Causal Modeling with Hidden Confounders

Michèle Sebag and Armand Lacombe

TAU, CNRS – INRIA – U. Paris-Saclay

Toulouse, March 8th, 2022

Credit for slides: Diviyan Kalainathan, David Blei

The AI wave faces a shock

Unexpected / unwanted results

- Prediction: Issues with accuracy and generality; adversarial examples, out-of-distribution pbs
- Decision: Issues with trust; this workshop.
- Intervention: Issues with efficiency

Wanted: An AI with common decency

- Fair — no biases
- Accountable — model can be explained
- Transparent — decisions can be explained
- Robust

The dark side of AI:

Zeynep Tufekci	We're building a dystopia just to make people click on ads
C. O'Neill	Weapons of Math Destruction
Timnit Gebru	`www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru`

The big data promise:

Knowledge → Prediction → Control

Savoir pour prévoir afin de pouvoir

Auguste Comte, 1798 – 1857

Tasks

- Predict
- Decide
- Intervene

If umbrellas then rain

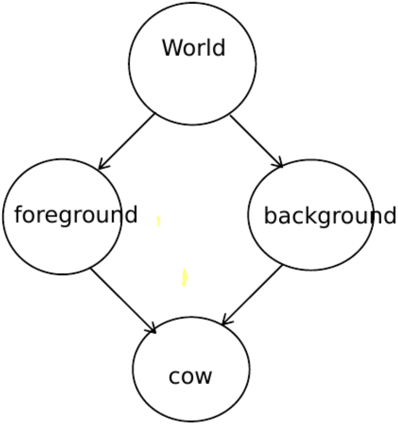
Issues with Prediction / Robustness

Perona, 2017

When it works

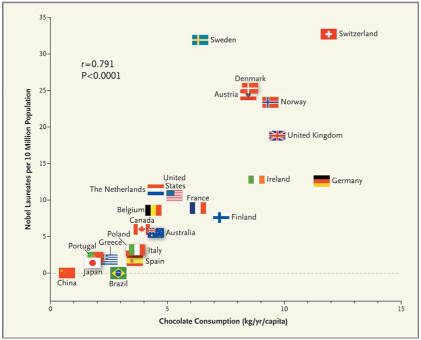
When it does not work

Tentative interpretation



Using Machine Learning models out of their scope

If you can predict...



F. H. Messerli: *Chocolate Consumption, Cognitive Function, and Nobel Laureates*, N Engl J Med 2012

... can you make things happen ?

Recommend people to eat more chocolate for the country to get more Nobel prizes.

The big data promise:

Knowledge $\rightarrow$ Prediction $\rightarrow$ Control

Savoir pour prévoir afin de pouvoir

Auguste Comte, 1798 – 1857

Interventions can only be based on causal models

Causal models will expectedly enable control:

- health and nutrition
- education
- economics/management
- climate

Causal models, formal background

Pearl 2003-2018; Peters et al., 2017

Definition 1: Intervention Pearl 2009

Intervention $do(X = x)$ forces variable X to value x

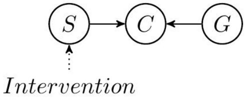
Definition 2: Direct cause $X_i \rightarrow X_j$

$$P_{X_j|\mathrm{do}(X_i=x,\mathbf{X}_{\setminus ij}=\mathbf{c})} \neq P_{X_j|\mathrm{do}(X_i=x',\mathbf{X}_{\setminus ij}=\mathbf{c})}$$

Example

C: Cancer, S : Smoking, G : Genetic factors

$P(C|do\{S = 0, G = 0\}) \neq P(C|do\{S = 1, G = 0\})$



Beware !

Intervening is *not* conditioning

- Conditioning: observing what happens for smokers
- Intervening: making everyone smoke; and observing what happens
  not ethical indeed; we'll come back to this

Causal Discovery: The royal road

Gold standard: Randomized controlled experiments

- Draw iid samples, form two subsets:
  - T=1: treatment group
  - T=0: control group
- Compute Average Treatment Effect (ATE)

Notations

- Y: outcome (survival)
- X: covariates (age, gender,...)
- T: treatment (0 or 1)
- $Y_i(0)$: outcome of the i-th sample if it does not get the treatment
- $Y_i(1)$: outcome of the i-th sample if it does get the treatment

Goal: estimate

$$ATE = \mathbb{E}[Y(1) - Y(0)]$$

Pb: only one out of $Y_i(0)$ and $Y_i(1)$ is known

Estimating ATE

Under assumptions, it works

$$
\begin{aligned}
ATE &= \mathbb{E}[Y(1) - Y(0)] \\
&= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\
&\qquad\qquad \textit{linearity of expectation} \\
&= \mathbb{E}_X[\mathbb{E}[Y(1)|X]] - \mathbb{E}_X[\mathbb{E}[Y(0)|X]] \\
&\qquad\qquad \textit{expectation over covariates}
\end{aligned}
$$

Estimating ATE

Under assumptions, it works

$$
\begin{aligned}
ATE &= \mathbb{E}[Y(1) - Y(0)] \\
&= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\
&\qquad \textit{linearity of expectation} \\
&= \mathbb{E}_X[\mathbb{E}[Y(1)|X]] - \mathbb{E}_X[\mathbb{E}[Y(0)|X]] \\
&\qquad \textit{expectation over covariates} \\
&= \mathbb{E}_X[\mathbb{E}[Y(1)|T=1, X]] - \mathbb{E}_X[\mathbb{E}[Y(0)|T=0, X]] \\
&\qquad \textit{no hidden confounder; no unbserved common causes} \\
&\qquad \textit{overlap assumption, } T{=}1 \textit{ and } T{=}0 \textit{ are observed in the data}
\end{aligned}
$$

Estimating ATE

Under assumptions, it works

$$
\begin{aligned}
ATE &= \mathbb{E}[Y(1) - Y(0)] \\
&= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\
&\qquad\qquad \textit{linearity of expectation} \\
&= \mathbb{E}_X[\mathbb{E}[Y(1)|X]] - \mathbb{E}_X[\mathbb{E}[Y(0)|X]] \\
&\qquad\qquad \textit{expectation over covariates} \\
&= \mathbb{E}_X[\mathbb{E}[Y(1)|T=1, X]] - \mathbb{E}_X[\mathbb{E}[Y(0)|T=0, X]] \\
&\qquad\qquad \textit{no hidden confounder; no unbserved common causes} \\
&\qquad\qquad \textit{overlap assumption, T=1 and T=0 are observed in the data} \\
&= \mathbb{E}_X[\mathbb{E}[Y|T=1, X]] - \mathbb{E}_X[\mathbb{E}[Y|T=0, X]] \\
&\qquad\qquad \textit{consistency: } Y_i(1) \sim Y|T=1, X=X_i
\end{aligned}
\tag{1}
$$

Randomized Controlled Experiments: Caveat

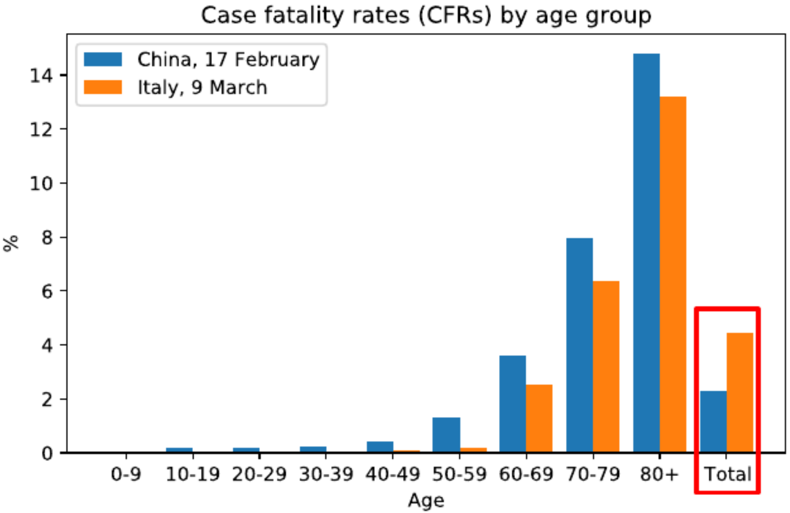
The Simpson paradox: comparing treatments A and B of kidney stones

Stone size	Treatment A	Treatment B
Small stones	93% (81/87)	87% (234/270)
Large stones	73% (192/263)	69% (55/80)
Total	78% (273/350)	83% (289/350)

Despite the global figures (bottom line), treatment A dominates treatment B on both groups of patients with large and small kidney stones. This paradox is explained as treatment A, known to be more efficient by the physician, is applied more frequently on (more difficult) large kidney stones cases.

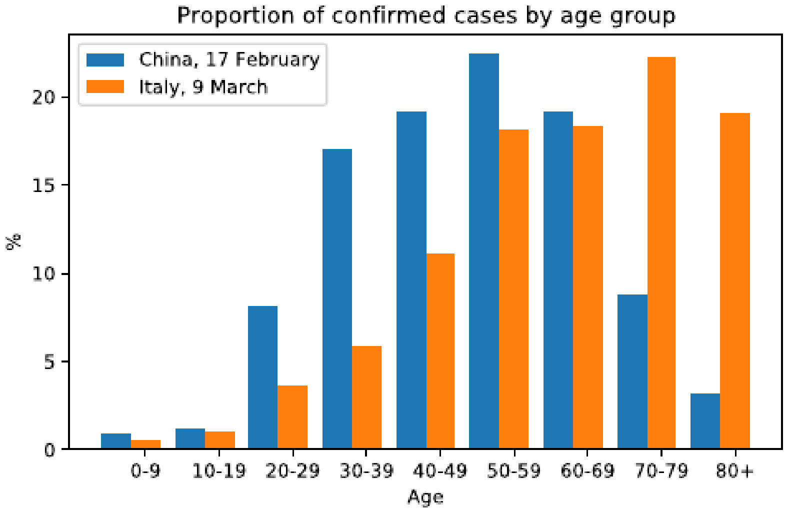
Simpson's paradox in Covid-19 case fatality rates

Kügelgen et al., 2020



Simpson's paradox in Covid-19 case fatality rates, 2

Kügelgen et al., 2020



Causal Discovery: The royal road

Why do we need an alternative to RCTs

- RCTs might be unethical (cannot make people smoke to see the effects)
- RCTs might be infeasible (no second planet to make experiments about climate)
- RCTs might be too costly (e.g. testing assumptions on economic rationality)

Alternative: Observational Causal Discovery

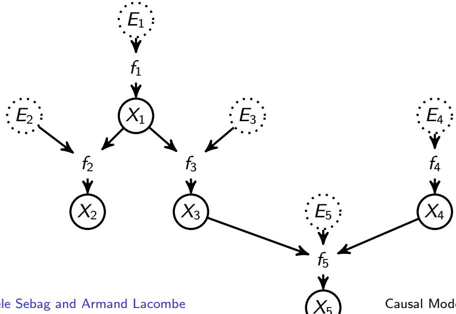
What is similar wrt ML

- Given data, infer causal models
- Challenges: data quality; data quantity; learning criterion...

What is different: Functional Causal Models (FCMs) Given $X_1, ..X_d$,

$$X_i = f_i(X_{\mathrm{Pa}(i;\mathcal{G})}, E_i), \forall i \in [1, d]$$

with $X_{\mathrm{Pa}(i;\mathcal{G})}$ the set of parents of X_i in $\mathcal{G}$ (= causes of X_i),
E_i a random independent noise variable modeling the unobserved other causes,
f_i a deterministic function: the causal mechanism



$$\begin{cases} X_1 = f_1(E_1) \\ X_2 = f_2(X_1, E_2) \\ X_3 = f_3(X_1, E_3) \\ X_4 = f_4(E_4) \\ X_5 = f_5(X_3, X_4, E_5) \end{cases}$$

Functional Causal Models, 2

Markov decomposition

$$P(X_1, \ldots, X_d) = \prod P(X_i | X_{\mathrm{Pa}(i;\mathcal{G})})$$

Usual Assumptions

Causal Sufficiency: no unobserved confounders

Causal Markov: all d-separations in the causal graph $\mathcal{G}$ imply conditional independences in the observational distribution P

Causal Faithfulness: all conditional independences in P imply d-separations in $\mathcal{G}$.

Motivations

Causal modelling

The confounders

The Deconfounder

Application to Human Resources

Limitations

Causal models are

- less accurate, prediction-wise (usually easier to predict causes from effects than the other way round)
- data hungry
  (variables independence tests: in d^2)
  (variables dependency tests conditionally to another variable, in d^3)
- subject to big assumptions !
- subject to even bigger requirements !!

Assumptions

- Causal Markov / causal faithfulness
  $\equiv$ model distribution $==$ empirical distribution $==$ true distribution.
- No unobserved confounders
  (remember Simpson.
  But confounders are all over the place).

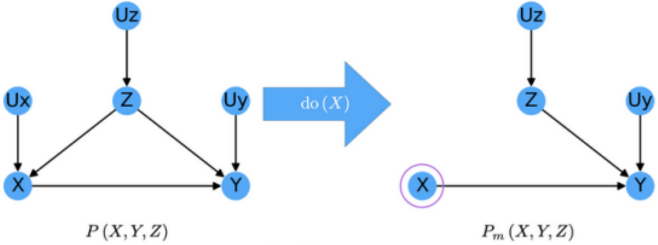
Requirement

- Identifiability

Blocking confounders

Pearl 2009

1. The adjusment effect



intervention

modified graph

$$P_m(X) = x$$
$$P_m(Z) = P(Z)$$
$$P_m(Y|X,Z) = P(Y|X,Z)$$

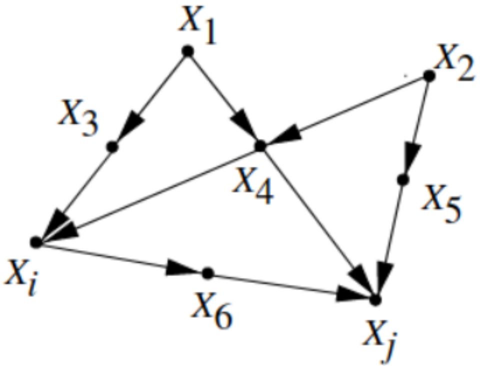
$$P(Y|do(X = x)) = \sum_z P(Y|x, Z = z)P(Z = z)$$

Blocking confounders, 2

2. The backdoor effect

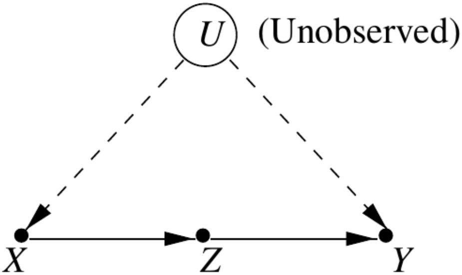
Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the backdoor criterion relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.

$$P(Y|do(X = x)) = \sum_z P(Y|x, Z = z)P(Z = z)$$



Blocking confounders, 3

3. The frontdoor effect



$$P(Y|do(X=x)) = \sum_z P(Y|Z=z)P(Z=z|X=x)$$

Observational causal modelling in real-world problems

1. **Confounders**
   - Known confounders: Front door and backdoor adjustments; require the causal graph to be known
   - Else, assume Causal Sufficiency (no hidden confounders)

2. **High dimensionality of data**
   - Hinders the discovery of models
   - Dimensionality reduction ?
   - But causal relations among constructed features are questionable

3. **Constructed features**
   - e.g. in economics: Investment, Investment / Salaries, Salaries, ...
   - Inflate the causal relationships

The deconfounder

Wang and Blei, 2019, 2020

Example

- Data about movies: casts and revenue
- Goal: Understand the **causal effect** of putting an actor in a movie
- Causal: “What will the revenue be if we make a movie with a particular cast?”

(David Blei, Oberwolfach 2019)

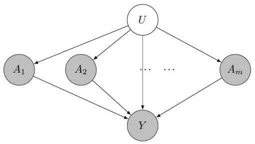
Causal inference vs prediction

- James Bond movies do well
- Cast: James Bond, M, Q, Ms Moneypenny
- M, Q, Ms Moneypenny only appear in James Bond movies
- (here we have a hidden confounder: the "James Bondedness"...)

(David Blei, Oberwolfach 2019)

The deconfounder

Wang and Blei, 2019, 2020



Notations

- A_i potential causes
- Y outcome
- U hidden confounders

Intuition

- Find a factor model: Z s.t.

$$P(A_1, \ldots A_n) = \prod_j P(A_j | Z)$$

- Z (ranging in $z_1, \ldots z_L$) is "Substitute Hidden Confounder"
- Informally, when $Z = z_\ell$, hidden confounders are assumed to be constant, too...

Assumptions

- Single ignorability: no Z causing a single A_j
- (Why ? If Z causes A_j that causes Y, one cannot separate the effects of Z and the effects of A_j.

The deconfounder, 2

Wang and Blei, 2019, 2020

Then, the average treatment effect can be computed as

$$\mathbb{E}[Y|do(A_1 = a)] = \sum \mathbb{E}_i[Y|(A_1 = a), Z = z_i)p(Z = z_i|A)$$

Informally: conditionally to $Z = z_i$, the hidden confounders are blocked.

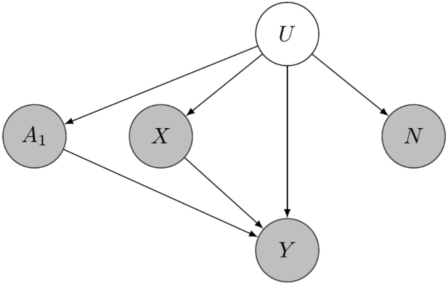
On-going strong debate

Damour 19, Athey et al 20, Imai et al 20, Grimmer et al. 20,...

- single ignorability untestable
- Z is not unique
- Pb if Z depends in a probabilistic way of X (this was fixed in Wang Blei 2020).

Fixing the Deconfounder

Wang and Blei, 2021



Glossary

- Proxy: an observed variable, child of the unobserved confounder
- Null Proxy: proxy with no effect on the outcome.
- Set of m treatments $A_1 \ldots A_m$, with a shared confounder U, partitioned into
  - A_C: treatments on which we intervene
  - A_X: treatments on which we don't intervene, used as proxy
  - A_N: a set of treatments with $Y \perp\!\!\!\perp f(A_N)|U, A_C, A_X$

Fixing the Deconfounder

Wang and Blei, 2021

An intervention distribution is identifiable iff it can be written as a function of the observed distribution.

Theorem

- If exists $f(A_N)$ s.t. $Y \perp\!\!\!\perp f(A_N) | U, A_C, A_X$
- $P(u|a_C, f(a_N))$ complete in $f(a_N)$ for almost all a_C
- $P(f(a_N)|a_C, a_X)$ complete in a_X for almost all a_C

Then

$$P(y|do(a_C)) = \int h(y, a_C, a_X) P(a_X) da_X$$

for h s.t.

$$P(y|a_C, f(a_N)) = \int h(y, a_C, a_X) P(a_X | a_C, f(a_N)) da_X$$

Definition

$P(x|y,z)$ complete wrt z iff for any square integrable g function,
$\int g(x,y) P(x|y,z) dx = 0$ for almost all $z \Rightarrow g(x,y) = 0$ for almost all x

Enquete Emploi

Statistical survey of the labor market

- Each EU member transmits the survey every 3 months to Eurostat
- Years 2017-2018: 110,945 individuals (1.5 year trajectories)
- Selected: unemployed people (5,009)

Features: 720

- age (average 39)
- gender (49% women),
- immigrant (16%)
- approximate income
- family status
- category of home location (12% in Quartier Prioritaire)
- health,
- level of studies (19% $>$ bacalaureat)
- search for jobs through: Public Agencies, Interim Agencies, social networks, etc

Counterfactual effects of features

The scientific question

Learn a causal model $P(finding.a.job)$.
How would this probability be modified if I were a woman, an immigrant, if I search a job primarily with a public agency, or my social network, or...

The methodology

- Pre-processing the data (most features are binary)
- Build P using Bayesian Logistic Regression
- Check the assumptions.
  Caveat Variability

Counterfactual effects of features

Results

Feature	ATE
is single	-12.2
parents not immigrants	2.19
not immigrant	6.49
no child in house	3.15
is woman	0.86
is from DOM	-16.9
lives in a sensitive urban area	3.02
lives in prioritary neighborhood	-6.56
asked to public agency	4.58
asked to interim agency	10.89
asked to relatives	2.33
asked to colleagues	3.80
asked on social networks	0.27
took a public exam	6.52
spontaneously application	4.87
published a classified ad	-2.20
answered a job offer	9.34
other methods	0.48

Limitations

- Data preparation
- Model stability
- Interpretations
- Setting: choice of means (combinatorial treatment)

As usual: confirm with field experiments.

Conclusion

Causal modelling

- The most exciting game !
- A most slippery game :-(
- When everything fails, use common sense

Perspectives

The bottlenecks

- Find a causal graph (d^3 independency tests)
- Data-hungry task !
- The causal ladder: predict; What if; What if not

On-going

- Structure-Agnostic Model — Kalainathan et al. 21
- We need changing representations
- How to enforce identifiability ?
- Generalized contrastive losses — Hyvarinen et al. 2019

Thanks!

Diviyan Kalainathan

Armand Lacombe

Philippe Caillou

Isabelle Guyon

Olivier Goudet