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

Overview of Course

Part 1: Causal Reasoning
— with directed acyclic graphs (DAGs)

Part 2: Estimation (Learning) of Causal Effects

Part 3: Causal Discovery — finding (potential) causes

Note: all of this is a *subjective* selection of material based on what I like and know (though I try to cover a variety of topical material).

Aim of Course

- Introduce basic concepts of causal learning (reasoning, modelling & inference)
- ... to enable you to read more advanced 'causal' papers
- Focus on:
  - formulating causal (research) questions
  - some basic methods
  - understanding sources of (avoidable and unavoidable) bias
- Mix of mathematics & stories/examples

Causality in Data Science

- 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
- Data science: very new field — but pretty much what many different fields have in common: data
- Causality very fundamental to many research questions in data science!

Part 1: Causal Reasoning

Preamble

- Causation / causality: philosophical, moral and other usages of the term — not what we are concerned with here
- *Today:* 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 $\Rightarrow$ Target Trial!

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

Research Questions

Descriptive / predictive:

“Is this patient at high risk of developing complications during surgery?”

Causal:

(A) “Which type of anaesthetic should this patient receive to reduce the risk of complications during surgery?”

(A’) “How does the amount of anaesthetic affect the risk of complications during surgery?”

(B) “What can be done to reduce the risk of complications during surgery for an average / a particular type of patient?”

Research Questions

Descriptive / predictive:

“Which type of client will buy which kind of product?”

Causal:

(A) “Should advert be at the top or bottom of website to increase the probability of viewing product?”

(A’) “How does the size of advert affect the probability of viewing product?”

(B) “How can I get a client to buy my product?”

Research Questions

Descriptive / predictive:

“Who is most likely to become long-term unemployed?”

Causal:

(A) “Will a minimum wage legislation increase the unemployment rate of a country?”

(B) “What can be done to prevent someone from becoming unemployed?”

Research Questions

Type-A causal questions: **Causal Effects**

“what is the causal effect of a ‘treatment’* ?”

“dose-response relation”

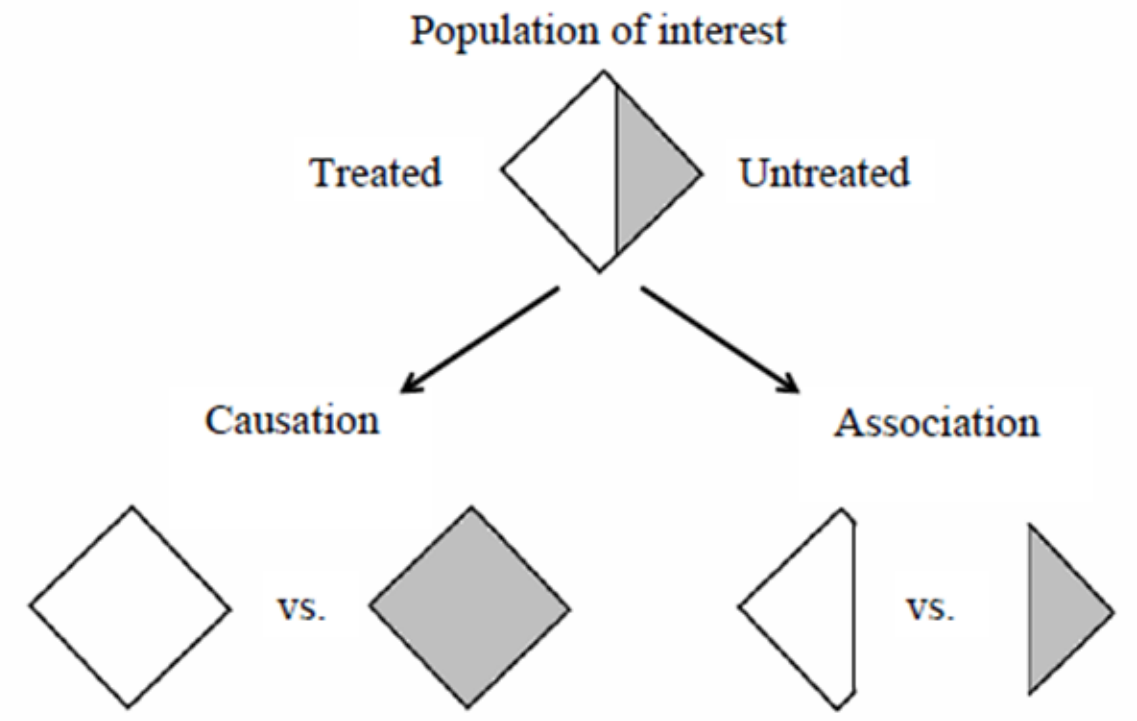
Type-B causal questions: **Causal Discovery**

“where can / should we possibly intervene?”

*Note: ‘treatment’ needs not be medical, could be: policy, teaching method, physical activity etc.

Causation versus Association

(Hernan & Robins, 2020 book)



Causal effect: contrast of outcome if 'everyone was treated' versus if 'no-one was treated'

Causal Models 1-0-1

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

Identifiability:
aspects of the interventional situation equal certain functions of the observational situation

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 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

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

Basic Causal Concepts

Formalisms to make interventions explicit:

do-notation / causal DAGs / decision theory

Potential Responses / counterfactuals

Structural equations / structural causal models: *not enough time to cover these*

do–Notation

(Pearl, 2000)

Judea Pearl introduced intuitve 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 $p(y \,|\, \text{see}(X = x)) = p(y \,|\, x)$

do–Intervention

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

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

$$p(y, x, c_1, c_2) = p(y|x, c_1, c_2)p(x|c_1, c_2)p(c_2|c_1)p(c_1)$$

May have reasons to believe that under intervention:

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

DAGs help to structure the *factorisation* so as to represent prior-causal knowledge

Identifiability

Will see that under **three structural assumptions** we have for suitable set C of covariates:

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

left: interventional distribution; right: observational distrib.

$\Rightarrow$ **non-parametrically identified**, i.e. without parametric assumptions like linearity, Gaussianity etc.

Potential Responses (PRs)

(Rubin, 1974; many others)

Consider binary 'treatment' $X^i \in \{0, 1\}$, individual i

$Y^i(0)$ = response under intervention setting $X^i = 0$

$Y^i(1)$ = response under intervention setting $X^i = 1$ for same subject (at the same time)

$\Rightarrow \{Y^i(0), Y^i(1)\}$ can *never be observed together*
$\Rightarrow$ potential responses (or potential outcomes).

Potential Responses

More generally, for arbitrary treatment type $X \in \mathcal{X}$

$Y^i(x)$ = response if we *set* $X^i = x$

Counterfactuals

Once a treatment has been realised, say $X^i = 1$,

then $Y^i(1)$ can be observed

and $Y^i(0)$ becomes *counterfactual* (and vice versa).

Approaches relying on assumptions / properties of the joint distribution of $(Y(0), Y(1))$ can be called counterfactual as these assumptions are never empirically verifiable.

Potential Responses and 'do'

Many approaches, in fact, do not rely on *joint* distribution of $(Y(0), Y(1))$, and could equivalently be expressed using do$(\cdot)$-notation.

(but PRs strong tradition in biomedical / econometric literature.)

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

But joint distribution of $(Y(0), Y(1))$ has no counterpart in do–notation.

$\Rightarrow$ Can express more (also more dubious) concepts with PRs

Causal Effects

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

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

or of $p(Y(x))$ versus $p(Y(x'))$

For instance: **Average Causal Effect**

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

Cause and Effect

Can now define:

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

$$p(y \,|\, \mathsf{do}(X = x)) \neq p(y \,|\, \mathsf{do}(X = x'))$$

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.

Counterfactual Prediction

Decision needed about treatment $X = 1$ or $X = 0$

Want to predict what happens with Y under either setting $X = 1$ or $X = 0$

$\Rightarrow$ Only one can be applied: counterfactual prediction.

Key Assumptions

for identifiability from observational data

Causal Consistency Assumption:

$$\text{if we observe } X = x \text{ then } Y = Y(x)$$

Positivity Assumption:

$$p(x \mid c) > 0 \text{ for all } x, c \quad (p(c) > 0)$$

Key Assumptions

Assumption of **no unmeasured confounding:**
(aka: random treatment assignment, or cond. exchangeability, ignorability, or ...)

Set C of observed (measured) pre–treatment covariates exists such that

$$Y(x) \perp\!\!\!\perp X \mid C$$

for all x to be considered as treatment values

Interpretation:
within values of C, can consider X like randomised wrt Y

Denote: C is *sufficient* to adjust (control) for confounding;
or 'valid adjustment set'

No-Unmeasured-Confounding

with do$(\cdot)$

Assumption of **no unmeasured confounding** & **'consistency'** with do–notation:

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

Interpretation: within values of C, whether $X = x$ obtained by intervention or observation makes no difference wrt. distribution of Y.

Identifiability Revisited

We consider $p(y \,|\, \text{do}(X = x))$ or equivalently $p(Y(x))$.

With the above assumptions:

$$p(Y(x)) \overset{(i)}{=} \sum_c p(Y(x)|c)p(c) \overset{(ii)}{=} \sum_c p(Y(x)|x,c)p(c)$$

$$\ldots \overset{(iii)}{=} \sum_c p(y|x,c)p(c)$$

(i) probability calculus
(ii) valid adjustment set
(iii) causal consistency & positivity

Adjustment / Standardisation

Consider the above result

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

- left = causal quantity; right = observational quantity
  $\Rightarrow$ identified if covariates C measured
- right hand side = identifying functional (under the assumptions)
- know as adjustment formula, or standardisation (to the marginal distribution of C)
- also: simplest case of so-called ‘g-formula’ (Robins, 1986)

Confounding

Above: confounding is present if

$$Y(x) \not\perp\!\!\!\perp X$$

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

Usually:

Confounding = some *(unobserved) common cause* of X and Y

$\Rightarrow$ Use causal DAGs to clarify!

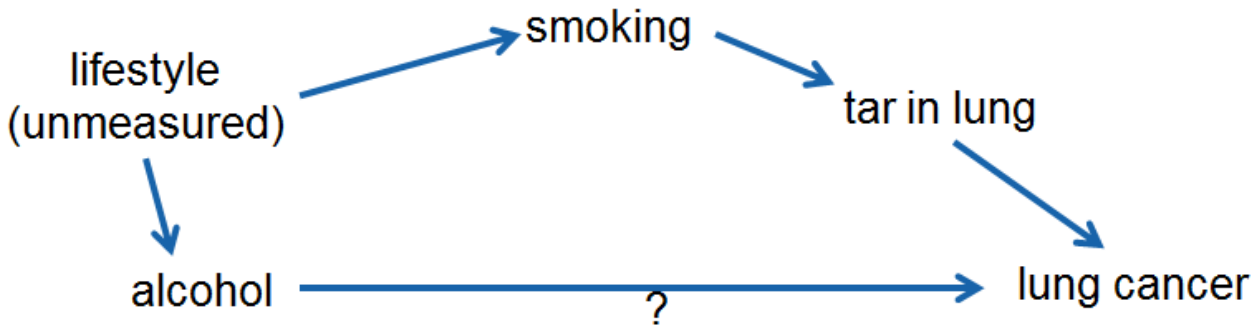
Quiz

Possible break for quiz.

Causal Directed Acyclic Graphs (DAGs)

Graphs — Terminology

Graph $G = (V, E)$
$V =$ vertices 0 nodes = variables / features
$E =$ edges = possible (causal) dependence
Non-edge = known (conditional) independence



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

(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)

Why Graphs?

Make explicit: underlying assumptions & required background knowledge!

Graphs: one way to *represent* & *organise* assumptions / prior knowledge

Here: will focus on bias sources related to

- confounding
- selection

Graphs for Identifiability

- Can we identify causal effects from observational data?
- ... for what do we need to adjust?
- ... for what must we not adjust?

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

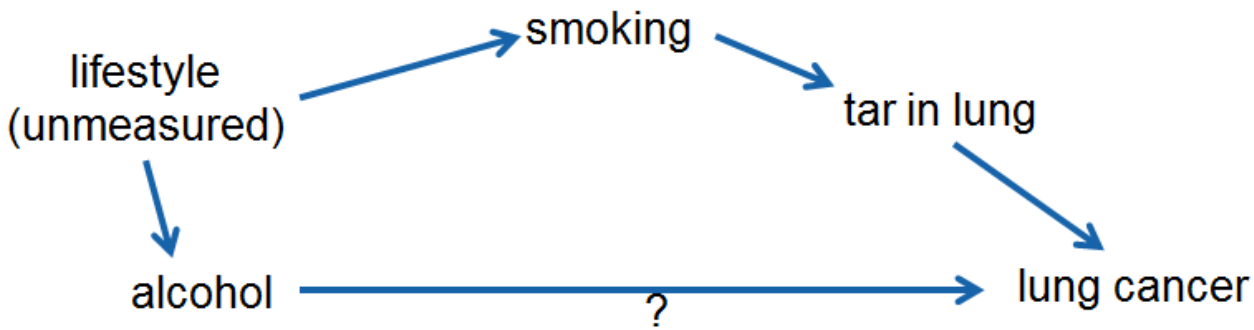
$\Rightarrow$ Formulating causal DAG for such situations: active research!

Graphs — Terminology

Graphical terms:

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

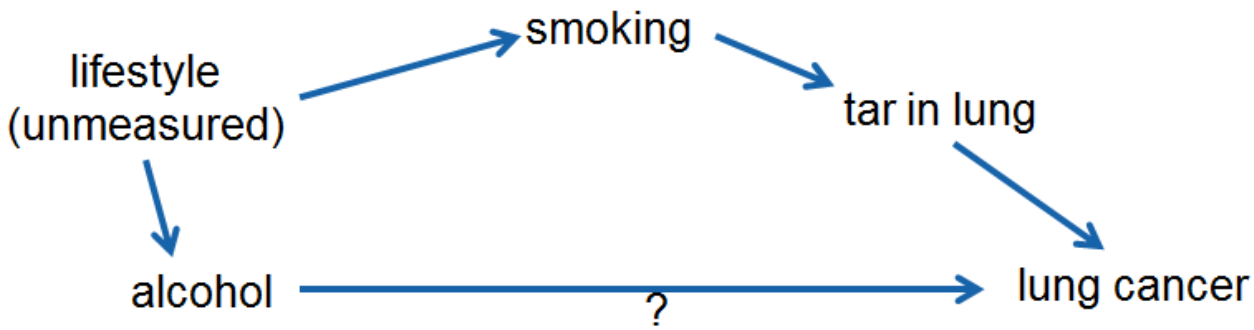
‘(directed) paths’, ‘(directed) cycles’



Graphs — Markov Property

with DAGitty

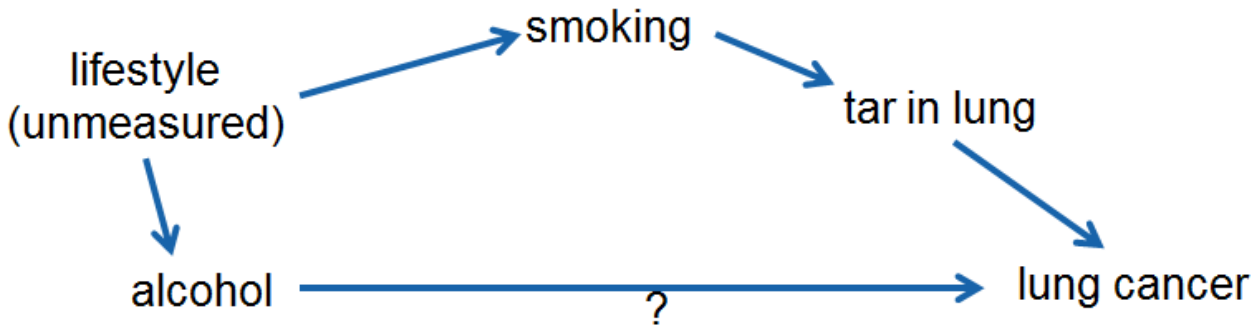
Observationally:
Absence of edges into outcome: if we know whether there is tar in the lungs and whether person drinks alcohol, then smoking status or any further information on lifestyle are non-informative for the probability of lungcancer.



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

Causal Markov Condition

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



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

Direct Causal Effect

Notion of **Controlled direct effect:**

For other parent node(s) Z of Y, what is the effect of intervening in X on Y while fixing Z by intervention?

Principle: block certain causal pathways by fixing Z; then assess remaining effect of X on Y.

$\Rightarrow$ contrast of

$$p(y \,|\, \text{do}(X = x, Z = z)) \quad \text{versus} \quad p(y \,|\, \text{do}(X = x', Z = z))$$

Causal Markov Condition

Axiom (Causal Markov Condition):

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

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

('direct' relative to other nodes)

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

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 cond. independencies using **d-separation** (later)

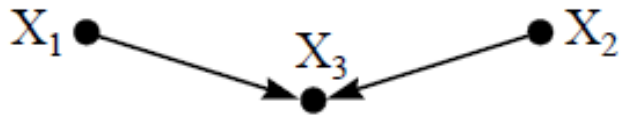
$\Rightarrow$ testable implications of DAG models!

Selection Effect

"collider bias"

Important for the interpretation:

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



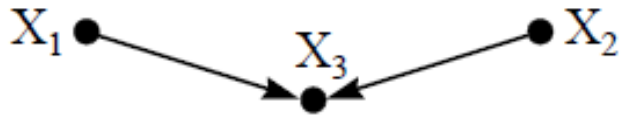
here: $X_1 \perp\!\!\!\perp X_2$ but $X_1 \not\perp\!\!\!\perp X_2 \mid X_3$

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

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

Selection Effect

"collider bias"



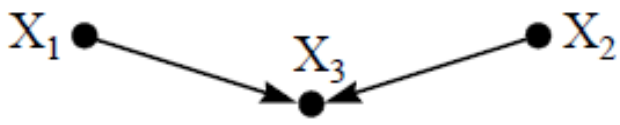
Example: some school admission process is such that pupils are admitted (X_3) if they are either good at maths (X_1) or good at sports (X_2).
Assume in population X_1 and X_2 are independent(!)

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)

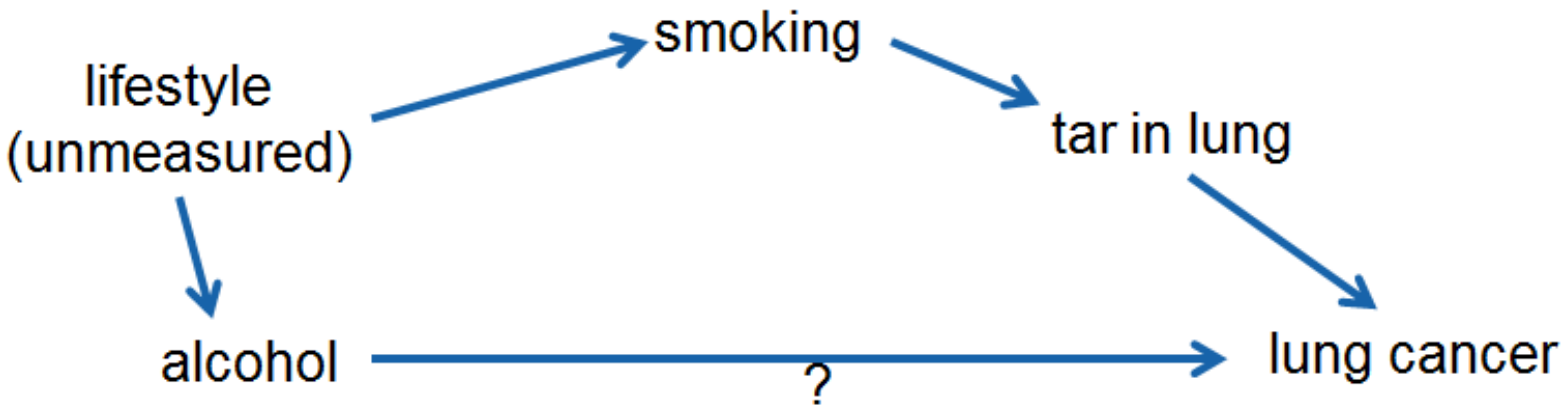
Given DAG $G = (V, E)$. A path between a and $b \in V$ is blocked by $S \subset V \backslash \{a, b\}$ if

(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

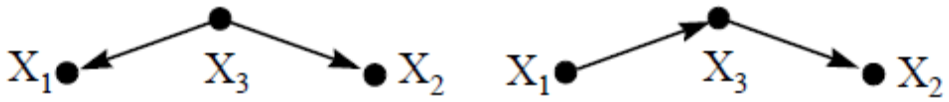
A and $B \subset V$ are **d–separated** by $S \subset V \backslash (A \cup B)$ if every path between A and B is blocked by S.

d–Separation — Quiz



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 data alone.

Causal DAG

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

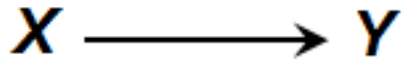
Additional **semantics** relating DAG to interventions:

- 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.

Example 1



This causal DAG expresses:

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

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

Example 1 ctd.

Moreover:

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

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

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.

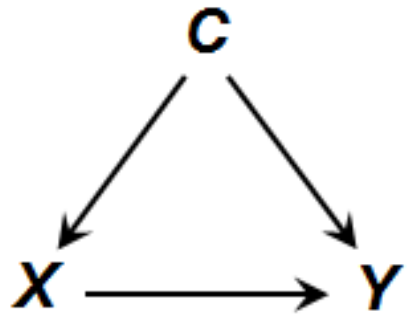
Finally:

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

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

- i.e. X is independent of Y under an intervention in Y.

Example 2



This causal DAG expresses:

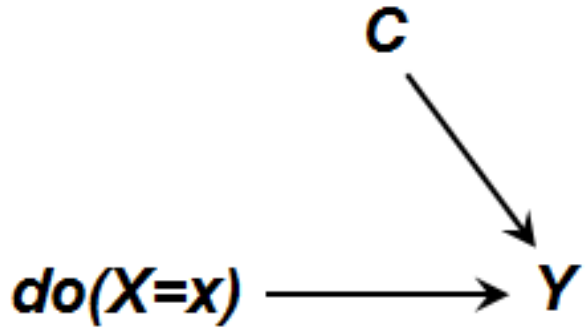
- an intervention in X can affect Y, *but not* C
- an intervention in C can affect X and Y
- an intervention in Y *cannot* affect X nor C.

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

Example 2 ctd.

Moreover:

- an intervention in 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 in X does not affect C and intervention in C does not affect X
- otherwise, regarding X, C, Y same as Example 2.

Example 4

This causal DAG expresses:

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

Example 4 ctd.

Moreover:

- an intervention in Z prevents and intervention in 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.

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')

Use of causal DAGs

"Draw your assumptions before your conclusoins!"

(Hernán)

- Make explicit your assumptions $\Rightarrow$ draw DAG based on backgroud knowledge (good thing in any case)
- Check if some implications of DAG can be verified empirically, e.g. implied conditional independencies
- Check if your desired target can be identified with the observable data
- Possibly: motivate sensitivity analysis by different competing causal DAGs reflecting uncertainty in subject matter knowledge.

Standardisation

Remember: identifying functional for the effect of X on Y

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

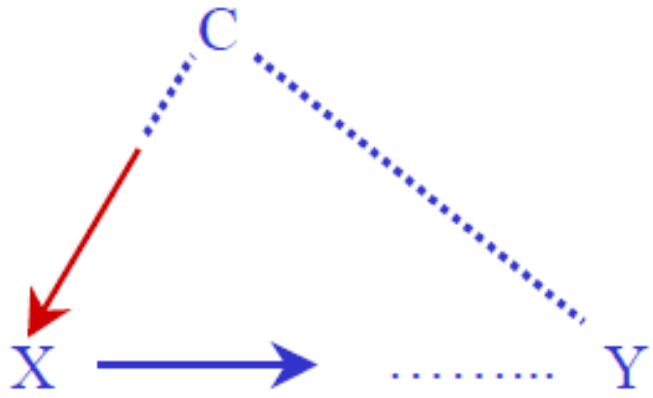
Requires assumption 'no-unmeasured confounding given C'.

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 is non-descendant of X and

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

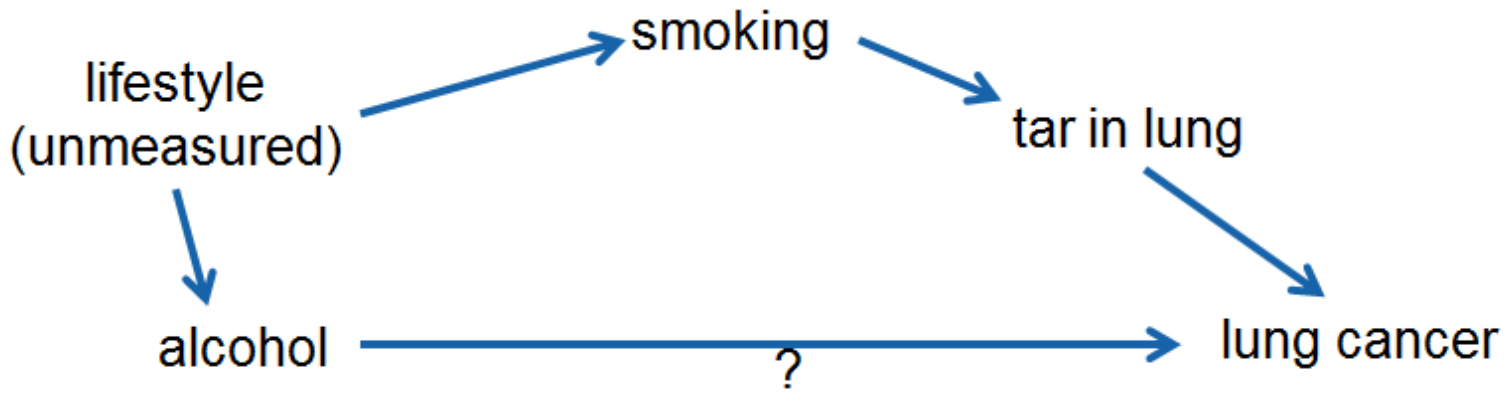
C is then *sufficient* adjustment set.

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

Back-door Criterion — Exercise

with `DAGitty`

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



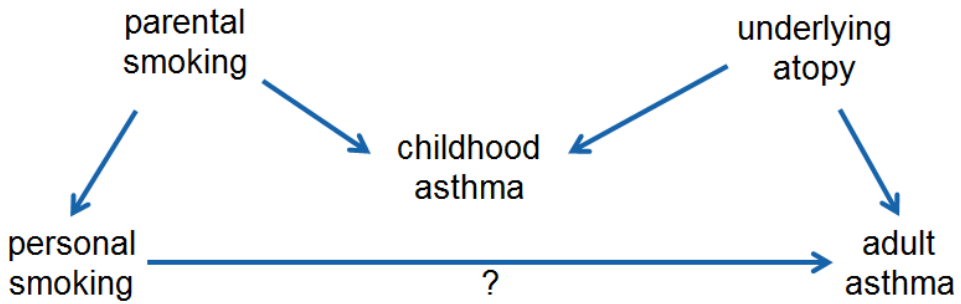
Sufficient set of covariates to identify the effect of X on Y?

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

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?

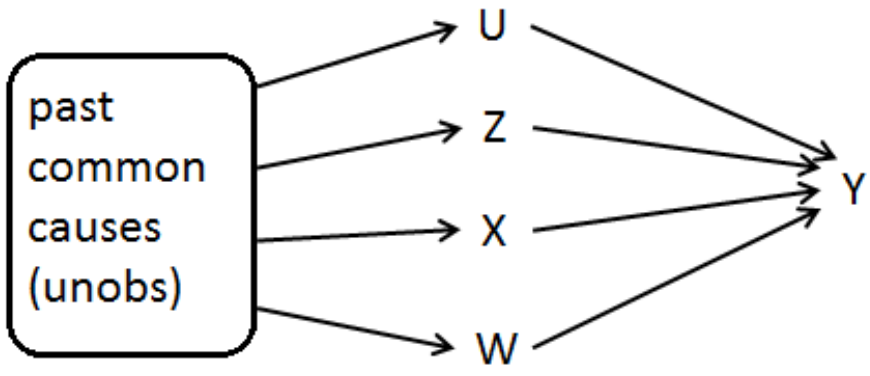
- Construct the DAG based on knowledge of
  - subject matter (basic biology etc.)
  - temporal ordering
  - study design
  - statistical evidence
  - justify all missing edges and absence of further hidden variables (i.e. include all common causes)

$\Rightarrow$ 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.

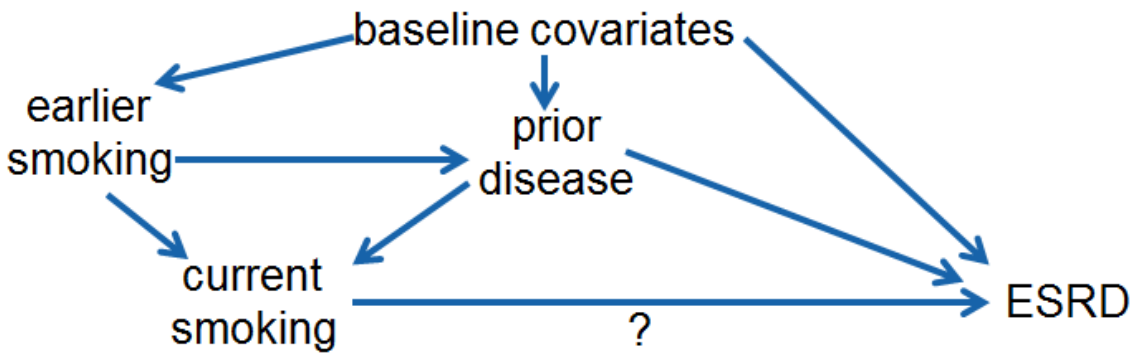
Question:
what happens if W and Y affected by unobserved factor?

Further Examples

do this with `DAGitty` !

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

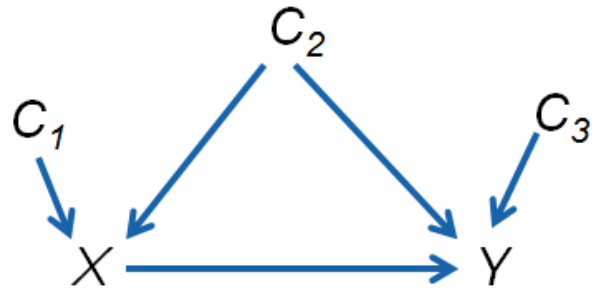
No data available on 'earlier smoking' – is this a problem?



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.

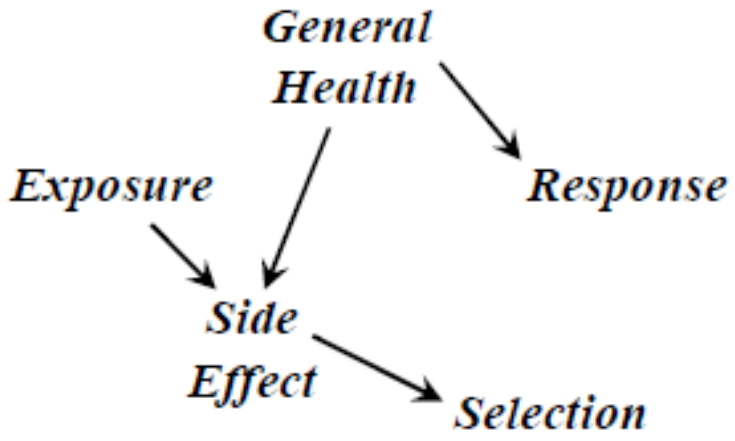
Assume set of covariates sufficient to adjust for confounding.

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



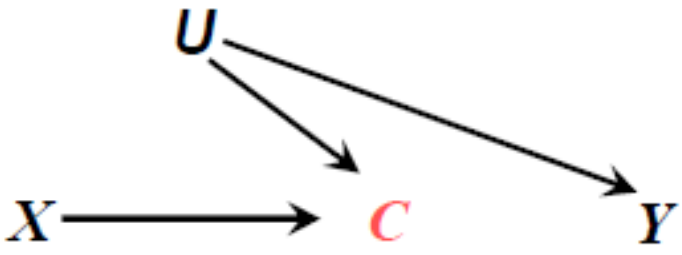
Let X= exposure, Y= response, E= side effect, S= selection (patients with bad side effects drop out of the study).

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

Post–Treatment Covariates

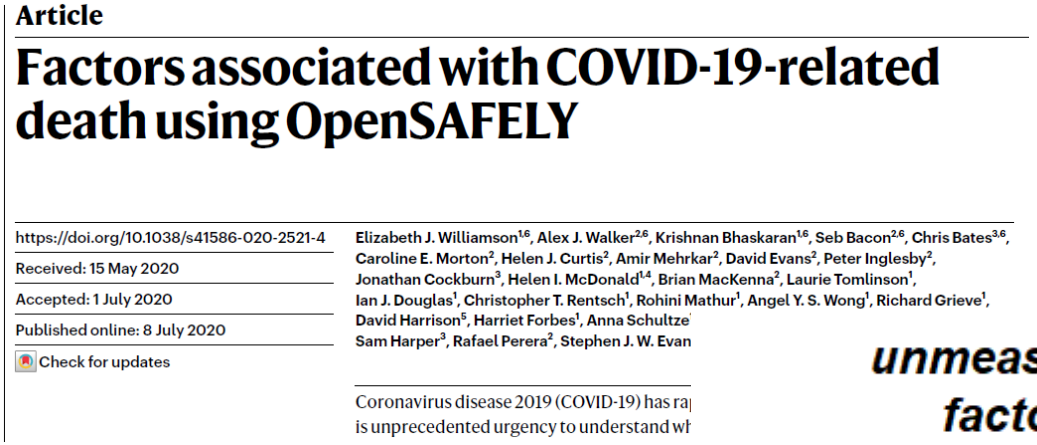
If C is post–treatment covariate (e.g. liver function after treatment) we typically do not adjust for it as we may find $Y \perp\!\!\!\perp X|C$ even when X has a causal effect (but mediated by C). But often done to find the 'direct effect' of X on Y.

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[1], Sam Harper[3], Rafael Perera[2], Stephen J. W. Evan[1]

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

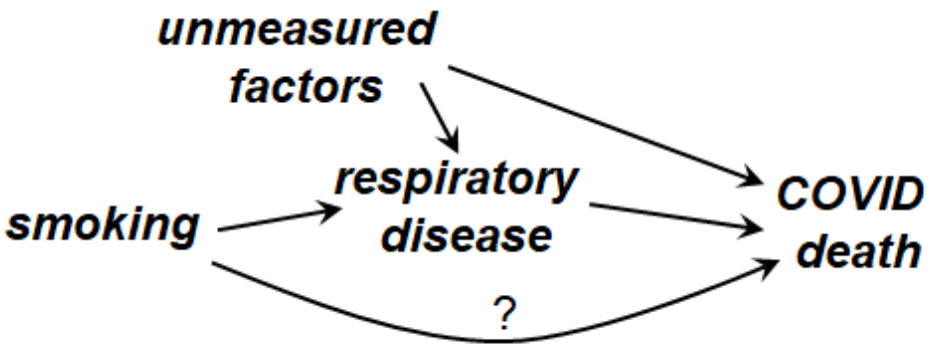


Table-2 Fallacy?

American Journal of Epidemiology
© The Author 2013. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
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.

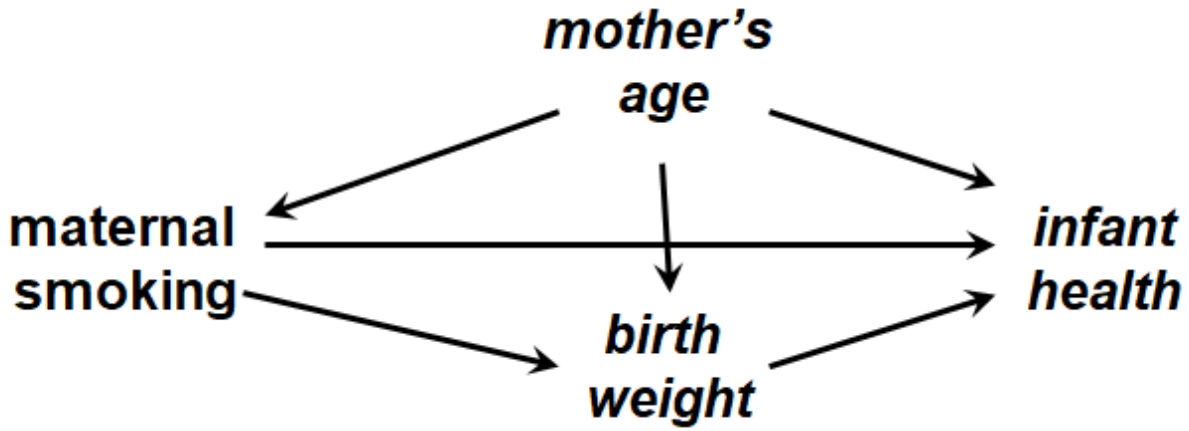
It is common to present multiple adjusted effect estimates from a single model in a single table. For example, a table might show odds ratios for one or more exposures and also for several confounders from a single logistic regression. This can lead to mistaken interpretations of these estimates. We 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!**

Causal DAGs

Summary

- Graphs are helpful to organise your causal reasoning / structuring of a given causal question with data at hand.
- Main purpose: can the causal effect be identified from the available data in the first place? Can we test for causal effect? Can we estimate the causal effect?
- 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?
- 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 1: Causal Reasoning — Appendix

Workflow of Causal Analysis

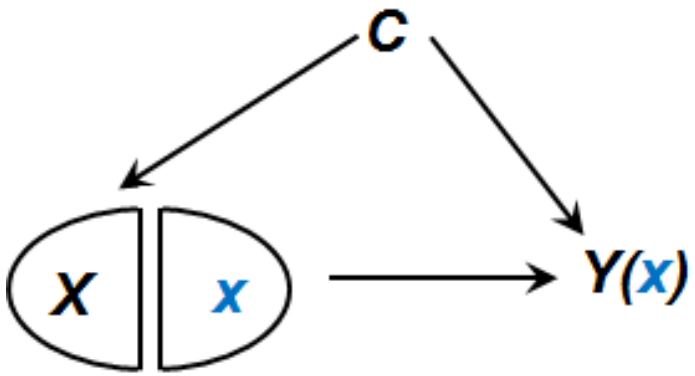
1. Formulate causal research question (e.g. target trial, decision problem)
2. Elicit (from domain experts) relevant quantities / variables / features and...
3. ... construct causal model reflecting plausible structural assumptions (mix of domain expertise and empiricism)
4. Formalise ‘target of inference’, 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.

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



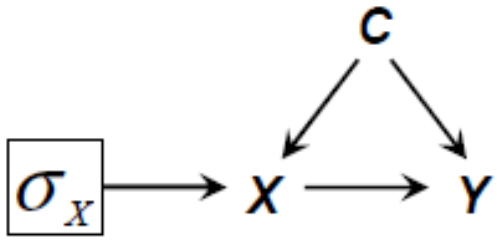
Can see: $X \perp\!\!\!\perp Y(x) \,|\, C$.

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 2: Estimating a Causal Effect

Basic Setting

$X =$ binary (point-)treatment

$Y =$ some (numeric) outcome

(not survival / duration — that's special)

$C =$ sufficient adjustment set of pre-treatment covariates

Keeping it simple to focus on principles!

REFERENCE: Goetghebeur, E, le Cessie, S, De Stavola, B, Moodie, EE, Waernbaum, I. Formulating causal questions and principled statistical answers. Statistics in Medicine. 2020; 39: 4922– 4948.

Defining '(Point-)Treatment'

Oscar-winners live longer

$X =$ binary (point-)treatment

Well-defined?

Beware of **immortal-time bias**:

$X =$ 'did patient ever receive drug ABC? (yes/no)'

$\longrightarrow$ not a point treatment!

Target trial:
define unique time of eligibility and treatment assignment!

(Total) Causal Effects

In words

Total marginal (or population) effect:

what is the overall effect of intervening in X on Y?

Target trial: randomise X, regression of Y on X

Contrast setting do$(X = 1)$ versus setting do$(X = 0)$ by some well-defined (but possibly hypothetical) intervention.

(Total) Causal Effects

Can consider subgroups

Total conditional (or subgroup) effect:

what is the overall effect of intervening in X on Y within a *subgroup*, e.g. women aged 50-60?

Target trial: restrict to subgroup, randomise X; regress Y on X.

Note: subgroups relevant if we expect effect heterogeneity
$\Rightarrow$ nothing to do with confounding!
Finding such subgroups: active research

(Total) Causal Effect

Will focus on:

Formally: average causal effect

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

or, with potential responses

$$ACE = E(Y(1) - Y(0)) = E(Y(1)) - E(Y(0))$$

aka: average treatment (ATE) or total causal effect (TCE), etc.

Key Assumptions

Consistency

Consistency Assumption:

If we observe $X^i = x$ then $Y^i = Y^i(x)$ (for individual i)

i.e. the outcome we observe under the observed treatment is the potential response had the treatment been *set* to what it was observed to be.

Violated if manipulation of X not well defined or so 'invasive' that observational setting not informative.

Example: X is 'BMI' — how to manipulate BMI itself?

Often: if violated, need more elaborate model and suitably detailed data.

Consistency

Under consistency and binary X:

$$Y^i = Y(1)^i X^i + Y(0)^i (1 - X^i)$$

Note:

consistency implicit in graphical / do$(\cdot)$ approaches

$\rightarrow$ invariance

No-Interference

Common assumption: **no-interference**:

Vector $\mathbf{x} =$ treatment values for **all** n units, then $Y^i(\mathbf{x}) = Y^i(x^i)$,

i.e. PR does not depend on treatment other units received.

Violation: e.g. vaccines, social networks.

Stable unit–treatment value (SUTVA):
consistency + no-interference.

No-Unmeasured-Confounding

Assumption of **no unmeasured confounding:**
(aka: random treatment assignment, or cond. exchangeability, ignorability, or ...)

Set C of observed (measured) pre–treatment covariates exists such that

$$Y(x) \perp\!\!\!\perp X \mid C$$

for all x to be considered as treatment values

Interpretation: within values of C, can consider X like randomised wrt Y

Denote: C is *sufficient* to adjust (control) for confounding;
or 'valid adjustment set'

Pre-Treatment Covariates?

What makes C **pre–treatment covariates**?

$\Rightarrow$ must be known not to be affected by intervention in treatment X!

Sufficient: C prior in time to X — but not necessary.

Often: C and X contemp. & share themselves common causes through past history, e.g. patient's medical history.

Graphically: C non-descendants of X.

(Overview: methods for causal covariate selection see Witte & Didelez, 2018)

Positivity Assumption (Overlap)

All methods for effect estimation essentially require

Assumption of positivity:

$$p(x \mid c) > 0 \text{ for all } x, c \quad (p(c) > 0)$$

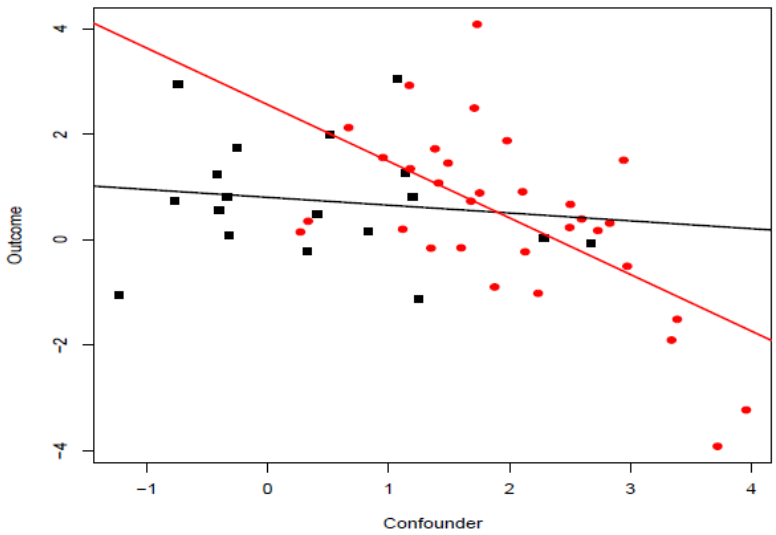
Interpretation: for all possible confounder values, it must be possible that a subject receives any value of treatment.

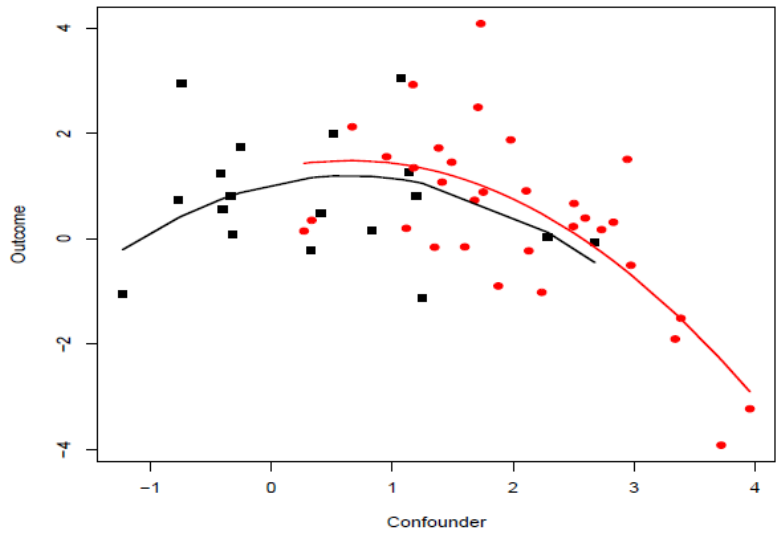
Positivity Assumption — Notes

- Positivity can be violated either by coincidence (small sample size), or **structurally**: certain combinations of C and X may not make sense!
- Target trial: can / should define eligibility criteria so that positivity is satisfied — requires domain knowledge
- (Lack of) positivity can be evaluated empirically (look at $p(x|c)$ or $p(c|x)$) — high-dim C becomes challenging Methods exist to characterise 'area of overlap' (Oberst et al, 2020)
- Do not include superfluous variables in C, especially: strong predictors of X that do not affect Y — can lead to apparent lack of positivity despite not being a problem

Positivity — Extrapolation

Regression-based approaches (based on fitting $p(y|x, c)$) may **mask** lack of positivity as regession models **extrapolate** $\Rightarrow$ can lead to vastly different causal effect estimates





Checking Assumptions?

- Consistency / no-interference:
  domain knowledge, study design
- No-unmeasured-confounding: compare analysis of observational data with actual randomised trial —
  Example: HRT-controversy
  Also: negative controls and similar designs
- Positivity (overlap) check:
  - basic: boxplot of each variable in C by treatment group;
  - advanced: consider **propensity score**, i.e. assess
    $P(X = 1|C = c)$
    obtain fitted values $\hat{p}^i = \hat{P}(X = 1|C = c^i)$ for each unit i,
    check $\hat{p}^i$ near zero in treated / controls, respectively.

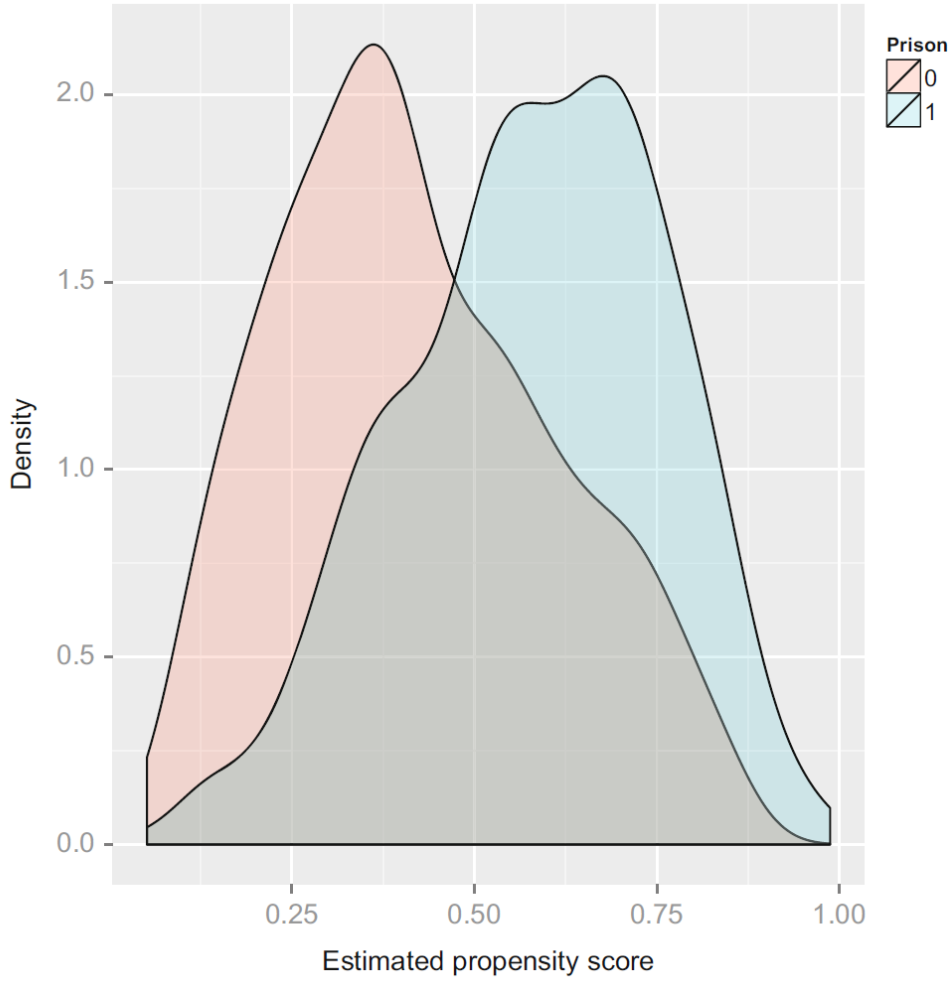
Propensity Score: Checking Positivity

Example:

$n = 1022$ offenders sentenced to either probation $X = 0$ or prison $X = 1$;
$C = 17$ covariates;
$Y =$ recidivism (yes/no);

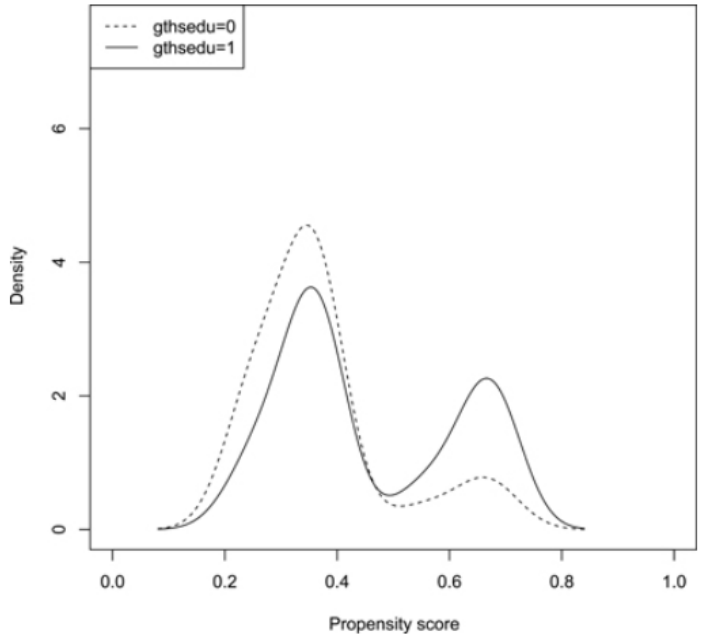
$\Rightarrow$ reasonable overlap.

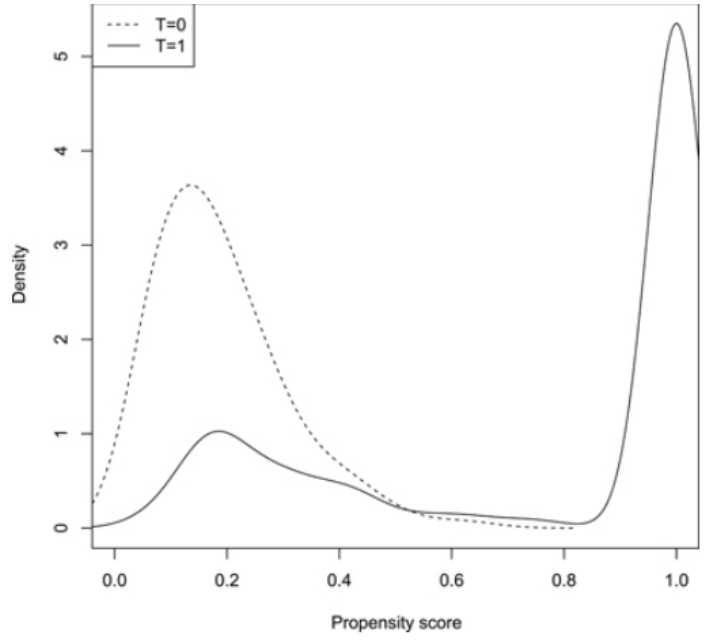
(Example taken from Guo et al., 2016)



Propensity Score: Checking Positivity

Which is good / bad overlap?

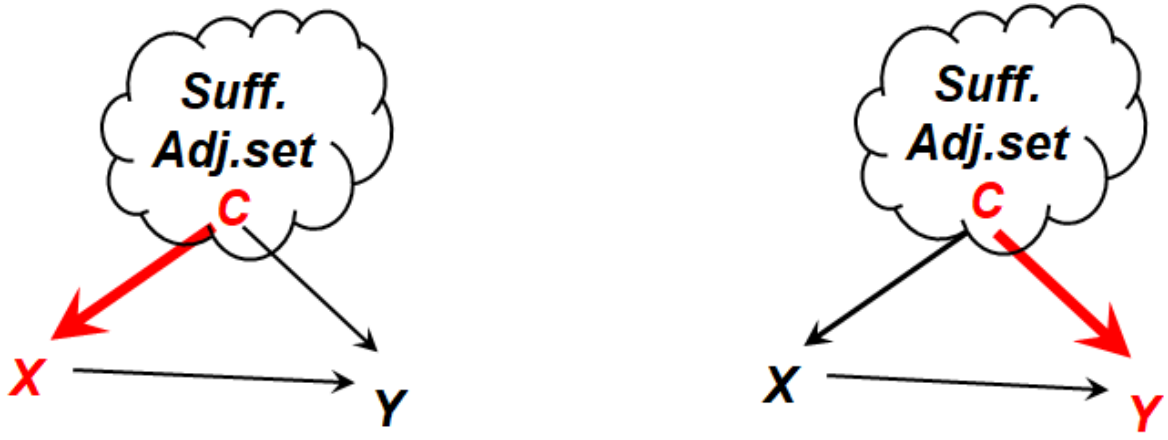




From: Brumback (2021, book)

Methods of Estimation

Based on treatment model $p(x|c)$ (propensity score)
or outcome model $p(y|x,c)$ or both



Principles:
– regression + standardisation,
– inverse-probability weighting (IPTW),
– stratification / matching,
– hybrid: doubly-robust estimation (double-ML)

Regression + Standardisation

aka: G-Formula *(Robins, 1986)*

Reminder: if C is sufficient set of covariates

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

An obvious way to use this is:

– fit flexible regression model for $E(Y|x, c; \beta)$ to data

– average over empirical C-distribution: $\sum_i E(Y|x, c^i; \hat{\beta})/n$

– R package `stdReg` (Sjolander and Dahlqwist, 2021)

Standardisation — Example

$Y =$ 'low birth weight' (binary); $X =$ 'mother smokes' (binary),
$C = \{$'age', 'race'$\}$ (Sjolander, 2016)

```
> fit2 <- glm(formula=lbw~(smoker+race+age)^2,
    family="binomial", data=clslowbwt)
```

flexible outcome model

Standardisation — Example

$Y =$ 'low birth weight' (binary); $X =$ 'mother smokes' (binary),
$C = \{$'age', 'race'$\}$ (Sjolander, 2016)

```
> fit2 <- glm(formula=lbw~(smoker+race+age)^2,
   family="binomial", data=clslowbwt)
> fit.std <- stdGlm(fit=fit2, data=clslowbwt, X="smoker",
   clusters="id")
> summary(fit.std)

  Estimate Std. Error lower 95 upper 95
0    0.279     0.0406    0.199    0.358
1    0.407     0.0555    0.298    0.516
```

flexible outcome model

standardised means
control / treatment groups

Standardisation — Example

$Y =$ 'low birth weight' (binary); $X =$ 'mother smokes' (binary),
$C = \{$'age', 'race'$\}$ (Sjolander, 2016)

```
> fit2 <- glm(formula=lbw~(smoker+race+age)^2,
   family="binomial", data=clslowbwt)
> fit.std <- stdGlm(fit=fit2, data=clslowbwt, X="smoker",
   clusters="id")
> summary(fit.std)

  Estimate Std. Error lower 95 upper 95
0    0.279     0.0406    0.199    0.358
1    0.407     0.0555    0.298    0.516

> summary(fit.std, contrast="difference", reference=0)

  Estimate Std. Error lower 95 upper 95
0    0.000     0.0000  0.00000    0.000
1    0.128     0.0681 -0.00544    0.262
```

flexible outcome model

standardised means
control / treatment groups

difference, i.e.
estimated ACE

Regression, but Why Standardisation?

Why not just look at coefficient ('effect') of X in a regression model for $E(Y|X, C)$?

- Marginal effect sensible summary also with arbitrary interactions / complex models
- Contrast of marginal $E(Y \,|\, \text{do}(X = x))$ corresponds to randomised trial where covariates C can be / are ignored
- Further issue: **non-collapsibility**! logistic regression / odds ratios not collapsible.
    - If set of sufficient covariates C not unique, cond. effects may depend on choice of C, but not marginal ones.

Regression + Standardisation

- Consistency (asy. unbiasedness) of estimation relies on correctly specified model for $p(y|x, c)$.
- Danger of extrapolation: it can happen that the regression relation $p(y|x, c)$ is determined primarily by treated subjects in one region of C and control subjects in another...
- ... should not happen under positivity — must be checked!

Regression + Standardisation

- The method is special case of **G–formula** for sequential treatments (Robins, 1986).
- Population effect $E(Y \mid \text{do}(X = x))$ depends on *distribution* $p(c)$ of covariates target population

  $\Rightarrow$ not necessarily the same in different populations (e.g. age distribution). If $p(y|x, c)$ regarded as 'stable' across populations, then can just replace $\hat{p}(c)$ in the above by different covariate distribution for different populations

  (e.g. UK versus USA covariate distribution).

  $\Rightarrow$ 'Transportability'

Partial Dependency Plots

(Zhao & Hastie, 2021)

- Close relation between Friedman's **partial dependence plot (PDP)** for visualising black-box prediction methods and back-door adjustment / standardisation
- Under the causal assumptions, can interpret PDP like ACE even for *continuous* treatment $\approx$ standardising over adjustment set C
- But: positivity hard to justify for entire range of a continuous treatment...
- PDP as basis for estimation of total causal effect very unstable and erratic asymptotic behaviour
  $\Rightarrow$ double-machine-learning! (later)

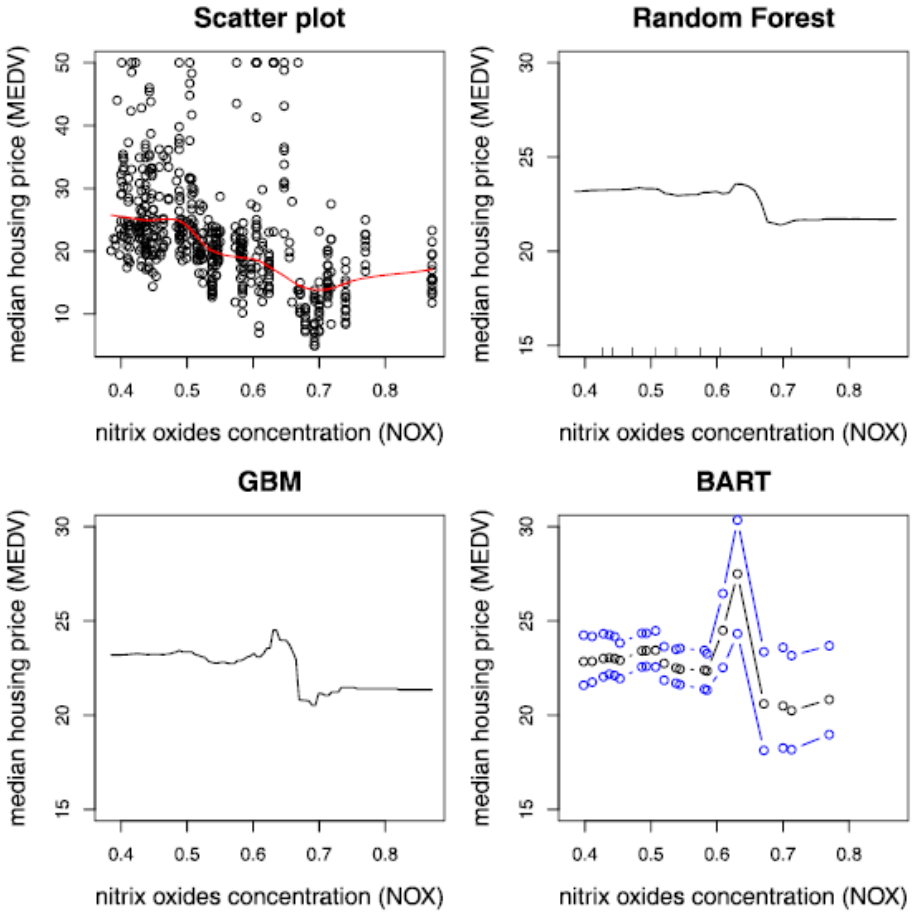
Partial Dependency Plots

(Zhao & Hastie, 2021)

Effect of nitrix oxides concentration on median value of owner-occupied homes

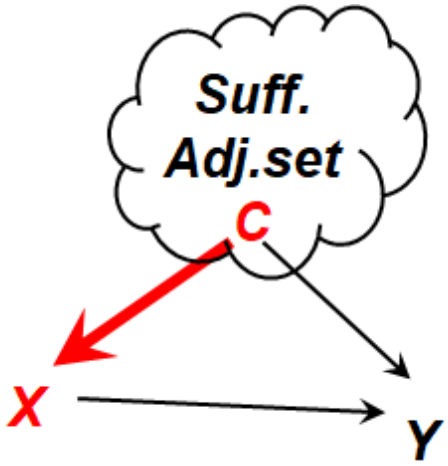
‘adjusted’ for: crime rate,
prop. residential/industrial zones,
av. # of rooms per dwelling,
age of the houses,
distance to city / highways,
pupil-teacher ratio,
% of blacks and
% of lower class

— valid adjustment set?



Treatment Modelling Approaches

The following methods are all based on models for X given C instead of modelling Y given C.



Marginal Structural Models (MSMs)

The adjustment formula

$$\sum_c p(y|x,c)p(c) \quad \text{or} \quad \sum_c E(Y|x,c)p(c)$$

might be 'awkward', $E(Y|x,c)$ non–linear with interactions, or C high dimensional and/or partly continuous.

$\Rightarrow$ Parameterise $E(Y \,|\, \text{do}(X = x))$ itself?!

$\Rightarrow$ **Marginal structural models (MSM)**

$\Rightarrow$ fitted by inverse probability of treatment weighting (IPTW)

Marginal Structural Models

(Hernán et al, 2001)

MSM: *semiparametric* model for

$p(y \mid \text{do}(X = x))$ or more typically $E(Y \mid \text{do}(X = x))$

e.g. linear, logistic, CoxPH, loglinear, probit etc.

Marginal: refers to time-varying covariates $\rightarrow$ not covered

Structural: model under intervention in X (not observational)

Note: term 'structural' is used in many different ways — here it always refers to modelling the underlying causal relationships.

Inverse Probability Weighting

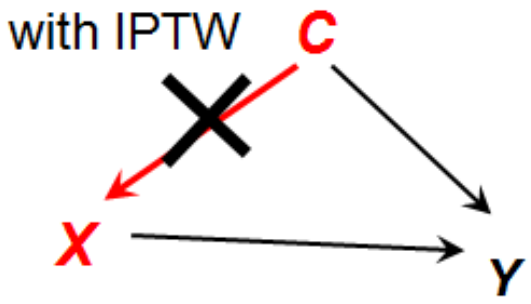
(IPTW)

Idea is based on:

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

$$= \sum_c p(y|x, c)p(c)$$

$$= \sum_c \frac{p(y,x,c)}{p(x|c)}$$



In words: the population is re-weighted so that X becomes independent of C.

Covariate balance check:
success of re-weighting can be assessed empirically in data!

Inverse Probability Weighting

(IPTW)

Idea is based on:

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

$\Rightarrow$ fit MSM with individuals' weights $w^i = p(x^i|c^i)^{-1}$

$\Rightarrow$ creates 'pseudo sample' in which C is not confounding

$\Rightarrow$ unbiased estimating equations for parameters of MSM $E(Y \,|\, \text{do}(X = x); \beta)$

Here, $w^i = p(x^i|c^i)^{-1}$ is the inverse of the probability that individual i receives 'treatment' x^i given they have covariates c^i.

IPTW Estimator

Define $\pi(c) = P(X = 1|C = c)$ — propensity score.

Can show (under our assumptions):

$$E\left(\frac{X}{\pi(C)}Y\right) = E(Y \mid \text{do}(X = 1))$$

and similarly

$$E\left(\frac{1-X}{1-\pi(C)}Y\right) = E(Y \mid \text{do}(X = 0))$$

Proof: iterated conditional expectation.

IPTW Estimator

With model $\pi(C;\alpha) \Rightarrow$ plug-in $\pi(C;\hat{\alpha})$

IPTW yields consistent estimator for ACE

- if $\pi(C;\alpha)$ correctly specified
- can obtain sandwich standard errors or bootstrap, or theoretical asymptotical standard errors
- IPTW often large variance, wide CIs — reflects lack of information in areas with 'extreme weights'
- ... extreme weights indicate possible near violation of positivitiy
- Solution: restriction of relevant population and / or truncation of weights (e.g. at 99%-percentile).

MSM / IPTW — Implementation

Easy to implement with standard software for regression models by specifying weights

```
p.i <- glm(x.trt~c1*c2*c3,family="binomial")$fitted
w.i <- 1/(x.trt*p.i+(1-x.trt)*(1-p.i))
msm.out <- glm(y.out~x.trt,family="binomial",weights=w.i)
Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -2.02055    0.01270 -159.118   <2e-16
x.trt        0.13202    0.01753    7.532    5e-14
cov.out <- sandwich(msm.out)
sqrt(cov.out[2,2])  [1] 0.03218826
```

Note: default standard errors ignore variability in (estimated!) weights

$\Rightarrow$ sandwich st.error

Notes on IPTW

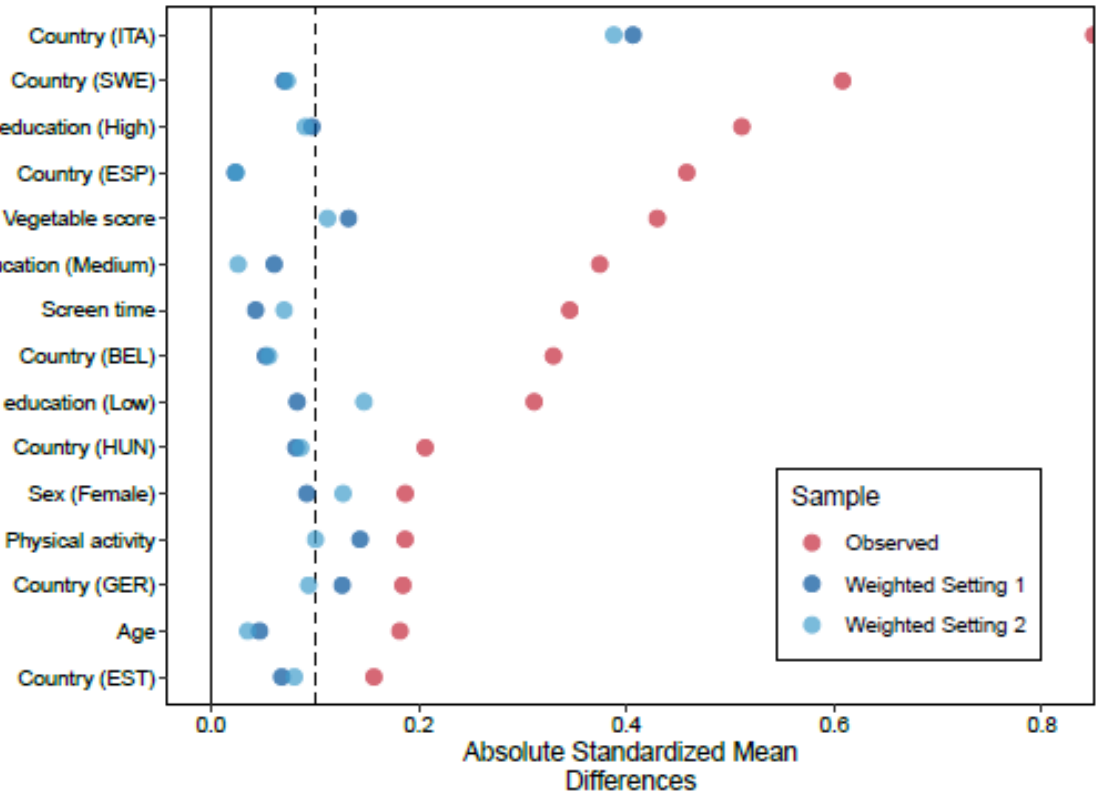
- Consistent when both, models for $E(Y \mid \text{do}(X = x))$ and $\pi(c)$ correctly specified...
- ... avoids modelling of Y-C relation
- Extension: 'overlap weights' — **new estimand** with most weights on subpopulations with higher overlap
- IPTW especially useful when study design (or other) supplies background knowledge to model weights $p(x|c)$.
- Problem: estimation of weights $p(x|c)^{-1}$ not obvious when X continuous.

Notes on IPTW

- MSM mimics a model for Y given X in the situation of a trial where X was randomised
- Even if X *actually* randomised, there may be changes to treatment status over time (non-adherence) — this then becomes a problem of **time-dependent treatment**
- MSMs with IPTW mostly used in longitudinal situations / time-dependent / sequential treatments with time-varying confounding
  $\Rightarrow$ 'marginal' over time-dep. confounders / covariates.

Checking Assumptions: Balance

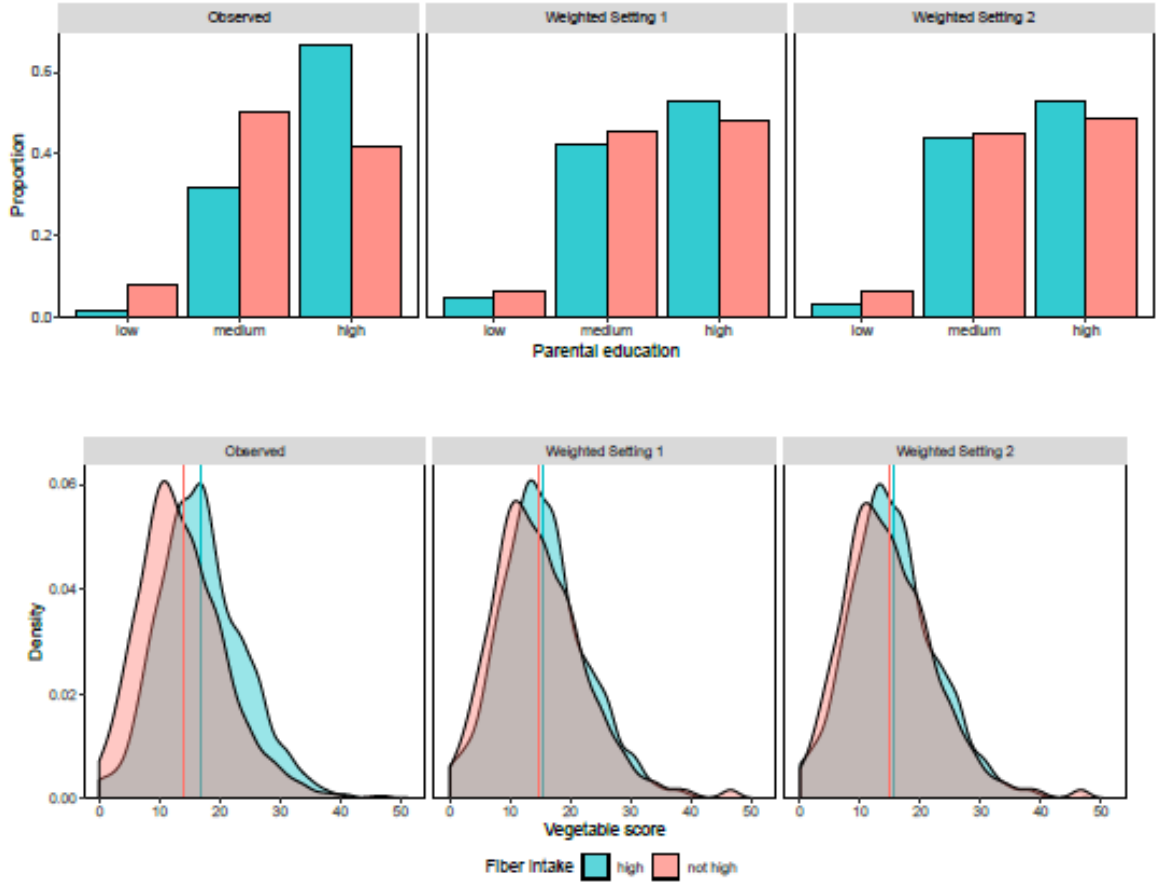
Example: causal effect of ‘fibre intake’ on children’s BMI; large adjustment set (country, parental edu, vege-score, etc.)
Plot: std.mean-diff b/w re-weighted treated / untreated sample



(Nöhren, 2021, MA thesis)

Checking Assumptions: Balance

Example(ctd.), checking balance of whole distribution of covariates



Double Robustness (DR)

(Robins & Rotnitzky, 2001)

- Regression-standardisation relies on correct outcome model
- MSM / IPTW relies on correct treatment model
- Danger (especially with high-dim C): models will be 'misspecified'

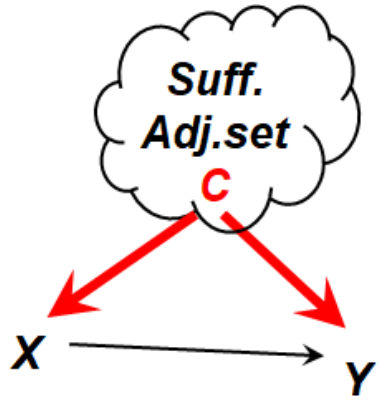
  $\Rightarrow$ want to fit them data-adaptively

  $\Rightarrow$ known to yield unstable (irregular) effect estimators!

  $\Rightarrow$ Better: **double-machine learning** of causal effects based on doubly-robust estimation.

Double Robustness (DR)

(Robins & Rotnitzky, 2001)



Can find class of augmented IPTW (AIP(T)W) estimators:

consistent for ACE if

- either model $\pi(C;\alpha) = p(X = 1 \mid C;\alpha)$ correctly specified

  or

- model for $p(y \mid x, c;\beta)$ correctly specified;
- but one of them can be wrong.

AIPTW Estimator

Basic Idea

Still binary exposure X
For *arbitrary* functions $m(C)$ and $\pi(C)$, define

$$\hat{\mu}_1 = m(C) + \frac{X}{\pi(C)}(Y - m(C)) = \frac{X}{\pi(C)}Y + \left[1 - \frac{X}{\pi(C)}\right]m(C)$$

Property (under our assumptions): if

either $m(C) = E(Y \mid X = 1, C)$ or $\pi(c) = P(X = 1|C = c)$, then

$$E(\hat{\mu}_1) = E(Y \mid \text{do}(X = 1))$$

Analogously for $\hat{\mu}_0$ and $E(Y \mid \text{do}(X = 0))$.

AIPTW Comments

- Kang and Shafer (2007) find: AIPTW with parametric models can be quite bad if *both models slightly misspecified* $\rightarrow$ much research on improvements
- **NEW:** Statistical properties of doubly-robust estimators allow the use of machine learning for treatment and outcome models
  - $\Rightarrow$ DR minimises slow convergence rates / overfitting typical for machine learning with sample-splitting or cross-fitting
- `AIPW` R package (Zhong et al., 2021) or `npcausal` R package (Kennedy, 2021)

'Demystifying statistical learning based on efficient influence functions', Hines et al (2021; arXiv:2107.00681)

Super Learner??

- To fit $m(C)$ and $\pi(c)$ can use data-adaptive methods developed for prediction!
  - `AIPW` package uses the Super Learner
  - ... an ensemble method allowing combination of several prediction algorithms into one
  - ... uses k-fold cross-validation to build the optimal weighted combination of predictions from a library of candidate algorithms
    – choice of library quite important (active research)
- double machine learning methods avoid strong modeling assumptions

  ... and can still achieve optimal $\sqrt{n}$ rate of convergence for causal effect estimation under *some* conditions

Propensity Score (PS) — Other Usage

(Rosenbaum & Rubin, 1983)

Have used $\pi(c) = P(X = 1|C = c)$

$\Rightarrow$ also known as **propensity score**.

Note: $\pi := \pi(C)$ is random variable.

MSM: used $\pi(C)$ for weighting.

But: can also use $\pi(C)$ for adjustment-type approaches, due to it being a balancing score...

Propensity / Balancing Score

(Still assuming: C sufficient set of covariates; X binary.)

Use of propensity scores (vs. IPTW) is based on

$$X \perp\!\!\!\perp C \,|\, \pi \qquad \text{i.e. } \pi \text{ balances } C$$

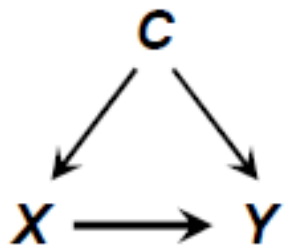
Hence (with properties of C):

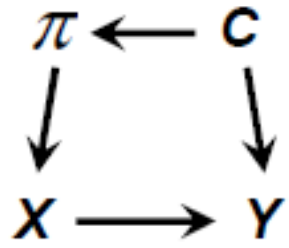
$$Y(x) \perp\!\!\!\perp X \,|\, \pi$$

making π a minimal sufficient reduction of C (univariate $\in [0, 1]$).

Propensity Score — Graphically

Propensity score $\pi := \pi(C) = P(X = 1|C)$ satisfies these conditional independencies:





Left: assumption of C being sufficient set of covariates.

Right: π is deterministic function of C and $X \perp\!\!\!\perp C \,|\, \pi$.

Propensity Score in Practice

- Estimate propensity score $\hat{\pi}$ with model for $\pi(C; \alpha)$.
- Required: correctly specified model $\pi(C; \alpha)$.
  Non-parametric approaches: random forests etc.
- Note: predictive quality of $\pi(C; \alpha)$ for X not important because need C to be (X, Y)-confounders, not nec. strong predictors of X.

  In fact: strong X-predictors $\Rightarrow$ bias amplification

  (Pearl, 2011).
- Check balancing property (see IPTW).
- Check positivity / overlap $\Rightarrow$ if necessary: restrict / prune!

Propensity Score in Practice

Methods for using PS (other than weighting):

- PS **stratification**: divide into strata (often quintiles) and fit $p(y|x)$ to each stratum separately.

  (Strata specific effects can reveal effect modification.)

  Then weighted average to obtain overall population effect.

- Alternative: **matching** on propensity score, i.e. match each treated with k untreated with similar propensity score — this estimates effect of treatment on the treated (ETT)!

- Sometimes: **PS adjustment** — specify model for $p(y|x, \pi)$ and fit with $\hat{\pi}$ plugged in.

- Extrapolation is automatically avoided.

Survival of Cancer Patients

Example

US National Cancer Institute's SEER data base; observational study.

Covariates: year of diagnosis, tumor size, geogr. registry, race, marital status

Propensity score	Treatment	No.	5-Year-Surv.	Difference
1st quintile	A	56	85.6%	
	B	1008	86.7%	-1.1%
2nd quintile	A	106	82.8%	
	B	964	83.4%	-0.6%
3rd quintile	A	193	85.2%	
	B	866	88.8%	-3.6%
4th quintile	A	289	88.7%	
	B	978	87.3%	1.4%
top quintile	A	462	89.0%	
	B	604	88.5%	0.5%

Overall estimated (weighted average) $ACE = -0.68$.

From strata specific results: slight suggestion that treatment B is better for those who are more likely to receive it.

Notes on Propensity Score

- Best with binary treatment / exposure.
- PS stratification consistent if $\pi(C;\alpha)$ correctly specified, but can be markedly biased due to residual confounding within strata possible.
- Consistency can be achieved by increasing number of strata when sample size is 'large' or by additional modelling of $E(Y|C=c;\mathrm{do}(X=x))$ within strata.
- PS popular especially for matching: π is 'one-dimensional reduction' of covariates — but at cost of first modelling / estimating $\pi = p(x|c;\alpha)$.
- PS matching / stratification not really suitable for sequential treatments.

Notes on Propensity Score

- Danger: modelling $\pi(C;\alpha)$ may focus on strong predictors of $X \Rightarrow$ can amplify bias! $\Rightarrow$ selection of C as adjustment set should be separate process from fitting $\pi(C;\alpha)$.
- Interpretation of PS-analyses sometimes regarded as difficult compared to actual covariate values.
- Simulations suggest that IPTW with π superior to stratification. (Lunceford & Davidian, 2004)
- Critique of PS-matching: King & Nielsen (working paper)

Estimating Causal Effects

Summary (no unobs. conf.)

Given suff. adjustment set C (& other structural asspts):

- Traditional: regression adjustment to be supplemented by...
- ... standardisation to obtain population effect (g-formula in time-varying context) – underused in practice
- or MSMs fitted by IPTW – easy to use, also with time-varying data – but can be inefficient
- propensity score methods (stratification / matching) – overused?
- Combination leads to *doubly robust* estimation procedures $\Rightarrow$ promising new methods use double-machine learning
- Always check positivity/overlap & balance with **all** methods!

Validation?

- Causal (counterfactual) conclusions from observational data cannot be validated on that same data!
  - Can check balance on *observed* confounders, but more important for *unobserved* factors
  - Need: experimental validation on different / new (ideally experimental) data $\longrightarrow$ some example tomorrow
- Recent studies: compare randomised trials with real-world (observational) studies

$\Rightarrow$ Often: evidence that much bias is due to inappropriate analysese, more than to lack of randomisation

Validation?

- Helpful: compare very different methods for estimating the same estimand

  $\Rightarrow$ if not in agreement some assumptions are violated
- Sometimes: different study designs can be used to check if same conclusion is obtained
  - Natural experiments, pragmatic trials ...
- **Negative controls:** similar exposure or similar outcome with same source of confounding but known zero-effect $\Rightarrow$ assess unobserved confounding
- Instrumental variables — topic of its own...

Quantitative Bias Analysis

- Can investigate:
  "How much would our conclusions change if there was an unobserved confounder with certain properties"
  $\Rightarrow$ Sensitivity / bias analysis (Lash et al, book)
- Formal approaches: based on Bayesian models (Greenland, Handbook Epidemiology chapter!)
- Ad-hoc method: **E-value** (Ding & Vanderweele)
  "How strongly must an unobserved confounder be associated with X and Y to explain away the causal effect (in the worst case)?"

Extensions / Outlook

- Multiple treatments (X has more than two levels), continuous treatments (positivity?)
- Different interventions: nudging / shift-interventions
- Different estimands: effect of treatment on the treated, (in)direct effects, 'principal-stratum' effect etc.
- Different outcomes: survival / time-to-event (censoring), multivariate outcomes
- Sequential / time-dependent treatments (dealing with 'switching', 'when-to-start?')
  $\longrightarrow$ time-dependent confounding!
- Effect heterogeneity, individualised / adaptive / optimal treatments $\longrightarrow$ (optimal) dynamic treatments

Appendix to Part 2:
Estimating a Causal Effect

Regression

Question: When, if at all, are coefficients (specifically coefficient β_X of exposure X) in regression models (linear, logistic, ...) estimating causal effects?

If at all, what causal effects are they estimating?

(How) Does the type of regression model matter?

Note: in general, *causal target* of inference does *not need to be and is not* a specific *parameter in a parametric model*.

Table-2 Fallacy?

American Journal of Epidemiology
© The Author 2013. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
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;

Regression ctd.

Assume GLM:

$$\eta(E(Y|X,C)) = \beta_0 + \beta_X X + \beta_C C$$

Note: here assumed no (X, C)-interaction term!

– linear regression: $\eta =$ identity function

– logistic regression: $\eta =$ logistic function

– etc.

Regression ctd.

Assume GLM: $\eta(E(Y|X,C)) = \beta_0 + \beta_X X + \beta_C C$.

If: C valid adjustment set (& positivity, consistency), model correctly specified, then

$$\beta_X = \eta(E(Y|C, \mathsf{do}(X = x+1))) - \eta(E(Y|C, \mathsf{do}(X = x)))$$

or

$$\eta^{-1}(\beta_X) = E(Y|C, \mathsf{do}(X = x+1)) - E(Y|C, \mathsf{do}(X = x))$$

– conditional effect — given C

– must take scale η into account

– more complicated when (X, C)-interaction terms.

Regression ctd.

Caveats:

- in linear no-interaction model: conditional = marginal causal effect; i.e. if η identity, then $\beta_X = E(Y|\text{do}(X = x + 1)) - E(Y|\text{do}(X = x)) = ACE$;
  $\Rightarrow$ here β_X **collapsible**;
- but β_X not collapsible in logistic model (or e.g. Cox model)
  $\Rightarrow$ effect conditional on C not the same as marginal;
- typically cannot make causal assumptions about (some or all elements of) $C \Rightarrow \beta_C$ has no causal interpretation;

Regression ctd.

Multiple exposures?

Often in applications not a clear distinction between single exposure and covariates used for adjustment.

First: be clear about causal question relating to multiple exposures — what would be your ideal target trial?

Causal Parameters with multiple exposures:

– (controlled) direct effects;

– joint effects;

– strategy for sequential treatments.

Regression ctd.

Controlled Direct Effect

Consider two 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:
presupposes that X_2 is possibly mediator for X_1 on Y effect.

Regression ctd.

Joint intereventions:

Consider two exposures X_1, X_2.

Joint interventions means: find effect of combination of X_1, X_2-values, e.g.

$$\text{parameters for } E(Y|C, \text{do}(X_1 = x_1, X_2 = x_2))$$

But:

'no-unobserved-confounding' assumption for direct effects, or general multiple / sequential interventions more complicated and not covered in detail here.

$\Rightarrow$ e.g. see time-varying confounding later.

Regression ctd.

Multiple exposures — some examples

Consider two exposures X_1, X_2 and linear model:

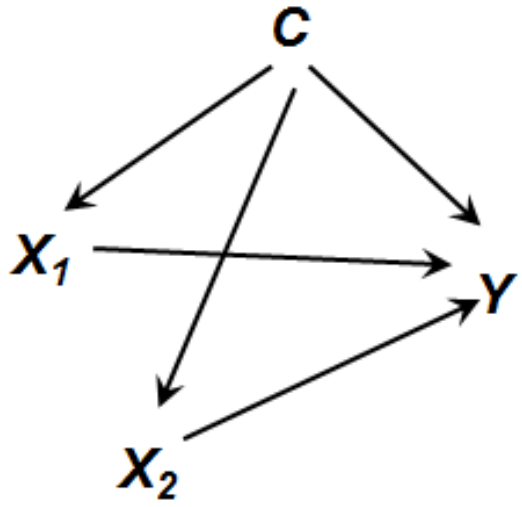
$$E(Y|X_1, X_2, C) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_C C$$

Caveat:

meaning of β_1, β_2 depends on
– causal structure between X_1 and X_2
– whether C suff. to adjust for confounding for both exposures.

Regression ctd.

Ideal situation (rare...):

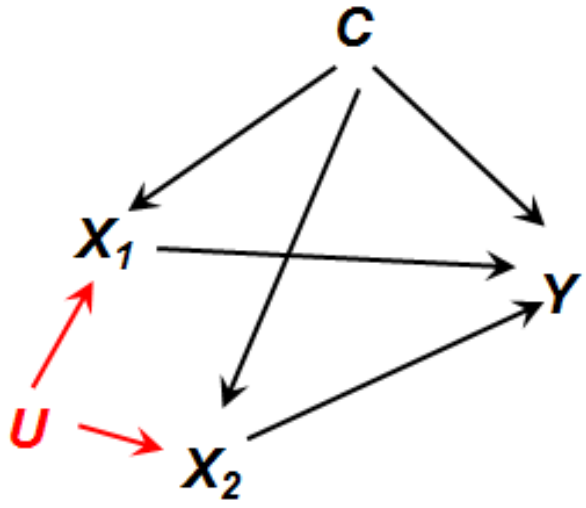


$\Rightarrow \beta_1, \beta_2$ represent total causal effects (individually or jointly) given C.

Note: here, C adjustment set for both X_1, X_2.

Regression ctd.

Associated exposures:

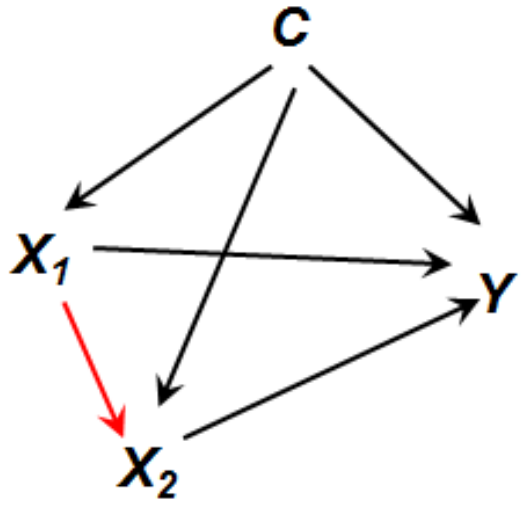


Still: β_1, β_2 represent total causal effects (individually or jointly) given the other exposure and C.

Note: here, (C, X_1) adjustment set for X_2 while (C, X_2) adjustment set for X_1.

Regression ctd.

Causally ordered exposures:



β_2 (total) causal effect conditional on C, X_1

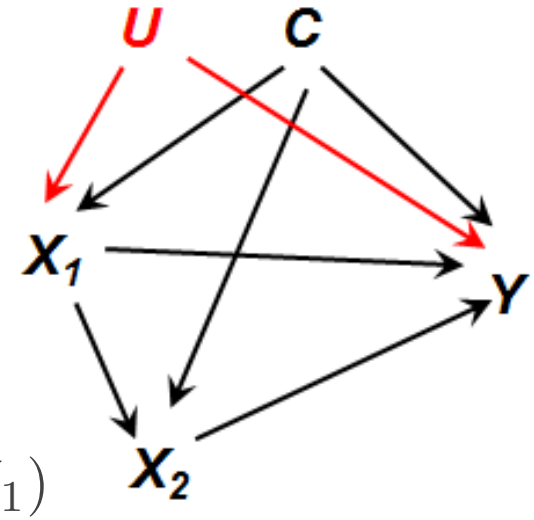
Note: here, (C, X_1) adjustment set for X_2

β_1 CDE of X_1 on Y while fixing X_2

Note: β_1, β_2 not the same kind of causal effect.

Regression ctd.

Some unobserved confounding:



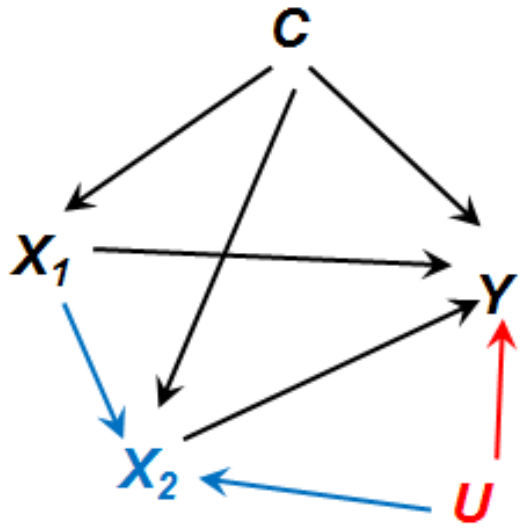
β_2 (total) causal effect conditional on (C, X_1)

Note: here, (C, X_1) adjustment set for X_2

But β_1 confounded! no causal interpretation.

Regression ctd.

Some unobserved confounding:



Neither β_1 nor β_2 causally meaningful!

β_2 obviously confounded.

β_1 affected by selection bias due to conditioning on collider X_2.

Note: Total effect of X_1 can be obtained as α_1 from a simple regression $E(Y|X_1, C) = \alpha_0 + \alpha_1 X_1 + \alpha_C C$ — not confounded!

Regression — Summary

Single exposure

- Under causal assumptions (all other variables in model must be pre-exposure and valid adjustment set etc.) ...
- ... then regression coefficient (suitably transformed) is total conditional effect (if no-interaction and model correct).
- Caveat: conditional not always the same as marginal effect.
- For comparison with RCT or for population effects: may prefer marginal effect $\Rightarrow$ later.

Regression — Summary

Multiple exposures:

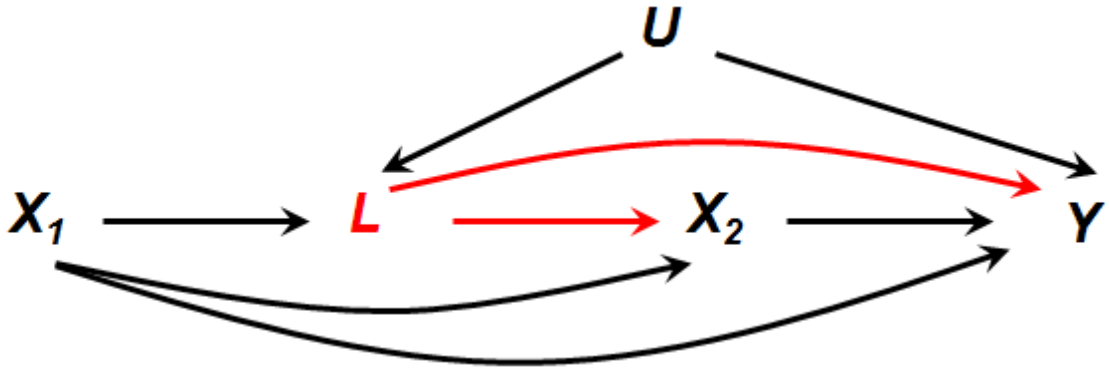
- Much care must be taken due to subtle issues relating to
  - causal ordering of exposures
  - unobserved confounding affecting some but not all exposures
  - interpretation in terms of joint, direct or total effects.
- More issues relating to the interpretation of interactions / effect modification etc. (not covered here).

$\Rightarrow$ See “Table-2 Fallacy” paper.

Note: Multiple regression with variable selection sometimes used to ‘discover’ (direct) causes of Y — all the above issues apply! (But see paper by Peters et al, 2015)

IPTW – Sequential Treatments

Time-varying confounding - Problem 1: L is confounder for X_2, so must adjust for L to obtain correct effect of X_2.



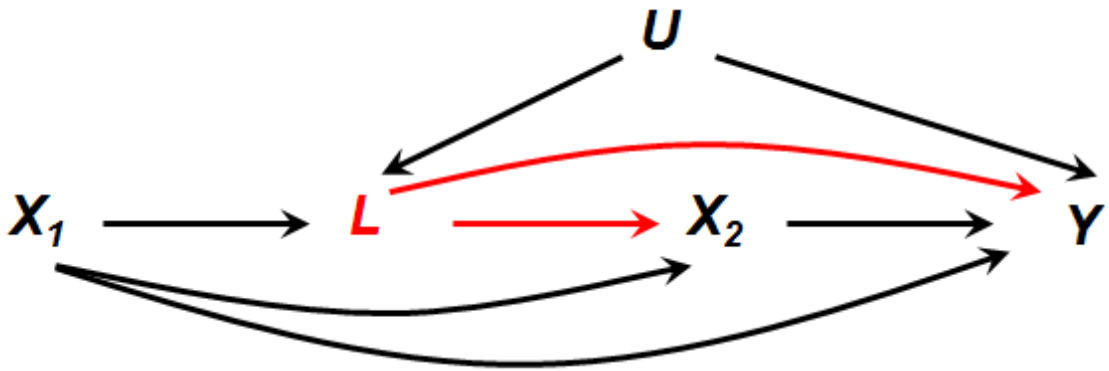
$\Rightarrow$ regression $Y \sim X_1 + X_2$ inappropriate

Time-Varying Confounding

Example: $X_1 =$ initial treatment (A or B);

$L =$ occurrence of side-effect / adverse reaction (cannot be manipulated / fixed);

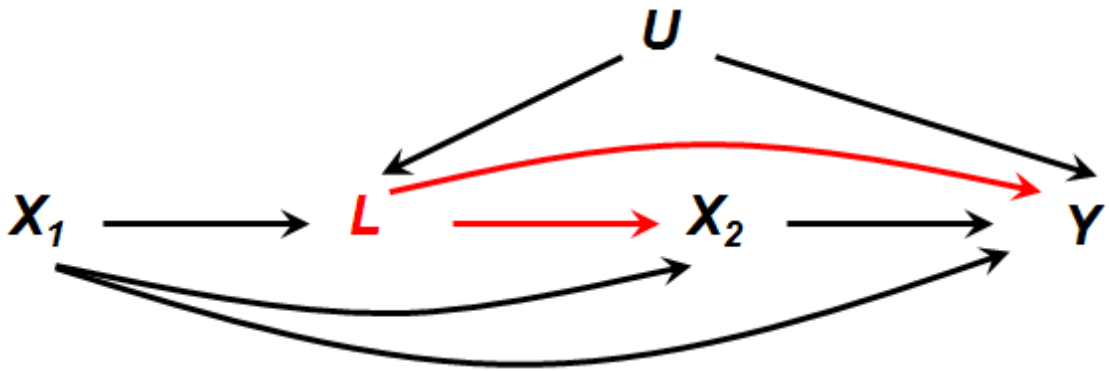
$X_2 =$ switching treatment (yes or no).



Time-Varying Confounding

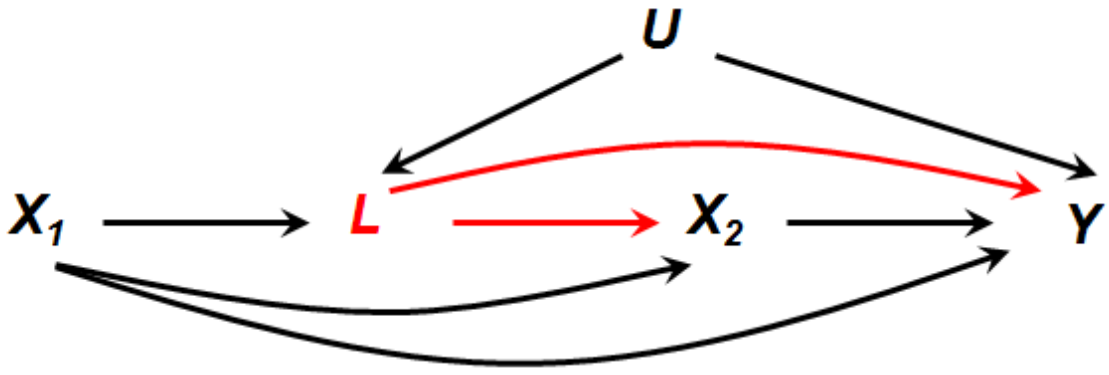
Wanted: causal effect in terms of e.g.
$E(Y \mid \text{do}(X_1 = x_1, X_2 = x_2))$

$\Rightarrow$ **no one regression involving observables appropriate!**



Time-Varying Confounding

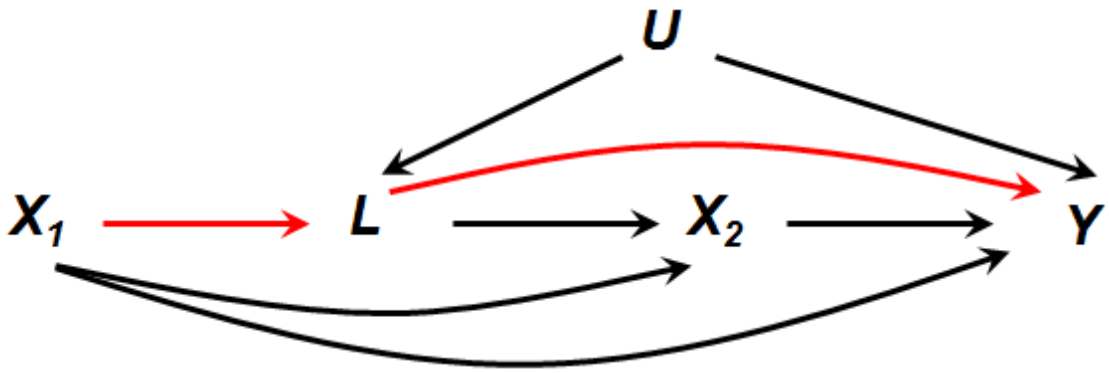
Problem 1: L is confounder for X_2, so must adjust for L to obtain correct effect of X_2.



$\Rightarrow$ regression $Y \sim X_1 + X_2$ inappropriate

Time-Varying Confounding

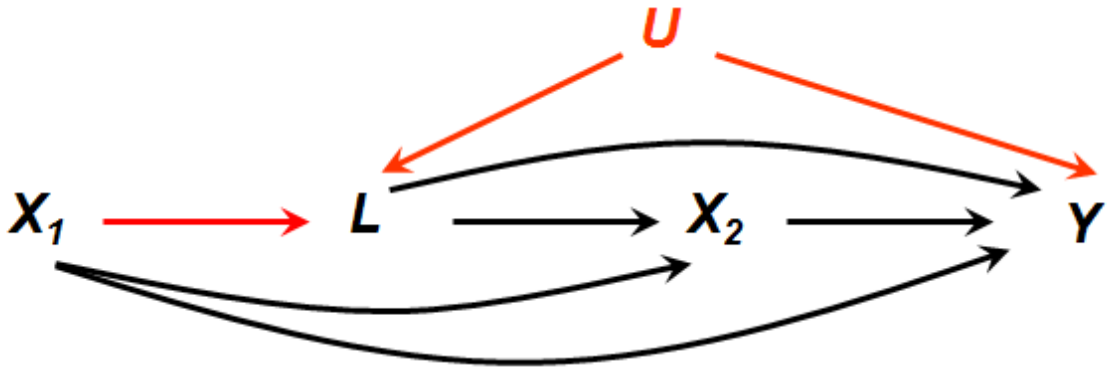
Problem 2: L is on causal pathway of X_1, so must not adjust for L for total effect of X_1.



$\Rightarrow$ regression $Y \sim X_1 + X_2 + L$ inappropriate

Time-Varying Confounding

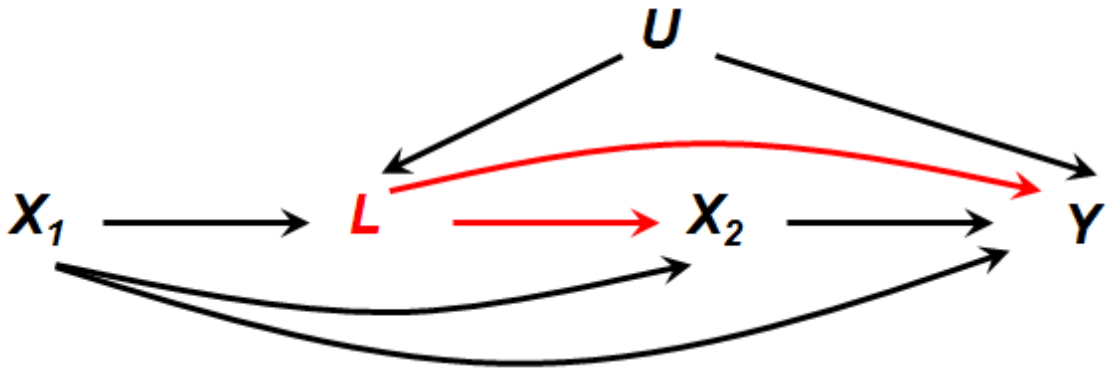
Problem 3: L is collider on path between X_1 and Y, so must again not adjust for L.



$\Rightarrow$ regression $Y \sim X_1 + X_2 + L$ inappropriate

Time-Varying Confounding

Despite three problems, $E(Y \mid \text{do}(X_1 = x_1, X_2 = x_2))$ is identified from data on X_1, X_2, Y, L in this situation
$\Rightarrow$ g-formula (Robins, 1986)



G-Formula

aka G-Computation etc.

In this example:

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

$$\sum_{l} E(Y \mid X_1 = x_1, L = l, X_2 = x_2) P(L = l \mid X_1 = x_1)$$

(possibly all conditional on pre-X_1 baseline covariates C)

Note: time-varying confounder L used, but integrated out!

Assumption: 'no unobserved time-varying confounding'

No Unobserved Time-Var. Confounding

Here, in our example (C baseline covariates):

$$Y(x_1, x_2) \perp\!\!\!\perp X_1 \mid C, \qquad Y(x_1, x_2) \perp\!\!\!\perp X_2 \mid (X_1 = x_1, L, C)$$

Can check graphically: Pearl & Robins (1995)

Decision theoretically (no counterfactuals): Dawid & Didelez (2010)

Methods – Overview

Methods for sequential (dynamic) treatments:

(e.g. Orellana et al, 2010)

- parametric g-formula
- inverse-probability of treament weighting – IPTW for MSMs;

  in survival, e.g. 'cloning' and artificial censoring

  (Gran et al, 2010)
- G-estimation e.g. for accelerated failure-time models
- double-robust estimation
- optimal dynamic treatments (Murphy, 2003)

MSMs for Sequential Treatments

Consider two time-ordered treatments / exposures X_1, X_2.

MSM: semiparametric model for

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

or for hazard function, e.g. prop. hazards MSM (Cox-MSM)

Marginal: over post-X_1 and pre-X_2 covariates L

Structural: model under intervention (not observational)

$\Rightarrow$ need time-dependent IPTW!

IPTW – Sequential Treatments

In 2-treatments setting, Y measured after (X_1, X_2):

IPTW for sequential treatments X_1, X_2 (let C baseline confounders)

$$\text{weights for } i \;=\; \frac{1}{p(x_1^i|c^i)}\,\frac{1}{p(x_2^i|l^i, x_1^i, c^i)}$$

$\Rightarrow$ weighted regression of Y on X_1, X_2.

IPTW – Sequential Treatments

Hazard / survival models, Y (possibly censored) survival time:

IPTW for time-varying weights (let C baseline confounders)

$$\text{weights at } t \;= \prod_{s=1}^{t} \frac{1}{p(x_s^i | \bar{l}_s^i, c^i)}$$

where $\bar{l}_s^i$ values of time-dep. obs. confounders before s.

$\Rightarrow$ regression of Y on $\bar{X}_t$ with time-varying weights.

Notes on IPTW

Similar to single exposure/treatment case

- Models for ($X_t \mid$ 'past') must be correctly specified;
- in practice use stabilised weights;
- generalisation to continuous time exist; (Røysland, 2011)
- also: optimal dynamic treatments with *Q-Learning*.
  (Chakraborty & Moodie, 2013)

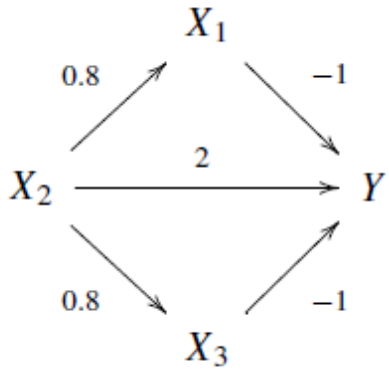
Part 3: Causal Discovery

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

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

$\Rightarrow X_2$ causally least important.

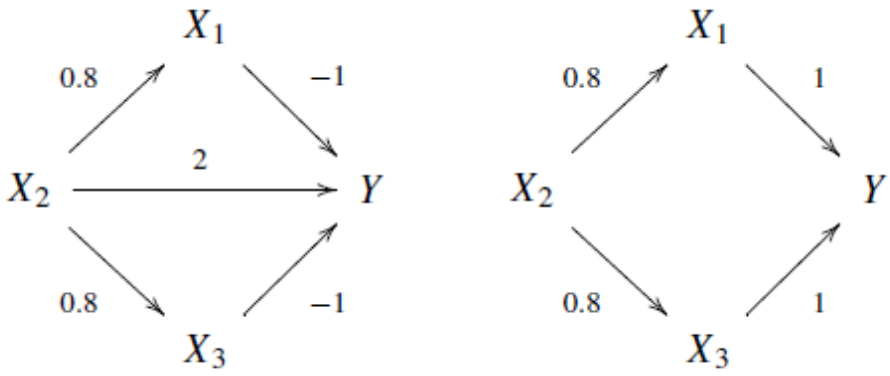
(Here linear structural equation models, LSEM)

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)

Causal Discovery

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

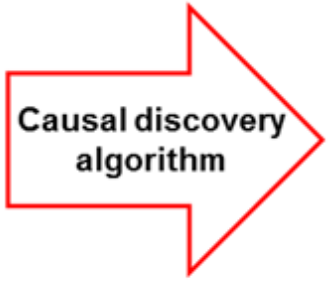
Causal discovery is about **finding** a causal graph when there is no (sufficient) causal background knowledge.

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

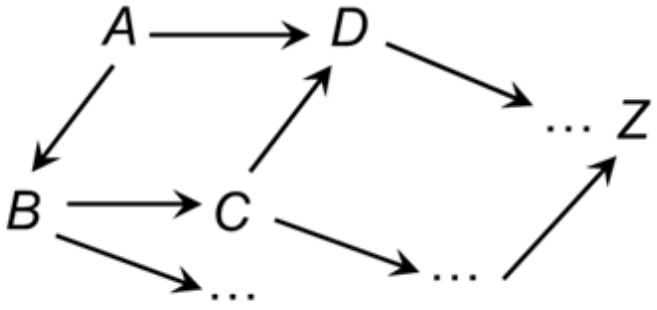
Causal Discovery

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**

$\Rightarrow$ 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. Maybe generate some causal hypotheses.

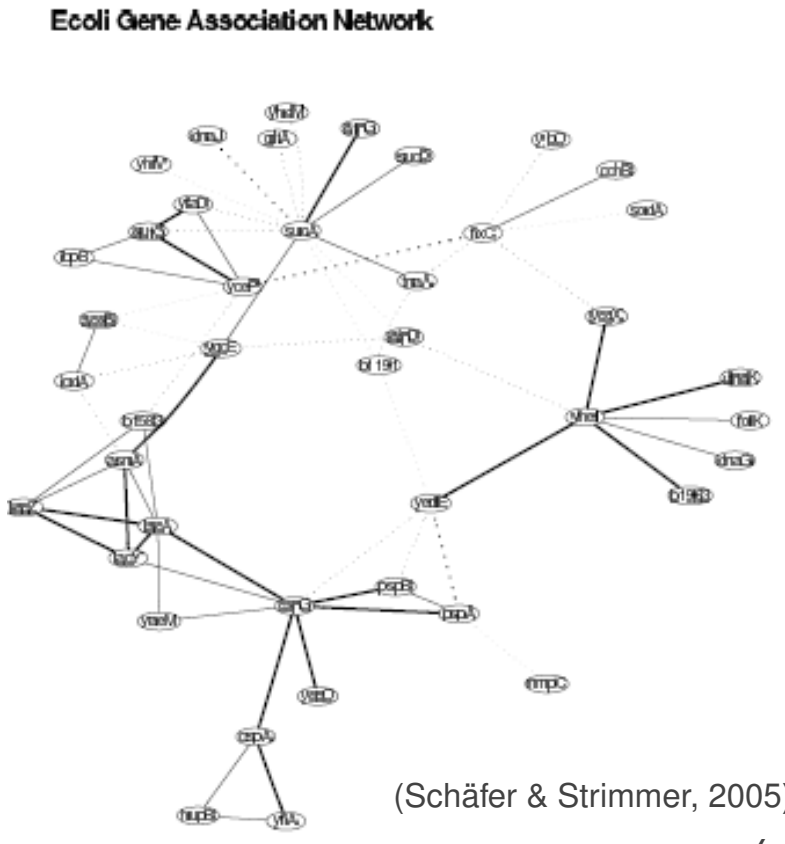
$\Rightarrow$ consider causal discovery as **exploratory** data analysis

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



(Schäfer & Strimmer, 2005)

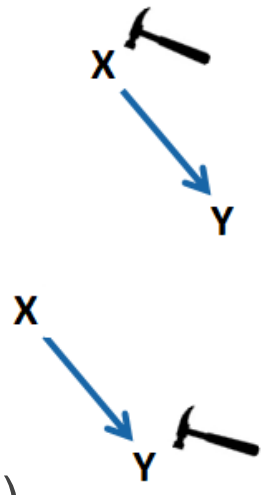
Causal Interpretation

... 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

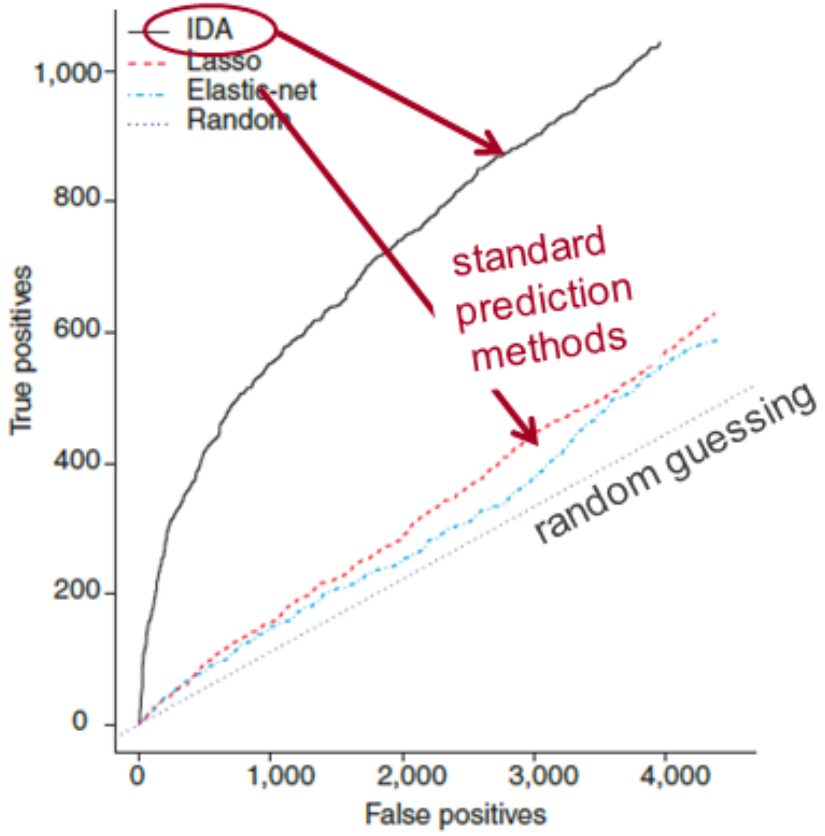
Motivation ctd: Gene Regulation

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

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

('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!

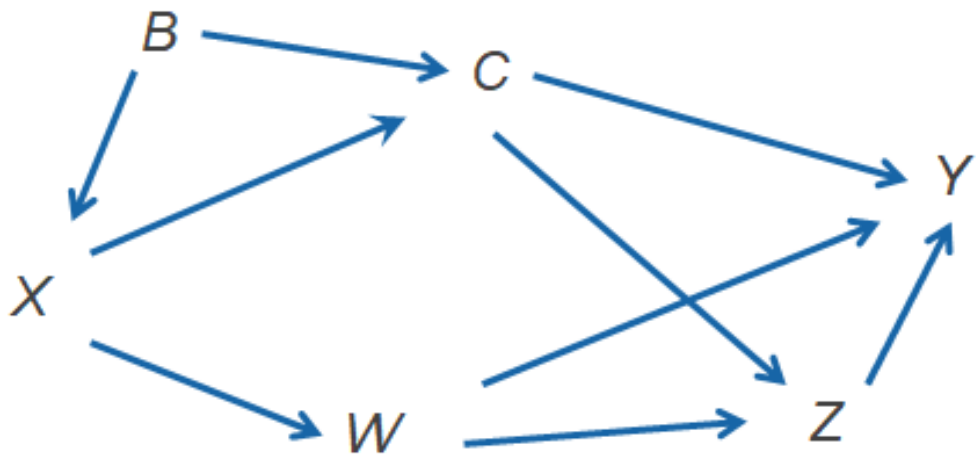
Separation and Independence

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

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**)

$\Rightarrow$ 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

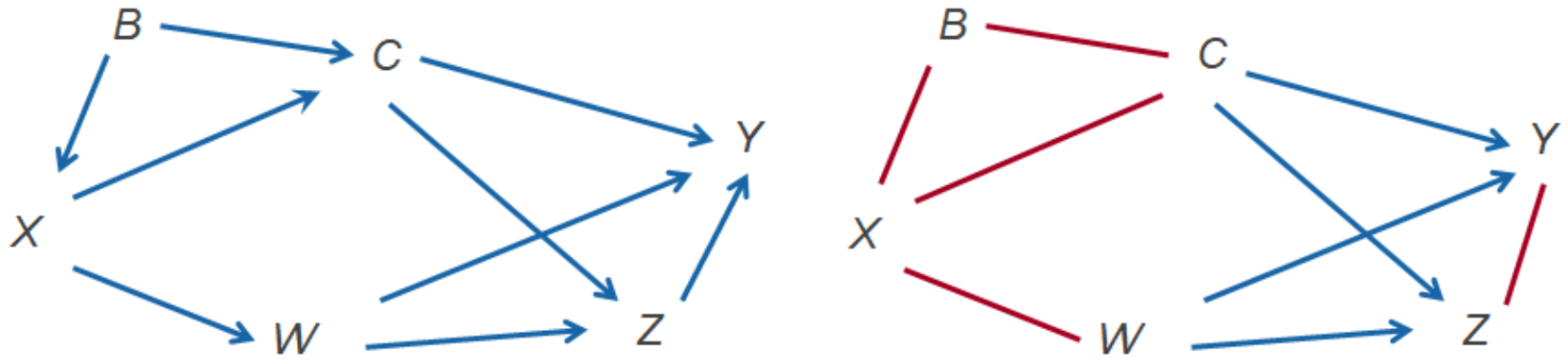
– some directed and some undirected edges

– undirected means: in class, both directions exist

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

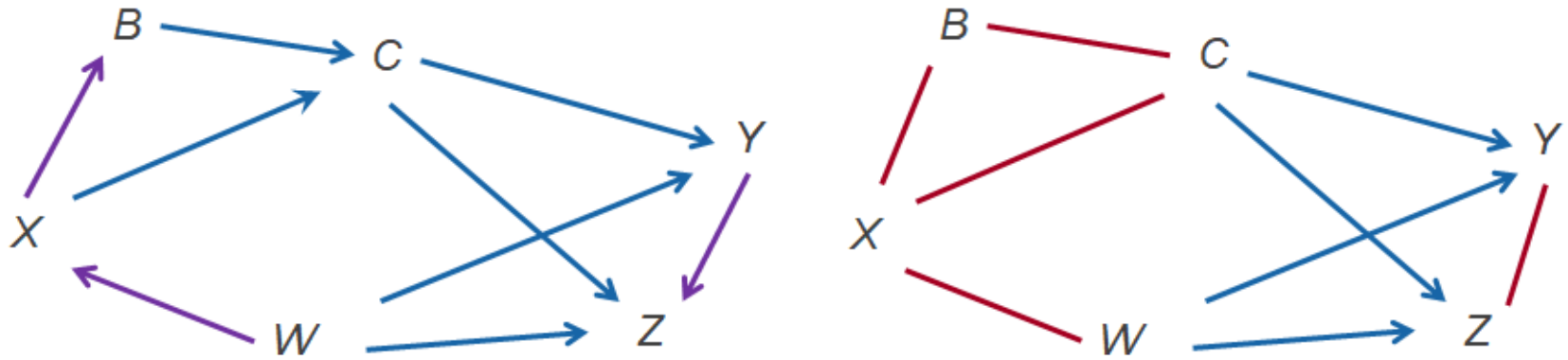
CPDAG Example

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 if possible.

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

Software: `TETRAD` 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.

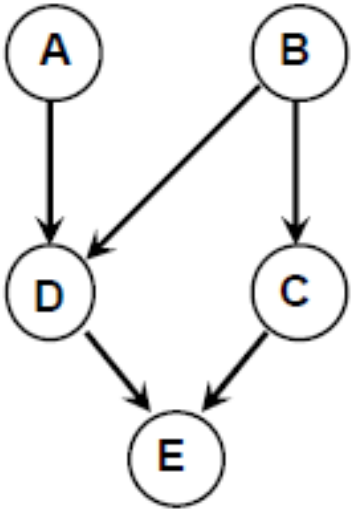
PC Algorithm — First Step

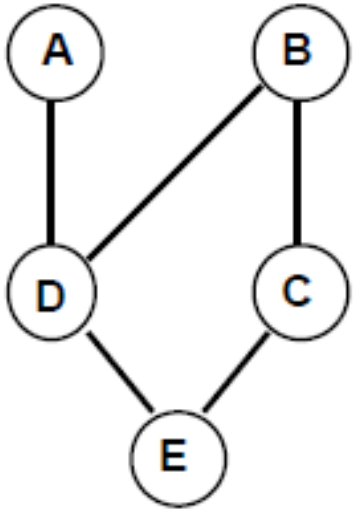
$V =$ set of nodes, and each node A has a set of adjacent nodes adj_A.

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

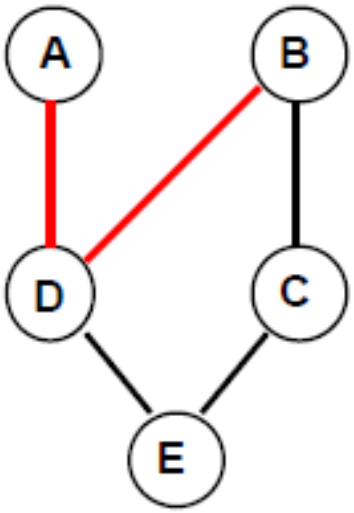
Identify V–structures

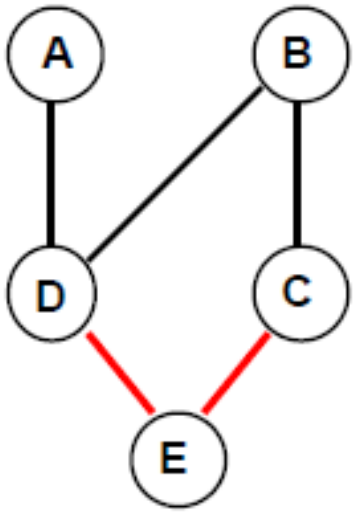
Procedure

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.



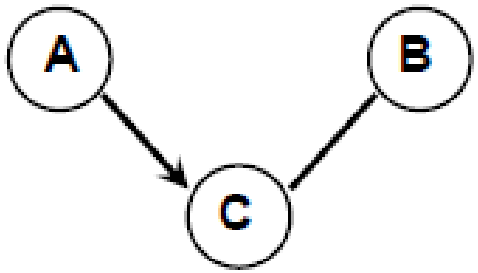


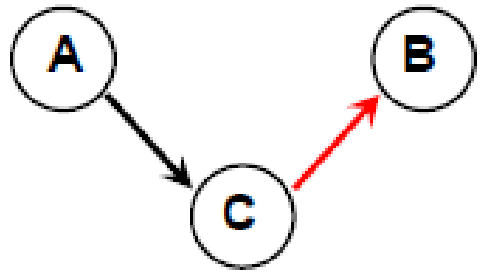
PC Algorithm — Third Step

Meek's Rules

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

Examples: constellations that can be oriented



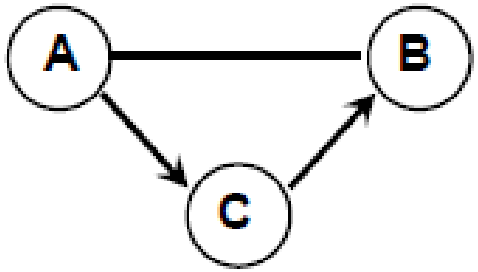


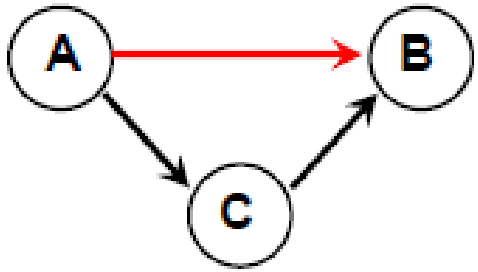
PC Algorithm — Third Step

Meek's Rules

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

Examples: constellations that can be oriented



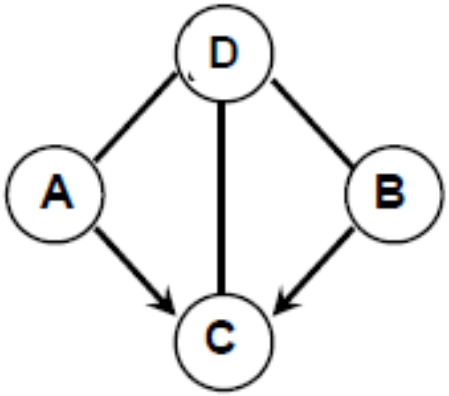


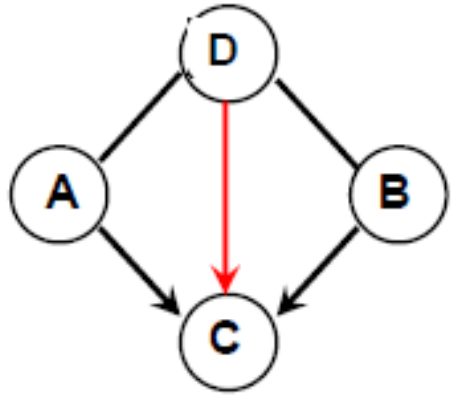
PC Algorithm — Third Step

Meek's Rules

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

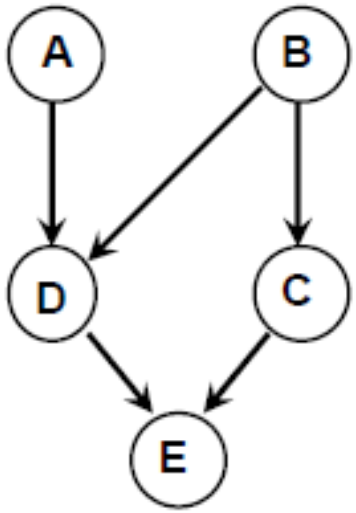
Examples: constellations that can be oriented

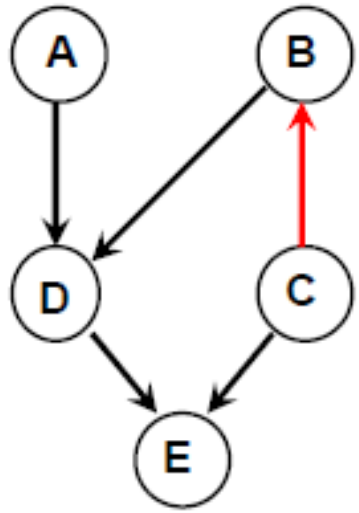




PC Algorithm — Finally

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

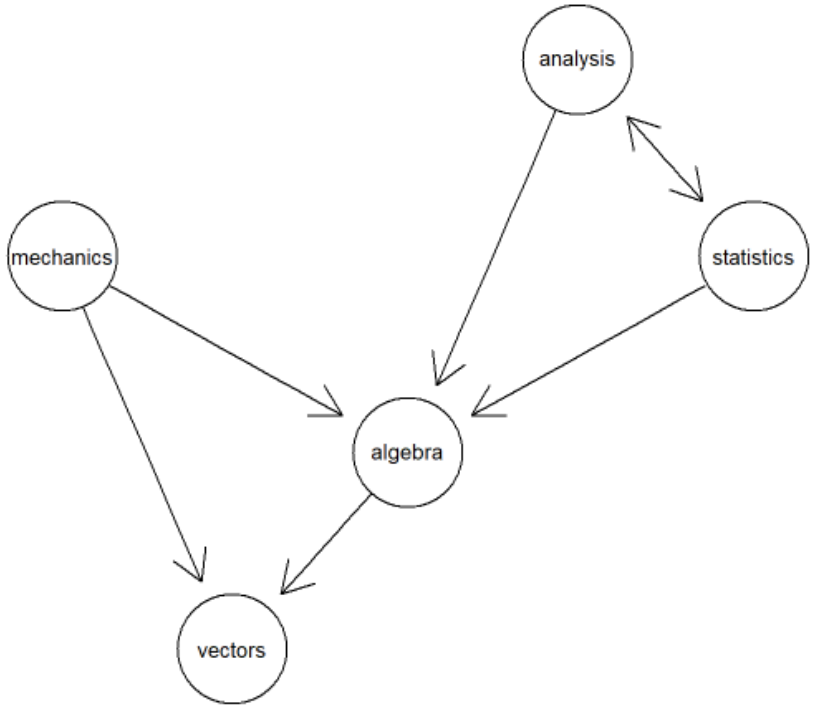




PC algorithm outputs *CPDAG* representing a Markov equivalence class of DAGs.

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!

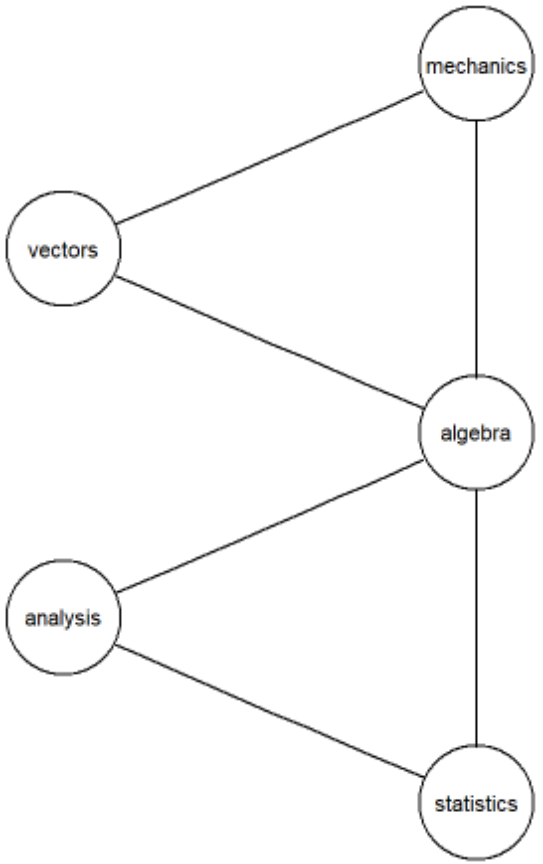
$\rightarrow$ outputs can be very unstable
$\rightarrow$ outputs may not be valid CPDAGs
$\Rightarrow$ should bootstrap results to assess variability of graph!

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

... 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

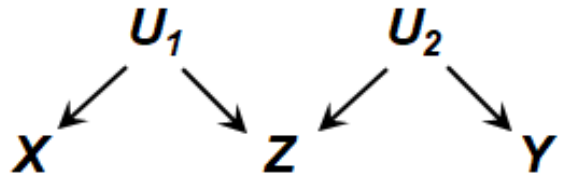
Relaxing Causal Sufficiency

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



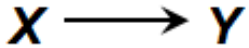
PC algorithm: wrong output

FCI algorithm: correct PAG

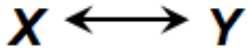
Interpretation: Edges in PAGs

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

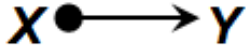
X cause of Y (ancestor)



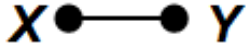
Y does not cause X nor vice versa, there may be a latent common cause



Y does not cause X

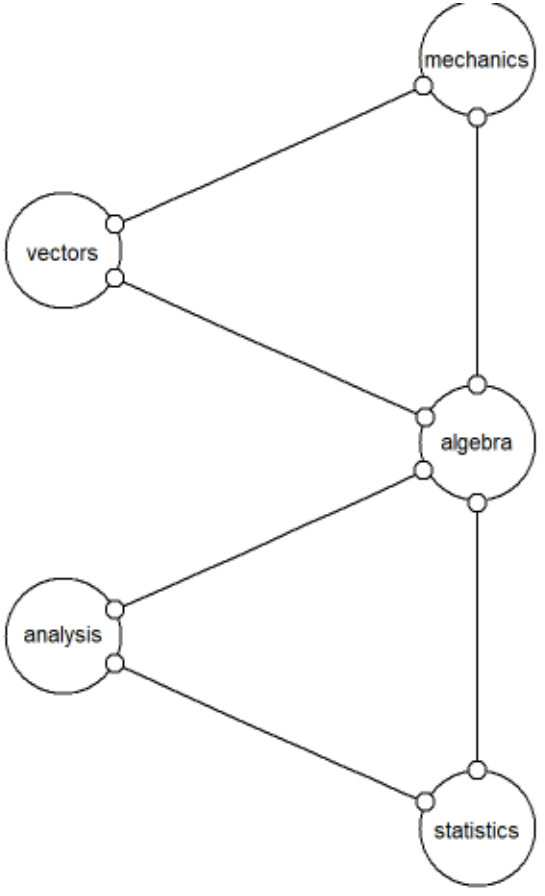


any of the above (and $X \leftarrow Y$) occur in equiv.class



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 $\approx$ 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

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

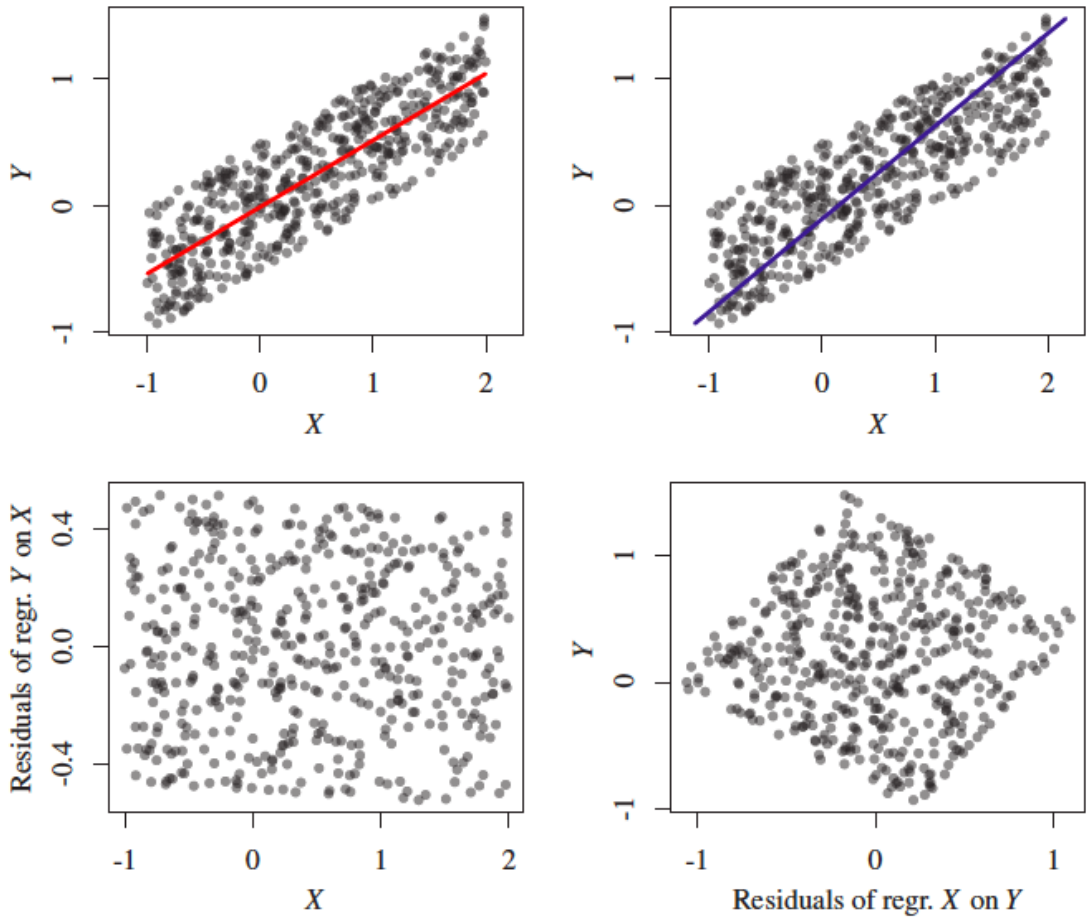
2) ε non-normally distributed (`lingam`)

$\Rightarrow$ 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

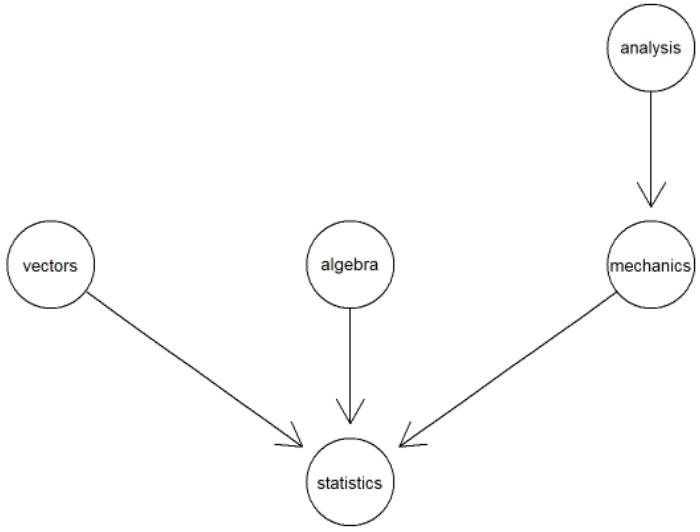
- GES with Gaussian BIC in `R` with `pcalg`:

  `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:

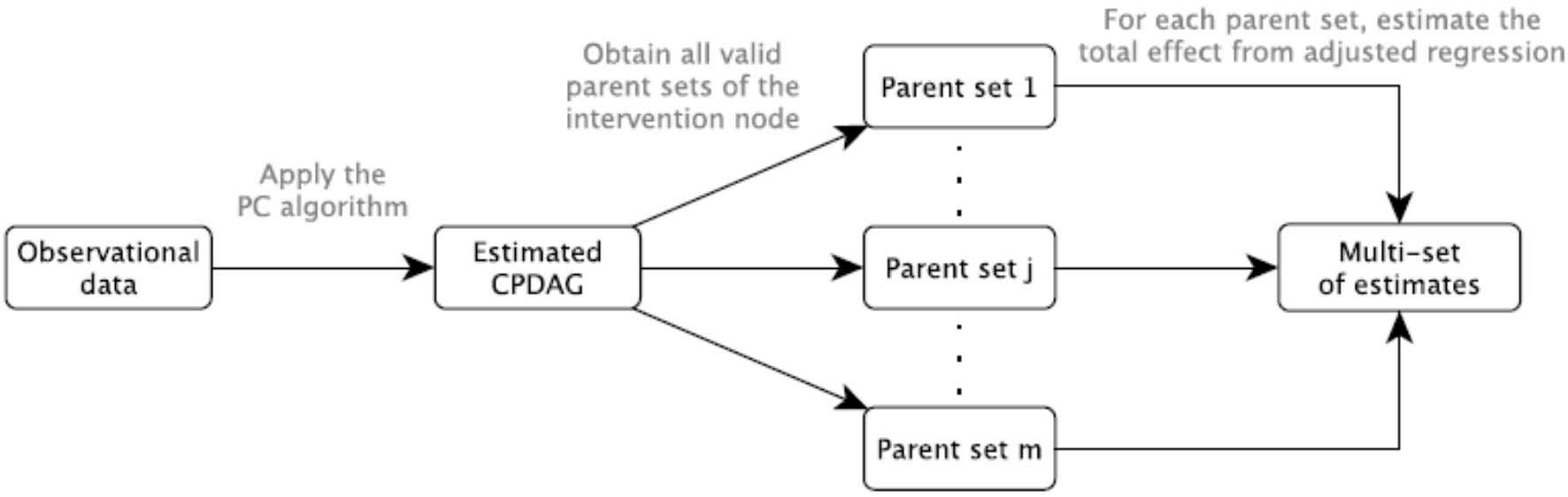
- enumerate all DAGs in CPDAG
- for each DAG and each X_i, Y_j pair determine sufficient adjustment set C — see yesterday!
- estimate causal effect
  if assume multivariate normal $\Rightarrow$ linear regression
  else: use other estimation method — see yesterday!

$\Rightarrow$ 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.



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

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.

Further references

References IV

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.

Aalen OO, Stensrud MJ, Didelez V, Daniel R, Røysland K, Strohmaier S (2019). Time-dependent mediators in survival analysis: Modelling direct and indirect effects with the additive hazards model. Biometrical Journal. (in print).

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

References V

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

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

References VI

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.

Dawid AP (2002): Influence diagrams for causal modelling and inference. *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 (2008). Identifying optimal sequential decisions. Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence, 113-120.

References VII

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.

Gran JM, Røysland K, Wolbers M, Didelez V, Sterne JA, Ledergerber B, Furrer H, von Wyl V, Aalen OO (2010). A sequential Cox approach for estimating the causal effect of treatment

References VIII

in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study. Statist. Med., 29: 2757-2768.

Guo, Dawid, Berzuini (2016). Sufficient covariate, propensity variable and doubly robust estimation. In: Statistical Causal Inferences and Their Applications in Public Health Research. Springer , p. 49-89.

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.

References IX

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

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, Brumback B, Robins JM (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. JASA, 96(454):440-448.

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

Hudgens, MH, Halloran, ME, (2008). Toward Causal Inference With Interference, JASA, 103:482, 832-842.

References X

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.

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`

References XI

Lunceford, Davidian (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects. Statistics in Medicine.

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.

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

References XII

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 (2011). Invited Commentary: Understanding Bias Amplification. American Journal of Epidemiology, 174(11).

Pearl J, Robins JM (1995). Probabilistic evaluation of sequential plans from causal models with hidden variables. In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence, 444–453. Morgan Kaufmann Publishers, San Francisco.

References XIII

Peters, J. , Bühlmann, P. and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. JRSSB, 78: 947-1012.

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

References XIV

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.

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

Robins JM, Hernán MA, Brumback B (2000). Marginal structural models and causal inference in epidemiology. Epidemiology, 11(5):550-560.

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.

References XV

Robins JM (2003). Semantics of causal DAG models and the identification of direct and indirect effects. In Highly Structured Stochastic Systems, P. Green, N.L. Hjort, S. Richardson, Editors. NY: Oxford University Press, pp. 70-81.

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

Robins JM, Wasserman L (1997). Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. Proceedings of the thirteenth conference annual conference on uncertainty in artificial intelligence (UAI-97). Morgan Kaufmann, San Francisco, pp 409–420.

References XVI

Rosenbaum P, Rubin D (1984). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55

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

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:

References XVII

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 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.

References XVIII

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.

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).

References XIX

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

Zivich, Breskin (2021): Machine Learning for Causal Inference: On the Use of Cross-fit Estimators, Epidemiology: Volume 32 - Issue 3 - p 393-401