of G

1. Join all pairs of vertices $(a, b) \in \boldsymbol{V}$ with an undirected 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 undirected 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 utilizes the fact that a collider structure (f igure 2d) 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 [38] modifies step 2 of the SGS algorithm to utilize 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 [38] (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 equivelence class of G

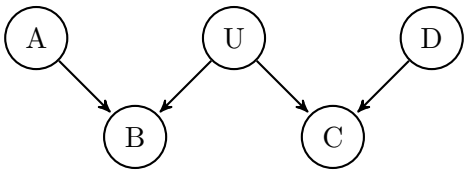
1. As for SGS
2. **for** each link $a - b$:
   $n = 0$
   $\boldsymbol{A}_{a,b} = \{A_1 ... A_j\}$ be the set of nodes adjacent to a and/or b
   while a and b are connected and $n < j$:
   if any subset of size n of $\boldsymbol{A}$ makes a and b conditionally independent:
   delete the link
   $n = n + 1$
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 with equation 17. We can then bound the true causal effect by combining the results for the all the networks. This procedure is the IDA algorithm [18] and has been found to outperform standard regularization techniques at finding causal effects in a high-dimensional yeast gene expression data set [17]. An implementation is available in the R package [15]

5.1.2 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 are

Figure 11: A distribution faithful to this DAG is not faithful to any DAG over the variables $\{A, B, C, D\}$ after marginalizing over U.



an infinity many possible structures to search over. In addition, the space of causal networks is not closed under marginalization. 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 marginalizing over $\boldsymbol{U}$, may not be faithful to any bayesian network (see figure 11).

The key to constraining the space of possible models is that many latent structures are equivalent (under transforms of the hidden variables). See example figure XXX.

Theorem 5.1 *[40] 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 marginalization and conditioning.

For any DAG with latent (and selection) variables there is a unique MAG [26]. This makes it possible to extend the PC algorithm to latent structures, resulting in the FCI algorithm [38]. The logic is very similar. Certain structures are ruled out by 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. Assuming there are no selection variables (see [**?**]xx), the PAG can contain four types of link:

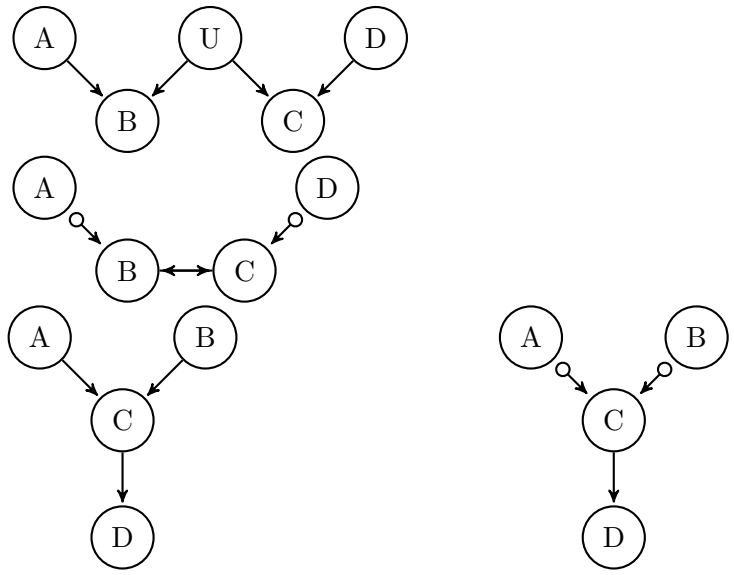
1. $X \rightarrow Y$, meaning X causes Y
2. $X \leftrightarrow Y$, meaning there is a latent variable that causes X and Y.
3. $X \circ\!\!\rightarrow Y$, either X causes Y or a latent variable causes both.
4. $X \circ\!\!-\!\!\circ Y$, either X causes Y or Y causes X or a latent variable causes both.

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 12).

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 [44]. More recently [3] have proposed the RFCI algorithm, which in some cases returns more ambiguous links than FCI but is substantially faster. [2] 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.

With latent variables we are not using all the information - so we could go further (to nested markov models and inequalities.) [34] [35]

Figure 12: A distribution faithful to this DAG is not faithful to any DAG over the variables $\{A, B, C, D\}$ after marginalizing over U.



These are all constraint based methods ... efficient because they stop early, but also may not be robust to errors early on.

A comparison of algorithms

Alg.	**Method**	**Scales (num.vars)**	**$\sim$ Vars**	**Latent**	**Reference**
IC/SGS	Constraint based	Exponential	10	No	Pearl(2000)/Sprites(2000)
PC	Constraint based	Worst case exponential, polynomial for sparse graphs	5000	No	Sprites(2000)
FCI	Constraint based	Worst case exponential, polynomial variant FCI+ for sparse graphs	30	Yes	Sprites(2000)
RFCI	Constraint based	?	500	Yes	Colombo(2012)
GES	Search & Score	Worst case exponential	50	No	Chickering(2002)
MMHC	Hybrid	?	5000	No	Tsamardinos(2006)

5.1.3 Doing conditional independence tests

The off-diagonal elements of the standardized inverse of the correlation matrix are the negatives of the partial correlation coefficients between the corresponding variables given the remaining variables (see e.g. Whittaker, 1990). Hence in the linear case, the independence graph can be efficiently constructed by placing an edge between A and B if and only if the entry in the standardized inverse correlation matrix is non-zero. In the discrete case, Fung and Crawford (1990) have recently proposed a fast algorithm for constructing an independence graph from discrete data. We have not tested their procedure as a preprocessor for the PC algorithm. (COPIED FROM SPRITES)

[46] Kernel independence tests

HSIC [7]

5.2 Discovery with functional models

All of 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 we have only two variables and the only 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 2.1). There are a number of assumptions about the form of the functions that can allow us to identify the causal direction: non-invertible functions, additive noise **??**, post-non-linear additive noise

Linear models with non-gaussian noise [10]

Allow us to make assumptions about the form of the relationships between variables. With just two variables:

Additive noise, [11]

post-non-linear additive noise [45],

causal anti-causal,

deterministic functions,

IGCI [13]

learning what causality looks like.

If we have a method to solve this case we can solve the more general case (with no hidden variables) Putting things together to learn the full network.

5.3 Scoring

6 Causality and prediction

Refresh the difference between a causal query and predictive query. Point out the importance of accurate prediction for causal inference. For example, if we can predict the control response, we can run experimental studies without controls ([1])

7 Concluding remarks

Section 3 tells empirical researchers how to formalize what they already do when forced to work with observational data. The results should not be taken to mean that experimentation should be abandoned. Similarly the fact that causal discovery from general assumptions is possible (section 5) does not mean we should throw away careful analysis of a specific problem in favour of simply throwing all the variables into a causal discovery algorithm. A human may likely still be the state of the art for constructing a qualitative network (or better set of plausible

networks) in fields like economics, ecology and social science where the variables are subject to human intuition and there exists some some theory to constrain models. However, we should not assume this will always be the case. As causal discovery methods are improved and the ability of AI to automatically incorporate large bodies of theory or prior knowledge into models develops computers should become better at science than us. Even if you don't believe in such dreams, there are already problems with large number of variables where we don't have a lot of theoretical understanding of the relationships between them. For these problems, even very naive causal discovery algorithms can improve on the alternative of standard feature selection.

References

[1] KH Brodersen, Fabian Gallusser, and Jim Koehler. Inferring Causal Impact Using Bayesian Structural Time-Series Models. (June):1–32, 2013.

[2] Tom Claassen, Joris Mooij, and Tom Heskes. Learning sparse causal models is not NP-hard. In *UAI*, 2013.

[3] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Stat.*, 40(1):294–321, February 2012.

[4] AP Dawid. Causal inference without counterfactuals. *J. Am. Stat. Assoc.*, 2000.

[5] AP Dawid. Statistical Causality from a Decision-Theoretic Perspective. *arXiv Prepr. arXiv1405.2292*, 2014.

[6] BA Golomb. Association between more frequent chocolate consumption and lower body mass index. *Arch. Intern. ...*, 172(6):2012–2014, 2012.

[7] Arthur Gretton, K Fukumizu, CH Teo, and L Song. A kernel statistical test of independence. pages 1–8, 2008.

[8] T Haavelmo. The statistical implications of a system of simultaneous equations. *Econom. J. Econom. Soc.*, 11(1):1–12, 1943.

[9] James Heckman. Econometric causality. *Int. Stat. Rev.*, 2008.

[10] Patrick Hoyer, A Hyvarinen, and Richard Scheines. Causal discovery of linear acyclic models with arbitrary distributions. *pre-print*, 2012.

[11] Patrick Hoyer, Dominik Janzing, and Joris Mooij. Nonlinear causal discovery with additive noise models. In *NIPS*, 2009.

[12] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Stat. Sci.*, 25(1):51–71, February 2010.

[13] Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31, May 2012.

[14] Dominik Janzing and Jonas Peters. On causal and anticausal learning. In *ICML*, 2012.

[15] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *JSS*, VV(Ii), 2012.

[16] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

[17] Marloes H. Maathuis, D Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nat. Methods*, 7(4):247–248, 2010.

[18] Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *Ann. Stat.*, 37(6A):3133–3164, December 2009.

[19] Stephen Morgan and Christopher Winship. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press, 2007.

[20] Jessica a Myers, Jeremy a Rassen, Joshua J Gagne, Krista F Huybrechts, Sebastian Schneeweiss, Kenneth J Rothman, Marshall M Joffe, and Robert J Glynn. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am. J. Epidemiol.*, 174(11):1213–22, December 2011.

[21] J Pearl. On a class of bias-amplifying variables that endanger effect estimates. *arXiv Prepr. arXiv1203.3503*, (July):417–424, 2012.

[22] Judea Pearl. Causal Diagrams for Empirical Research. *Biometrika*, 82(4):669, December 1995.

[23] Judea Pearl. *Causality: models, reasoning and inference*. MIT Press, Cambridge, 2000.

[24] Judea Pearl. Interpretation and Identification of Causal Mediation. *Psychol. Methods*, June 2014.

[25] Jonas Peters, Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *JMLR*, 15:2009–2053, 2014.

[26] T Richardson and P Spirtes. Ancestral graph Markov models. *Ann. Stat.*, 30(4):962–1030, 2002.

[27] Thomas S. Richardson and James M. Robins. Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. 2013.

[28] PR Rosenbaum. The consquences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Stat. Soc. Ser. A (...*, 1984.

[29] PR Rosenbaum and DB Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[30] DB Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, 1974.

[31] DB Rubin. Bayesian inference for causal effects: The role of randomization. *Ann. Stat.*, 1978.

[32] DB Rubin. Causal Inference Using Potential Outcomes. *J. Am. Stat. Assoc.*, 100(469):322–331, March 2005.

[33] DB Rubin. For objective causal inference, design trumps analysis. *Ann. Appl. Stat.*, 2(3):808–840, September 2008.

[34] I Shpitser and TS Richardson. Parameter and structure learning in nested Markov models. *pre-print*, 2012.

[35] Ilya Shpitser, Robin J. Evans, Thomas S. Richardson, and James M. Robins. Introduction To Nested Markov Models. *Behaviormetrika*, 41(1):3–39, 2014.

[36] Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. *arXiv Prepr. arXiv1206.6876*, 2012.

[37] Arvid Sjölander. Propensity scores and M-structures. *Stat. Med.*, 28(9):1416–20; author reply 1420–3, April 2009.

[38] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.

[39] Tyler J VanderWeele and Sonia Hernández-Diaz. Is there a direct effect of pre-eclampsia on cerebral palsy not through preterm birth? *Paediatr. Perinat. Epidemiol.*, 25(2):111–5, March 2011.

[40] TS Verma. Graphical aspects of causal models. Technical report, 1993.

[41] Deena S Weisberg and Alison Gopnik. Pretense, counterfactuals, and Bayesian causal models: why what is not real really matters. *Cogn. Sci.*, 37(7):1368–81, 2013.

[42] J Wooldridge. Should instrumental variables be used as matching variables. Technical Report September 2006, Michigan State University, MI, 2009.

[43] S Wright. Correlation and causation. *J. Agric. Res.*, 1921.

[44] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, November 2008.

[45] Kun Zhang and A Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. *NIPS 2008 Work. Causality. URL http//www ...*, 2008.

[46] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *pre-print*, 2012.