

D. Do-calculus

Causal questions...

- What are the predicted results of some given action? ('Effects of causes', EoC)

I have a headache. Will taking aspirin help?

- 'type' causation, decision theory

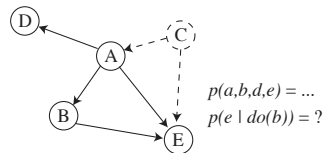
- What was the cause of some given event? ('Causes of effects', CoE)

My headache is gone. Is it because I took aspirin?

- 'token' causation, 'counterfactuals'

In this section:

- Let's assume that we know the structure of the causal network. When and how can we compute the answers to $P(y \mid do(x))$ -type queries, from the joint distribution of the observed variables?



DAG model: either causal Bayesian network
or functional causal model

- discrete case
- linear/continuous

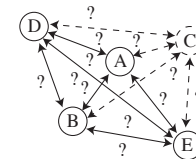
Later in the course:

- Assuming that we know the full causal model, when and how can we compute the answers to 'counterfactuals' such as: "what would have happened had I chosen differently?"

- discrete case
- linear/continuous

...and still later in the course:

- Can we, under some assumptions, say anything about the structure of the causal model from purely observational data?



$p(a,b,d,e) = \dots$

What is the structure of the data generating model?

- discrete case
- linear/continuous

Estimation of do(x) probabilities, direct approach:

- If one wants to estimate $P(y | do(x))$ -type probabilities, the most straightforward approach is of course to manipulate (i.e. 'do') x and then observe y . (e.g. 'randomized controlled trials')
 - Definitely the preferred approach, if possible (examples: medicine, natural sciences)
 - However, this approach can in many cases be...
 - Expensive
 - Ethically questionable
 - Technically impossible

Calculation of do(x) queries, when the full model is known...

- Assume that we are given a causal Bayesian network
- How can we calculate regular conditional/marginal probabilities?

[Note: efficient 'message-passing' methods ('belief propagation', 'sum-product') available for large networks!]

We can first generate the full joint probability (multidimensional array), and then conditionalize/marginalize in the normal way:

(these are specified in the model)

$$P(x, y, z) = P(x)P(y | x)P(z | x, y)$$

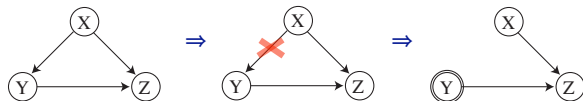
(the structure is specified in the model)

$$P(z | y) = \sum_x P(x, z | y) = \sum_x \frac{P(x, y, z)}{P(y)}$$

$$= \sum_x \frac{P(x, y, z)}{(\sum_{x,z} P(x, y, z))}$$

- How can we calculate do(y)-conditional probabilities?

Cut all arrows pointing into the 'do' node...



...and then calculate conditionals in the 'standard' way:

$$P_{\hat{y}}(x, y, z) = P(x)P_{\hat{y}}(y)P(z | x, y)$$

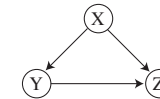
$$P(z | do(y)) = P(z | \hat{y}) = \sum_x P(x, z | \hat{y}) = \sum_x \frac{P_{\hat{y}}(x, y, z)}{P_{\hat{y}}(y)}$$

$$= \sum_x P(x)P(z | x, y)$$

($P(x)$ and $P(z | x, y)$ specified, and $P_{\hat{y}}(y) = \begin{cases} 1, & \text{when } y = \hat{y} \\ 0, & \text{otherwise} \end{cases}$)

What if the structure of the model is known, but not the complete model (i.e.: the marginal and conditional probabilities not given)?

- If there are no hidden variables (that is we can observe all the variables) then the whole model is possible to estimate:



- We can estimate the joint distribution $P(x, y, z)$ from the data (note: with a finite amount of data may be difficult in practice!)
- From it we can calculate the required marginal and conditional distributions $P(x)$, $P(y | x)$ and $P(z | x, y)$, and so we have an estimate for the complete model

- What if there are **hidden variables** in the network?

- If all the parents of the do-variable(s) have been observed, then calculation is straightforward:

$$P(x_1, \dots, x_n \mid \text{do}(x'_i)) = \begin{cases} \frac{P(x_1, \dots, x_n)}{P(x'_i \mid \text{pa}(x_i))} & \text{if } x_i = x'_i, \\ 0 & \text{if } x_i \neq x'_i. \end{cases}$$

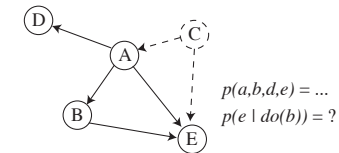
$$= \begin{cases} \frac{P(x_1, \dots, x_n)P(\text{pa}(x_i))}{P(x'_i, \text{pa}(x_i))} & \text{if } x_i = x'_i, \\ 0 & \text{if } x_i \neq x'_i. \end{cases}$$

$$= \begin{cases} P(x_1, \dots, x_n \mid x'_i, \text{pa}(x_i))P(\text{pa}(x_i)) & \text{if } x_i = x'_i, \\ 0 & \text{if } x_i \neq x'_i. \end{cases}$$

$$\Rightarrow P(y \mid \text{do}(x'_i)) = \sum_{\text{pa}(x_i)} P(y \mid x'_i, \text{pa}(x_i))P(\text{pa}(x_i))$$

('adjustment for direct causes', Pearl theorem 3.2.2)

- Example:

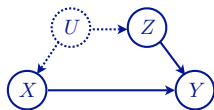


$$P(e \mid \text{do}(b)) = \sum_a P(e \mid a, b)P(a)$$

(on the right side of the equation we have only 'regular' probabilities which can be estimated from observed uncontrolled data!)

- But what if some of the relevant parents are hidden?

Example:



Can we derive $P(y \mid \text{do}(x))$ from $P(x, y, z)$ given this causal graph?

$$P(y \mid \text{do}(x)) = \sum_u P(y \mid x, u)P(u) \quad (\text{adjustment for direct causes})$$

$$= \sum_u \sum_z P(y, z \mid x, u)P(u)$$

$$= \sum_z \sum_u P(y \mid z, x, u)P(z \mid x, u)P(u)$$

$$= \sum_z \sum_u P(y \mid z, x)P(z \mid u)P(u) \quad (\text{using indep. given by graph})$$

$$= \sum_z P(y \mid z, x) \sum_u P(z, u)$$

$$= \sum_z P(y \mid z, x)P(z)$$

Identifiability of causal effects

Definition: (Pearl, def 3.2.4)

The causal effect of X on Y is **identifiable** in a graph G if $P(y \mid \hat{x})$ can be computed uniquely from any positive distribution over the observed variables.

I.e. this guarantees that we can estimate $P(y \mid \hat{x})$ from two sources of information:

- **Passive observations**, from which we can estimate $P(v)$
- **A causal graph** G , which gives a (qualitative) description over which variables causally affect which others

The condition $P(v) > 0$ guarantees that we can observe all the contexts which might occur when X is set arbitrarily.

Back-door criterion

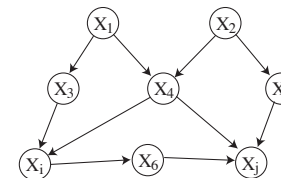
Definition (Pearl def 3.3.1)

A variable set Z satisfies the back-door criterion relative to the ordered pair of variables (X_i, X_j) in a DAG G if:

- i. none of the variables in Z are descendants of X_i , and
- ii. Z blocks (d-separates) all paths between X_i and X_j which include an arrow into X_i

If X and Y are sets of variables, then Z satisfies the back-door criterion relative to the ordered pair (X, Y) if and only if it satisfies the criterion relative to all pairs (X_i, Y_j) such that $X_i \in X$ and $Y_j \in Y$.

Example (back-door)



In this graph $Z = \{X_3, X_4\}$ satisfies the back-door criterion relative to (X_i, X_j) . Also $Z = \{X_4, X_5\}$ satisfies the criterion. However, $Z = \{X_4\}$ does not, since it does not block the path $X_i - X_3 - X_1 - X_4 - X_2 - X_5 - X_j$

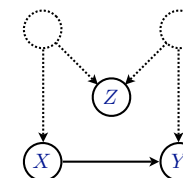
Back-door-adjustment (Pearl theorem 3.3.2)

If the variable set Z satisfies the back-door criterion relative to (X, Y) then the causal effect of X on Y is identifiable and is obtained as

$$P(y | \hat{x}) = \sum_z P(y | x, z)P(z)$$

Note: this result can be derived from the fact that when the back-door criterion is satisfied, we have in this context that $see(x)$ and $do(x)$ are equivalent

Note: Cannot condition on all observed variables prior to the potential cause 'just to be safe':

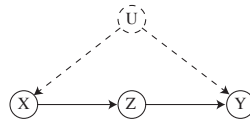


Say we are interested in the causal effect of X on Y . We have also measured an additional variable Z which gets its value prior to X . Should we condition on it? Answer: No!

$$P(y | do(x)) = P(y | x) \neq \sum_z P(y | x, z)P(z)$$

Can **intermediate variables** (i.e. which are affected by the ‘treatment’ variable) **help identify causal effects?**

Example:



We have...

$$P(x, y, z, u) = P(u)P(x | u)P(z | x)P(y | z, u)$$

...giving the post-intervention distribution:

$$P(y, z, u | \hat{x}) = P(y | z, u)P(z | x)P(u)$$

Summing over z and u gives:

$$P(y | \hat{x}) = \sum_z P(z | x) \sum_u P(y | z, u)P(u)$$

Note that:

$$P(u | z, x) = P(u | x)$$

$$P(y | x, z, u) = P(y | z, u)$$

This yields...

$$\begin{aligned} \sum_u P(y | z, u)P(u) &= \sum_x \sum_u P(y | z, u)P(u | x)P(x) \\ &= \sum_x \sum_u P(y | x, z, u)P(u | x, z)P(x) \\ &= \sum_x P(y | x, z)P(x) \end{aligned}$$

allowing the causal query to be answered:

$$P(y | \hat{x}) = \sum_z P(z | x) \sum_{x'} P(y | x', z)P(x')$$

So variable Z helped identify the causal effect! (However, note the **missing edges** in the graph! They make this possible.)

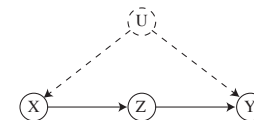
Front-door criterion

Definition (Pearl def 3.3.3)

A variable set Z satisfies the front-door criterion relative to the ordered variable pair (X_i, X_j) in a DAG G if and only if:

- i. every directed path from X_i to X_j contains a variable in Z , and
- ii. there is no back-door path from X_i to Z , and
- iii. X_i blocks all back-door paths from Z to X_j

Prototypical front-door example:



In this graph Z satisfies the front-door criterion with respect to (X, Y)

Front-door adjustment (Pearl theorem 3.3.4)

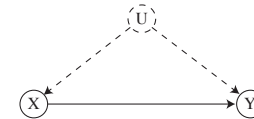
If a variable set Z satisfies the front-door criterion with respect to (X, Y) , and $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is obtained as

$$P(y | \hat{x}) = \sum_z P(z | x) \sum_{x'} P(y | x', z) P(x')$$

(It should be noted that the front-door criterion is a sufficient, but not necessary, criterion for the above equation to hold. We will soon describe the do-algebra which is a general tool that can handle all graphs)

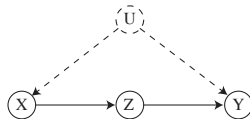
Example: Smoking, cancer, and the gene hypothesis

- Does smoking cause lung cancer?
 - Health authorities: "Large correlation between smoking and lung cancer!"
 - Tobacco-companies: "Some gene might both make people addicted, and cause cancer?"



X : smoking, Y : lung cancer, U : gene
Cannot determine $P(y | \hat{x})$ without controlled experiments

- What if we could measure a variable Z : the amount of tar deposits in the lungs, which has the following properties
 - smoking causes lung cancer only through tar deposits
 - the gene does not directly influence tar deposits, only through the effect of smoking
 - no other hidden variable affects both smoking and tar deposits, or directly both tar deposits and lung cancer
 - all possible combinations of (smoking, tar) can be seen in the data, even though some may be rare



- The collected data might look like this (purely fictional!):

	group	$P(x, z)$ size of group (% of population)	$P(Y = 1 x, z)$ % of cancer in each group
$X = 0, Z = 0$	no smoking, no tar	47,5	10
$X = 1, Z = 0$	smoking, no tar	2,5	90
$X = 0, Z = 1$	no smoking, tar	2,5	5
$X = 1, Z = 1$	smoking, tar	47,5	85

Using the formula given above, we get...

$$P(Y = 1 | \text{do}(X = 1)) = 0.4525$$

$$P(Y = 1 | \text{do}(X = 0)) = 0.4975$$

So, if the assumptions and the data held true, it would be beneficial to smoke! (In reality probably the other way around though... Note also the importance of the assumptions!)

do-calculus

- Preliminary notation: Let X and Y be distinct subsets of variables in a DAG G :
 - We denote by $G_{\overline{X}}$ the graph obtained from G when we remove all the arrows pointing into X
 - We denote by $G_{\underline{X}}$ the graph obtained from G when we remove all the arrows pointing out of X
 - We denote by $G_{\overline{X}\underline{Y}}$ the graph obtained from G when we remove all the arrows pointing into X and all the arrows pointing out of Y
- We next give **three rules** which are **necessary and sufficient** to calculate all identifiable effects in a causal model based on a DAG... (At the time of writing his book, Pearl had not proved that all identifiable models could be solved using these. This has recently been proved.)

Rules of do-calculus: (Pearl theorem 3.4.1)

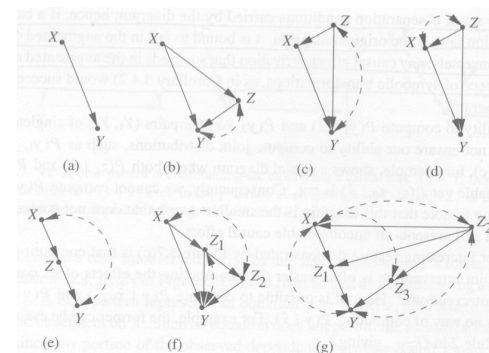
1. (insertion/deletion of observations)
 $P(y | \hat{x}, z, w) = P(y | \hat{x}, w)$ if $(Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}}}$
2. (action/observation change)
 $P(y | \hat{x}, \hat{z}, w) = P(y | \hat{x}, z, w)$ if $(Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}\underline{Z}}}$
3. (insertion/deletion of actions)
 $P(y | \hat{x}, \hat{z}, w) = P(y | \hat{x}, w)$ if $(Y \perp\!\!\!\perp Z | X, W)_{G_{\overline{X}\underline{Z}\underline{W}}}$

where $Z(W)$ is the set of Z -nodes that are not ancestors of any W -node in $G_{\overline{X}}$

do-calculus rules intuitively:

1. Inserting or removing an observation does not affect anything if the variables are conditionally independent (d-separation)
2. Back-door: If all back-door paths are blocked, then observations and actions are interchangeable
3. Making an intervention (action) which is 'after' the sought effect does not change anything

Examples of identifiable cases



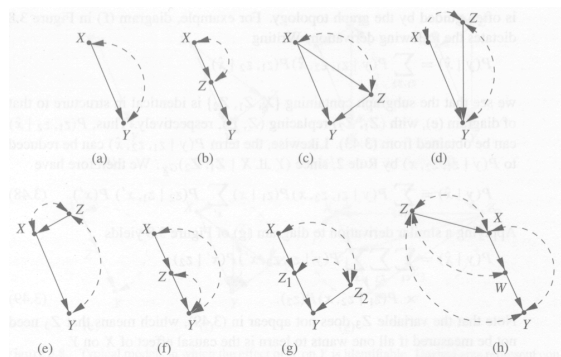
In all the above cases $P(y | \hat{x})$ is computable from the joint distributions over the observed variables. (The two-headed dashed edges can be replaced by a hidden common cause with arrows to both of the endpoints.)

- Note:

- Removing arrows can only help, so all edge subgraphs are also identifiable. Note also that adding observed variables can only help, never impede, identification.
- All the graphs in the figure are maximal such that adding any additional arrow (or a dashed arc) makes $P(y | \hat{x})$ no longer identifiable
- Although most of the diagrams contain bow patterns, none of these patterns emanates from X onto any ancestor of Y .
- (a) ja (b): no backdoor path from X to Y , so $P(y | \hat{x}) = P(y | x)$
- (c) ja (d): Z satisfies the back-door criterion

- In all cases (a-g) we can calculate $P(y | \hat{x})$ using do-calculus
- In cases (e-g) identifiability is rendered possible by variables which are causally affected by X . 'This stands contrary to the warning – repeated in most of the literature on statistical experimentation – to refrain from adjusting for concomitant observations that are affected by the treatment.'
- In cases (b), (c) and (f) Y has a parent whose effect on Y is not identifiable. This shows that local identifiability is not a necessary condition for global identifiability. In other words, to identify the effect of X on Y we need not insist on identifying each and every link along the paths from X to Y

Examples of non-identifying cases:



In all of the above graphs $P(y | \hat{x})$ is not uniquely computable from the joint distribution over the observed variables. (The two-headed dashed edges can be replaced by a hidden common cause with arrows to both of the endpoints.)

- Note...

- All cases (a-h) contain a back-door path which cannot be blocked by observed variables (this is a necessary but not a sufficient test for non-identifiability, cf. graph (e) in earlier figure)
- A sufficient condition for non-identifiability is the existence of a confounding path between X and any of its children on a path from X to Y , as in (b) and (c)
- Graph (g) shows that local identifiability is not sufficient for global identifiability: We can identify $P(z_1 | \hat{x})$, $P(z_2 | \hat{x})$, $P(y | \hat{z}_1)$, and $P(y | \hat{z}_2)$ but not $P(y | \hat{x})$

Required reading

- Chapter 3 (but not section 3.6) from Pearl's book:
<http://bayes.cs.ucla.edu/BOOK-2K/ch3-pref.pdf>
<http://bayes.cs.ucla.edu/BOOK-2K/ch3-1.pdf>
<http://bayes.cs.ucla.edu/BOOK-2K/ch3-2.pdf>
<http://bayes.cs.ucla.edu/BOOK-2K/ch3-3.pdf>
<http://bayes.cs.ucla.edu/BOOK-2K/ch3-4.pdf>
<http://bayes.cs.ucla.edu/BOOK-2K/ch3-5.pdf>