Causal Learning for Data Science

Vanessa Didelez
with much help from Ronja Foraita and Christine W Bang

Leibniz Institute for Prevention Research & Epidemiology – BIPS

December 2021

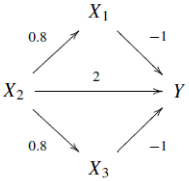
Data Train — UBRA Bremen

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

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

$\Rightarrow X_2$ causally least important.

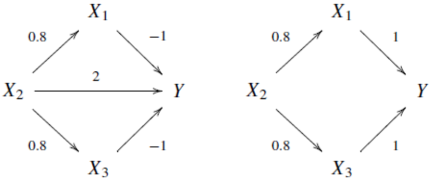
(Here linear structural equation models, LSEM)

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

Causal effects: $\theta_1 = \theta_3 = 1$ but $\theta_2 = 1.6$

$\Rightarrow X_2$ causally most important.

(Here linear structural equation models, LSEM)

Part 3: Causal Discovery

Causal Discovery

So far: causal graph (DAG) **given** based on causal background knowledge. ⇒ Can query the graph as to whether observed (conditional) associations can have causal interpretation.

Causal discovery is about **finding** a causal graph when there is no (sufficient) causal background knowledge.

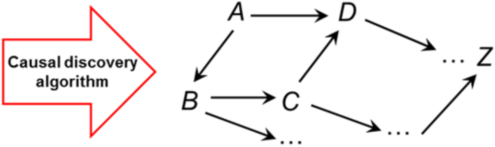
aka: causal search, (causal) structure learning, (causal) graph estimation, network inference ...

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**

⇒ 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. Maybe generate some causal hypotheses.

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

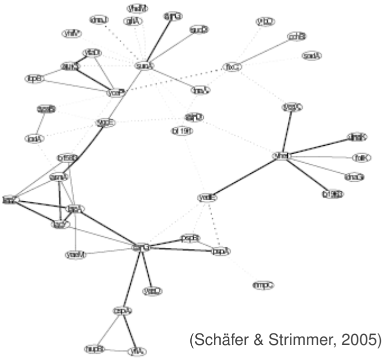
Motivation: Gene Regulation

Causal interpretation of gene networks: not so obvious as interventions rather unusual.

Earlier: do(X) to denote intervention on X

- e.g. by knock–out / inhibition / activation
- genetic changes = random process $\Rightarrow$ could have been different
- ‘causal pathways’ similar to mechanistic description

Ecoli Gene Association Network

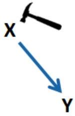


(Schäfer & Strimmer, 2005)

Causal Interpretation

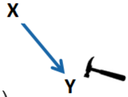
... for gene regulation?

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)

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: expression measurements 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), then estimate possible effects

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

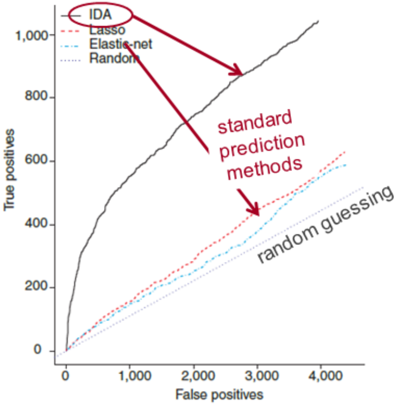
Motivation ctd: Gene Regulation

Maathuis et al (2010)

# true vs false positives
for top 5000 effects predicted
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)



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

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

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)

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!

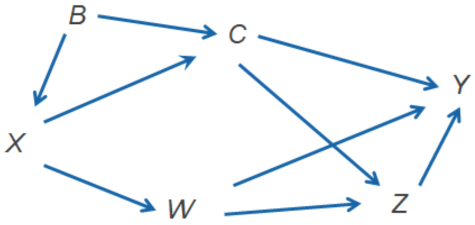
Separation and Independence

Reminder

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. $p(x, y|s) = p(x|s)p(y|$

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



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 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 $X \perp\!\!\!\perp Y|Z$, 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**)

⇒ 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

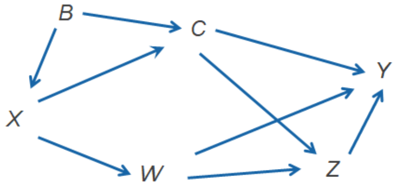
– some directed and some undirected edges

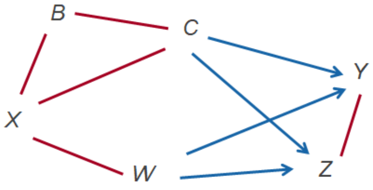
– undirected means: in class, both directions exist

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

CPDAG Example

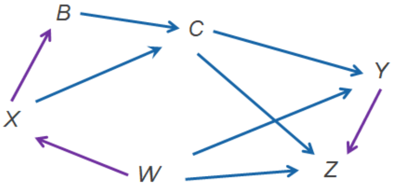
CPDAGs are mixed graphs with...
undirected edges if either direction occurs at least once in the equivalence class

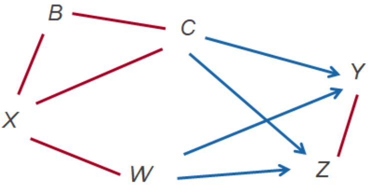




CPDAG Example

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 if possible.

Note: this is the **simplest** constraint-based discovery algorithm;
assumptions: 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.

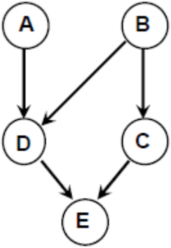
PC Algorithm — First Step

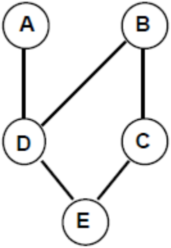
V = set of nodes, and each node A has a set of adjacent nodes adj_A.

```
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

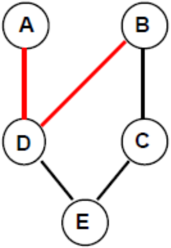
Identify V–structures

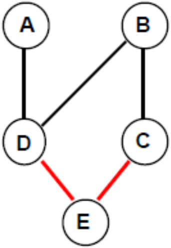
Procedure

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.



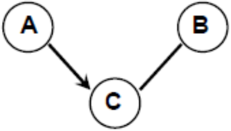


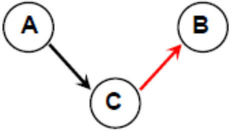
PC Algorithm — Third Step

Meek's Rules

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

Examples: constellations that can be oriented



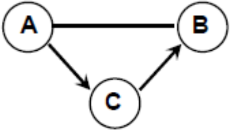


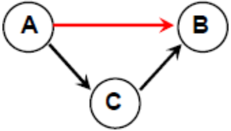
PC Algorithm — Third Step

Meek's Rules

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

Examples: constellations that can be oriented



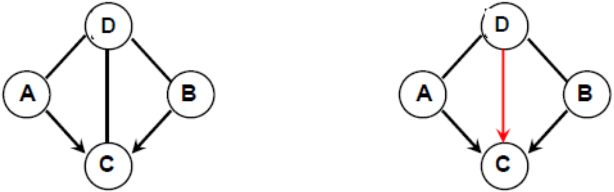


PC Algorithm — Third Step

Meek’s Rules

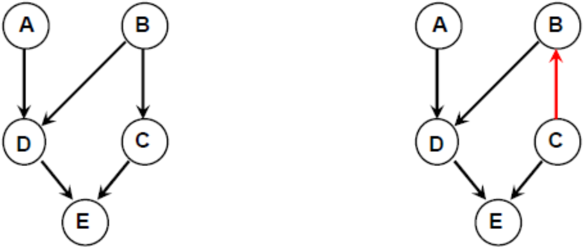
Orient remaining edges such that
– cycles are avoided
– no new V–structures are created.

Examples: constellations that can be oriented



PC Algorithm — Finally

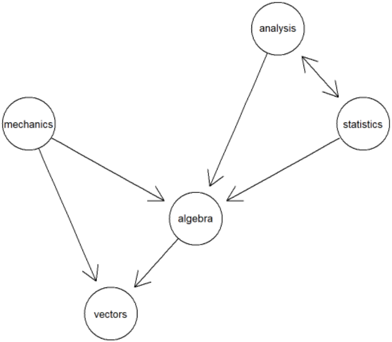
In original example: cannot orient B–C edge as both graphs are Markov equivalent.



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 ⇒ 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!

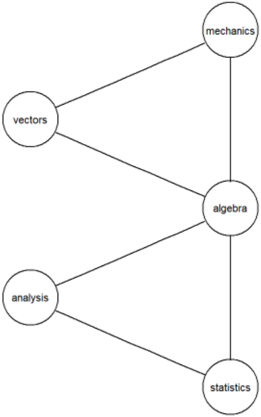
→ outputs can be very unstable
→ outputs may not be valid CPDAGs
⇒ should bootstrap results to assess variability of graph!

→ outputs may depend on order of input variables

... to avoid in `pcalg`
— 'stable' skeleton search
— '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 — 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)

FCI Algorithm

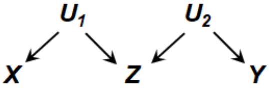
Relaxing Causal Sufficiency

FCI = ‘fast causal inference’ — but algorithm actually quite slow

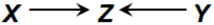
Allowing latent (unmeasured) variables: much more complicated equivalence class!

$\longrightarrow$ partial ancestral graph (PAG)

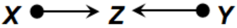
True DAG: latent U_1, U_2



PC algorithm: wrong output



FCI algorithm: correct PAG



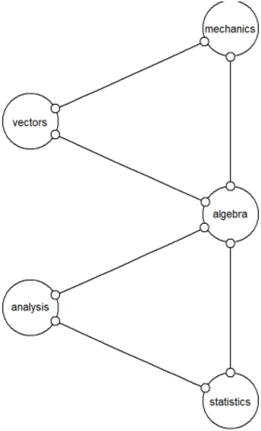
Interpretation: Edges in PAGs

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

— 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 ≈ 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

Exploiting structural asymmetries

Additive Noise Models

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

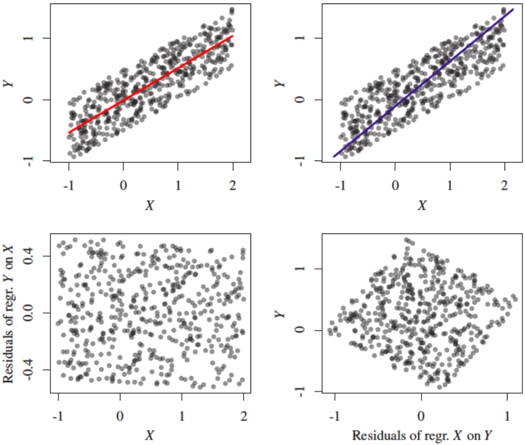
2) ε non-normally distributed (`lingam`)

$\Rightarrow$ orient edges in Markov equivalent graphs

Note: purely mathematical definition of asymmetry — may or may not coincide with causal direction — additional 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`:

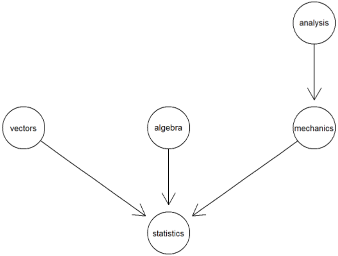
  ```
  ges(new("GaussL0penObsScore", mathmark))
  ```

  $\Rightarrow$ here, same result as PC algorithm

- LiNGAM in `R` chooses everything for you

  ```
  lingam(mathmark)
  ```

  but needs to be
  transformed into a DAG...



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:

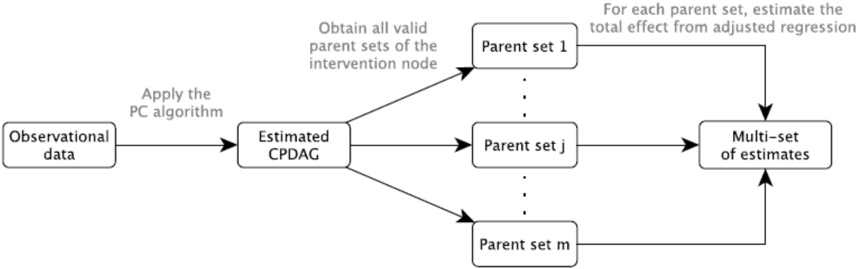
- enumerate all DAGs in CPDAG
- for each DAG and each X_i, Y_j pair determine sufficient adjustment set C — see yesterday!
- estimate causal effect
  if assume multivariate normal $\Rightarrow$ linear regression
  else: use other estimation method — see yesterday!

$\Rightarrow$ 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.



Observational data
Apply the PC algorithm
Estimated CPDAG
Obtain all valid parent sets of the intervention node
Parent set 1
Parent set j
Parent set m
For each parent set, estimate the total effect from adjusted regression
Multi-set of estimates

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

- Originally: IDA only for linear causal models, but can be generalised using estimation methods from Part 2
- 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

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.

- 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
- **Assess uncertainty in selected graph:** use bootstrap or similar methods
- **Deep-learning approaches:** many recent proposals — still need thorough ‘testing’ on real data

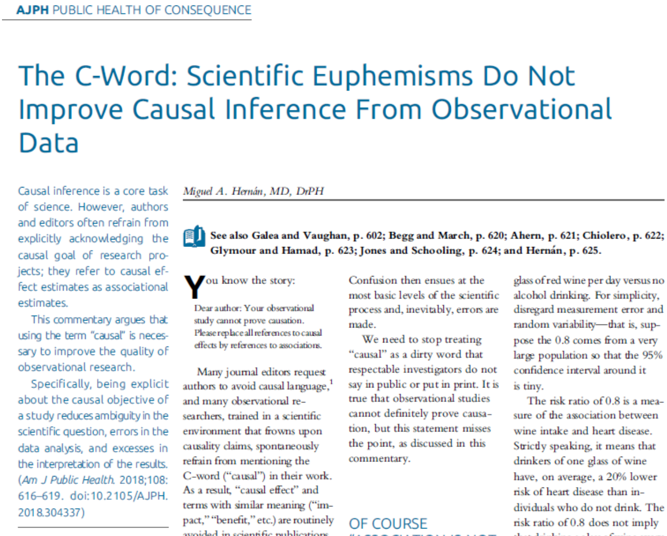
Last Words

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 avoided in scientific publications

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