Causal discovery

Szymon Jaroszewicz

Instytut Podstaw Informatyki, PAN
Wydział MiNI, Politechnika Warszawska

Zebranie plenarne KI PAN, 09.11.2022

Overview

- Causal discovery:
  - what is it?
  - why is it important? (statistically and socially)
  - why we all should be more aware?
- Tools: randomized experiments
- Tools: non-randomized experiments
  - classical: propensity scores, instrumental variables
  - newer: machine learning approaches
- Advantages and limitations

Motivation

- Classical machine learning is correlation based
- We make predictions based on observed values of other variables

- Such prediction is not always useful
  - does not explain the underlying mechanism
  - the model will not help us take appropriate action

Correlation based models – simple example

- Shoe size is a good predictor of mathematical abilities in children

Correlation based models – simple example

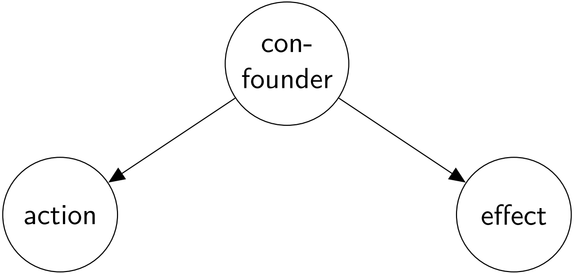
- Shoe size is a good predictor of mathematical abilities in children

The example is comical but illustrates the point:

- we can use the model for prediction
- we can not use it to act:
  - giving children larger shoes will not improve math skills
- Correlation is a result of a confounding variable: *age*

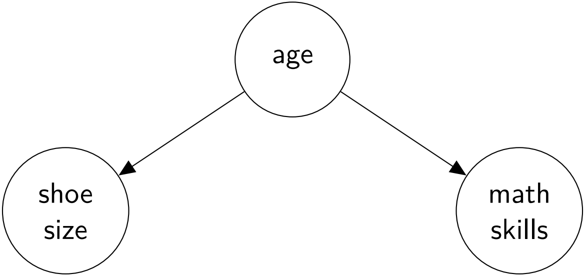
Confounders

- The main reason for lack of causal meaning are confounders



- They introduce correlations which we confuse with causal relations

Confounders

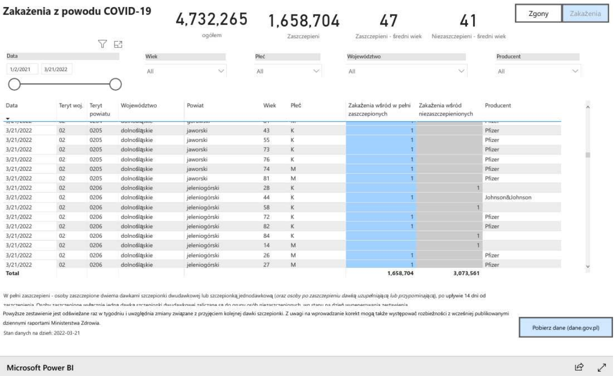


- Confounder: age
- If we look only at 6-graders correlation disappears

More realistic example: COVID vaccine effectiveness

- Main source of COVID data in Poland is *Baza Analiz Systemowych i Wdrożeniowych*

  `https://basiw.mz.gov.pl/index.html#/visualization?id=3761`



Zakażenia z powodu COVID-19

4,732,265 ogółem | 1,658,704 Zaszczepieni | 47 Zaszczepieni - średni wiek | 41 Niezaszczepieni - średni wiek

Zgony | Zakażenia

Data: 1/2/2021 – 3/21/2022 | Wiek: All | Płeć: All | Województwo: All | Producent: All

Data	Teryt woj.	Teryt powiatu	Województwo	Powiat	Wiek	Płeć	Zakażenia wśród w pełni zaszczepionych	Zakażenia wśród niezaszczepienionych	Producent
3/21/2022	02	0205	dolnośląskie	jaworski	43	K	1		Pfizer
3/21/2022	02	0205	dolnośląskie	jaworski	55	K	1		Pfizer
3/21/2022	02	0205	dolnośląskie	jaworski	73	K	1		Pfizer
3/21/2022	02	0205	dolnośląskie	jaworski	76	K	1		Pfizer
3/21/2022	02	0205	dolnośląskie	jaworski	74	M	1		Pfizer
3/21/2022	02	0205	dolnośląskie	jaworski	81	M	1		Pfizer
3/21/2022	02	0206	dolnośląskie	jeleniogórski	28	K		1	
3/21/2022	02	0206	dolnośląskie	jeleniogórski	44	K	1		Johnson&Johnson
3/21/2022	02	0206	dolnośląskie	jeleniogórski	58	K		1	
3/21/2022	02	0206	dolnośląskie	jeleniogórski	72	K	1		Pfizer
3/21/2022	02	0206	dolnośląskie	jeleniogórski	82	K	1		Pfizer
3/21/2022	02	0206	dolnośląskie	jeleniogórski	84	K		1	
3/21/2022	02	0206	dolnośląskie	jeleniogórski	14	M		1	
3/21/2022	02	0206	dolnośląskie	jeleniogórski	26	M	1		Pfizer
3/21/2022	02	0206	dolnośląskie	jeleniogórski	27	M	1		Pfizer
Total							**1,658,704**	**3,073,561**	

W pełni zaszczepieni - osoby zaszczepione dwiema dawkami szczepionki dwudawkowej lub szczepionką jednodawkową (*oraz osoby po zaszczepieniu dawką uzupełniającą lub przypominającą*), po upływie 14 dni od

Powyższe zestawienie jest odświeżane raz w tygodniu i uwzględnia zmiany związane z przyjęciem kolejnej dawki szczepionki. Z uwagi na wprowadzanie korekt mogą także występować rozbieżności z wcześniej publikowanymi dziennymi raportami Ministerstwa Zdrowia.

Stan danych na dzień: 2022-03-21

Pobierz dane (dane.gov.pl)

Microsoft Power BI

More realistic example: COVID vaccine effectiveness

- The database is used to study vaccine effectiveness
- Used by both sides of the debate

More realistic example: COVID vaccine effectiveness

https://mobile.twitter.com/MZ_GOV_PL/status/1487078020322152450

More realistic example: COVID vaccine effectiveness

Alicja
@alutk_a

Rozpoczynamy kolejną falę #covid w Polsce, więc to dobry moment, żeby podsumować poprzednią (IV).

Przedstawiam analizę zgonów #covid od 1.10.2021-17.01.2022, wraz z uproszczoną symulacją ile osób zmarłoby gdyby nikt nie był zaszczepiony, a ile gdyby byli wszyscy dorośli.

			w pełni zaszczepieni: 2 dawki lub 1 dawka J&J				niezaszczepieni			
upa ekowa	Ludność	Liczba zgonów od 01.10.2021 do 17.01.2022	Procent zaszczepionych	Ludność zaszczepiona	Liczba zgonów od 01.10.2021 do 17.01.2022	Zgonów na 100 tys.	Ludność niezaszczep.	Liczba zgonów od 01.10.2021 do 17.01.2022	Zgonów na 100 tys.	szczepienie obni ryzko zgonu [raz
8-24	2 688 690	28	46,78%	1 257 710	3	0.24	1 430 980	25	1.75	7,3
5-49	14 216 985	1 180	52,77%	7 502 584	136	1.81	6 714 401	1 044	15.55	8,6
0-59	4 605 466	1 641	63,39%	2 919 426	288	9.86	1 686 040	1 353	80.25	8,1
0-69	5 185 843	5 141	71,78%	3 722 265	1348	36.21	1 463 578	3 793	259.16	7,2
0-79	2 930 420	7 231	84,01%	2 461 808	2296	93.26	468 612	4 935	1053.11	11,3
80+	1 683 970	11 788	63,83%	1 074 903	3061	284.77	609 067	8 727	1432.85	5,0
UMA	31 311 374	27 009			7 132			19 877		

Symulacja: 100% zaszczepionych			
Grupa wiekowa	Ludność	Zgonów na 100 tys.	Szacowana liczba zgonów
18-24	2 688 690	0,24	6
25-49	14 216 985	1,81	258
50-59	4 605 466	9,86	454
60-69	5 185 843	36,21	1 878
70-79	2 930 420	93,26	2 733
80+	1 683 970	284,77	4 795
SUMA			10 125

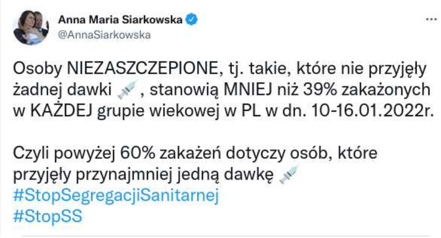
Symulacja: 0% zaszczepionych			
Grupa wiekowa	Ludność	Zgonów na 100 tys.	Szacowana liczba zgonów
18-24	2 688 690	1,75	47
25-49	14 216 985	15,55	2 211
50-59	4 605 466	80,25	3 696
60-69	5 185 843	259,16	13 440
70-79	2 930 420	1053,11	30 861
80+	1 683 970	1432,85	24 129
SUMA			74 382

or:
lutk_a

Źródło:
procent zaszczepionych na dzień 01.10.2021 (39 tydzień): ECDC https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea
liczba ludności: GUS https://stat.gov.pl/obszary-tematyczne/ludnosc/ludnosc
liczba zgonów COVID: https://dane.gov.pl/pl/dataset/2582/resource/35906/table

https://mobile.twitter.com/alutk_a/status/1484471204907999232

More realistic example: COVID vaccine effectiveness



Anna Maria Siarkowska
@AnnaSiarkowska

Osoby NIEZASZCZEPIONE, tj. takie, które nie przyjęły żadnej dawki 💉, stanowią MNIEJ niż 39% zakażonych w KAŻDEJ grupie wiekowej w PL w dn. 10-16.01.2022r.

Czyli powyżej 60% zakażeń dotyczy osób, które przyjęły przynajmniej jedną dawkę 💉
#StopSegregacjiSanitarnej
#StopSS

https://twitter.com/AnnaSiarkowska/status/1484170230448414724

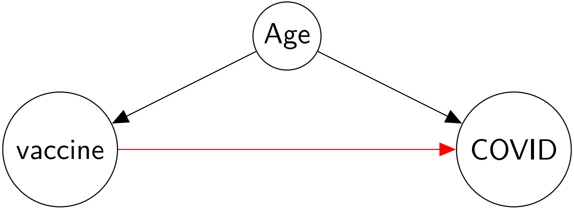
More realistic example: COVID vaccine effectiveness

- Public debate is based on correlation-type analysis
- But causal conclusions are implicitly made
  - vaccine will cause higher resistance to COVID
  - estimates of effectiveness

- Can we really make such conclusions?
- Let's look at possible confounders

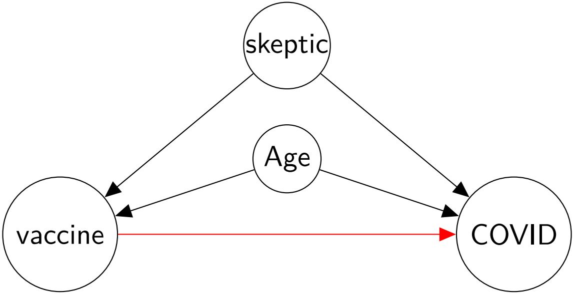
More realistic example: COVID vaccine effectiveness

- Age (most analyses take it into account)



More realistic example: COVID vaccine effectiveness

- Attitude towards COVID
- Skeptics
  - don't wear face-masks
  - meet people as usual
  - don't vaccinate
- Believers
  - wear face-masks
  - isolate
  - vaccinate
- Hard to measure

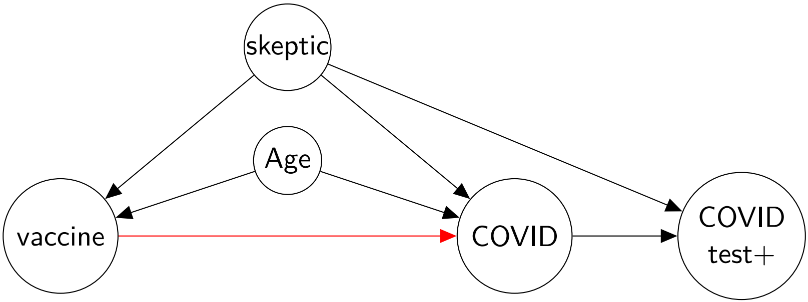


More realistic example: COVID vaccine effectiveness

- rural/city location (access to vaccines)
- prior disease
- ...

More realistic example: COVID vaccine effectiveness

- More problems: we only observe test results
- If a person does not get tested, we assume no COVID



Confounders

- Difference between correlation and causation seems obvious
- ... but most people forget about it
- Public discussion ignores it almost completely

- But it is very important
  - are COVID vaccines effective after all?
  - how effective they really are?

- How to make our discoveries causal?
  - old school: no causation without experimentation
  - newer discoveries: we can sometimes discover casual relationships from observational data

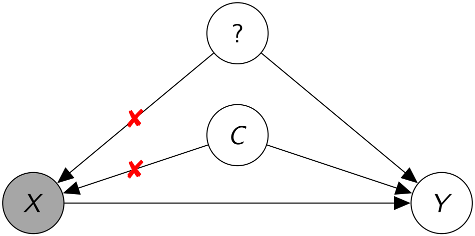
Randomized trials

Randomized trials – details

- Assign subjects randomly to either
  - Treatment group
  - Control group

Randomized trials

- In many cases the only way to guarantee the discovery is causal
- Assigning treatment at random
  - cuts influence of all confounders
  - removes effect of unobserved and even unknown confounders
- Not many additional assumptions needed but care should still be taken

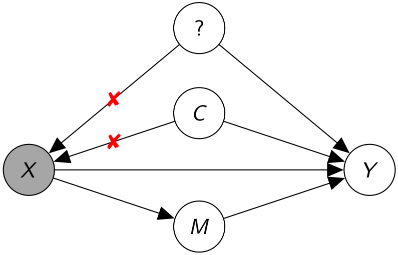


Randomized trials – applications

- Known for years in medicine
- Now popular in marketing / advertising
  - control groups commonly to assess effectiveness
  - A/B testing
- Problems:
  - mediators
  - noncompliance
  - ethics
  - cost

Randomized trials – problems

- Placebo effect
- Doctor's bias when examining the outcome
- Unwanted mediator!



- Solution: doubly blind trial

Randomized trials – noncompliance

- Exposure: has the subject really been subjected to the action (exposed)
- Example: a patient may be to sick to receive the assigned treatment, gets no treatment
- Should he/she be moved to the control group?

Randomized trials – noncompliance

- Exposure: has the subject really been subjected to the action (exposed)
- Example: a patient may be to sick to receive the assigned treatment, gets no treatment
- Should he/she be moved to the control group?
- NO!: very sick patients would be removed from treatment and added to control group

Intention to treat

The treatment assigned is considered the true treatment

Randomized trials – ethics

- Is it ethical to convince a random group of people to smoke?
- A real study of cardiac stent:
  - all patients get inserted the catheter (treatment and control)
  - only treatment cases get the stent
- Is it ethical?

Non-randomized trials

Non-random (biased) treatment assignment

- Randomized trials may be
  - costly
  - unethical
  - unavailable
- But plenty of observational data available: e.g. patient records

- Causal discovery with biased treatment assignment is very popular now
  - online advertising
  - marketing
  - medicine

Non-random (biased) treatment assignment

Problem

- Lack of randomization means backdoor paths are not closed
- Need another mechanism to achieve causal discovery
- Need additional untestable assumptions

- Many approaches available
- Classical
  - matching
  - propensity scores (most popular)
  - instrumental variables
- New
  - neural architectures
  - GANs
  - latent variable models

The potential outcomes framework

- Neyman (1923), Rubin
- Every case
  - received a treatment ($t_i = 1$)
  - was a control ($t_i = 0$)
- For every case x_i we have two potential outcomes

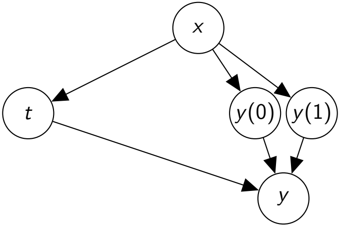
$$y_i(0) \text{ and } y_i(1)$$

  1: treatment, 0:control
- Every case has an observed outcome y_i

$$y_i = \begin{cases} y_i(0) & \text{if } t_i = 0 \\ y_i(1) & \text{if } t_i = 1 \end{cases}$$

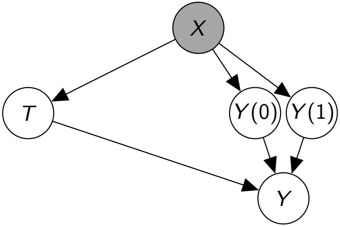
The potential outcomes framework

- Variables x_i
  - are all pre-treatment variables
  - change the potential outcomes $y_i(0), y_i(1)$
  - affect how treatment is selected (confounders)



- Idea: close all confounding paths through x_i
- Need additional assumptions

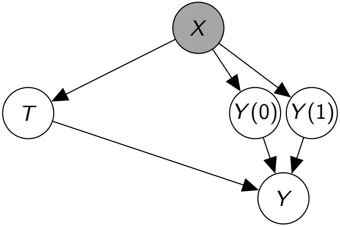
Strong ignorability



Strong ignorability

$$(Y(0), Y(1)) \perp T|\mathbf{X}$$

Strong ignorability



Strong ignorability

$$(Y(0), Y(1)) \perp T | \mathbf{X}$$

- Assume all confounding paths pass through X
- Assumption: no unmeasured confounders

Strong ignorability

- Cannot be verified from data
- In practice often hard to satisfy
- Example (not satisfied):
  - medical treatment based on doctor's decision
  - e.g. unknown genetic factors influencing condition and treatment outcome
- Example (satisfied):
  - online ads shown based on a model (model's outcome is fully understood)
  - medical treatment given based on strict recommendations
- Can't condition on all available variables (colliders)

Positivity

- Need a second assumption

Positivity

for all T and X

$$P(T|X) > 0$$

Treatment and control groups need to overlap

Inverse propensity score weighting

- Most popular method
- Propensity score

$$e_i = P(t_i = 1 | x_i)$$

- Inverse propensity score weighting IPTW
- i-th data record gets a weight

$$\frac{1}{P(t_i|x_i)} = \begin{cases} \frac{1}{e_i} & \text{if } t_i = 1 \\ \frac{1}{1-e_i} & \text{if } t_i = 0 \end{cases}$$

- Properties:
  - removes bias: expectation is the same as if no confounding was used
  - problem: increased variance

Matching

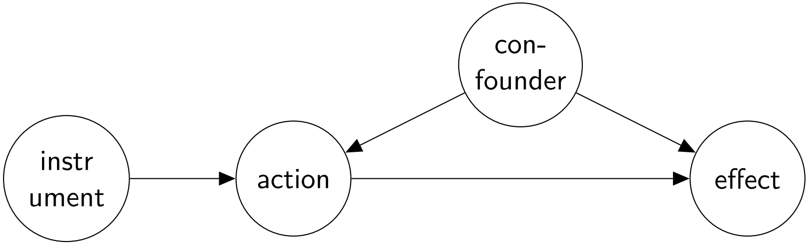
- Simplest method, historically first
- Idea: for every treatment case find a most similar control case and pair them

Matching

- Simplest method, historically first
- Idea: for every treatment case find a most similar control case and pair them
- Covid vaccine case could match on
  1. age
  2. region
  3. town
  4. gender
- Advantage: simple, implementations available
- Disadvantage: hard to match on many variables

Instrumental variables

- Instrumental variable:
  - influences treatment
  - independent from confounders



- Two stage estimation:
  1. estimate surrogate action $\hat{t}$ based on instrument
  2. build model of effect using $\hat{t}$ as predictor

Instrumental variables

- Example:
  - a complex medical treatment
  - instrument: living close to a big hospital
- Justification:
  - choice of place to live rarely affected by nearby hospitals
  - living close to a hospital increases chance of receiving the treatment

- An instrument might not exist
- An instrument needs to be justified

Regression discontinuity design

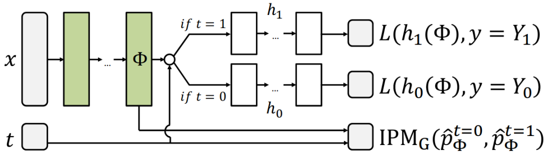
- Suppose we want to assess the effect of passing an exam on future earnings
- Randomized experiment not possible

- Pass exam $\Leftrightarrow$ points $>$ threshold
- Assume: small variations in # points are random
- Groups:
  - Treatment: just above threshold
  - Control: just below threshold

Neural causal models

- Deep learning is very popular now
- Several deep learning approaches to causal discovery have been proposed
  - representation learning
  - using GANs to generate counterfactual responses

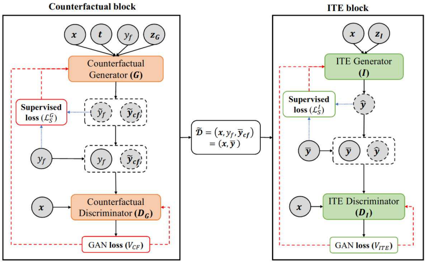
Representation learning



(source: Shalit, Johannson)

- Bottom layers compute representation Φ
- Representation is learned such that it has the same distribution in treatment and control groups
- Additional loss term is used for that
- The representation is then used to build a double model

GANITE



(source: Yoon, Jordon, van der Schaar)

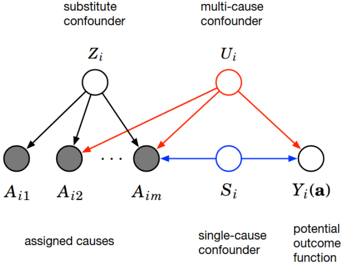
- Try to generate coutnerfactual outcomes using a Generative Adversarial Network
    - Discriminator tries to distinguish true from simulated outcome
- We have both potential outcomes $y_i(0)$ and $y_i(1)$: can predict $y_i(1) - y_i(1)$
- My worry: how do we know the GAN really converged? (generalization)

Neural causal models

- Hard to determine if predictions are really causal
- E.g. has the GAN converged sufficiently?
- No unmeasured confounders still necessary

Models discovering hidden confounders

- Wang, Blei. *The Blessings of Multiple Causes*, 2019



source: Wang, Blei, 2019

  - allows for detecting hidden confounders when multiple causes are present
  - assumptions
    1. no single cause confounders
    2. substitute Z needs to be perfect
- CEVAE: use an autoencoder to construct *ersatz* latent causes

Summary

- So, what do we have?
- Randomized trials
  - can guarantee causality, often only way to really convince
  - cost, ethics problems
- Causal discovery methods
  - need additional assumptions
  - often easily defeatable $\Rightarrow$ not entirely convincing

Way forward?

- Care about finding the truth
- Learn to do randomized trials cheaper/faster
  - adaptive designs
  - randomized instrumental variables: e.g. send randomized incentives to vaccinate
- Carefully choose a method for observational data