On Causality

Simon Dirmeier

05 August, 2018

Contents

This text collects some notes on causality. I took most of the content from the papers, books and lecture notes below. Consequently, some passages might sound familiar to you. If so, these are probably inspired, or taken directly, from the following sources:

- Bottou et al. (2013)
- Hauser and Bühlmann (2012)
- Heinze-Deml, Maathuis, and Meinshausen (2018)
- Koller and Friedman (2009)
- Maathuis et al. (2009)
- Nicolai Meinshausen's lecture notes
- Cosma Shalizi's draft of Advanced Data Analysis from an Elementary Point of View
- Chickering (2002)
- Pearl (2009b)
- Pearl (2009a)
- Shimizu (2014)
- Spirtes et al. (2000)
- Bühlmann et al. (2016)
- Morgan and Winship (2015)
- Peters et al. (2014)

Notation: You find the used notation at the end of this document.

I do not take warranty for the correctness or completeness of this document. Furthermore, if there's typos, mistakes or errors, please let me know.

1 *Introduction*

Graphical models represent a joint probability distribution of a multivariate random vector X. Causal models are a special class that in addition specify the distribution under interventions, i.e. when we do *surgery* on the graph. That allows us to infer cause-effect relationships. But what is causal inference? Basically we can distinguish between:

- estimate the effect of one variable on another if the causal graph is known,
- given data estimate the causal graph.

Answering causal claims is inherently more difficult than establishing statistical associations, because we often cannot answer them from observational data, even at the population level. When we talk about causal effects, we usually mean the distribution $P(Y \mid \text{do}(x))$ of some variable of interest Y when we have intervened on another variable X, i.e. setting some variables artificially to $X = x$. We will see

later what that means in particular. **Note that this is different from** $P(Y \mid x)$. Thus, while we are often interested in the interventional distribution $\tilde{\mathbb{P}}$, only $\mathbb{P}$ is available.

The easiest way to compute $P(Y \mid \text{do}(x))$ would be to just make an experiment. For various reasons, be it ethical, financial or whatsoever, this is often not possible. However, with certain assumptions we are still able to infer causal relations, for instance using adjustment, e.g. with back-door and front-door criteria, do-calculus, instrumental variables, etc. If we observe all relevant variables, and assume no latent confounders, then causal inference is the same as standard statistical inference (if correctly adjusted for observed confounding variables).

Before we delve into the myths of *causal inference*, we define some concepts:

- **Causal effect**: $\frac{\partial}{\partial x} E(Y \mid do(X_i = x)) \mid_{x=x'}$ (this is sometimes the regression coefficient adjusted for confounders).
- **Confounding effect**: whenever we are observing $P(Y \mid \text{do}(x)) \neq P(Y \mid x)$.
- **Causal sufficiency**: refers to assuming to latent confounding variables.
- **Markov**: $\mathbb{P}^X$ is said to be *Markov w.r.t the DAG* $\mathcal{G}$ if A, B d-separated by $C \Rightarrow A \perp\!\!\!\perp B \mid C$ for all disjoint sets A, B, C.
- **Faithfulness**: $\mathbb{P}^X$ is said to be *faithful to the DAG* $\mathcal{G}$ if A, B d-separated by $C \Leftarrow A \perp\!\!\!\perp B \mid C$ for all disjoint sets A, B, C.
- **Causal minimality**: a distribution satisfies *causal minimality w.r.t the DAG* $\mathcal{G}$ if it is Markov w.r.t. $\mathcal{G}$, but not any proper subgraph of $\mathcal{G}$. So deleting an edge would result in a new conditional independence that does not hold in the distribution.
- **Markov blanket**: the Markov blanket of a node X is the set of parents, children and co-parents.

The goal of this document is explaining how to compute an interventional distribution $\tilde{\mathbb{P}}^X$ from the observational distribution $\mathbb{P}$.

2 *Observations vs. Experiments*

Observational and experimental data can be distinguished by their sources of origins. This example is taken from the review of Marloes Maathuis and Preetam Nandy in Bühlmann et al. (2016).

Suppose we are observing a group of prisoners participating *voluntarily* in a rehabilitation program and analyse the probability of rearrest, i.e. the prisoners are released and we track if they are rearrested or remain on free foot. The data under analysis would be called *observational*, because the subjects choose their own treatment.

Confounding factors, such as intrinsic motivation, might influence the analysis. On the other hand, when we *randomly assign* prisoners to the rehabilitation program or not, we deal with *experimental* data (and the experiment is randomized controlled). We see later what the random assignment entails on a causal graph.

Estimating causal effects from experimental data, for instance via randomized controlled trials, is relatively straightforward, while it is hard for observational data. In fact, it is generally impossible without causal assumptions.

3 *SEMs and path diagrams*

A structural equation model (SEM) is a tuple $\mathcal{S} := (S, \mathbb{P}^{\epsilon})$:

$$S_i : \; X_i = f_i(PA_i, \epsilon_i),$$

where $\mathbb{P}^{\epsilon} = \mathbb{P}^{\epsilon_1,\dots,\epsilon_p}$ is a distribution of noise variables which are *jointly independent*. The exogeneous variables ϵ are usually a collection of extrinsic factors. The value of these is outside the model and their data-generation processes are defined by the modeller. The variables on the right hand side are *causing* the effect on the left side, so the graph underlying the structural equation model is called a *causal graph* or *path diagrams*. Drawing them follows trivially from the assignment operator. In that sense edges represent direct causal effects. SEMs can be used not only for observational, but also for the definition of interventional distributions. Later we will see how to infer the causal structure of a SEM, which turns out to be quite difficult (to say the least). We will only be able to do that with further assumptions (again). The SEM is called *structural*, because of the generating mechanism for each X_i.

Alternatively we can define a SEM to be a quadruple of endogenous variables, exogenous variables, structural equations and probability distributions of the exogenous variables.

In case of path diagrams that are *acyclic directed graphs* (DAGs), we can always find an ordering $\pi(X_i)$ of the variables, called a *causal order*, such that

$$\pi(X_i) < \pi(X_j) \text{ if } X_j \in PA_i.$$

Example Let's have a look at a simple SEM with five structural equations.

$$V \leftarrow \epsilon_V \tag{1}$$
$$W \leftarrow 2V + \epsilon_W \tag{2}$$
$$X \leftarrow -W + \epsilon_X \tag{3}$$
$$Y \leftarrow \alpha W + \epsilon_Y \tag{4}$$
$$Z \leftarrow -2V + 3X + 5Y + \epsilon_Z \tag{5}$$
$$\tag{6}$$

And here's the respective R code. The result of a simulation of 100 draws can be seen in Figure 1.

```
V <- function(n = 100) rnorm(n)
W <- function(n = 100, v) 2 * v + rnorm(n)
X <- function(n = 100, w) -w + rnorm(n)
Y <- function(n = 100, alpha = 2, w) alpha * w +
    rnorm(n)
Z <- function(n = 100, v, x, y) -2 * v + 3 * x +
    5 * y + rnorm(n)

sem <- function(alpha = 2, n = 100, v = NULL,
    w = NULL, x = NULL, y = NULL, z = NULL) {
    v <- V(n)
    w <- W(n, v)
    x <- X(n, w)
    y <- Y(n, alpha, w)
    z <- Z(n, v, x, y)

    data.frame(v = v, w = w, x = x, y = y, z = z)
}

simu <- sem()
```

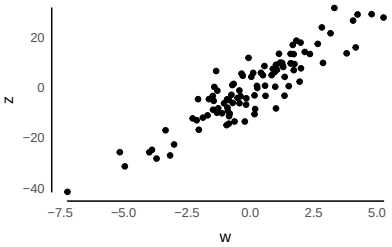

Figure 1: Simulation of two variables of a SEM.

Linear structural model A linear structural causal model over a multivariate random variable $X = (X_1, \ldots, X_p)$ with noise $\epsilon = (\epsilon_1, \ldots, \epsilon_p)$ is defined as:

$$X_j \leftarrow \sum_i^p \beta_{j,k} X_k + \epsilon_k.$$

In matrix notation that is:

$$X \leftarrow BX + \epsilon.$$

Before we finish this section, let us introduce *Reichenbach's common cause principle. When two variables are dependent, there must be a causal explanation, either:*

- X is (in-)directly causing Y or
- Y is (in-)directly causing X,
- there is a confounder Z that causes both.

3.1 *Properties*

Before we continue we interventions and counterfactuals, let's define some concepts about distributions, and causal graphs which will be helpful later. Conditional independencies are essential properties of distributions and understanding them is crucial for causal inference (and inference on graphical models in general).

In the simplest case, two variables X, Y are dependent, if they are connected with an edge in the causal graph. But how about indirect connections?

Markov property and d-separation Figure~3.1 shows four triplets of nodes. The three black graphs encode the same conditional indepence relationships, the red one encodes the *reverse*:

Triplet	$X \perp\!\!\!\perp Y \mid Z$	$X \perp\!\!\!\perp Y$
Causal trail: $X \to Z \to Y$	true	false
Evidential trail: $X \leftarrow Z \leftarrow Y$	true	false
Common causal: $X \leftarrow Z \to Y$	true	false
Common effect: $X \to Z \leftarrow Y$	false	true

When influence can flow from X to Y via Z, we say that the path is *active*, otherwise it is *blocked*. For longer paths $X_1 \to \cdots \to Y$, we require that at least one triplet blocks the path in order to block the influence from X to Y. Or more generally:

D-separation A set of nodes $\mathcal{C}$ *blocks* a path p from X to Y if either of the two conditions hold

- there is at least one node $c \in \mathcal{C}$ that emits an arrow,
- there is at least one node $c \notin \mathcal{C}$ that is a collider and has no descendants that in $\mathcal{C}$.

If all paths between X and Y are blocked by $\mathcal{C}$ we say that the set *d-separates* the two nodes and write $X \perp\!\!\!\perp Y \mid \mathcal{C}$.

This brings us to the Markov properties. For a DAG $\mathcal{G}$ and random vector X a joint probability distribution (JPD) $\mathbb{P}^X$ is said to satisfy:

- the **global Markov** property w.r.t. $\mathcal{G}$ if A, B d-separated by $C \Rightarrow A \perp\!\!\!\perp B \mid C$ for all disjoint sets A, B, C,

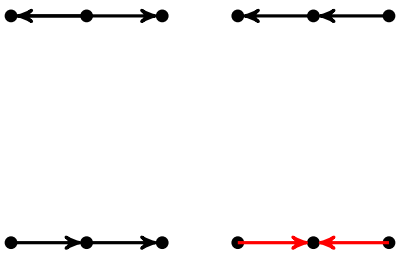

Figure 2: Four DAGs. The three black ones are Markov equivalent. The red one is a v-structure and encodes other conditional independencies.

- the **local Markov** property w.r.t $\mathcal{G}$ if each variable is independent of its non-descendants given its parents, and
- the **Markov factorization** property if $P(X) = \prod_j P(X_j \mid PA_j)$.

Denote with $\mathcal{M}(\mathcal{G})$ the set of distributions that are Markov w.r.t $\mathcal{G}$: $\mathcal{M}(\mathcal{G}) := \{\mathbb{P} : \mathbb{P} \text{ satisfies the global Markov property w.r.t } \mathcal{G}\}$. Two DAGS are *Markov equivalent* if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$. This is true iff both DAGS encode the same set of d-separations, i.e. the same set of CI conditions. The set of all DAGs in an equivalence class is represented by a completed partially directed acyclic graphs: a *CPDAG*. Two DAGs are Markov equivalent iff they have the same skeleton and the same v-structures. A *v-structure*, or *immorality*, is triplet of nodes with the following orientation of edges: $X \rightarrow Z \leftarrow Y$, i.e. the *common effect* example from above. The center node is called *collider*. We will later come back to the concept of CPDAGs when we try to learn the structure of the causal graph from data.

Faithfulness A distribution $\mathbb{P}^X$ is said to be *faithful to the DAG* $\mathcal{G}$ if A, B d-separated by $C \Leftarrow A \perp\!\!\!\perp B \mid C$ for all disjoint sets A, B, C. This is the reverse implication if the Markov property. It turns out that we can find SEMs that generate the same observational distributions $\mathbb{P}^X$. In biology this case is probably rather pathological, because exact cancellation of effects rarely occurs.

Remarks

- If $\mathbb{P}^X$ is faithful and Markov, causal minimality is satisfied. In most model classes, identifiability is impossible to obtain without causal minimality.
- SEMs and graphs with the Markov condition are equivalent. Counterfactual statements can be implied in both formulations.
- SEMs and their underlying path diagrams can be used for several things:
  - describe the observational distribution,
  - describe all interventional distributions (i.e. when a structural equation is modified),
  - do counterfactual statements. Note that in this case, we implicitely assume that the noise would not have changed, i.e. we put point mass on ϵ. For interventions, we would not need this assumption, but rather be ok, if the noise changed.

4 *Interventions*

Next we define the target we are *actually* interested in, the interventional distribution $\tilde{\mathbb{P}}^X$.

Interventional Distribution Replacing a structural equation in $\mathcal{S}$ results in a new SEM $\tilde{\mathcal{S}}$ for which we call its new generating distributions *interventional distribution*s. Variables with replaced structural equations have been *intervened on*. Denote the new distribution

$$\mathbb{P}^X_{\tilde{\mathcal{S}}} = \mathbb{P}^{X \mid \mathrm{do}(X_j = \tilde{f}(\widetilde{PA}_j, \tilde{\epsilon}_j))}_{\mathcal{S}}.$$

We intervene on a variable using the *do*-operator. Note that $\tilde{\mathcal{S}}$ now contains *old* error terms ϵ_j and new, intervened ones $\tilde{\epsilon}_j$. If $\tilde{f}(\widetilde{PA}_j, \tilde{\epsilon}_j)$ puts a point mass on some value a, we simply write $\mathbb{P}^{X \mid \mathrm{do}(X_j = a)}_{\mathcal{S}}$ and call it *perfect* intervention. An intervention with $\widetilde{PA}_j = PA_j$ is called *imperfect*. An intervention is *stochastic* if the marginal distribution of the intervened variable has a positive variance. We use several notations for the same thing:

$$\mathbb{P}^{Y|\mathrm{do}(x)}_{\mathcal{S}} = P_{\mathcal{S}}(Y \mid \mathrm{do}(x)) = P_{\mathcal{S},\mathrm{do}(x)}(Y).$$

Example Two variables X, Y with independent standard normal noise ϵ are described by:

$$X \leftarrow \epsilon_X \tag{7}$$
$$Y \leftarrow 2X + \epsilon_Y. \tag{8}$$

Then $\mathbb{P}^Y_{\mathcal{S}} = \mathcal{N}(0,5)$, while $\mathbb{P}^{Y|\mathrm{do}(X=2)}_{\mathcal{S}} = \mathcal{N}(4,1) = \mathbb{P}^{Y|X=2}_{\mathcal{S}}$. So, in this setting the intervention distribution is identical to the conditional distribution. However $\mathbb{P}^X_{\mathcal{S}} = \mathcal{N}(0,5)$. However, $\mathbb{P}^{X \mid \mathrm{do}(Y=2)}_{\mathcal{S}} = \mathbb{P}^X_{\mathcal{S}} \neq \mathbb{P}^{X \mid Y=2}_{\mathcal{S}}$. So, when we intervene on the structural equation S_Y, we break the dependence to Y's parents. If we intervene on X this dependence is not broken.

Example Other than using the *do*-operator (do-intervention), i.e. setting a variable to a value or giving it a new noise distribution, we can also intervene with a *shift*-intervention, which adds noise. For instance, in the SEM below

$$X \leftarrow \beta PA + \epsilon \tag{9}$$
$$X \leftarrow x_0 \text{ do-intervention} \tag{10}$$
$$X \leftarrow \tilde{\epsilon} \text{ do-intervention} \tag{11}$$
$$X \leftarrow \epsilon + Z \text{ shift intervention.} \tag{12}$$

4.1 *Truncated factorization*

Consider the SEM that evolves from $\mathcal{S}$ after $do(X_k = \tilde{\epsilon}_k)$:

$$P_{\mathcal{S},do(X_j=\epsilon_k)}(X) = \tilde{P}(X_j)\prod_{i\neq j} P_{\mathcal{S}}(X_i \mid pa_i).$$

A perfect do-intervention $do(X_j = x_0)$ replaces the (pre-intervention) conditional density $P(X_j \mid PA_j)$ with a scalar x_0 **that has point mass**. For Markovian models, the generated distributed is given by the *truncated factorization* formula (Pearl 2009b), also known as *g-formula* or *manipulation theorem*:

$$P_{\mathcal{S},do(X_j=x)}(X) = \begin{cases} \prod_{i,i\neq j}^{p} P(X_i \mid PA_i)\mid_{x_j=x'} & \text{if } x_j = x', \\ 0 & \text{else} \end{cases} \tag{13}$$

It follows that intervening with do and conditioning becomes equivalent for any variable X_1 that does not have parents:

$$P_{\mathcal{S}}(X_2, \ldots, X_p \mid X_1 = a) = P_{\mathcal{S},do(X_1=a)}(X_2, \ldots, X_p),$$

The factorization after intervening on the variables replaces **only** the conditional densities of the respective variables. That means a structural equation is *invariant* to possible changes in other structural equations:

$$P_{\mathcal{S}}(X_k \mid PA_k) = P_{\mathcal{S}}(X_k \mid PA_k),$$

for any SEM $\tilde{\mathcal{S}}$ that is constructed from $\mathcal{S}$ by replacing some S_j with $j \neq k$. We also call the structural equations to be *autonomous*. Assuming structural invariance (autonomy) is crucial for causal inference, as it allows derivation of the distribution of X under intervention.

Example An intervention on a variable X replaces the structural equation. For the causal graph that means, we cut down the causal effect of the parents of X. Consider Figure~3 for this. One the left side, we have the un-intervened SEM $\mathcal{S}$. On the right the modified SEM $\tilde{\mathcal{S}}$.

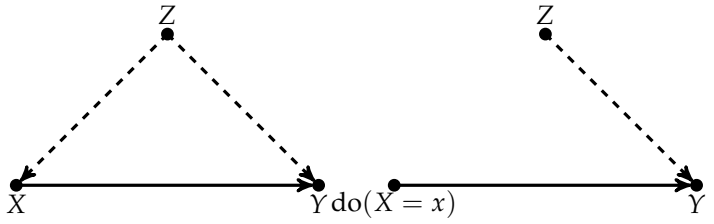

Figure 3: Original SEM and SEM with replaced structural equation after do$(X = x)$

Example Let's have a look at a simple SEM with five structural equations from the last section (which I write down again for convencience). This time we intervene with $\text{do}(Y = 2)$.

$$V \leftarrow \epsilon_V \tag{14}$$
$$W \leftarrow 2 * V + \epsilon_W \tag{15}$$
$$X \leftarrow -W + \epsilon_X \tag{16}$$
$$Y \leftarrow 2 \tag{17}$$
$$Z \leftarrow -2V + 3X + 5Y + \epsilon_Z \tag{18}$$
$$\tag{19}$$

And here's the respective R code again. The result of a simulation of 100 draws can be seen in Figure 4.

```
V <- function(n = 100) rnorm(n)
W <- function(n = 100, v) 2 * v + rnorm(n)
X <- function(n = 100, w) -w + rnorm(n)
Y <- function(n = 100, alpha = 2, w) 2
Z <- function(n = 100, v, x, y) -2 * v + 3 * x +
    5 * y + rnorm(n)

sem <- function(alpha = 2, n = 100, v = NULL,
    w = NULL, x = NULL, y = NULL, z = NULL) {
    v <- V(n)
    w <- W(n, v)
    x <- X(n, w)
    y <- Y(n, alpha, w)
    z <- Z(n, v, x, y)

    data.frame(v = v, w = w, x = x, y = y, z = z)
}

simu <- sem()
```

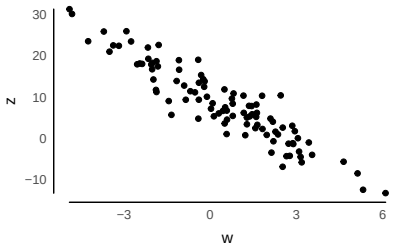

Figure 4: Simulation of two variables of a SEM wiht intervention $do(Y = 2)$.

4.2 *Conditional vs interventional distributions*

It should be clear by now that conditional (observational) and interventional distributions are *usually* two different things. While a conditional distribution $P(Y \mid x)$ *subsets* a joint distribution of a random vector (X, Y), an interventional distribution $P(Y \mid \text{do}(x))$ *creates a new one*.

Example Figure 4.2 shows an example where $P(Y \mid x) \neq P(Y \mid \text{do}(x))$, due to the confounding effect of U_X. Setting $X = x$ cuts off the influence of the parent U_X.

On the other hand, Figure 4.2 gives an example where the conditional distribution $P(Y \mid x)$ coincides with the inverventional distribution $(Y \mid \text{do}(x))$. Sometimes, we thoughtlessly assume orthogonality $\text{Cov}(X, U_Y) = 0$, such that we can, for instance, model $P(Y \mid do(x))$ as $P(Y \mid x)$. To cite Pearl (Pearl 2009a): *[. . .] the celebrated orthogonality condition in linear models* $\text{Cov}(X, U_Y) = 0$, *[. . .], [. . .] has been used routinely, often thoughtlessly, to justify the estimation of structural coefficients by regression techniques.*

Figure 5: A simple structure equation model for which $P(Y \mid x) \neq P(Y \mid \text{do}(x))$.

Figure 6: A simple structure equation model for which $P(Y \mid x) = P(Y \mid \text{do}(x))$.

Surgery What is *surgery*, or *intervention*, in particular? This process is shown in the path diagram in Figure 6.1, where we removed the effect of X_i on X_j by intervening on X_j.

Basically, we are doing the following steps:

1) eliminate edges that go into X,
2) fix the value $X = x$,
3) compute the **new** distribution that arises. We do **not** condition, but really create an entirely new subpopulation.

4.3 Remarks

- A SEM is a correct model for a latent data generating process if the observational distribution is correct and all interventional distributions $\mathbb{P}_{\mathcal{S}}^{X \mid X_j = \tilde{\epsilon}_j}$ correspond to distributions that we obtain from randomized experiments. Therefore an SEM is falsifiable.
- In RCTs a treatment T is randomly assigned to patient P according to $\tilde{\epsilon}_T$. In the SEM this is modelled as $\mathbb{P}_{\mathcal{S}}^{X \mid \text{do}(T = \tilde{\epsilon}_T)}$. If we still find a dependency between T and recovery, we find T to have a causal effect.
- Structural equation models are algebraic objects. As long as the causal graph remains acyclic, algebraic manipulations are interpreted as interventions giving rise to an interventional distribution $\tilde{\mathbb{P}}$. Bayesian networks represent joint probability distributions, and, as such, do not support interventions.

5 Counterfactuals

I am so tired today. I should've gone to bed earlier yesterday. Amiright? Would we be less tired if we went earlier? Observing both of these scenarios is apparently impossible, answering them isn't. The first statement I've made here is a *factual* statement (it aligns with a fact), while the second one contradicts the fact, thus it's a *counterfactual*.

A counterfactual SEM **replaces the distribution of noise variables** as

$$\mathcal{S}_{X=x} := (S, \mathbb{P}^{\epsilon}_{\mathcal{S},X=x}),$$

where $\mathbb{P}^{\epsilon}_{\mathcal{S},X=x} = \mathbb{P}^{\epsilon \mid X=x}$. Thus we always need to have some observed data before formulating a *counterfactual statement*. **We formulate the counterfactual statement in the same context as the factual statement**, i.e. we leave every variable with the value it had before, except the variable we are intervening on. That makes sense given the name alone: you cannot to a counterfactual without facts to contrast to. Note that the new set of noise variables does not need to be mutually independent any more. A *counterfacutal statement* can now be seen as a *do*-statement in the new counterfactual SEM.

Let's look at two examples, taken from Nicolai Meinshausens' lecture notes.

Example Suppose we are looking at the following SEM:

$$X \leftarrow \epsilon_X \tag{20}$$
$$Y \leftarrow X^2 + \epsilon_Y \tag{21}$$
$$Z \leftarrow 2Y + X + \epsilon_Z, \tag{22}$$

with independent standard normal noise and observations $X = 1$, $Y = 2$, $Z = 4$. Then $\mathbb{P}^{\epsilon}_{\mathcal{S},X=1,Y=2,Z=4}$ puts point mass on $\epsilon = (1, 1, -1)$ (we just need to do the math here and solve for the variables). So if we now do a *counterfactual statement* it would be in the context of the data ($X = 1, Y = 2, Z = 4$). For instance: Z would be 11, had X been 2. Mathematically formulated, we do an intervention on some variable, which translates as a *had been* statement and point mass is put on the noise variables: $\mathbb{P}^{Z|\mathrm{do}(X=x)}_{\mathcal{S},X=1,Y=2,Z=4}(Z = 11) = 1$ has point mass.

Example Suppose we are looking at the following SEM, where T represents a patient taking a specific treatment and B where the patient goes blind.

$$T \leftarrow \epsilon_T \tag{23}$$
$$B \leftarrow T \cdot \epsilon_B + (1 - T) \cdot (1 - \epsilon_B). \tag{24}$$

Said patient comes into the hospital where a doctor suggests treating him $T = 1$. The poor fellow afterwards goes blind $B = 1$. Was this related to the doctor's decision, i.e. would the patient have gone

blind if he hadn't taken the treatment? We can strutinize this with counterfactuals. We are observing $B = 1, T = 1$ and want to know the distribution of B given *no* treatment:

$$\mathbb{P}^{B|\text{do}(T=0)}_{\mathcal{S},B=1,T=1}.$$

Doing the same math as before (i.e. plugging the values into the SEM and computing the error point mass), gives us $\epsilon_T = 1$, $\epsilon_B = 1$. That means

$$\mathbb{P}^{B|\text{do}(T=0)}_{\mathcal{S},B=1,T=1} = \text{Bernoulli}(0),$$

which gives us $\mathbb{P}_{\mathcal{S},B=1,T=1}(B = 0 \mid \text{do}(T = 0)) = 1$. So the patient would not have gone blind in this scenario.

5.1 Remarks

Sometimes two different SEMs that **encode the same observational and interventional distributions** give different counterfactual statements. We must be cautios to reliable infer the correct SEM, if we want to predict counterfactual statements.

6 Computing causal effects from observational data

When we want to compute causal effects from observational data, the simplest case is that we have a known acyclic path diagram with independent exogenous variables (which means we don't have latent confounders). The scenario where $\text{Cov}(X, U_Y) = 0$ is especially easy (Figure 6), because here $\mathbb{E}(Y \mid (do(X = x))) = \mathbb{E}(Y \mid x)$ and we can utilize standard regression models for inference of a causal effect.

Generally, we need to come up with other approaches to compute interventional distributions. In this section we will assume the causal graph is known. If we additionally assume there are no latent confounders, we can determine which causal effect is identifiable from observational data. Next we will introduce what a *causal effect* is and discuss some approaches to compute these.

Figure 7: A simple structure equation model for which $P(Y \mid x) \neq P(Y \mid \text{do}(x))$.

6.1 Causal Effects

Given an SEM $\mathcal{S}$, there is a (total) causal effect from X to Y iff

$$X \not\perp\!\!\!\perp Y \text{ in } \mathbb{P}^{Y \mid \text{do}(X=\tilde{\epsilon}_X)}_{\mathcal{S}},$$

for some variable $\tilde{\epsilon}_X$. Alternatively we can recognize a causal effect when

$$\mathbb{P}^{Y \mid \text{do}(X=x_0)}_{\mathcal{S}} \neq \mathbb{P}^{Y \mid \text{do}(X=x_1)}_{\mathcal{S}}$$

for some x_0, x_1, which we can for instance quantify with the averge causal effect

$$\mathbb{E}(Y \mid do(X = x_1)) - \mathbb{E}(Y \mid do(X = x_2)),$$

Total effect We define the **total effect** of X on Y as:

$$\frac{\partial}{\partial x}\mathbb{E}(Y \mid \mathrm{do}(X = x)) \mid x = x'$$

Note The causal effect can for instance be computed using covariate adjustment, instrumental variables or inverse probability weighting. If we use a linear SEM the total effect can even be computed with linear regression while adjusting for the parents of X (or in fact any sufficient set).

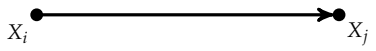

Figure 8: A simple causal model of two variables.

Example Take a linear SEM with path diagram in Figure 6.1

$$X_i \leftarrow \epsilon_i \tag{25}$$
$$X_j \leftarrow \beta_{j,i} X_i + \epsilon_j, \tag{26}$$

where β are the path coefficients. We then do surgery on the path diagram and set $do(X_j = x)$ giving the modified SEM $\tilde{\mathcal{S}}$

$$X_i \leftarrow \epsilon_i \tag{27}$$
$$X_j \leftarrow x. \tag{28}$$
$$\tag{29}$$

Thus the average causal effect of $do(X_j = x)$ on X_i is:

$$\mathbb{E}(X_i \mid do(X_j = x_1)) - \mathbb{E}(X_i \mid do(X_i = x_2)) \tag{30}$$
$$= \mathbb{E}(e_i) - \mathbb{E}(e_i) = 0. \tag{31}$$

If we turn this around, we get:

$$\mathbb{E}(X_j \mid do(X_i = x_1)) - \mathbb{E}(X_j \mid do(X_i = x_2)) \tag{32}$$
$$= \mathbb{E}(\beta_{j,i} x_1 + \epsilon_j) - \mathbb{E}(\beta_{j,i} x_2 + \epsilon_j) \tag{33}$$
$$= \beta_{j,i}(x_1 - x_2). \tag{34}$$

Example Consider the SEM

$$Z_1 \leftarrow \epsilon_{Z_1} \tag{35}$$
$$X \leftarrow 2Z_1 + \epsilon_X \tag{36}$$
$$Z_2 \leftarrow 3X + \epsilon_{Z_2} \tag{37}$$
$$Y \leftarrow 2Z_2 + Z_1 + \epsilon_Y, \tag{38}$$

where ϵ_i are mutually independent arbitray mean zero distributions and respective path diagram in Figure~6.1. Now we intervene on X with $\text{do}(X = x)$, which means we replace the original structural equation with $X \leftarrow x$. Due to autonomy, the other structural equations do not change. Hence

$$Y \leftarrow 2Z_2 + Z_1 + \epsilon_Y = 6x + \epsilon_{Z_1} + 2\epsilon_{Z_2} + \epsilon_Y \tag{39}$$
$$\mathbb{E}(Y \mid \text{do}(X = x)) = 6x \tag{40}$$
$$\frac{\partial}{\partial x}\mathbb{E}(Y \mid \text{do}(X = x)) = 6 \tag{41}$$

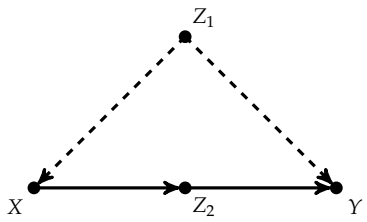

Figure 9: Another SEM

6.2 *Adjusting*

When the causal graph is known and all variables are measured, it is sufficient to find a *adjustment set* of variables to compute the interventional distribution $P(Y \mid x)$, i.e. a set of variables that blocks every back-door path between X and Y.

Valid adjustment set: A set $Z \subseteq V \setminus \{X, Y\}$ is a valid adjustment set for the ordered pair (X, Y) if

$$P_{\mathcal{S},\text{do}(X=x)}(Y) = \sum_z P_{\mathcal{S}}(Y \mid x, z) P_{\mathcal{S}}(z).$$

This is the case for sets Z where:

- no element of Z is a descendant of X
- Z blocks all back-door paths (paths with an arrow into X).

In general if

$$P_{\mathcal{S},\text{do}(X=x)}(y \mid x, z) = P_{\mathcal{S}}(y \mid x, z)$$

and

$$P_{\mathcal{S},do(X=x)}(z) = P_{\mathcal{S}}(z),$$

then Z is a sufficient set for the pair (X, Y). A sufficient, but not necessarily minimal set, are the parents of X: PA_X. The causal effect, given a variable's parents, is then:

$$P_{\mathcal{S},\text{do}(X=x)}(Y) = \sum_{pa_X} P_{\mathcal{S}}(Y \mid x, pa_X) P_{\mathcal{S}}(pa_X).$$

In the continuous case, when we choose the parents of X as adjustment set:

$$P_{\mathcal{S},\text{do}(X=x)}(Y) = \int P_{\mathcal{S}}(Y \mid x, pa_X) P_{\mathcal{S}}(pa_X)\ dpa_X,$$

for which the average causal effect would be (Maathuis et al. 2009):

$$\mathbb{E}_{\mathcal{S},\text{do}(X=x)}(Y) = \int \mathbb{E}(Y \mid x, pa_X) P_{\mathcal{S}}(pa_X)\ dpa_X$$

Example Consider the SEM $\mathcal{S}_X$ that is obtained by replacing the stuctural equation for X using $do(X = x)$. We can estimate $\mathbb{P}_{\mathcal{S}_X}(Y = 1) = \mathbb{P}_{\mathcal{S}}(Y = 1 \mid \text{do}(X = x))$ like this:

$$\begin{aligned}
\mathbb{P}_{\mathcal{S},\text{do}(x)}(Y = 1) &= \sum_z \mathbb{P}_{\mathcal{S}_X}(Y = 1, x, z) && (42)\\
&= \sum_z \mathbb{P}_{\mathcal{S}_X}(Y = 1 \mid x, z)\mathbb{P}_{\mathcal{S}_X}(x, z) && (43)\\
&\overset{\text{Point mass}}{=} \sum_z \mathbb{P}_{\mathcal{S}_X}(Y = 1 \mid x, z)\mathbb{P}_{\mathcal{S}_X}(z) && (44)\\
&\overset{\text{Invariance}}{=} \sum_z \mathbb{P}_{\mathcal{S}}(Y = 1 \mid x, z)\mathbb{P}_{\mathcal{S}}(z) && (45)
\end{aligned}$$

Example Consider the SEM with path-diagram 6.2 (from Pearl (2009a)). We are interested in computing the causal effect $P(Y \mid \text{do}(x))$. The pre-interventional distribution factorizes as:

$$\begin{aligned}
&P(X, Y, Z_1, Z_2, Z_3) && (46)\\
&= P(Z_1)P(Z_2)P(Z_3 \mid Z_1, Z_2)P(X \mid Z_1, Z_3)P(Y \mid X, Z_2, Z_3). && (47)
\end{aligned}$$

Next we introduce the intervention $\text{do}(X = x)$:

$$\begin{aligned}
&P(Y, Z_1, Z_2, Z_3 \mid do(X = x)) && (48)\\
&= P(z_1)P(z_2)P(z_3 \mid z_1, z_2)P(Y \mid x, Z_2, Z_3). && (49)
\end{aligned}$$

Note how we used the autonomy of the other stuctural equations. We *only* replaced the equations (conditional distributions) that involved X. Now, we are ready to compute the causal effect $P(Y \mid \text{do}(x))$, i.e. we by marginalization over Z_1, Z_2, Z_3:

$$P(Y \mid do(x)) = \sum_{z_1,z_2,z_3} P(z_1)P(z_2)P(z_3 \mid z_1, z_2)P(Y \mid x, z_2, z_3)$$

We can also easily intervene on two variables:

$$P(Y, Z_2, Z_3 \mid do(X = x, Z_1 = z_1)) = P(Z_2)P(Z_3 \mid z_1)P(Y \mid x, Z_2, Z_3)$$

In that case the causal effect is:

$$P(Y \mid do(X = x, Z_1 = z_1)) \tag{50}$$
$$= \sum_{z_2, z_3} P(z_2)P(z_3 \mid z_1)P(Y \mid x, z_2, z_3). \tag{51}$$

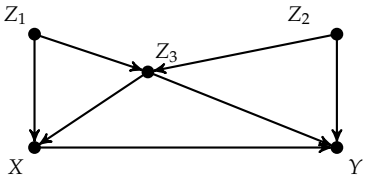

Figure 10: Path diagram of five variables.

Characterization of sufficient sets Valid adjustment sets can be characterized by one of these three criteria (Shpitser, Van der Weele, and Robins 2010):

- **Parent adjustment** $Z := PA_X$ is a valid adjustment set for (X, Y) for any $Y \notin \{X, PA_X\}$
- **Backdoor criterion** Any Z with a) Z contains no descendant of X and b) Z blocks all paths from $X \to Y$ entering through the backdoor, is a valid set for (X, Y) for any $Y \notin \{X, PA_X\}$
- **Towards necessity** Any Z with a) Z contains no descendant of any node on a directed path from $X \to Y$ and b) Z blocks all non-directed paths from $X \to Y/$

Not all sets are valid adjustment sets, see for instance Berkson's paradox. **The back-door criterion lets us select variables for adjustment in order to compute causal effects**. Thus we do not need to intervene on the graph directly, since we can use regression techniques for the estimation of $P(Y \mid x)$.

6.3 Causal effects for multivariate Gaussians

In general, the **causal effect** of X_i on Y is given by $\beta_{i|pa_i}$, where for each set $S \subseteq \{X_1, \ldots, X_p, Y\} \setminus \{X_i\}$:

$$\beta_{i|S} = \begin{cases} 0 & \text{if } Y \in S, \\ \text{coefficient of } X_i \text{ in } Y \sim X_i + S & \text{else} \end{cases} \tag{52}$$

Hence in the Gaussian case the causal effect does not depend on the value of x_i' and can be interpreted as

$$E(Y \mid do(X_i = x_i' + 1)) - E(Y \mid do(X_i = x_i')) \tag{53}$$

Thus **for jointy Gaussian variables the causal effect can be computed using LR** given the DAG. This is precicely the average causal effect

$$\frac{\partial}{\partial x} \mathbb{E}_{\mathcal{S}, \mathrm{do}(X = x)}(Y)$$

for a continuous variable X. Generally the expectation is a function on x, but in the Gaussian case it is a constant.

Example Consider a jointly Gaussian random vector $X_1, \ldots, X_p, Y$. We are interested in estimating the causal effect of X_i on Y as described above. Since Gaussianity implies

$$\mathbb{E}(Y \mid x_i', pa_i) = \beta_0 + \beta_i x_i + \beta_{PA_i}^T pa_i,$$

we find the computation of the causal effect especially easy. By marginalizing out the parents, we get the causal effect, which in the Gaussian case is only the regression coefficient of X_i. However, **intervention calculus might still give different results**, because the variables we control for are different. For illustration consider this SEM (and its path diagram in Figure 6.3:

$$X_2 \leftarrow \epsilon_2 \tag{54}$$
$$X_1 \leftarrow 0.8X_2 + \epsilon_1 \tag{55}$$
$$X_3 \leftarrow 0.8X_2 + \epsilon_3 \tag{56}$$
$$Y \leftarrow -X_1 + 2X_2 - X_3 + \epsilon \tag{57}$$

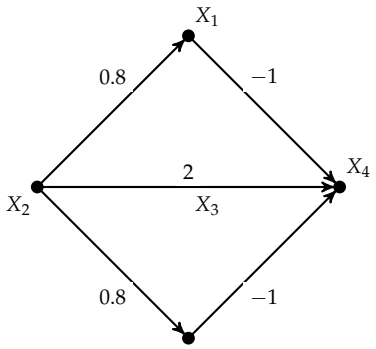

Figure 11: A simple structural equation model with causal effect $X_i \to X_j$.

Regression Y on X_1, X_2, X_3 would give us regression coefficients $\beta_2 = 2$ making X_2 the variable with the largest effect. If we compute the *intervention effect* as above we get:

$$\theta_2 = \beta_{2|\emptyset} = \frac{\partial}{\partial x_2}\mathbb{E}\left[-(0.8x_2 + \epsilon_1) + 2x_2 - (0.8x_2 + \epsilon_3) + \epsilon\right] = 0.4.$$

Here, X_2 would have the lowest effect. We can easily verify this using R. First we simulate a data set according to the SEM:

```
n <- 10000
x2 <- rnorm(n)
x1 <- 0.8 * x2 + rnorm(n)
x3 <- 0.8 * x2 + rnorm(n)
y <- -x1 - x3 + 2 * x2 + rnorm(n)
```

Then compute the regression coefficients, i.e. the **direct effect**:

```
lm(y ~ x2 + x1 + x3)
```

```
##
## Call:
## lm(formula = y ~ x2 + x1 + x3)
##
## Coefficients:
## (Intercept)           x2           x1
##     0.01732      1.98867     -0.99268
##          x3
##    -0.99733
```

Compare this to the overall causal effect, i.e. **total effect**:

```
lm(y ~ x2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Coefficients:
## (Intercept)           x2
##    0.006472     0.409023
```

Let's also check the equality of direct effect $\beta_1 = -1$ and total effect $\theta_1 = \beta_{1|2} = -1$ of X_1. These two should be the same:

```
lm(y ~ x2 + x1)
```

```
##
## Call:
## lm(formula = y ~ x2 + x1)
##
## Coefficients:
## (Intercept)           x2           x1
##     0.02559      1.19223     -0.99056
```

Both of them are almost equal in the two models.

6.4 Remarks

- Keep in mind that a causal effect coefficient θ is not *interpreted*, but **proven** and will its retain its causal interpretation regardless of how X is selected.
- We can distinguish between intervention effect and regression coefficient as follows. The regression coefficient β_i estimates the **direct** effect on the response. The intervention/causal effect $\theta_i = \beta_{i|S}$ measures the **total** effect.

6.5 Do-calculus

Sometimes we need to compute interventional distributions other than by adjustment, for instance, when not all variables are measured. A causal effect $P(Y \mid \text{do}(x))$ is thus sometimes not *identifiable*, i.e. we cannot compute it from the graph structure and an observational distribution. An interventional distribution is identifiable if there is a valid adjustment set Z for (X, Y).

Judea Pearl's *do*-calculus (Pearl 2009b) has three rules. Denote $\mathcal{G}_{\overline{X}}$ the graph obtained by deleting all incoming edges to X and $\mathcal{G}_{\underline{X}}$ the one obtained by deleting outgoing edges.

- **Rule 1** Insertion/deletion of observations: $p_{\mathcal{S},do(X=x)}(y \mid z, w) = p_{\mathcal{S},do(X=x)}(y \mid w)$ if Y d-separates Z given X, W in $\mathcal{G}_{\overline{X}}$.
- **Rule 2** Action/observation exchange: $p_{\mathcal{S},do(X=x,Z=z)}(y \mid w) = p_{\mathcal{S},do(X=x)}(y \mid z, w)$ if Y d-separates Z given X, W in $\mathcal{G}_{\overline{X},\underline{Z}}$.
- **Rule 3** Insertion/deletion of actions: $p_{\mathcal{S},do(X=x,Z=z)}(y \mid w) = p_{\mathcal{S},do(X=x)}(y \mid w)$ Y d-separates Z given X, W in $\mathcal{G}_{\overline{X},\overline{Z(W)}}$. $Z(W)$ is the set of nodes in Z that are not ancestors of any node in W in graph $\mathcal{G}_{\overline{X}}$

Applying these three rules all identifiable intervention distributions can be computed (see Meinshausen's lecture notes for more detail and Pearl (2009b) Theorem 3.4.1).

Example Consider the SEM in Figure 6.5. If we do not observe U, we do not have a valid adjustment set. However, using the *front-door criterion* (from the rules above) we can still infer $P(Y \mid \mathrm{do}(X = x))$:

Using the *do*-calculus the interventional distribution can be computed as:

$$P(Y \mid \mathrm{do}(x)) = \sum_m P(m \mid x) \sum_{x'} P(Y \mid x', m) P(x')$$

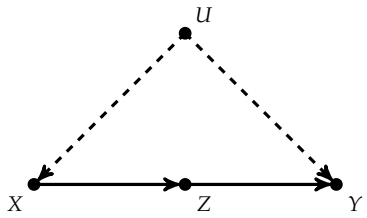

Figure 12: Graph of a SEM with no valid adjustment sets.

6.6 *Instrumental variables*

A variable Z is *instrumental* if it has no parents and the intervened variable X is a descendant. To make the causal effect $P(Y \mid do(x))$ *not* identifiable, we additionally assume a latent confounder U, such that

$$X = \beta Z + \gamma U + \epsilon_X,$$

and the effect $X \to Z$ *is* estimateable. The causal effect $P(Y \mid x)$ can be estimated using *two stage least squares*. First we regress

$$X = Z\beta + \epsilon_X$$

and then **the predicted values**

$$Y \sim Z\hat{\beta} + \epsilon_Y.$$

In order to make the instrumentable variable work, we need to make **stringent assumptions**, i.e.

- linear SEMs,
- non-zero β,
- $U \perp\!\!\!\perp Z$,
- no direct influence $Z \to Y$.

Shalizi's mentions that many instruments are *weak*, i.e. they only have little influence on on X and a small covariance with it. Weak instruments can lead to noisy and imprecise estimation of causal effects.

7 Structure learning

Structure learning concerns itself with computation of the causal graph from observational data. It turns out that identiying causal DAGs is in most cases not possible if we do not use stringent assumptions. There are several approaches how to estimate the causal graphs. Here, I introduce these two:

- constraint-based (e.g. PC-algorithm and FCI (Spirtes et al. 2000))
- score-based (e.g. GES (Chickering 2002) or GIES (Hauser and Bühlmann 2012))

Other approaches include SEMs with additional restrictions (e.g. LINGAM (Shimizu et al. 2006)) or hybrid approaches (MMHC (Tsamardinos, Brown, and Aliferis 2006)).

7.1 Identifiability

Consider a distribution $\mathbb{P}^X$ with strictly positive density w.r.t to a Lebesgue measure that is Markov w.r.t to a DAG $\mathcal{G}$. Then there exists a SEM $\mathcal{S} = (S, \mathbb{P}^\epsilon)$ with DAG $\mathcal{G}$ that generates $\mathbb{P}^X$. This holds in particular for all fully connected graphs. That means several different SEMs can generate our target distribution. In order to obtain identifiablity results, we thus need more assumptions.

Markov equivalence Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are *Markov equivalent* if they have the **same skeleton and immoralities**. A Markov equivalence class (i.e. a set of DAGs in a Markov equivalence class) can be completely represented by a *complete partially diracted acyclic graph* (CPDAG). The CPDAG has a directed edge $X_i \to X_j$ if the edge is found in *every* DAG in the equivalence class. Otherwise we use an undirected edge, because we are uncertain about its direction.

Faithfulness Assume $\mathbb{P}^X$ is Markov and faithful w.r.t $\mathcal{G}^0$. Then, for each graph $\mathcal{G} \in CPDAG(\mathcal{G}^0)$, we find a SEM that generates the distribution. Furthermore $\mathbb{P}^X$ is **not** faithful and Markov to any other graph $\mathcal{G} \notin CPDAG(\mathcal{G}^0)$.

If $\mathbb{P}^X$ is Markov and faithful, we have a one-to-one correspondence between conditional independence relationships and d-separation statements. So we do not need to consider other graphs outside the Markov equivalence class, because they impose conditional independences that do not hold in $\mathbb{P}^X$. That means we need only to be able to distinguish graphs in the Markov equivalence class. As it stands, **in general we can only identify the CPDAG from observational data and the empirical distribution** $\mathbb{P}^X$.

Example Assuming independent multivariate Gaussian errors ϵ, a causal DAG $\mathcal{G}$ and linear structural equations, we are generally not able to infer the underlying causal model when we are only given observational data. However, we are usually perfectly able to identify the Markov equivalence class of DAGs that encode the same conditional independencies (or d-separation relationships). This holds, because from observational data we are only able to infer immoralities and the skeleton of the causal model.

Note From observational data, we can compute a lower bound on the causal effect $X \to Y$, e.g. by computing the causal effects of all DAGs in a Markov equivalence class and take the lowest value. This only holds if the distribution is faithful to the true unknown DAG and no latent confounders are present.

For learning, we will assume an underlying DAG with indepenent multivariate noise. If we assume Gaussianity for the noise variables we actually limit ourselves in the possibiilty of estimating the causal model. However, the normal assumption is the most common. Likewise linear models are usually harder to estimate than their non-linear counterparts, but require less data for estimation.

Even with the Markov and faithfulness assumptions we still in many cases accont uniquely identify the causal structure of a linear acyclic SEM with no latent confounders and Gaussian noise.

Additive noise models SEMs of the form

$$S_j : \; X_j \leftarrow f_j(PA_j) + \epsilon_j,$$

are called *additive noise model*s, i.e. when the noise acts additively. Those models are **causally minimal** if the functions f are not constant. That means we can always find two values x_0 and x_1 for *every* parent $X_i \in PA_j$ of *every* X_j in the graph for which

$$f_j(x_{PA_j \setminus \{X_i\}}, x_0) \neq f_j(x_{PA_j \setminus \{X_i\}}, x_1).$$

Often we assume causal minimality, because generally we are not able to detect if a variable contributes to the value of another in a constant way. The class of additive noise models we just described are in general not identifiable (see the linear Gaussian case), but identifiable if the functions are non-linear.

Identifiable Gaussian models The linear Gaussian case is however rather the exception, because it turns out we can identify almost all other combinations of funtions and distributions. All non-identifiable cases have already been characterized (Peters et al. 2014, Zhang and

Hyvärinen (2009)). So, which models are identifiable and which not? Let's have a look for cases with $\epsilon_i \sim \mathcal{N}(\cdot, \cdot)$:

- The general SEM ($X_i \leftarrow f_i(PA_i, \epsilon_i)$) is **not identifiable**, due to the fact that multiple DAGs encode the same CI relations.
- The additive noise model $X_i \leftarrow f_i(PA_i) + \epsilon_i$ for *non-linear*, *non-constant* functions f_j is **identifiable**.
- The causal additive model $X_i \leftarrow \sum_{X_j \in PA_i} f_{ik}(X_k) + \epsilon_i$ for *non-linear*, *non-constant* functions f_j is **identifiable**.
- The linear Gaussian $X_i \leftarrow \sum_{X_j \in PA_i} \beta_{ik}(X_k) + \epsilon_i$ is **not identifiable**

Linear non-Gaussian acyclic models **Proposition** Let X and Y be two random variables for which

$$Y \leftarrow \phi X + \epsilon,\ X \perp\!\!\!\perp \epsilon,\ \phi \neq 0,$$

holds. Then we can reverse the process, i.e. there exists some $\psi \in R$ and noise $\tilde{\epsilon}$, such that

$$X \leftarrow \psi Y + \tilde{\epsilon},\ Y \perp\!\!\!\perp \tilde{\epsilon},$$

iff X and ϵ are normally distributed (see also the Theorem Dermois-Skitovic). That would imply that the graph is identifiable in the non-Gaussian case?

Theorem Assume a SEM with graph $\mathcal{G}$

$$X_j \leftarrow \sum_{X_k \in PA_j} \beta_{jk} X_j + \epsilon_j,$$

with **mutually independent**, **non-Gaussian** ϵ with **strictly positive** density an $\beta \neq 0$. Then $\mathcal{G}$ is identifiable from $\mathbb{P}^X$. The model is called *LInear Non-Gaussian Ayclic Model* (LINGAM, Shimizu et al. (2006), Shimizu (2014)) .

Nonlinear Gaussian additive noise models For distributions $\mathbb{P}^X$ generated by $X_j \leftarrow f_j(PA_j) + \epsilon_j$ with marginal, normal, zero mean noise and three-times diffable non-linear functions fs, the graph is identifiable.

The special case $X_j \leftarrow \sum_{X_k \in PA_j} f_{j,k}(X_k) + \epsilon_j$ with normally distributed zero-mean noise is known as **causal additive model**. Again we assume three-times diff-able non-linear functions f. In both cases the correct graph $\mathcal{G}$ can be identified from $\mathbb{P}^X$.

7.2 *Independence-based methods*

Independence-based methods **assume faithfulness** and therefore **estimate the underlying CPDAG from conditional independencies**

in $\mathbb{P}^X$. That means we are interested in finding the skeleton, then orient the edges.

Lemma Two nodes X, Y are adjacent iff they cannot be d-separated by any set of nodes $S \subseteq V \setminus \{X, Y\}$. If two nodes are not adjacent, then they are d-separated by either PA_X or PA_Y.

That means if two variables are dependent no matter on which variables we condition on, then they are neighbors. This reasoning is used in Pearl's **Inductive Causation** algorithm (Pearl 2009b), in the **SGS**-algorithm (Spirtes, Glymoir and Scheines, Spirtes et al. (2000)) or in the **PC**-algorithm (Peter and Clark, Spirtes et al. (2000)). The SGS-algorithm proceeds like this:

1) start with an complete, undirected graph,
2) for each pair of variables examine all possible conditioning sets,
3) test every of these triplets for CI and remove an edge if the Null is accepted,
4) if we have an *Oracle* test, the output will be the true skeleton of the true DAG.

The PC-algorithm builds on this and tries to avoid conditioning on all subsets. It tests all pairs of variables with the empty conditioning set and removes edges that are independent. Then it recursively adds edges to the conditioning sets which quickly gives sparse graphs. In the population version of PC algorithm the conditional dependencies are *known*, so the estimated CPDAG is correct. In the *sample version* we need to estimate conditional dependencies using hypothesis tests. In both cases we can direct edges from immoralities $X \to Z \leftarrow Y$. If X, Y are independent for some d-separation set S and adding Z to the set introduces depence, we know the triple is an immorality.

Independence tests are in general not easy to conduct, especially in the non-Gaussian case. In the Gaussian case we just test for partial correlation of zero:

$$H_0 : \rho_{X,Y|S} = 0, \tag{58}$$

$$H_1 : \rho_{X,Y|S} \neq 0 \tag{59}$$

Using Fisher's Z-transform, the test statistic has the follow distribution under the Null:

$$\hat{Z}_{X,Y|S} = \frac{1}{2} \log \left(\frac{1 + \tilde{\rho}_{X,Y|S}}{1 - \tilde{\rho}_{X,Y|S}} \right) \sim \mathcal{N}(0, (n - |S| - 3)^{-1}),$$

if $\rho_{X,Y|S} = 0$. Thus for a specified α we conclude that the true partial correlation $\rho_{X,Y|S} \neq 0$ if

$$|\hat{Z}_{X,Y|S}|\sqrt{n-|S|-3} > \phi^{-1}(1-\alpha/2) \tag{60}$$

where ϕ^{-1} is the inverse of the normal CDF.

The partial correlations can be computed via regression by first regressing $S \to X$ and $S \to Y$ and then computing the correlation of the residuals. Alternatively we compute the empirical correlation $\hat{\Sigma}$ of X, Y, S: $\hat{\Sigma}_{X,Y,Z}$ and take the inverse $\hat{\Lambda} = (\hat{\Sigma}_{X,Y,Z})^{-1}$. The partial correlation is then

$$\hat{\rho}_{X,Y|S} = -\frac{\hat{\Lambda}_{X,Y}}{\sqrt{\hat{\Lambda}_{X,X}\hat{\Lambda}_{Y,Y}}}.$$

This gives us an estimated CPDAG $\hat{\mathcal{G}}$.

The PC-algorithm however has problems when

- latent confounders are present.
- conflicting immoralities.

7.3 Score-based methods

If we try identifying the DAG from observational data, another alternative is to not rely on conditional independence statements, but on *scores*, i.e. we want to find the graph with the highest score.

$$\mathcal{G} := \arg\max_{\mathcal{G} \text{ over } X} S(\mathcal{D}, \mathcal{G}),$$

where $\mathcal{D}$ are observations of the random vector X. Usually we assume X to be Gaussian or multinomial, which introduces a set of parameters θ that we need to estimate.

Penalized likelihood One possibility is to use the likelihood as a score S. Since the maximum likelihood estimate would be the complete graph we additionally add a penalty, like the BIC, such that

$$S(\mathcal{D}, \mathcal{G}) = \log P(\mathcal{D} \mid \hat{\theta}, \mathcal{G}) - \frac{\#\text{ parameters}}{2} \log n,$$

where n is the number of samples. Since the DAG-space grows super-exponentially in te number of variables, we need to find a clever solution to estimate as many DAGs as possible. Traversing the DAG-space is done by, for instance, removing/adding/rewiring edges. Since in many cases (e.g. Gaussian) we are not able to find the true DAG anyways, but only the CPDAG, we often are interested in only identifying this graph. For instance, *greedy-equvalence-search* (GES, Chickering (2002)) works like this:

- start with the empty graph,

- add edges until a local maximum is reached (i.e. the score is not getting better any more),
- remove edges until a local maximum is reached.

In a Bayesian formulation we would put priors on the graph structure $P(\mathcal{G})$ and the parameters $P(\theta)$ and consider the log-posterior as score

Two graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ are distribution equivalent if for each parameter θ_1 there is a corresponding parameter θ_2, such that the distributions obtained from $\mathcal{G}_1$ and θ_1 are the same from $\mathcal{G}_2$ and θ_2. In the linear Gaussian case, two graphs are distribution equivalent iff they are Markov equivalent.

8 Notation

Symbol/Term	Meaning
$(\Omega, \mathcal{F}, \mathbb{P})$	probability space
Ω	sample space, e.g. for a die $\Omega = \{1, \dots, 6\}$
$\mathcal{F}$	σ-algebra on Ω, e.g. $\mathcal{F} = \{\emptyset, \{1,2,3\}, \{4,5,6\}\}$
$\mathbb{P} : \mathcal{F} \to [0,1]$	probability measure
$\mathbb{P}^X$	distribution of p-dimensional random vector X
$X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$	measurable function with respect to a Borel σ-algebra (a random variable)
$P(X = x)$ or $P(x)$	Probability of X having the value x
$P(Y \mid \mathrm{do}(X = x))$,$P(Y_x)$ or $(Y \mid \mathrm{do}(x))$	Probability of Y having setting $X = x$
Immorality/v-structure	A triple of nodes with two edges colliding on one node $X_i \to X_k \leftarrow X_j$
PDAG	graph without directed cycle
$An(X)$	Ancestral set of the set X
PA_j or PA_{X_j}	parents of random variable/node X_j in a graph
$\mathbb{P}^X_{\mathcal{S}}$	distribution of p-dimensional random vector X given by SEM $\mathcal{S}$

References

Bottou, Léon, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. "Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising." *The Journal of Machine Learning Research*.

Bühlmann, Peter, Petros Drineas, Michael Kane, and Mark van der Laan. 2016. *Handbook of Big Data*. CRC Press.

Chickering, David Maxwell. 2002. "Optimal Structure Identification with Greedy Search." *Journal of Machine Learning Research*.

Hauser, Alain, and Peter Bühlmann. 2012. "Characterization and Greedy Learning of Interventional Markov Equivalence Classes of Directed Acyclic Graphs." *Journal of Machine Learning Research*.

Heinze-Deml, Christina, Marloes H Maathuis, and Nicolai Meinshausen. 2018. "Causal Structure Learning." *Annual Review of Statistics and Its Application*.

Koller, Daphne, and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*.

Maathuis, Marloes H, Markus Kalisch, Peter Bühlmann, and others. 2009. "Estimating High-Dimensional Intervention Effects from Observational Data." *The Annals of Statistics*.

Morgan, Stephen L, and Christopher Winship. 2015. *Counterfactuals and Causal Inference*. Cambridge University Press.

Pearl, Judea. 2009a. "Causal Inference in Statistics: An Overview." *Statistics Surveys*.

———. 2009b. *Causality*.

Peters, Jonas, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. 2014. "Causal Discovery with Continuous Additive Noise Models." *The Journal of Machine Learning Research*.

Shimizu, Shohei. 2014. "LiNGAM: Non-Gaussian Methods for Estimating Causal Structures." *Behaviormetrika*.

Shimizu, Shohei, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. "A Linear Non-Gaussian Acyclic Model for Causal Discovery." *Journal of Machine Learning Research*.

Shpitser, Ilya, Tyler Van der Weele, and James M Robins. 2010. "On the Validity of Covariate Adjustment for Estimating Causal Effects." *Conference on Uncertainty in Artificial Intelligence*.

Spirtes, Peter, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. 2000. *Causation, Prediction, and Search*.

Tsamardinos, Ioannis, Laura E Brown, and Constantin F Aliferis. 2006. "The Max-Min Hill-Climbing Bayesian Network Structure

Learning Algorithm." *Machine Learning*.

Zhang, Kun, and Aapo Hyvärinen. 2009. "On the Identifiability of the Post-Nonlinear Causal Model." In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*.