Introduction to Causal Discovery

Vanessa Didelez
with much help from Ronja Foraita and Christine W Bang
Leibniz Institute for Prevention Research & Epidemiology – BIPS

July 2022
Statistical Computing — Reisensburg

Overview of Course

Part 1: Introduction to Causal Concepts

Part 2: (Causal) Directed Acyclic Graphs, DAGs

Part 3: Causal Discovery — finding (poss.) causal structures

Note: The course does not cover methods for estimating causal effects (e.g. g-computation, propensity scores, IPTW, double-robust estimation etc.).

Presented material is a *subjective* selection of material based on what I like and know — though I try to cover a variety of topical material.

Aims of Course

- Introduce basic concepts of causal learning
- ... to enable you to read more advanced ‘causal’ papers
- Focus on:
  - basic causal notions, DAGs
  - some basic methods of causal discovery
  - understanding sources of (avoidable and unavoidable) bias
- Mix of mathematics & stories/examples

Motivating Example
(causal structure concealed)

Outcome Y,

three (possibly) explanatory variables X_1, X_2, X_3

$\Rightarrow$ regression analyses with flexible model, no misspecification

Best prediction: use **all three** variables, no subset is as good!

Causal interpretation of regression?

$Y \mid X_1 * X_2 * X_3 \Rightarrow$ no ‘causal’ coefficients

$Y \mid X_1 * X_2 \Rightarrow$ no ‘causal’ coefficients

...

$Y \mid X_2 * X_3 \Rightarrow$ conditional (X_2), direct (X_3) causal effect

$Y \mid X_2 \Rightarrow$ not ‘causal’

$Y \mid X_3 \Rightarrow$ total causal effect

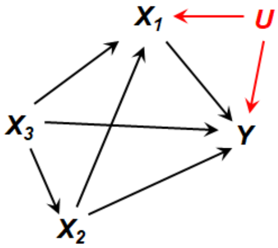
$\Rightarrow$ meaning of each analysis depends on causal structure

Motivating Example

(causal structure *revealed*)

What is going on?

Possible story:
Y = infant health,
X_1 = birth weight,
X_2 = maternal smoking (pregnancy),
X_3 = maternal education,
U = unknown genetic predisp.



$\Rightarrow$ statistical analysis must account for causal structure!

Audience?

Who thinks they know ...

- the difference between association and causation?
- the **formal (math.)** difference between association and causation?
- what confounding is?
- what a ‘counterfactual’ is?
- what a graphical model is?
- what a causal DAG is?

Causal Inference — History

- Causality / causal inference very broad topic!
- Has developed and evolved quite separately in different fields: philosophy, sociology, epidemiology, econometrics, computer science, (statistics), mathematics ...
- Different terminology, approaches, accepted assumptions, designs / types of data sources
- Last few (only!) years: some convergence has emerged across fields
- Causality very fundamental to many research questions in many fields / data science!

Part 1

Introduction to Causal Concepts

Preamble

- Causation / causality: philosophical, moral and other usages of the term — not what we are concerned with
- *Here:* particular (narrow) view of causality most relevant for scientific enquiries: **causality we can implement**
- "Causal effect" a difference in outcomes between (hypothetical) experiments we might do, i.e. effect of **(hypothetical) interventions**

Preamble

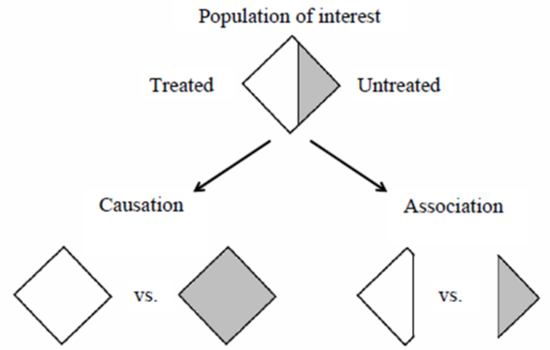
To obtain a causal answer, **start with a causal question!**

Describe the ideal (hypothetical) experiment with which you could investigate your research question (target trial)

Or: describe the decision problem you would like to solve.

Causation versus Association

(Hernan & Robins, 2020 book)

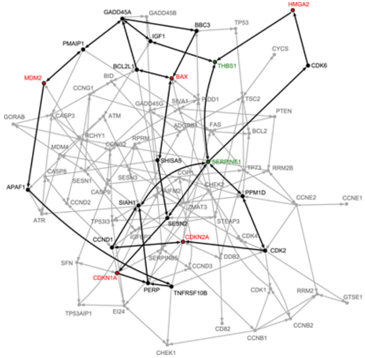


Causal effect: contrast of outcome if ‘everyone was treated’ versus if ‘no-one was treated’ (= intervention effect, “counterfactual”)

Motivation: Gene Regulation

Causal interpretation of gene networks: interventions have become increasingly feasible

- e.g. by knock–out / inhibition / activation
- ‘causal pathways’ similar to mechanistic description
- potential targets for drug development
- Example on HNSCC: causal role of HMGA2 gene on p53 signalling pathway?



(Foraita et al., 2020, JRSSA)

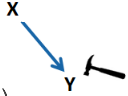
Causal Interpretation

... for gene regulation?

Gene expressions of X and Y are associated
– i.e. X is predictive of Y and Y is predictive of X

But: inhibition of X affects Y

while inhibition of Y does not affect X



Formally: distinguish ‘seeing’ and ‘doing’ (intervention)

Notation: $P(Y|\text{see}(X = x))$ versus $P(Y|\text{do}(X = x))$

Motivation ctd: Gene Regulation

Maathuis et al (2010)

Question: predict the effect of single-gene deletion from wild-type cultures?
— gene expression profiles of Saccharomyces cerevisiae

Observational data:
expressions of 5361 genes for 63 wild-type cultures

— Predict effect of interventions (234 deletions) on rem. genes
— Method: Intervention when the DAG is Absent (IDA)
— first find (all plausible) DAG(s) = causal discovery
then estimate possible effects

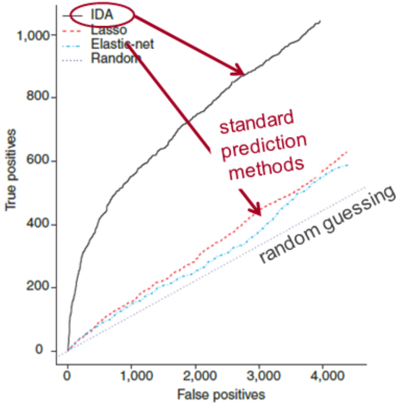
Interventional data (for validation): 234 single-gene deletion mutant strains of the same 5361 genes

Motivation ctd: Gene Regulation

Maathuis et al (2010)

true vs false positives
for top 5000 largest pred. eff.
from observational data

Compare top 10% of true
. . . with top 5000 predicted
effects



Many extensions of IDA since
(e.g. Witte et al., 2020, JMLR)

Causal Models

Here: all models **probabilistic**!

Causal model:
describes situation (distribution) under **(hypothetical) interventions** / manipulations / changes

... needs to be related to:
observational (no intervention / ‘natural’ / ‘idle’) situation (distribution) generating our data

$\Rightarrow$ **Identifiability**

Basic Causal Concepts

Formalisms to make interventions explicit:

do-notation / causal DAGs

Not enough time to cover:

Potential Responses / counterfactuals

Structural equations / structural causal models

do–Notation

(Pearl, 2000)

Notation to distinguish association and causation: 'do' and 'see'

$$p(y \,|\, \text{intervene to set } X = x) = p(y \,|\, \text{do}(X = x))$$

and

$$p(y \,|\, \text{observe } X = x) = p(y \,|\, \text{see}(X = x))$$

$\Rightarrow$ **do–calculus** / **axioms** / directed acyclic graphs (DAGs).

Usually $p(y \,|\, \text{see}(X = x)) = p(y \,|\, x)$

Maths: P_{do} diff. probability measure, some common aspects with $P = P_{\text{see}}$

do–Intervention

$p(y \mid \text{do}(X = x))$ denotes point-intervention on wider system.

Consider: Y, X, C_1, C_2 such that *observationally ('see')*:

$$p(y, x, c_1, c_2) = p(y \mid x, c_1, c_2)p(x \mid c_1, c_2)p(c_2 \mid c_1)p(c_1)$$

May have reasons to believe that under intervention:

$$p(y, c_1, c_2 \mid \text{do}(X = \tilde{x})) = p(y|\tilde{x}, c_1, c_2)p(c_2|c_1)p(c_1)$$

Note: can be obtained by reweighting

DAGs help to structure the *factorisation*

Identifiability

Under suitable **structural assumptions** we have for certain sets C of covariates:

$$p(y \mid \mathsf{do}(X = x)) = \sum_c p(y \mid x, c) p(c)$$

left: interventional distribution; right: observational distrib.

$\Rightarrow P_{\text{do}}$ **non-parametrically identified**, i.e. not using parametric assumptions like linearity, Gaussianity etc.

Causal Effects

Total Causal Effect

Note: no such thing as *'the'* causal effect
— always need to choose what to contrast with what and how

Causal effects:
typically formulated as contrasts of some aspect of

$$p(y \mid \text{do}(X = x)) \quad \text{versus} \quad p(y \mid \text{do}(X = x'))$$

or of $p(Y(x))$ versus $p(Y(x'))$ poss. conditional on further variables

For instance: **Average Causal Effect** (total / pop. causal effect)

$$ACE = E(Y \mid \text{do}(X = 1)) - E(Y \mid \text{do}(X = 0))$$

Cause and Effect

Can now define:

X is a **cause** of Y and Y is an effect of X if for some $x \neq x'$

$$p(y \mid \text{do}(X = x)) \neq p(y \mid \text{do}(X = x'))$$

or $p(Y(x)) \neq p(Y(x'))$

i.e. if (hypothetically) intervening in X setting it to different values changes some aspect of the distribution of Y

Note: this corresponds to how we check causation in a basic randomised experiment

Other Causal Effects

Conditional Causal Effect

... or subgroup causal effect

Let $S = s$ characterise subset of population, e.g. age group

Conditional causal effect of X on Y given $S = s$:

$$E(Y|S = s; \text{do}(X = 1)) - E(Y|S = s; \text{do}(X = 0))$$

Note: S must not itself be causally affected by X, i.e. not be post-treatment

Other Causal Effects

Joint Causal Effect

Consider two (possibly sequential) exposures X_1, X_2.

The joint (total) causal effect of X_1 and X_2 on Y is

$$E(Y|\text{do}(X_1 = x_1, X_2 = x_2)) - E(Y|\text{do}(X_1 = x_1', X_2 = x_2'))$$

Note: potential issue here: ‘time-dependent’ confounding

Other Causal Effects

Controlled Direct Effect

Consider again two sequential exposures X_1, X_2

Controlled direct effect of X_1 while controlling X_2 means:
hold fixed do$(X_2 = 0)$ and contrast different values for X_1, e.g.

$$CDE = E(Y|\text{do}(X_1 = x, X_2 = 0)) - E(Y|\text{do}(X_1 = x', X_2 = 0))$$

Note:
'direct' presupposes that X_2 is possibly mediator for effect of X_1 on Y

Further Causal Effects

"Individual Causal Effect": requires counterfactual concepts

"Effect of treatment on the treated (ETT)"

Other interventions:
– dynamic interventions / adaptive treatments: adapt dosage to previous observations
– shift / random interventions: add a constant or noise to the 'treatment'

"Principal Stratum Causal Effects": requires counterfactual concepts

Potential Outcomes / “Counterfactuals”

Potential outcomes:

Y^x outcome (rand. var.) under an intervention setting $X = x$

Cannot observe both (Y^1, Y^0) — when X is realised, one PO becomes counterfactual

Individual causal effect for unit j: $ICE = Y_j^1 - Y_j^0$

Identifiability of certain causal quantities require assumptions about joint distribution $P(Y^1, Y^0)$ — will not consider these in this course.

Terminlogy: ‘counterfactual’ often (inappropriately) used for quantities that are just ‘hypothetical’ or conditional on many covariates

Part 2

(Causal) Directed Acyclic Graphs — DAGs

Two Uses of Graphs

Association $\neq$ causation but:

lack of (conditional) association "$\Rightarrow$" no causation

Graphs
(1) to represent conditional independencies
(2) can be supplemented with causal semantics

Basic Concepts

Conditional Independence

$P(Y = y)$, $p(y)$ etc. probability / density / prob.mass function

Conditional independence:

X and Y are conditionally independent given Z,
write $Y \perp\!\!\!\perp X \mid Z$, if

$$P(Y = y, X = x \mid Z = z) = P(Y = y \mid Z = z)P(X = x \mid Z = z)$$

for all x, y, z s.t. $p(z) > 0$. Or, equivalently if:

$$P(Y = y \mid X = x, Z = z) = P(Y = y \mid Z = z)$$

or $p(y|x, z) = p(y|z)$ — relate this to regression models!

Basic Concepts

Conditional Independence

In words: if we already know (observed) the value of Z then *additionally* knowing the value X is not informative with respect to the distribution (prediction) of Y

Example:

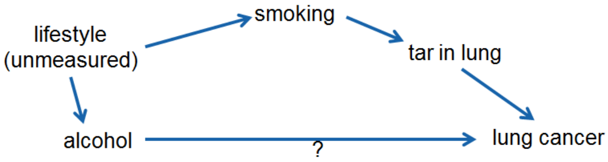
- while knowing (only) that some-one has tar-stained fingers is informative to predict if they will develop lung-cancer...
- ... once we also know that they are a smoker, the information on their tar-stained fingers becomes irrelevant

$$\text{lung-cancer} \perp\!\!\!\perp \text{tar-fingers} \mid \text{smoking-status}$$

Note: association (or lack of) is symmetric

Graphs — Terminology

Graph $G = (V, E)$
V = vertices & nodes = variables / features
E = edges = possible (causal) dependence
Non-edge = (conditional) independence



Note: nodes shown as ‘events’ represent binary indicator variables, e.g. ‘lung cancer’ $\in \{0, 1\}$ for ‘no’ / ‘yes’.

Nodes / Variables

The typical / traditional approach assumes one **already** has access to **variables which represent high-level semantic concepts**

This may not be the case when learning from raw video or imaging data, for example

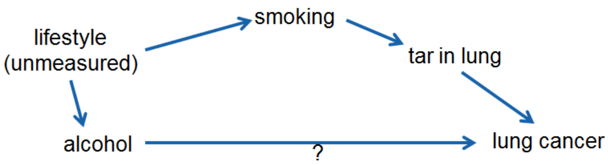
$\Rightarrow$ Formulating causal DAG for such situations: active research!

Graphs — Terminology

Graphical terms:

‘parents’, ‘children’, ‘ancestors’, ‘(non-)descendants’ etc.

‘(directed) paths’, ‘(directed) cycles’



Factorisation for DAGs

Factorisation: a distribution P (with pdf/pmf p) factorises according to a DAG G and is called **G–Markov** iff

$$p(\mathbf{x}) = \prod_{i=1}^{K} p(x_i | \mathbf{x}_{\mathsf{pa}(i)})$$

Note: the above factorisation is **equivalent** to

$$X_i \perp\!\!\!\perp \mathbf{X}_{\mathsf{nd}(i)\setminus\mathsf{pa}(i)} \mid \mathbf{X}_{\mathsf{pa}(i)} \text{ for every } i \in V$$

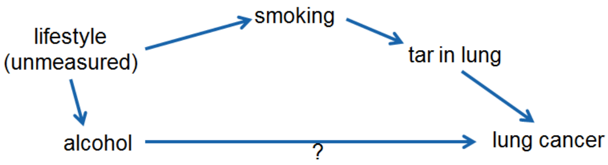
Rule: read off all implied cond. independencies using **d-separation**

$\Rightarrow$ testable implications of DAG models

Graphs — Markov Property

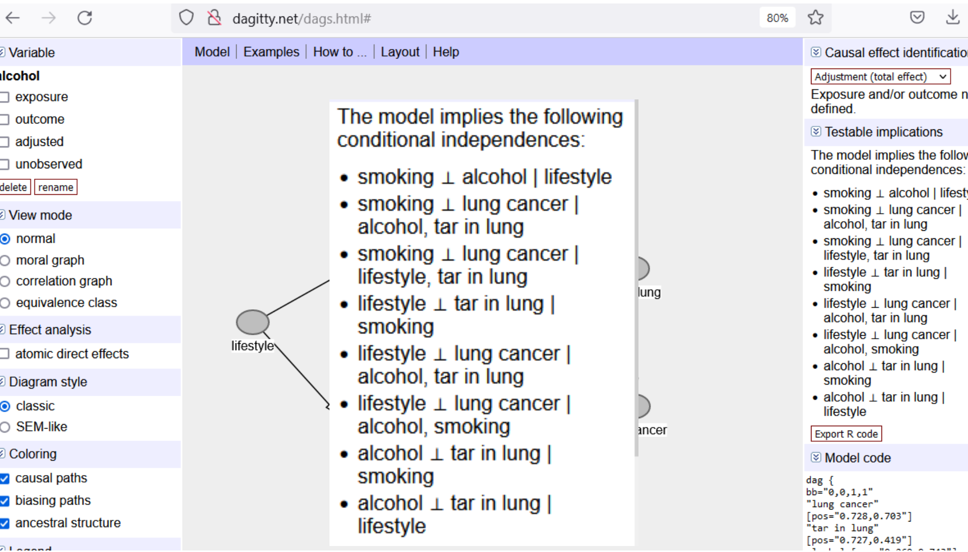
with `DAGitty`

Observationally:
Absence of edges into outcome: if we know whether there is tar in the lungs and whether person drinks alcohol, then smoking status or any further information on lifestyle are non-informative for the probability of lungcancer.



Check with `DAGitty`, software for querying DAGs (Textor et al, 2016)!

DAGitty Screen Shot

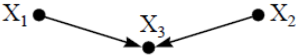


Selection Effect

“collider bias”

Important for the interpretation:

Conditioning on common child (selection) $\Rightarrow$ dependence



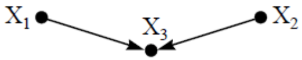
here: $X_1 \perp\!\!\!\perp X_2$ but $X_1 \not\perp\!\!\!\perp X_2 \mid X_3$

$$p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_1, x_2)$$

does not generally imply $X_1 \perp\!\!\!\perp X_2 \mid X_3$

Selection Effect

"collider bias"



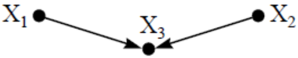
Example: some school admission process is such that pupils are admitted (X_3) if they are either good at maths (X_1) or good at sports (X_2).
Assume in population X_1 and X_2 are independent(!)

If we randomly draw a pupil from this school, $X_3 = 1$, and find this pupil is no good at sports, $X_2 = 0$, then we know s/he must be good at maths, $X_1 = 1$!

In other words, given X_3, X_2 becomes informative for X_1.

Separation in DAGs

Motivated by selection effect: want general rule to describe "separation"



Here: $\emptyset$ separates X_1 and X_2

but X_3 does not separate X_1 and X_2.

d–Separation in DAGs

(Pearl, 1988)

Given DAG $G = (V, E)$. A path between a and $b \in V$ is blocked by $S \subset V \backslash \{a, b\}$ if

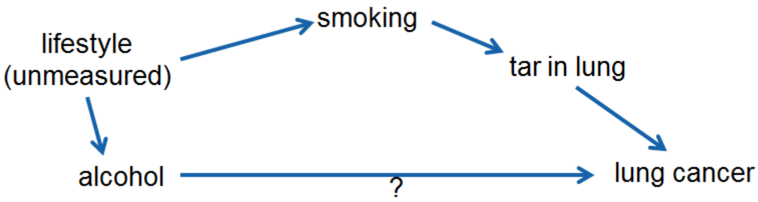
(i) it contains a non-collider $\longleftarrow z \longrightarrow$ or $\longleftarrow z \longleftarrow$ and $z \in S$ or

(ii) it contains a collider $\longrightarrow z \longleftarrow$ and neither z nor any descendants of z are elements of S

A and $B \subset V$ are **d–separated** by $S \subset V \backslash (A \cup B)$ if every path between A and B is blocked by S.

Theorem: Factorisation $\Leftrightarrow$ every d-sep. implies a cond.indep.

d–Separation — Quiz



How many paths between ‘smoking’ and ‘alcohol’ are blocked (by the empty set)?

How many paths between ‘lifestyle’ and ‘lung cancer’ are blocked by ‘smoking’?

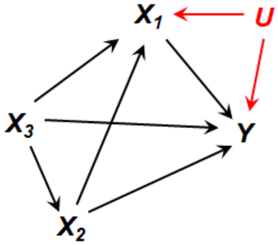
d–Separation

'Collider-Stratification Bias'

Earlier example

Here, d-separation shows $Y \not\perp\!\!\!\perp (X_2, X_3) \,|\, X_1$

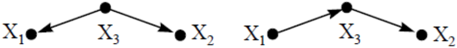
Possible story:
Y = infant health,
X_1 = birth weight,
X_2 = maternal smoking (pregnancy),
X_3 = maternal education,
U = unknown genetic predisp.



$\Rightarrow$ Careful with regression as model for $P(Y|X_1, X_2, X_3)$

Markov Equivalence of DAGs

Marginalizing w.r.t. common parent (confounder) or intermediate variables $\Rightarrow$ dependence



Here: $X_1 \perp\!\!\!\perp X_2 \mid X_3$, but $X_1 \not\perp\!\!\!\perp X_2$

Markov equivalence:
different DAGs imply same conditional independencies!

Implication:
cannot distinguish between equivalent DAGs from obs. data

(Causal) Graphs

aka: (causal) DAGs / diagrams / Bayesian networks

A causal graph is a (probabilistic) model for a set of random variables imposing

- restrictions on conditional independencies within the observational distribution

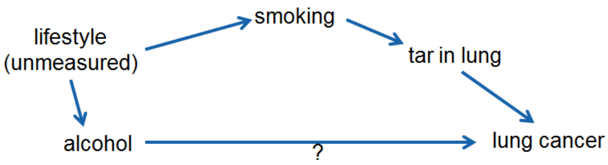
  and

- restrictions on conditional independencies within the distribution under hypothetical interventions

- ‘non-parametric’: graph contains no information on the functional shape of relations between variables (nor on strength / size of dependencies)

Causal Markov Condition

Causally:
An edge represents a possible ‘controlled direct effect’
e.g., if we fix ‘tar’ and vary ‘alcohol’ by an intervention then this will possibly change the probability for ‘lung cancer’



Notes:
– ‘direct effect’ relative to nodes included
– better: **absence** of edge guarantees **no direct effect**

Causal Markov Condition

Axiom (Causal Markov Condition):

if neither X *direct* cause of Y nor vice versa

$$\Rightarrow \text{ there exists a set } S \text{ s.t. } X \perp\!\!\!\perp Y \,|\, S$$

('direct' relative to other nodes)

('direct' while 'controlling' other parents)

Graphical: every variable is cond. independent of its non-effects (descendants) given its direct causes (parents).

Causal DAG

So what makes a DAG into a **causal DAG**?

Additional **semantics** relating DAG to interventions:

- effects of interventions follow direction of edges, i.e. can affect all descendants, but cannot affect non-descendants

  $\Rightarrow$ `DAGitty` depicts 'causal paths' and 'non-causal' paths inducing associations

- intervention distribution corresponds to DAG-model after removing edges into the intervened node.

Causal DAG

(for the mathematically interested)

Definition:
DAG G, distribution P is G–Markov. Then, G causal wrt $B \subset V$ if for any $A \subset B$

$$p(\mathbf{x}_V \mid \text{do}(A = a))) = \prod_{i \in V \setminus A} p(x_i \mid \mathbf{x}_{\text{pa}(i)}) \Big|_{\mathbf{x}_A = a}$$

in words:

- P describes 'behaviour' under observation, factorises
- under intervention, do$(A = a)$, the variables in $\mathbf{X}_A$ are simply fixed to a when appearing in $\mathbf{X}_{\text{pa}(i)}$
- and all conditional specifications on $V \setminus A$ remain the same ('invariance')

Example 1

$$X \longrightarrow Y$$

This causal DAG expresses:

- an intervention on X can affect Y
- an intervention on Y *cannot* affect X

Note: The DAG expresses no (cond.) independencies.

Example 1 ctd.

$$do(X{=}x) \longrightarrow Y$$

Moreover:

- an intervention on X removes arrows into X (here: none)
- the intervention distribution is identical to the (observational) conditional distribution

$$p(y \,|\, \text{do}(X = x)) = p(y \,|\, x)$$

Note: the latter reflects that the DAG expresses the assumption of no common causes for X and Y.

This would be plausible if X was known to be randomised.

Example 1 ctd.

X $\qquad\qquad$ $do(Y=y)$

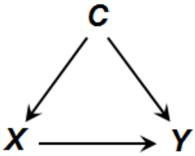
Finally:

- an intervention on Y removes arrows into Y
- the intervention distribution is identical to the (observational) marginal distribution

$$p(x \,|\, \text{do}(Y = y)) = p(x)$$

- i.e. X is ‘independent’ of (the value of) Y under an intervention on Y.

Example 2



This causal DAG expresses:

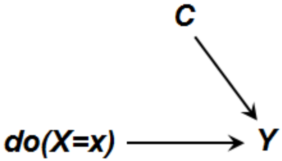
- an intervention on X can affect Y, *but not* C
- an intervention on C can affect X and Y
- an intervention on Y *cannot* affect X nor C.

Note: The DAG expresses no (cond.) independencies.

Example 2 ctd.

Moreover:

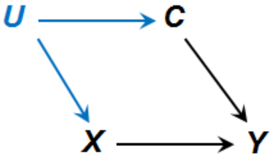
- an intervention on X removes arrows into X
- the intervention distribution is identical to the (observational) conditional distribution $p(y, c \,|\, \text{do}(X = x)) = p(y \,|\, c, x)p(c)$ and hence (standardisation again!)

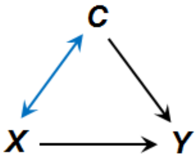


$$p(y \,|\, \text{do}(X = x)) = \sum_c p(y \,|\, c, x)p(c)$$

Note: because of the assumption of a common cause C, the formula for $p(y \,|\, \text{do}(X = x))$ is now different than in Example 1.

Example 3





Assume U unobserved (often represented by bi-directed edge)

- X and C are not independent (due to common cause U)
- but intervention on X does not affect C and intervention on C does not affect X
- otherwise, regarding X, C, Y same as Example 2.

Example 4

$$X \longrightarrow Z \longrightarrow Y$$

This causal DAG expresses:

- an intervention on X can affect Z and Y
- an intervention on Z can affect Y, but not X
- an intervention on Y cannot affect X nor Z

Example 4 ctd.

$$X \qquad do(Z{=}z) \longrightarrow Y$$

Moreover:

- an intervention on Z prevents and intervention on X having any effect on Y

$\Rightarrow$ relative to the considered set of variables:
Z is a direct cause of Y, X is an indirect cause of Y

$\Rightarrow$ the direct effect of X on Y controlling for Z is null

Standardisation

Remember: identifying functional for the interventional distribution of Y under intervention on X

$$p(y \mid \text{do}(X = x)) = \sum_c p(y \mid x, c) p(c)$$

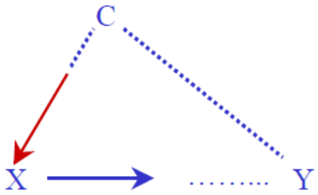
Requires assumption 'no-unmeasured confounding given C'.

Graphical formulation:
C must 'block all back-door paths' from X to Y...

Back-Door Path

Definition

A back-door path from X to Y starts with an edge $X \leftarrow \cdots Y$.



Back-Door Criterion

(Pearl, 1995)

Theorem

Given a DAG G on V, causal wrt. $X \in V$. Then $C \subset V \backslash \{X, Y\}$ identifies causal effect of X on Y if

(i) C contains no descendant of X and

(ii) all 'back-door' paths from X to Y are blocked by C

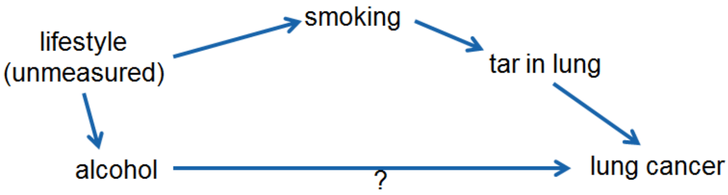
C is then *sufficient* adjustment set.

Note: C not unique; *minimal* C not unique.

Back-door Criterion — Exercise

with `DAGitty`

Note: lifestyle is *the* confounder (common cause), but unobserved!

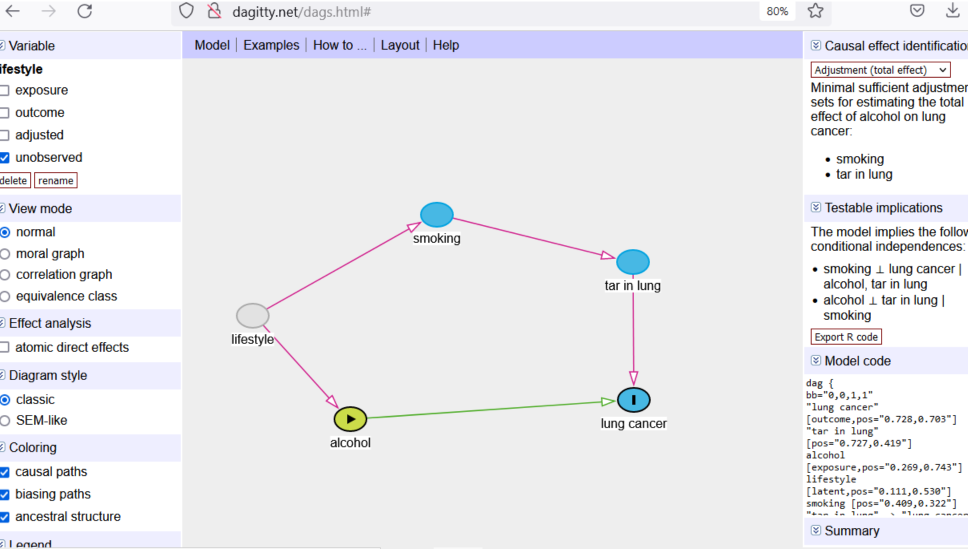


Sufficient set of covariates to identify the effect of X on Y?

$X =$ alcohol consumption, $Y =$ lung cancer

DAGitty Screen Shot

Markov properties



Causal DAGs

Summary

- Graphs are helpful to organise your causal reasoning / structuring of a given causal question with data at hand
- Confounding: which covariates do we have to take into account? $\Rightarrow$ Back–door criterion
- Selection- / collider-bias: which covariates should we not condition on?

$\Rightarrow$ **Recommended:** always *draw* your assumptions before your conclusions! (Hernán)

Further Topics

- **Sofware:** `DAGitty` — R package or online.
  Carries out queries on DAGs, e.g. find all minimal sufficient adjustment sets.
- Other identification criteria exist: e.g. Front-door criterion.
  Complete identification algorithm due to Shpitser (2006) available in software `ananke` (Python)
- Causal DAGs also used for:
  – decide transportability of inference across populations
  – identifiability with missing values
  – expert systems etc.

Further Topics

Appendix

- Workflow of causal analysis?
- Further examples for adjustment
- Single world intervention graphs (SWIGs)
  link between potential responses and graphs
- Alternative (niche): influence diagrams
- Structural equation models $\longrightarrow$ impose most structure
- Other interventions: nudging / shifting / stochastic
  interventions — active research

Part 3

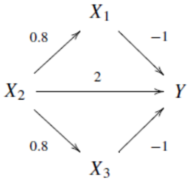
Causal Discovery

Recap: Causation versus Prediction

(Maathuis et al, 2009, 2010)

$Y \sim X_1 + X_2 + X_3$

Causal structure can for instance be chosen such that:



Example 1:

Regression coefficients: $\beta_1 = \beta_3 = -1$, $\beta_2 = 2$

Total causal effects: $\theta_1 = \theta_3 = -1$ but $\theta_2 = 0.4$

$\Rightarrow X_2$ causally least important.

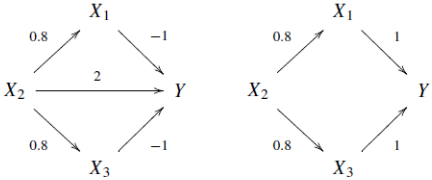
(Here linear structural equation models, LSEM)

Recap: Causation versus Prediction

(Maathuis et al, 2009, 2010)

$Y \sim X_1 + X_2 + X_3$

Causal structure can for instance be chosen such that:



Example 2:

Regression coefficients: $\beta_1 = \beta_3 = 1$, $\beta_2 = 0$

Total causal effects: $\theta_1 = \theta_3 = 1$ but $\theta_2 = 1.6$

$\Rightarrow X_2$ causally most important.

(Here linear structural equation models, LSEM)

Causal Discovery

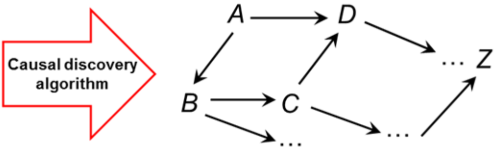
aka: causal search, (causal) structure learning, (causal) graph estimation, network inference ...

Causal Discovery

Input: data

A	B	C		Z
0.3	12	0	…	140
0.2	13	0		287
0.7	21	1		876
0.6	10	0		326
…	…	…		…

Output: causal DAG



Only with quite **strong assumptions**

$\Rightarrow$ carefully evaluate plausibility

Causal Discovery

Caveats

DAGs for 10 variables $> 4 \times 10^{18}$

Number of DAGs superexponential in number of nodes

$\Rightarrow$ cannot evaluate all possible DAGs!

There is no free lunch! — all methods rely on strong assumptions

More modest: interpret graph in terms of conditional (in)dependencies / associations; generate some causal hypotheses; absence of edge still absence of (direct) causation

$\Rightarrow$ consider causal discovery as **exploratory** data analysis

Types of Algorithms

(1) Constraint-based

- find (conditional) independencies (= constraints) in data
- construct graph to satisfy these constraints

(2) Score-based

- define a score for fit between data and causal graph (often: likelihood-based)
- optimise the score over space of graphs
- includes Bayesian approaches

Types of Algorithms (2)

(3) Exploiting structural asymmetries

- various ‘modelling’ assumption render $X \longrightarrow Y$ observationally different from $X \longleftarrow Y$

(4) Reformulation as continuous optimisation problems

- with smooth acyclicity constraints
- combine with black-box machine learning approaches
- *I would say: still work in progress...*

Constraint-Based Causal Learning

some principles

Causal Markov Condition: causal DAG implies conditional (in)dependencies

Let's turn this around and find conditional (in)dependencies from data, then construct DAG that implies these

Note: will need more assumption!

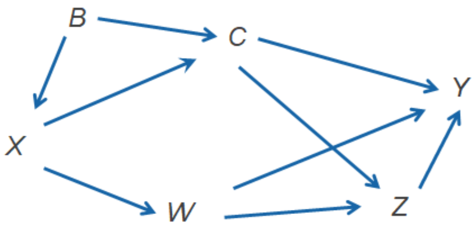
Authors: Spirtes, Glymour, Scheines (book, 1993, 2000) and much work since

Separation and Independence

Theorem: if X and Y are d-separated by S (i.e. every path between X and Y is blocked by S), then X and Y are conditionally independent given S.

Write $X \perp\!\!\!\perp Y|S$,
i.e. $p(x, y|s) = p(x|s)p(y|s$

Example:
$W \perp\!\!\!\perp C \mid B$??
$W \perp\!\!\!\perp C \mid (X, Z)$??



From Association to Causation

Consider: in large data set X and Y are **not associated**.

$\Rightarrow$ seems safe to assume that there is no causal relation.

But careful: could be that for $Z = 1$, X has positive effect on Y, and for $Z = 0$, X has negative effect on Y, so that the effects cancel each other out — unlikely but possible.

Faithfulness assumption: every (conditional) independence in the population ($\approx$ large data set) corresponds to a missing edge in the underlying causal DAG.

From Association to Causation

Consider: in large data set we find X and Y are **associated** (e.g. with standard test for correlation or χ^2–test).

Problem: many compatible causal structures
– X causes Y, or Y causes X or
– they are confounded or
– there is a selection effect or
– coincidence (less likely the larger the data set)

$\Rightarrow$ include more variables, e.g. to rule out confounding; include temporal information if possible.

Often: assume **causal sufficiency**, i.e. all common causes have been observed $\Rightarrow$ no unobserved confounding.

From Association to Causation

Consider: in large data set we find $X \perp\!\!\!\perp Y|Z$, i.e. X and Y are independent conditionally on Z, but no other independencies.

Problem: again, more than one compatible causal structure
– effect of X on Y is mediated by Z
– effect of Y on X is mediated by Z
– Z is a common cause of X and Y

$$X \longrightarrow Z \longrightarrow Y \qquad X \longleftarrow Z \longleftarrow Y \qquad X \longleftarrow Z \longrightarrow Y$$

These DAGs are **Markov equivalent** because they correspond to the same conditional independencies.

$\Rightarrow$ from observational data can only learn equivalence classes of DAGs — CPDAGs (completed partially directed DAGs).

From Association to Causation

Consider: in large data set we find $X \perp\!\!\!\perp Y$ but $X \not\perp\!\!\!\perp Y | Z$ and no other independencies.

Assuming causal sufficiency and faithfulness, there is only one causal structure compatible with this finding:

Z is a common effect of X and Y

$$X \longrightarrow Z \longleftarrow Y$$

(called **V–structure**)

$\Rightarrow$ will see that these are the most revealing structures.

Equivalence Class: CPDAG

Equivalent DAGs: iff same skeleton and same V-structures.

CPDAG (completed partially directed acyclic graph):

– mixed (types of edges) graphs

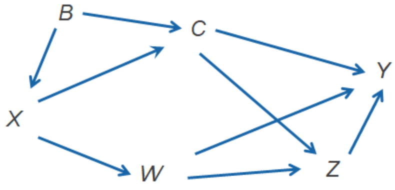
– some directed and some undirected edges

– undirected means: in class, both directions exist

– DAGs in class found by orienting undirected edges without creating cycles / V-structures

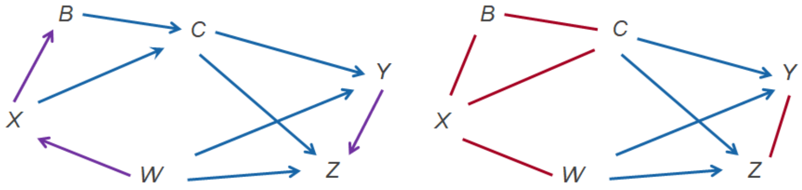
CPDAG Example

CPDAGs are mixed graphs with...
undirected edges if either direction occurs at least once in the equivalence class



CPDAG Example

CPDAGs are mixed graphs with...
undirected edges if either direction occurs at least once in the equivalence class



Attention:
software often outputs undirected edges as bi-directed edges!!

PC Algorithm

(Spirtes, Glymour & Scheines, 1993 & 2000)

Now: general procedure to construct DAG from conditional (in)dependencies on set of variables.

PC Algorithm basic procedure

1) Find undirected graph showing where edges should (not) be

2) Identify V–structures

3) Orient remaining edges (logical) if possible

Note: this is the **simplest** constraint-based discovery algorithm;
Assumes: causal sufficiency and faithfulness.

Software: `TEDRAD` Project (stand-alone) and *numerous* others!

PC Algorithm — First Step

Note: if A and B are not connected by an edge in a DAG then there exists some set S (possibly empty) such that $A \perp\!\!\!\perp B|S$.

$\Rightarrow$ check this for each pair of nodes, starting with *small* separating sets first and then moving to larger ones, i.e. check all S with $|S| = \emptyset$, then with $|S| = 1$ etc.

$\Rightarrow$ keep undirected edges A—B if they are not conditionally independent for any S.

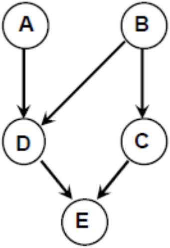
PC Algorithm — First Step

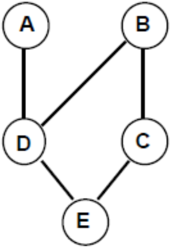
V = set of nodes, and each node A has a set of adjacent nodes adj_A.

1. Start with complete undirected graph G on V.
2. $i = 0$ (size of separating set)
3. Repeat
    4. For each $A \in V$
        5. For each $B \in adj_A$
            6. check if there is $S \subset adj_A \backslash B$ with $|S| = i$ and $A \perp\!\!\!\perp B | S$
            7. if yes then
                8. store $sep_{AB} = S$
                9. remove A—B edge from $\mathcal{G}$
    10. $i = i + 1$
11. Until $|adj_A| < i$ for all nodes A

PC Algorithm — First Step

Example: oracle (left) first step terminates with undirected graph (right) — no further conditional independencies to be found





Have to remember separating sets: $sep_{AB} = sep_{AC} = \emptyset$, $sep_{CD} = \{B\}$, and $sep_{AE} = sep_{BE} = \{C, D\}$.

PC Algorithm — Second Step

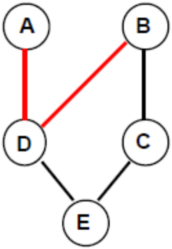
Identify V–structures

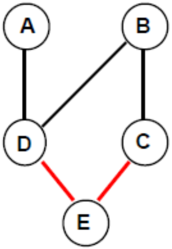
Procedure

1. For each constellation A—C—B (no edge linking A and B!)
    2. if $C \notin sep_{AB}$
        3. orient edges as $A \longrightarrow C \longleftarrow B$.

PC Algorithm — Second Step

We find that $D \notin sep_{AB} = \emptyset$ and that $E \notin sep_{CD} = \{B\}$, so can orient the corresponding edges such that D and E are colliders.



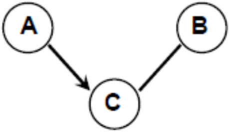


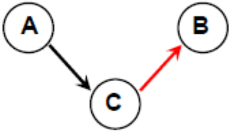
PC Algorithm — Third Step

Meek's Rules

Orient remaining edges such that
– cycles are avoided
– no new V–structures are created.

Examples: constellations that can be oriented



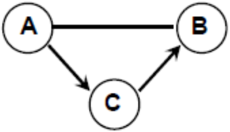


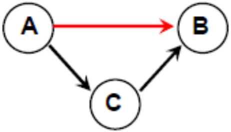
PC Algorithm — Third Step

Meek's Rules

Orient remaining edges such that
– cycles are avoided
– no new V–structures are created.

Examples: constellations that can be oriented



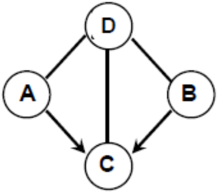


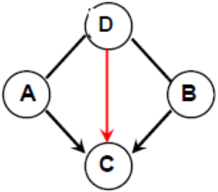
PC Algorithm — Third Step

Meek's Rules

Orient remaining edges such that
– cycles are avoided
– no new V–structures are created.

Examples: constellations that can be oriented





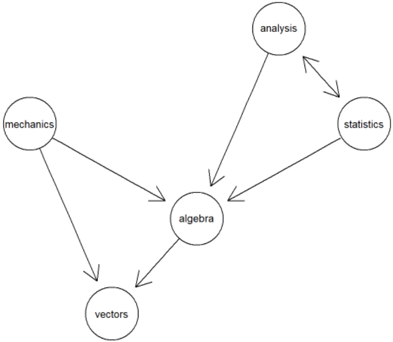
PC Algorithm — Finally

In original example: cannot orient B–C edge as both graphs are Markov equivalent.

PC algorithm outputs *CPDAG* representing a Markov equivalence class of DAGs.

PC Algorithm with `pcalg`

```
pc(suffStat = list(C = cor(mathmark),
  n = dim(mathmark)[1]),
  indepTest = gaussCItest,
  alpha = 0.05)
```



PC Algorithm — Properties

- It is relatively fast!
- If the underlying structure is indeed a causal DAG (& under causal sufficiency and faithfulness) and there are no errors in assessing the conditional independencies, then this algorithm is *exact*
- Can be adapted to case where some prior knowledge is available, e.g. time ordering / presence or absence of edges (`tPC`, Witte et al, 2021)

PC Algorithm — Properties

No distributional / parametric assumption as such

But in practice: need to choose a statistical tests for conditional independence — typically implies a distribution

- Popular (for continuous variables): Fisher's z-Test based on partial correlations (implicit: linearity / Gaussianity)
- All variables discrete: G^2 or similar — non-parametric (beware: low cell-frequencies)
- Wanted: non-paramteric but also high power!
  Sample size too small $\Rightarrow$ quite empty graph...

PC Algorithm — Properties

- A general non-parametric level-α statistical test cannot exist (Peters & Shah, 2020)
  But nearly non-parametric:

  — permutation-based kernel conditional independence test (Doran et al, 2014)

  — generalised covariance measure (Peters & Shah, 2020)

  — some more...
- In R package `pcalg`, can implement your own test or decision rule

PC Algorithm — Problems

In practice: statistical tests for conditional independence make type I & II errors!

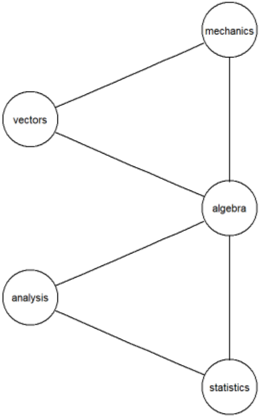
$\rightarrow$ outputs can be very unstable
$\rightarrow$ outputs may not be valid CPDAGs
$\Rightarrow$ should bootstrap results to assess variability of graph!

$\rightarrow$ outputs may depend on order of input variables

... avoided in `pcalg` by
— 'stable' skeleton search (default)
— 'solve.confl' leaves conflicting edges un-oriented

PC Algorithm with `pcalg`

```
pc(suffStat = list(C = cor(mathmark),
  n = dim(mathmark)[1]),
  indepTest = gaussCItest,
  alpha = 0.05,
  maj.rule = TRUE,
  solve.confl = TRUE,
  u2pd = "relaxed")
```



PC Algorithm — Problems

In practice: also need to choose nominal α-level for tests

- number of tests unknown, so no simple multiplicity correction
- regard α as tuning parameter
  e.g. choose so as to obtain desired number of edges
- inclusion of edge depends on power of test

PC Algorithm — High Dim

(Kalisch & Bühlman, 2007)

- PC algorithm has been adapted to gene network applications, especially when the sample size is smaller than the number of nodes and when graphs are sparse
- Uniform consistency for very high-dimensional, sparse DAGs
- Consistency carries over to Gaussian copula or nonparanormal models (Harris & Drton, 2013)

PC Algorithm for Cohort Data

(Witte, Foraita, Didelez et al., 2022)

- Adapted to take time structure into account
- Can now be combined with multiple imputation of missing data
- Automatic selection of tests for mixed variable scales
- Bootstrapping of output to assess uncertainty in graph selection
- Packages `micd` and `tpc`

FCI Algorithm

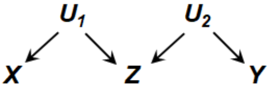
Relaxing Causal Sufficiency

FCI = 'fast causal inference' — but algorithm actually quite slow

Allowing latent (unmeasured) variables: much more complicated equivalence class!

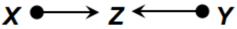
$\longrightarrow$ partial ancestral graph (PAG)

True DAG: latent U_1, U_2



PC algorithm: wrong output

FCI algorithm: correct PAG



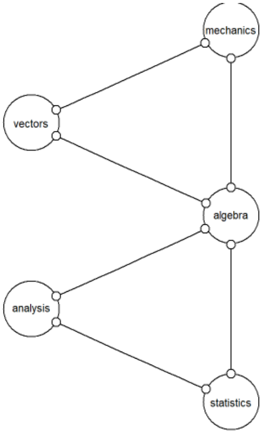
Interpretation: Edges in PAGs

PAG: some X-Y edge iff conditionally dependent given set S for all subsets S of the observed variables

X cause of Y (ancestor)	$X \longrightarrow Y$
Y does not cause X nor vice versa, there may be a latent common cause	$X \longleftrightarrow Y$
Y does not cause X	$X \bullet\!\!\longrightarrow Y$
any of the above (and $X \leftarrow Y$) occur in equiv.class	$X \bullet\!\!-\!\!\!-\!\!\bullet Y$

FCI Algorithm with `pcalg`

```
fci(suffStat = list(C = cor(mathmark),
  n = dim(mathmark)[1]),
  indepTest = gaussCItest,
  alpha = 0.05,
  labels = colnames(mathmark),
  maj.rule = TRUE,
  selectionBias = FALSE)
```



Note: non-edges!

Score-based Search

Score: define a measure $\mathcal{S}(G)$ for fit b/w a (CP)DAG and data
— typically: (penalized) log-likelihood, e.g. BIC
— penalising for complexity of graph

$\Rightarrow$ Goal:

$$\hat{G} = \text{argmax}_{G \in \mathcal{G}} \mathcal{S}(G)$$

$\mathcal{G}$ space of DAGs or better of CPDAGs

Need: some heuristic to search through space of graphs

Note: Bayesian approaches (with priors on graphs) are special case of score-based search.

Score-based Search

Greedy Equivalence Search (GES)

Score: should be (Chickering, 2002)

— score equivalent, i.e. same for Markov-equivalent graphs

— decomposable (every {node+parents} separately)

— consistent

Search: greedy grow-shrink algorithm with forward (adding edge) and backward phase (deleting edge)

GES guarantee: selection-consistent if:

— score equivalent, decomposable and consistent

— e.g. BIC for multiv. Gaussian / multinomial distributions

Compare: PC/FCI vs GES

Non-parametric?

- PC/FCI can be used with any desired conditional independence test, no (other) distributional assumption
- GES requires $\approx$ likelihood, so (fully) specified distribution

Output?

- PC/FCI output not always valid CPDAG / PAG (for finite samples)
- GES always outputs CPDAG

With/out causal sufficiency?

- GES near infeasible without causal sufficiency (i.e. with latent nodes)
  – equivalence class of PAGs very complicated
  – likelihood-based scores not decomposable

Orient All Edges?

The following approaches make more **structural assumptions** in order to orientate all edges

Alternatively, design efficient informative experiments
$\longrightarrow$ not covered here

Exploiting Structural Asymmetries

Additive Noise Models *(Peters et al., 2014)*

Assume **additive noise**: can distinguish $X \leftarrow Y$ from $X \rightarrow Y$ if

$$Y = f(X) + \varepsilon$$

and either

1) $f(\cdot)$ non-linear (`GeneralisedCovarianceMeasure`)

or

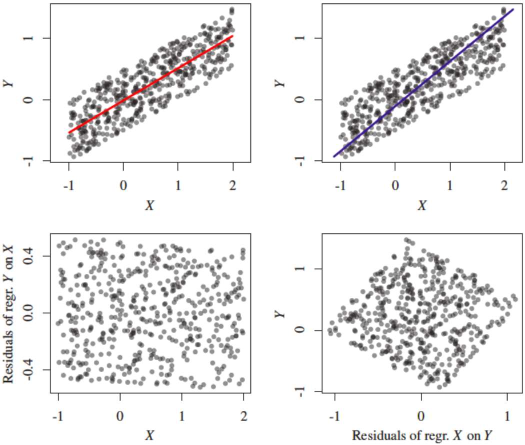
2) ε non-normally distributed (`lingam`)

$\Rightarrow$ orient edges in Markov equivalent graphs

Note: purely mathematical definition of asymmetry — may or may not coincide with causal direction — but: information geometric argument

Exploiting Structural Asymmetries

Illustration



Example: linear with uniform noise

residuals for $X \to Y$ and $X \leftarrow Y$.

Software

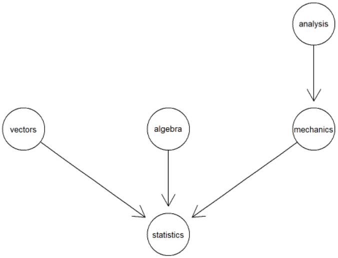
- GES with Gaussian BIC in `R` with `pcalg`:

  `ges(new("GaussL0penObsScore", mathmark))`

  $\Rightarrow$ here, same result as PC algorithm
- LiNGAM in `R` chooses everything for you

  `lingam(mathmark)`

  but needs to be
  transformed into a DAG...



DAG Search: Continuous Optimisation?

(Zheng et al., 2018, 2020)

Derive score based on fairly general models for factors

$$E(X_i \mid \mathbf{X}_{\mathsf{pa}(i)}) = g_i(f_i(\mathbf{X})) \quad E(f_i(\mathbf{X}_{\mathsf{pa}(i)})) = 0$$

g_i known (id., logit...), f_i const. on non-parents & to be learned

Then

$$\min_G \sum_{i \in V} \ell(x_i; f_i(\mathbf{x})) \text{ subject to } G \in \text{ DAG}$$

ℓ loss function ($\approx$ score)
How to enforce $G \in$ DAG 'continuously'?

DAG Search: Continuous Optimisation?

If (gen.) linear model $f_i(\mathbf{X}) = w_j^\top \mathbf{X}$,

let W be matrix of corresponding weights ($w_{ki} = 0 =$ non-edge)

Key insight:
Let $h(W) =$tr$(\exp\{W \circ W\}) - p$

$\Rightarrow$ smooth constraint $h(W) = 0$ enforces acyclicity!

Algorithm **NOTEARS**: Non-combinatorial Optimization via Trace Exponential and Augmented lagRangian for Structure learning

DAG Search: Continuous Optimisation?

Non-linear models? use partial derivatives of f_i

Identifiability: similar assumptions needed as in structural asymmetry case & f_i sufficiently smooth

- authors use NN to approximate more general functions f_i
  must reformulate as finite optimisation problem
  e.g. using 'multi-layer perceptrons'
- `https://github.com/xunzheng/notears`
- output always DAG, not an equivalence class

Discovery + Estimation $\Rightarrow$ IDA

IDA – Algorithm

(Maathuis et al, 2009)

Motivation

- PC (or other algorithms) only deliver an equivalence class of DAGs (CPDAG)
- May also want to quantify causal effects for manipulation of set of nodes $X_1, \ldots, X_p$ on $Y_1, \ldots, Y_m$
- Note: effects may vary with elements of CPDAG!
  $\Rightarrow$ can determine *set* of causal effects, one for each element in CPDAG class
- Maathuis et al. (2009, 2010) propose IDA algorithm ...

IDA – Algorithm

Intervention when the DAG is absent (IDA) – in principle:

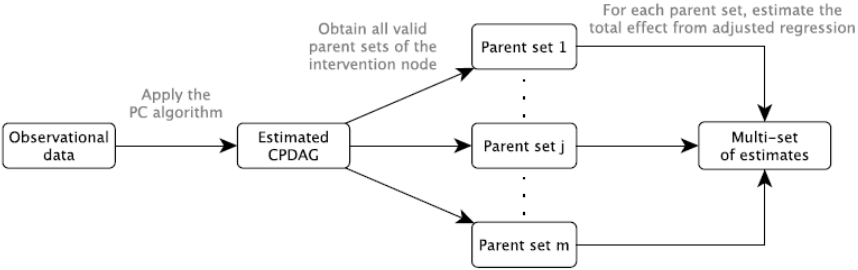
- enumerate all DAGs in CPDAG
- for each DAG and each X_i, Y_j pair determine sufficient adjustment set C — see back-door criterion
- estimate causal effect
  if assume multivariate normal $\Rightarrow$ linear regression
  else: use other estimation method — not covered here

$\Rightarrow$ obtain *multiset* of estimates for each X_i, Y_j pair.

Problem: enumerating all DAGs in CPDAG is *time-consuming*!

IDA – Algorithm

Note: in each DAG, pa(X_i) is a sufficient adjustment set (but not optimal)



IDA – Algorithm

- Can show: only need neighbourhood of X_i to determine sufficient adjustment sets for all possible DAGs

$\Rightarrow$ obtain *set* of estimates, but loose information on multiplicity

- Alternative: find *optimal* adjustment set instead of (inefficient) parent-set (Witte et al., 2020)
  optimal adjustment: estimator with smallest variance among all valid adjustment sets

- **Caution:** post-selection inference issues here! no valid standard errors / conf.intervals

Applied Example: Idefics Cohort Data

See

`https://bips-hb.github.io/ccg-childhood-obesity/`

Causal Discovery — Conclusions

- Searching for underlying graphical structure is in general a difficult task and very active area of research — the space of graphs is too large to be tractable explicitly, and many different proposals to approximate solutions are 'on the market'.

- Must look to exploit additional information: natural experiments / any possibility of randomisation; time-order; domain knowledge on presence / absence / directionality of some edges.

Causal Discovery — Conclusions

- Comparative (simulation) studies between different methods as well as different types of graphs show severe limitations of *all* methods with observational data
- More promising results can be found when using experimental data where perturbations / interventions have actually been carried out — design of experiments: active area of research
- Consider causal discovery as exploratory or hypothesis-generating data analytic method

Further Topics

- **Data integration:** combine / exploit *different* data sets, possibly obtained under different observational / experimental conditions
- **Bayesian methods:** good principle — much computational effort (Moffa & Kuipers)
- **Assess uncertainty in selected graph:** use bootstrap or similar methods
- **Deep-learning approaches:** many recent proposals, e.g. DAGs with NOTEARS — still need thorough ‘testing’ on real data

Last Words

Causal discovery:
aim to find causal structures purely from data...

... have seen that we always need some (empirically untestable) assumptions!

"No causality in, no causality out!" (Nancy Cartwright)

Last Words

AJPH PUBLIC HEALTH OF CONSEQUENCE

The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data

Causal inference is a core task of science. However, authors and editors often refrain from explicitly acknowledging the causal goal of research projects; they refer to causal effect estimates as associational estimates.

This commentary argues that using the term "causal" is necessary to improve the quality of observational research.

Specifically, being explicit about the causal objective of a study reduces ambiguity in the scientific question, errors in the data analysis, and excesses in the interpretation of the results. (*Am J Public Health*. 2018;108: 616–619. doi:10.2105/AJPH. 2018.304337)

Miguel A. Hernán, MD, DrPH

See also Galea and Vaughan, p. 602; Begg and March, p. 620; Ahern, p. 621; Chiolero, p. 622; Glymour and Hamad, p. 623; Jones and Schooling, p. 624; and Hernán, p. 625.

You know the story:

> Dear author: Your observational study cannot prove causation. Please replace all references to causal effects by references to associations.

Many journal editors request authors to avoid causal language,[1] and many observational researchers, trained in a scientific environment that frowns upon causality claims, spontaneously refrain from mentioning the C-word ("causal") in their work. As a result, "causal effect" and terms with similar meaning ("impact," "benefit," etc.) are routinely

Confusion then ensues at the most basic levels of the scientific process and, inevitably, errors are made.

We need to stop treating "causal" as a dirty word that respectable investigators do not say in public or put in print. It is true that observational studies cannot definitely prove causation, but this statement misses the point, as discussed in this commentary.

OF COURSE

glass of red wine per day versus no alcohol drinking. For simplicity, disregard measurement error and random variability—that is, suppose the 0.8 comes from a very large population so that the 95% confidence interval around it is tiny.

The risk ratio of 0.8 is a measure of the association between wine intake and heart disease. Strictly speaking, it means that drinkers of one glass of wine have, on average, a 20% lower risk of heart disease than individuals who do not drink. The risk ratio of 0.8 does not imply

Part 1 & 2: Appendix

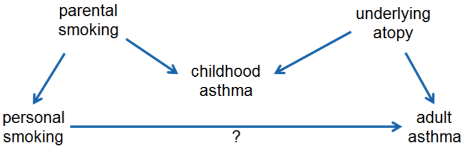
Workflow of Causal Analysis

1. Formulate causal research question (e.g. target trial, decision problem)
2. Elicit (from domain experts) relevant quantities / variables / features and...
3. ... construct causal model reflecting plausible structural assumptions (mix of domain expertise and empiricism)
4. Formalise ‘target of inference’, aka ‘causal estimand’
5. Assess identifiability of target as function of observable information (based on assumed causal model and available / observable data)
6. If identified, apply suitable statistical / data analytic method, e.g. for estimation of target
7. Check (testable implications of) assumptions and carry out sensitivity analyses for untestable assumptions.

M-Bias

(more generally: *collider-Bias*)

Example (simplified from Williamson et al., 2014): want effect of smoking on adult asthma; know that childhood asthma is associated with smoking and with adult asthma.
Is "childhood asthma" sufficient to adjust for confounding?



Note: it is impossible to define or empirically check for 'confounding' in terms of associations!

Always need prior structural knowledge.

Back-door Criterion

How can we use the Back-door Criterion in practice?

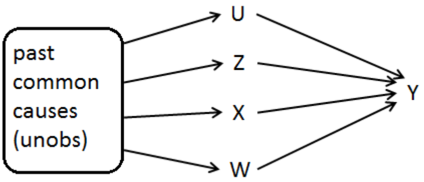
- Construct the DAG based on knowledge of
  - subject matter (basic biology etc.)
  - temporal ordering
  - study design
  - statistical evidence
  - justify all missing edges and absence of further hidden variables (i.e. include all common causes)

$\Rightarrow$ Causal DAG will typically include unobservable variables!

- check for which choice of C (if any) properties (i) and (ii) of Theorem hold $\rightarrow$ check for separations

Association due to Past

Common situation might be: associations between exposure X and other covariates are due to common past history, e.g. past life-style / disease process etc.



$\Rightarrow$ need all of U, Z, W to identify effect of X on Y.

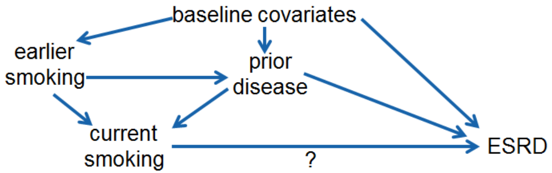
Question:
what happens if W and Y affected by unobserved factor?

Further Examples

do this with `DAGitty` !

Wanted: effect of current smoking on end-stage renal disease (ESRD) (Staplin et al., 2016)

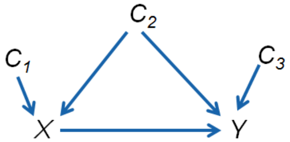
No data available on 'earlier smoking' – is this a problem?



Question: what if 'prior disease' and ESDR affected by further unobserved factors?

Further Examples

Some covariates are unnecessary: here C_1, C_3 not required to adjust for confounding, C_2 is sufficient.



But: while it can improve efficiency to include C_3 as additional predictor of outcome Y, it can be inefficient and even harmful to include C_1...

Bias amplification: can show that if there is some small residual unobserved confounding (e.g. C_2 measured with error), then including variables like C_1 will increase the bias.

Confounding

some misconceptions

- Confounding is a causal concept
- ...a definition of confounding in terms of associations is impossible (wrong in many textbooks)
- ‘associations’ cannot be confounded, only causal relations can be confounded
- notion of ‘confounder’ problematic — often better: ‘deconfounder’ = variables that are useful for reducing bias

Selection Effect (or Bias)

Traditional meaning
Potential to induce bias regarding causal inference through the way how the sample is selected.

Formally
Assume causal effect identified from marginal (observational) distribution of (X, Y, C), then selection effect occurs if it is not necessarily identified from $(X, Y, C|Sel = 1)$ (i.e. given selection).

More general meaning
Some form of *collider-bias:* potential to induce bias regarding causal inference by *conditioning / stratifying* on covariates $\approx$ opposite of confounding.

Selection Effect — Graphically

Let DAG represent background knowledge on conditional independencies and *causal order wrt.* X.

> i.e. variables known not to be affected by an intervention in X must not be descendants of X.

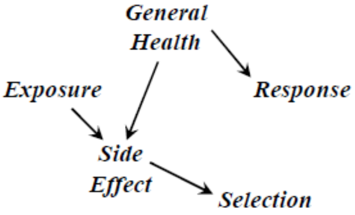
Assume set of covariates sufficient to adjust for confounding.

Trick: draw graph under null–hypothesis of no causal effect

$\Rightarrow$ check if exposure$\perp\!\!\!\perp$response$|$(selection, covariates)

If above check fails, then inference will typically be biased (even if there is a causal effect, i.e. not under null).

Graphical Check — Exercise



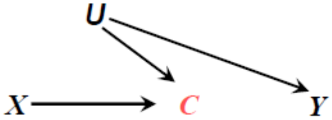
Let X= exposure, Y= response, E= side effect, S= selection (patients with bad side effects drop out of the study).

Exercise: Can we test the null–hypothesis of no causal effect from the patients remaining in the study?

Post–Treatment Covariates

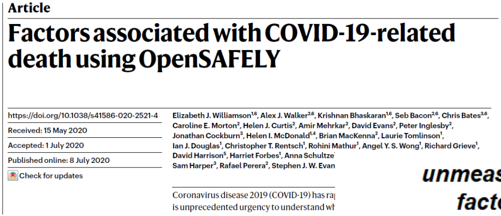
If C is post–treatment covariate (e.g. liver function after treatment) we typically do not adjust for it as we may find $Y \perp\!\!\!\perp X|C$ even when X has a causal effect (but mediated by C). But often done to find the 'direct effect' of X on Y.

Less well known: This can lead to $Y \not\!\perp\!\!\!\perp X|C$ even when X has no causal effect (direct or indirect) on Y! See DAG below...



Selection Bias in COVID Research?

nature paper found ‘protective effect’ of smoking on COVID-19 death


Article

Factors associated with COVID-19-related death using OpenSAFELY

https://doi.org/10.1038/s41586-020-2521-4
Received: 15 May 2020
Accepted: 1 July 2020
Published online: 8 July 2020
Check for updates

Elizabeth J. Williamson[1,6], Alex J. Walker[2,6], Krishnan Bhaskaran[1,6], Seb Bacon[2,6], Chris Bates[3,6], Caroline E. Morton[2], Helen J. Curtis[2], Amir Mehrkar[2], David Evans[2], Peter Inglesby[2], Jonathan Cockburn[3], Helen I. McDonald[1,4], Brian MacKenna[2], Laurie Tomlinson[1], Ian J. Douglas[1], Christopher T. Rentsch[1], Rohini Mathur[1], Angel Y. S. Wong[1], Richard Grieve[1], David Harrison[5], Harriet Forbes[1], Anna Schultze
Sam Harper[3], Rafael Perera[2], Stephen J. W. Evan

Coronavirus disease 2019 (COVID-19) has ra
is unprecedented urgency to understand wh

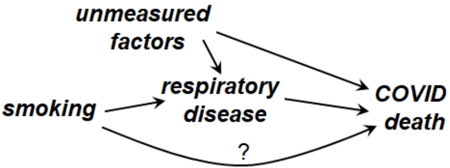


Table-2 Fallacy?

American Journal of Epidemiology

Vol. 177, No. 4
DOI: 10.1093/aje/kws412
Advance Access publication:
January 30, 2013

Commentary

The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients

Daniel Westreich* and Sander Greenland

* Correspondence to Dr. Daniel Westreich, Department of Obstetrics and Gynecology, Duke Global Health Institute, Duke University, DUMC 3967, Durham, NC 27710 (e-mail: daniel.westreich@duke.edu).

Initially submitted January 13, 2012; accepted for publication October 11, 2012.

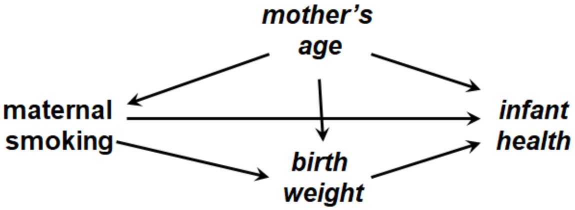
It is common to present multiple adjusted effect estimates from a single model in a single table. For example, a table might show odds ratios for one or more exposures and also for several confounders from a single logistic regression. This can lead to mistaken interpretations of these estimates. We use causal diagrams to display the sources of the problems. Presentation of exposure and confounder effect estimates from a single model may lead to several interpretative difficulties, inviting confusion of direct-effect estimates with total-effect estimates for covariates in the model. These effect estimates may also be confounded even though the effect estimate for the main exposure is not confounded. Interpretation of these effect estimates is further complicated by heterogeneity (variation, modification) of the exposure effect measure across covariate levels. We offer suggestions to limit potential misunderstandings when multiple effect estimates are presented, including precise distinction between total and direct effect measures from a single model, and use of multiple models tailored to yield total-effect estimates for covariates.

causal diagrams; causal inference; confounding; direct effects; epidemiologic methods; mediation analysis;

Prediction — Causation

Prediction of infant health: use *all* available information

Causal effect of maternal smoking on infant health: ignore birth-weight



Selection Effect in Longitudinal / Duration Studies

Problem
more potential for selection effect by inadvertently conditioning on information that occurs later in time.

Chance
time ordering is explicit and potential for selection effect easier to detect.

If time: simulated example

Causal DAG Construction?

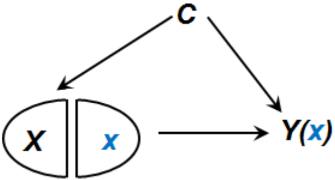
- Domain knowledge (check literature etc.) — talk a lot with subject matter experts!
- Include relevant unmeasured nodes (common causes) & justify absence of further edges and further nodes
- Can empirically assess *some* cond. indep. implications but key assumption of no unmeasured confounding cannot be tested...
- **Can do sensitivity analyses with multiple DAGs if uncertain!**

Single World Intervention Graphs

(Robins and Richardson, 2013)

To see relation with potential outcomes: single world intervention graphs

Node-splitting: X random value, x fixed value by intervention



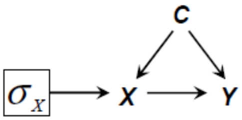
Can see: $X \perp\!\!\!\perp Y^x \,|\, C$.

Influence Diagrams

(Dawid 2002, 2003)

Alternative: influence diagrams
include node σ to indicate where intervention takes place.

$\Rightarrow$ more explicit, but rarely used in practice...



See also: 'Decision-Theoretic' Approach to Causal Inference
Dawid (2002, 2003), Dawid and Didelez (2010), Dawid (2012, 2015)

Structural Models

Models for Y^x or $p(y \,|\, \text{do}(X = x))$ or $E(Y^x)$ etc. are called structural models.

$\Rightarrow$ they model how Y depends on X 'causally', not 'associationally', i.e. how Y depends on an intervention in X.

Warning: Some, *but not all* structural models make assumptions about joint distribution of $\{Y^x, x \in \mathcal{X}\}$

Structural Equations Models (SEMs)

What makes them *structural*?

$$\text{output} \leftarrow f(\text{input})$$

function $f(\cdot)$ is invariant to how the 'input' is chosen / generated, e.g. observed or manipulated.

Warning: strong modelling assumption — system considered essentially a 'machine' with some random noise.
$\Rightarrow$ allows 'cross-world' assumptions (like counterfactuals)

$\Rightarrow$ see single world intervention graphs SWIGs as alternative
(Richardson & Robins, 2013a,b)

Non-Parametric SEMs

(NPSEMs-IE) *(Pearl, 2000)*

X = treatment / exposure, Y = response, C = covariate

Structural equation model (SEM) — ingredients:

- Directed acyclic graph (DAG) defines 'parents' $=$ inputs;
- equations: $X := f_X(\text{pa}(X), U_X)$

  $Y := f_Y(\text{pa}(Y), U_Y)$

  $C := f_C(\text{pa}(C), U_C)$

  where f_X, f_Y, f_C describe 'stable' functional relations
- probability distribution on (U_X, U_Y, U_C)

$\Rightarrow$ induce probability distribution on (X, Y, C).

Often: (U_X, U_Y, U_C) mutually independent $\Rightarrow$ NPSEM-IE

NPSEMs and PRs

With NPSEM-IE we have

$$Y^0 = f_Y(\text{pa}(Y) \backslash X, X = 0, U_Y)$$
$$Y^1 = f_Y(\text{pa}(Y) \backslash X, X = 1, U_Y)$$

with the same U_Y

$\Rightarrow$ distribution on (U_X, U_Y, U_C) also induces a probability distribution on (Y^0, Y^1, X, Y, C)

... in particular a joint distribution for (Y^0, Y^1)!

Example: linear case $Y := \alpha + \beta x + U_Y$

$\Rightarrow Y^i(0) = \alpha + u_Y^i$ and $Y^i(1) = \alpha + \beta + u_Y^i$
$\Rightarrow$ individual causal effect: $Y^i(1) - Y^i(0) = \beta$

Known as **treatment–unit additivity** assumption.

Appendix: Probabilistic Models and Conditional Independence

Probabilistic Modelling

Will use probabilistic models throughout!

- Random variables, e.g. Y, X, Z — “features”
- Distributions / probabilities / densities: $P(Y = y)$

Conditional probabilities:

$$P(Y = y \mid X = x) = \frac{P(Y = y \wedge X = x)}{P(X = x)}$$

in words: probability for event $Y = y$ given we already know $X = x$

‘Conditioning’ $\approx$ ‘stratifying’ $\approx$ ‘selecting’ $\approx$ ‘subgroups’

Independence (no association), write $Y \perp\!\!\!\perp X$:

$$P(Y = y \wedge X = x) = P(Y = y)P(X = x)$$

Conditional Independence

Y, X, Z random variables

Informally: Y is **conditionally independent** of X given Z if once we know/observe Z additional knowledge of X is not helpful in predicting Y

Y conditionally independent of X given $Z \Leftrightarrow Y \perp\!\!\!\perp X|Z$

Symmetry: $Y \perp\!\!\!\perp X|Z \Leftrightarrow X \perp\!\!\!\perp Y|Z$

Conditional Independence

More formally:

Y, X, Z random variables with joint distribution P (pdf/pmf p)

$Y \perp\!\!\!\perp X | Z \Leftrightarrow$

$$P(Y = y \mid X = x, Z = z) = P(Y = y \mid Z = z) \qquad \text{for all } y, x, z$$

Note: if $Z = \emptyset$ then $Y \perp\!\!\!\perp X$ marginal independence.

If $Y \not\perp\!\!\!\perp X | Z$ or $Y \not\perp\!\!\!\perp X |$, then Y, X (conditionally) **associated**.

Conditional Independence

Modelling?
for instance: (linear) regression model (supervised learning)

$$Y \sim a_0 + a_1 X + a_2 Z + \epsilon$$

ϵ independent error term

If* $a_1 = 0 \Rightarrow P(Y|X, Z) = P(Y|Z)$, i.e. $Y \perp\!\!\!\perp X|Z$.

* *and* model correctly specified

Conditional Independence

Conditional independence

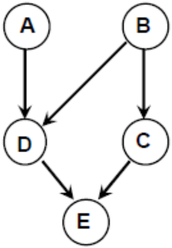
- can be verified empirically by larger variety of statistical tests
- marginal independence much easier to test than *conditional* independence
- a fully non-parametric test for $H_0 : Y \perp\!\!\!\perp X | Z$ does not exist (Peters & Shah, 2020)
- cond. independencies are the *testable* implications of causal models.

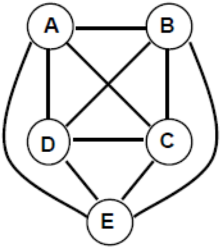
Part 3: Appendix

PC Algorithm — First Step

more details

Example: left = true causal DAG; want to reconstruct this just based on the conditional independencies it implies.
Check whether any pairs of nodes are separated by $S = \emptyset$.



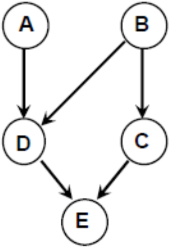


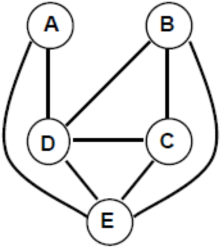
Find $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp C \Rightarrow sep_{AB} = sep_{AC} = \emptyset \Rightarrow$ delete those edges in the right graph.

PC Algorithm — First Step

more details

Check whether any pairs of (still adjacent) nodes are separated by $|S| = 1$ (i.e. by one other node).



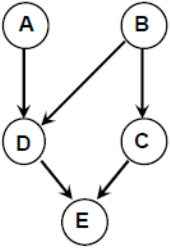


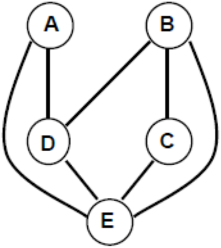
We only find $C \perp\!\!\!\perp D | B \Rightarrow sep_{CD} = \{B\} \Rightarrow$ delete the edge in the right graph.

PC Algorithm — First Step

more details

Check whether any pairs of (still adjacent) nodes are separated by $|S| = 2$ (i.e. by two other nodes).





We find $A \perp\!\!\!\perp E|(C, D)$ and $B \perp\!\!\!\perp E|(C, D)$

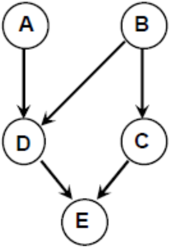
$sep_{AE} = sep_{BE} = \{C, D\}$

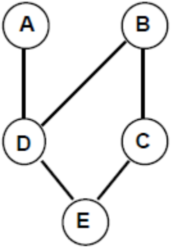
$\Rightarrow$ delete these edges in the right graph.

PC Algorithm — First Step

more details

Terminate, no further conditional independencies can be found.





Have to remember separating sets: $sep_{AB} = sep_{AC} = \emptyset$, $sep_{CD} = \{B\}$, and $sep_{AE} = sep_{BE} = \{C, D\}$.

References I

Introductory / overview texts

Brumback (2021): *Fundamentals of Causal Inference with R.* Taylor & Francos.

Dawid (2005): Fundamentals of Causal Inference. Technical Report, UCL.

Didelez (2018). Causal concepts and graphical models. In: Maathuis M, Drton M, Lauritzen SL, Wainwright M. Handbook of graphical models. Boston: CRC Press. S.353-380.

Didelez, Evans (2018) Causal inference from case-control studies. In: Breslow N, Borgan O, Chatterjee N, Mitchell G, Scott A, Wild C. Handbook of statistical methods for case-control studies. Florida: Chapman & Hall/CRC.

Glymour C, Zhang K, Spirtes P (2019): Review of causal discovery methods based on graphical models. Frontiers in Genetics, 10(524).

References II

Hernán MA, Hsu J, Healy B (2019): A second chance to get causal inference right: A classification of data science tasks. CHANCE, 32(1): 42–49.

Hernán, Robins (2020). Causal Inference. Freely available at: `www.hsph.harvard.edu/faculty /miguel-hernan/causal-inference-book/`

Lauritzen (2000): Causal inference from graphical models. In *Complex Stochastic Systems,* Eds. OE Bardorff-Nielsen, DR Cox, C Klüppelberg, pp. 63-107. CRC Press, London.

Pearl (2000, 2009): *Causality – models, reasoning and inference.* Cambridge University Press.

Pearl (2003): Statistics and causal inference: a review. *Test,* 12, 2, pp. 281-345.

References III

Peters J, Janzing D, Schölkopf B (2017): Elements of causal inference – Foundations and learning algorithms. MIT Press, Cambridge, MA.

Schölkopf B (2019): Causality for machine learning. arXiv preprint arXiv:1911.10500.

Spirtes, Glymour & Scheines (1993, 2000). Causation, Prediction, and Search (1st and 2nd ed.). MIT Press, Cambridge, MA.

Tennant, Murray, Arnold, Berrie, Fox, Gadd, Harrison, Keeble, Ranker, Textor, Tomova, Gilthorpe, Ellison (2021): Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations, International Journal of Epidemiology, 50(2) 620-632.

References IV

Further references

Aalen, Røysland, Gran, Kouys, Lange (2014). Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms. Statistical Methods in Medical Research, 25(5), 2294-2314.

Ankan, A., Wortel, I. M. N., Textor, J. (2021): Testing graphical causal models using the R package “dagitty”. Current Protocols, 1, e45. doi: 10.1002/cpz1.45

Cartwright (2003). Nature’s Capacities and Their Measurement. Oxford Scholarship Online

Chakraborty, B, Moodie, E.M. (2013). Statistical Methods for Dynamic Treatment Regimes Reinforcement Learning, Causal Inference, and Personalized Medicine. Springer.

References V

Chernozhukov, Newey, Robins (2018). Double/De-Biased Machine Learning Using Regularized Riesz Representers. (Online: arXiv:1802.08667)

Chickering DM (2002): Learning equivalence classes of Bayesian-network structures. Journal of Machine Learning Research, 2(2): 445–498.

Colombo D, Maathuis MH (2014). Order-Independent Constraint-Based Causal Structure Learning. Journal of Machine Learning Research 15, 3921-3962

Daniel RM, Stavola BLD, Cousens SN (2011). `gformula`: estimating causal effects in the presence of time-varying confounding or mediation using the g-computation formula. Stata J. ,11, 479–517.

Dawid AP (2000): Causal Inference without Counterfactuals, *JASA*, 95:450, 407-424.

References VI

Int. Stat. Review, 70, 2, pp. 161-189

Dawid AP (2015). Statistical Causality from a Decision-Theoretic Perspective. Annual Review of Statistics and Its Application 2:1, 273-303

Dawid AP, Didelez V (2010). Identifying the consequences of dynamic treatment strategies: A decision theoretic overview. Statistics Surveys, 4, 184-231

Didelez V, Kreiner S, Keiding N (2010). Graphical models for inference under outcome dependent sampling. Statistical Science, 25, 368-387

Foraita, R, Friemel, J, Gunther, K, Behrens, T, Bullerdiek, J, Nimzyk, R, Ahrens, W and Didelez, V (2020): Causal discovery of gene regulation with incomplete data. J. R. Stat. Soc. A, 183: 1747-1775.

References VII

Foraita, Witte, ... Pigeot, Didelez, V. (2022): A longitudinal causal graph analysis investigating modifiable risk factors and obesity in a European cohort of children and adolescents. medRxiv

Gustafson P, McCandless LC, Levy AR, Richardson S (2010). Simplified Bayesian sensitivity analysis for mismeasured and unobserved confounders. Biometrics 66, 1129-1137.

Hansen, Sokol (2014). Causal interpretation of stochastic differential equations.

Hernán MA (2010). The Hazards of Hazard Ratios. Epidemiology, 21(1):13-15.

Hernán MA (2018). The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data. American Journal of Public Health.

References VIII

Hernán MA, Alonso A, Logan R, et al. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology, 19: 766–79.

Hernán MA, Hernández-Diaz S, Robins JM (2004). A structural approach to selection bias. Epidemiology, 15:615-625.

Jung Y, Tian J, Bareinboim E (2021): Estimating identifiable causal effects on Markov equivalence class through double machine learning. Proceedings of the 38th International Conference on Machine Learning, 139: 5168–5179.

Kang, JDY, Schafer, JL (2007) Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Statist. Sci. 22, no. 4, 523-539.

References IX

Kennedy EH, Ma Z, McHugh MD, and Small DS (2017): Non-parametric methods for doubly robust estimation of continuous treatment effects. Journal of the Royal Statistical Society: Series B (Statistical Methodology); 79:1229-1245.

King G, and Nielsen, R. Why Propensity Scores Should Not Be Used for Matching. Working Paper. Copy at `http://j.mp/2ovYGsW`

Maathuis, M., Colombo, D., Kalisch, M. et al. (2010): Predicting causal effects in large-scale systems from observational data. Nat Methods 7, 247-248.

Maathuis MH, Kalisch M, Buhlmann P (2009): Estimating high-dimensional intervention effects from observational data. The Annals of Statistics, 37(6A): 3133–3164.

References X

Mitrovic J, Sejdinovic D, Teh YW (2018): Causal inference via kernel deviance measures. Advances in Neural Information Processing Systems, 31: 5206.

Mooij JM, Peters J, Janzing D, Zscheischler J, Schölkopf B (2016). Distinguishing Cause from Effect Using Observational Data. Journal of Machine Learning Research 17, 1-102

Naimi, Mishler, Kennedy (2020): Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms. arXiv:1711.07137

Pearl J (1995): Causal diagrams for empirical research (with discussion). *Biometrika,* 82, 669-710.

Pearl J (2001). Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, 411–420. San Francisco, CA.

References XI

Peters J, Mooij JM, . . . Schölkopf B (2014): Causal discovery with continuous additive noise models. Journal of Machine Learning Research, 15: 2009–2053.

Richardson TS, Robins JM (2013a). Single World Intervention Graphs (SWIGs) : A Unification of the Counterfactual and Graphical Approaches to Causality. Working paper Number 128, Center for Statistics and the Social Sciences, University of Washington,

Richardson TS, Robins JM (2013b). Single world intervention graphs: a primer. Second UAI Workshop on Causal Structure Learning, Bellevue, Washington

Robins JM (1986): A new approach to causal inference in mortality studies with sustained exposure periods — application to control of the healthy worker effect. *Mathematical Modelling,* 7,pp. 1393-1512.

References XII

Robins JM (2001): Data, design and background knowledge in etiologic inference. *Epidemiology,* 12, 3, pp. 313-320.

Robins JM, Hernán MA, Siebert U (2004). Effects of multiple interventions. In: Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors Vol I. Ezzati M, Lopez AD, Rodgers A, Murray CJL, eds. Geneva: World Health Organization.

Robins JM, Richardson TS (2011). Alternative graphical causal models and the identification of direct effects. In: Causality and psychopathology: finding the determinants of disorders and their cures. Oxford University Press

Rubin, D. (1974). Estimating causal effexts of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66:688-701.

References XIII

Rubin, D. (1978). Bayesian inference for causal effects:the role of randomization. Annals of Statistics, 6:34-58.

Shah RD, Peters J (2020): The hardness of conditional independence testing and the generalised covariance measure. The Annals of Statistics, 48(3): 1514–1538.

Shpitser I, Pearl J (2006). Identification of joint interventional distributions in recursive semi-Markovian causal models. In: Proceedings of the Twenty-First National Conference on Artificial Intelligence. Menlo Park: AAAI Press, 1219–1226

Staplin N, Haynes R, Herrington WG, Reith C, Cass A, Fellstrom B, Jiang L, Kasiske BL, Krane V, Levin A, Walker R, Wanner C, Wheeler DC, Landray MJ, Baigent C, Emberson J (2016). Smoking and Adverse Outcomes in Patients With CKD: The Study of

References XIV

Heart and Renal Protection (SHARP). American Journal of Kidney Diseases, 68, 371-380

Textor J, van der Zander B, Gilthorpe MK, Liskiewicz M, Ellison GTH (2016). Robust causal inference using directed acyclic graphs: the R package 'dagitty'. International Journal of Epidemiology 45(6):1887-1894.

Verma T & Pearl J (1990): Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence,* pp. 220-227.

Vowels MJ, Camgoz NC, Bowden R (2021): D'ya like DAGs? A survey on structure learning and causal discovery. arXiv preprint arXiv:2103.02582.

References XV

Williamson EJ, Aitken Z, Lawrie J, Dharmage SC, Burgess JA, Forbes AB (2014). Introduction to causal diagrams for confounder selection. Respirology, 19, 303-311

Witte J, Didelez V (2018). Covariate selection strategies for causal inference: classification and comparison. Biometrical Journal. 2019; (Epub 2018 Oct 10).

Witte J, Henckel L, Maathuis M, Didelez V (2020): On efficient adjustment in causal graphs. Journal of Machine Learning Research, 21(246): 1-45.

Zhao, Hastie (2021): Causal Interpretations of Black-Box Models, Journal of Business & Economic Statistics, 39:1, 272-281

Zheng, Aragam, Ravikumar, Xing (2018) DAGs with NO TEARS: Continuous Optimization for Structure Learning. In Advances in Neural Information Processing Systems, 2018.