Artificial Intelligence & Causal Modeling

Michèle Sebag

TAU

CNRS – INRIA – LRI – Université Paris-Saclay

CREST Symposium on Big Data – Tokyo – Sept. 25th, 2019

Artificial Intelligence & Causal Modeling

Michèle Sebag

Tackling the Underspecified

CNRS – INRIA – LRI – Université Paris-Saclay

CREST Symposium on Big Data – Tokyo – Sept. 25th, 2019

Artificial Intelligence / Machine Learning / Data Science

A Case of Irrational Scientific Exuberance

- Underspecified goals — Big Data cures everything
- Underspecified limitations — Big Data can do anything (if big enough)
- Underspecified caveats — Big Data and Big Brother

Wanted: An AI with common decency

- Fair — no biases
- Accountable — models can be explained
- Transparent — decisions can be explained
- Robust — w.r.t. malicious examples

ML & AI, 2

In practice

- Data are ridden with biases
- Learned models are biased (prejudices are transmissible to AI agents)
- Issues with robustness
- Models are used out of their scope

More

- C. O'Neill, *Weapons of Math Destruction*, 2016
- Zeynep Tufekci, *We're building a dystopia just to make people click on ads*, Ted Talks, Oct 2017.

Machine Learning: discriminative or generative modelling

Given a training set iid samples $\sim P(X, Y)$

$$\mathcal{E} = \{(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathbb{R}^d, i \in [[1, n]]\}$$

Find

- Supervised learning: $\hat{h} : X \mapsto Y$ or $\widehat{P}(Y|X)$
- Generative model $\widehat{P}(X, Y)$

Predictive modelling might be based on correlations

If umbrellas in the street, Then it rains

The implicit big data promise:

If you can predict what will happen,

then how to make it happen what you want ?

Knowledge $\rightarrow$ Prediction $\rightarrow$ Control

ML models will be expected to support *interventions*:

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

Intervention Pearl 2009

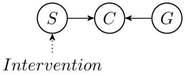
Intervention $do(X = a)$ forces variable X to value a

Direct cause $X \rightarrow Y$

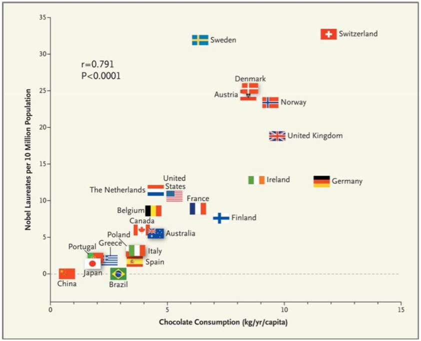
$$P_{Y|\mathrm{do}(X=a,\mathbf{Z}=\mathbf{c})} \neq P_{Y|\mathrm{do}(X=b,\mathbf{Z}=\mathbf{c})}$$

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

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



Correlations do not support interventions



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

Causal models are needed to support interventions

Consumption of chocolate enables to predict # of Nobel prizes
but eating more chocolates does not increase # of Nobel prizes

An AI with common decency

Desired properties

- Fair — no biases
- Accountable — models can be explained
- Transparent — decisions can be explained
- Robust — w.r.t. malicious examples

Relevance of Causal Modeling

- Decreased sensitivity wrt data distribution
- Support interventions — clamping variable value
- Hopes of explanations / bias detection

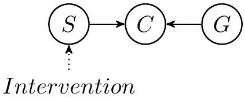
Causal modelling, Definition 1

Based on interventions Pearl 09, 18

X causes Y if setting $X = 0$ yields a Y distribution; and setting $X = 1$ ("everything else being equal") yields a different distribution for Y.

$$P(Y|\mathrm{do}(X = 1), \ldots Z) \neq P(Y|\mathrm{do}(X = 0), \ldots Z)$$

Example C: Cancer, S : Smoking, G : Genetic factors
$P(C|do\{S = 0, G = 0\}) \neq P(C|do\{S = 1, G = 0\})$



Causal modelling, Definition 1, follow'd

The royal road: randomized controlled experiments

Duflot Bannerjee 13; Imbens 15; Athey 15

But sometimes these are

- impossible — climate
- unethical — make people smoking
- too expensive — e.g., in economics

Causal modelling, Definition 2

Machine Learning alternatives

- Observational data
- Statistical tests
- Learned models
- Prior knowledge / Assumptions / Constraints

The particular case of time series and Granger causality

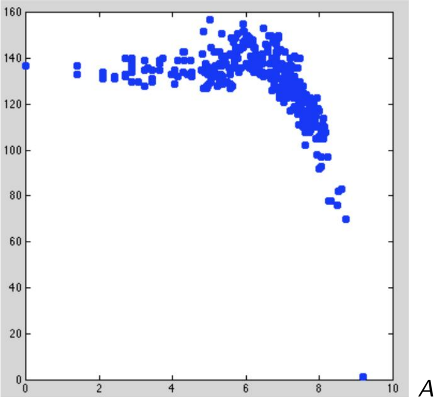
A "causes" B if knowing $A[0..t]$ *helps predicting* $B[t+1]$

More on causality and time series:

- J. Runge et al., *Causal network reconstruction from time series: From theoretical assumptions to practical estimation*, 2018

Causality: What ML can bring ?

Each point: sample of the joint distribution $P(A, B)$.
Given variables A, B



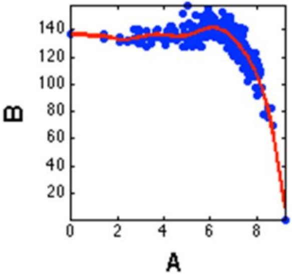
Causality: What ML can bring, follow'd

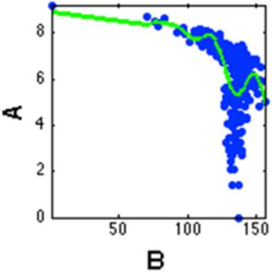
Given A, B, consider models

- $A = f(B)$
- $B = g(A)$

Compare the models

Select the best model: $A \to B$





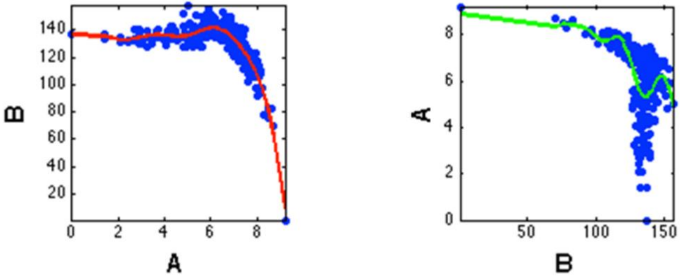
Causality: What ML can bring, follow'd

Given A, B, consider models

- $A = f(B)$
- $B = g(A)$

Compare the models

Select the best model: $A \rightarrow B$



A: Altitude, B: Temperature
Each point = (altitude, average temperature of a city)

Causality: A machine learning-based approach

Guyon et al, 2014-2015

Pair Cause-Effect Challenges

- Gather data: a sample is a pair of variables (A_i, B_i)
- Its label ℓ_i is the "true" causal relation (e.g., age "causes" salary)

Input

$$\mathcal{E} = \{(A_i, B_i, \ell_i), \ell_i \text{ in } \{\rightarrow, \leftarrow, \perp\!\!\!\perp\}\}$$

Example A_i, B_i	Label ℓ_i
A_i causes B_i	$\rightarrow$
B_i causes A_i	$\leftarrow$
A_i and B_i are independent	$\perp\!\!\!\perp$

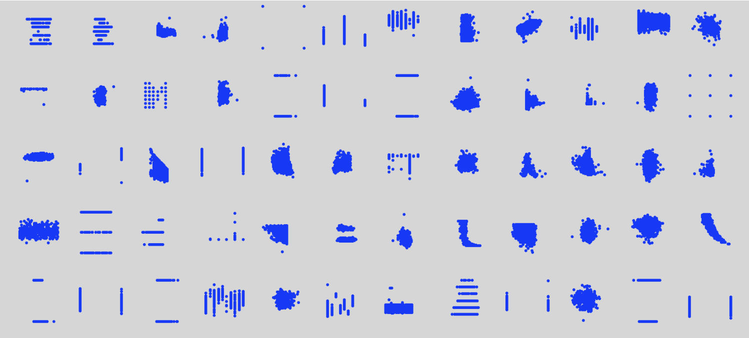
Output

using supervised Machine Learning

$$\text{Hypothesis } : (A, B) \mapsto \text{ Label}$$

Causality: A machine learning-based approach, 2

Guyon et al, 2014-2015



The Cause-Effect Pair Challenge

Learn a causality classifier (causation estimation)

- Like for any supervised ML problem from images — ImageNet 2012

More

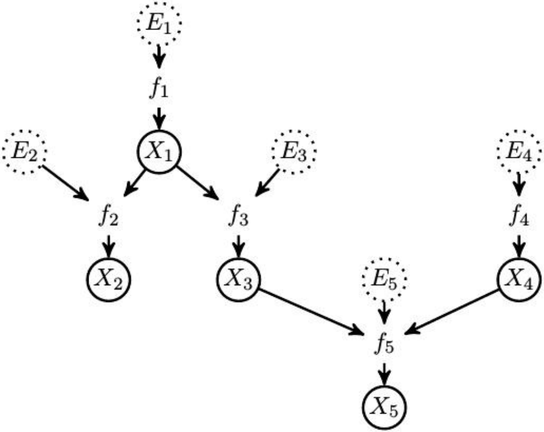
- Guyon et al., eds, *Cause Effect Pairs in Machine Learning*, 2019.

Functional Causal Models, a.k.a. Structural Equation Models

Pearl 00-09

$$X_i = f_i(\mathrm{Pa}(X_i), E_i)$$

$\mathrm{Pa}(X_i)$: Direct causes for X_i E_i: noise variables, all unobserved influences



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

Tasks

- Finding the structure of the graph (no cycles)
- Finding functions (f_i)

Conducting a causal modelling study

Spirtes et al. 01; Tsamardinos et al., 06; Hoyer et al. 09
Daniusis et al., 12; Mooij et al. 16

Milestones

- Testing bivariate independence (statistical tests)
  find edges $X - Y; Y - Z$
- Conditional independence
  prune the edges $X \perp\!\!\!\perp Z | Y$
- Full causal graph modelling
  orient the edges $X \rightarrow Y \rightarrow Z$

Challenges

- Computational complexity **tractable approximation**
- Conditional independence: data hungry tests
- Assuming causal sufficiency **can be relaxed**

$X - Y$ independance

$$P(X,Y) \stackrel{?}{=} P(X).P(Y)$$

Categorical variables

- Entropy $H(X) = -\sum_x p(x) log(p(x))$

  x: value taken by X, $p(x)$ its frequency

- Mutual information $M(X,Y) = H(X) + H(Y) - H(X,Y)$
- Others: χ^2, G-test

Continuous variables

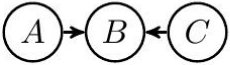
- t-test, z-test
- Hilbert-Schmidt Independence Criterion (HSIC) Gretton et al., 05

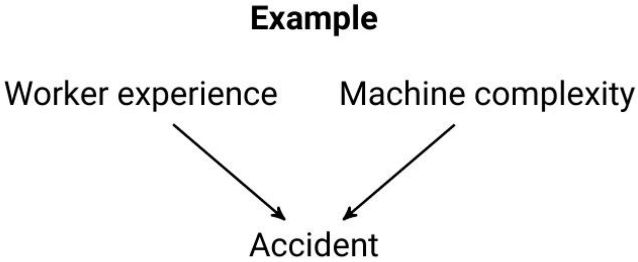
$$Cov(f,g) = \mathbb{E}_{x,y}[f(x)g(y)] - \mathbb{E}_x[f(x)]\mathbb{E}_y[g(y)]$$

  - Given $f : X \mapsto \mathbb{R}$ and $g : Y \mapsto \mathbb{R}$
  - $Cov(f,g) = 0$ for all f, g iff X and Y are independent

Find V-structure: $A \perp\!\!\!\perp C$ and $A \not\perp\!\!\!\perp C \mid B$

Explaining away causes



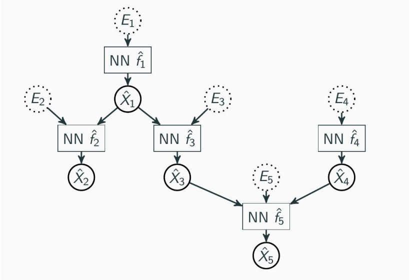


Causal Generative Neural Network

Goudet et al. 17

Principle

- Given skeleton given or extracted
- Given X_i and candidate $Pa(i)$
- Learn $f_i(\mathrm{Pa}(X_i), E_i)$ as a generative neural net
- Train and compare candidates based on scores



NB

- Can handle confounders (X_1 missing $\rightarrow$ ($E_2, E_3 \rightarrow E_{2,3}$))

Causal Generative Neural Network (2)

Training loss

- Observational data $\mathbf{x} = \{[x_1, \ldots, x_n]\}$ $\qquad x_i$ in $\mathbb{R} * *d$
- $(\mathrm{Graph}, \hat{f})$ $\hat{\mathbf{x}} = \{[\hat{x}_1, \ldots, \hat{x}_{n'}]\}$ $\qquad \hat{x}_i$ in $\mathbb{R} * *d$
- Loss: Maximum Mean Discrepancy $(\mathbf{x}, \hat{\mathbf{x}})$ (+ parsimony term), with k kernel (Gaussian, multi-bandwidth)

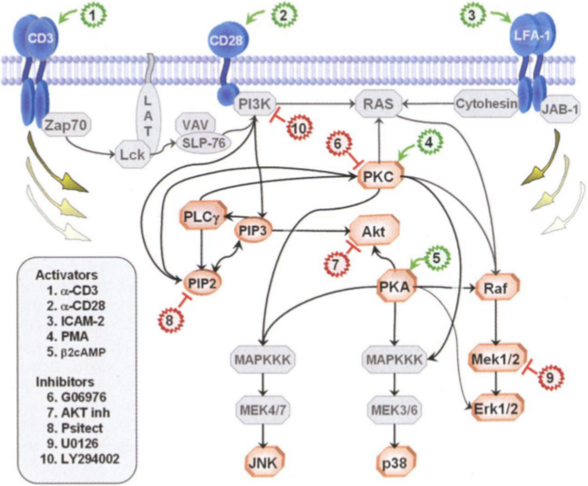
$$\mathrm{MMD}_k(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n^2} \sum_{i,j}^{n} k(x_i, x_j) + \frac{1}{n'^2} \sum_{i,j}^{n'} k(\hat{x}_i, \hat{x}_j) - \frac{2}{n \times n'} \sum_{i=1}^{n} \sum_{j=1}^{n'} k(x_i, \hat{x}_j)$$

- For $n, n' \to \infty$ $\qquad$ Gretton 07

$$\mathrm{MMD}_k(\mathbf{x}, \hat{\mathbf{x}}) = 0 \Rightarrow \mathcal{D}(\mathbf{x}) = \mathcal{D}(\hat{\mathbf{x}})$$

Results on real data: causal protein network

Sachs et al. 05



Edge orientation task

All algorithms start from the skeleton of the graph

method	AUPR	SHD	SID
Constraints			
PC-Gauss	0.19 (0.07)	16.4 (1.3)	91.9 (12.3)
PC-HSIC	0.18 (0.01)	17.1 (1.1)	90.8 (2.6)
Pairwise			
ANM	0.34 (0.05)	8.6 (1.3)	85.9 (10.1)
Jarfo	0.33 (0.02)	10.2 (0.8)	92.2 (5.2)
Score-based			
GES	0.26 (0.01)	12.1 (0.3)	92.3 (5.4)
LiNGAM	0.29 (0.03)	10.5 (0.8)	83.1 (4.8)
CAM	0.37 (0.10)	8.5 (2.2)	78.1 (10.3)
CGNN ($\widehat{\text{MMD}}_k$)	$\underline{0.74}$* (0.09)	$\underline{4.3}$* (1.6)	$\underline{46.6}$* (12.4)

AUPR: Area under the Precision Recall Curve
SHD: Structural Hamming Distance
SID: Structural intervention distance

CGNN

Goudet et al., 2018 **Limitations**

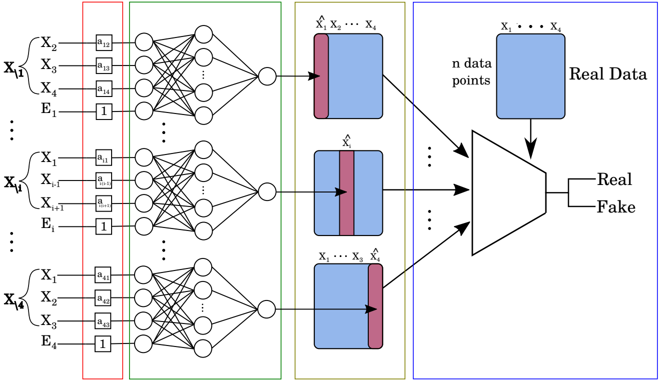
- Combinatorial search in the structure space
- Retraining fully the NN for each candidate graph
- MMD Loss is $O(n^2)$
- Limited to DAG

Structure Agnostic Modeling

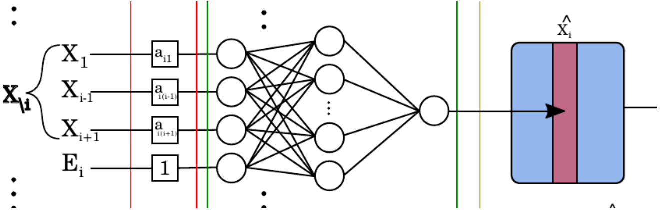
Kalainathan et al. 18

Goal: A generative model

+ Does not require CPDAG as input
+ Avoids combinatorial search for structure
− Less computationally demanding



Structure Agnostic Modeling, 2



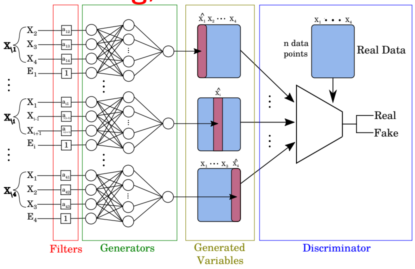
The *i*-th neural net

- Learns conditional distribution $P(X_i|X_{\backslash i})$ as $\hat{f}_i(X_{\backslash i}, E_i)$
- Filter variables $a_{i,j}$ are used to enforce sparsity (Lasso-like, next slide)
- 1st non-linear layer builds features $\phi_{i,k}$, 2nd layer builds linear combination of features:

$$f_i(X_{\backslash i}, E_i) = \sum \beta_{i,k}\phi_{i,k}(a_{i,1}X_1, \ldots, a_{i,d}X_d, E_i)$$

In the large sample limit, $a_{i,j} = 1$ iff $X_j \in MB(X_j)$ Yu et al. 18

Structure Agnostic Modeling, 3



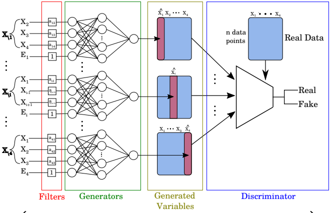
Given observational data $\{x_1, \ldots, x_n\} \sim P(X_1, \ldots, X_d)$ $\qquad x_i$ in $\mathbb{R}^d$

Adversarial learning

- Generate $\{\tilde{x}_i^{(j)}\}$ with j-th component of $\tilde{x}_i^{(j)}$ set to $\hat{f}_i(x_i, \epsilon)$, $\epsilon \sim \mathcal{N}(0,1)$
- Discriminator D among observational data $\{x_i\}$ and generated data $\{\tilde{x}_i^{(j)}, i = [[1,n]], j = [[1,d]]\}$
- Learning criterion (adversarial + sparsity)

$$\min \left(\text{Accuracy } (D) + \lambda \sum_{i,j} |a_{i,j}| \right)$$

Structure Agnostic Modeling, 4



Learning criterion $\min \left(\text{Accuracy}\ (D) + \lambda \sum_{i,j} |a_{i,j}|\right)$

Competition between discriminator and sparsity term $\sum \|\mathbf{a}\|_1$

- Avoids combinatorial search for structure
- Cycles are possible
- DAGness achieved by enforcing constraints on trace of $A = (a_{i,j})$ and A^k

Quantitative benchmark - artificial DAG

Directed **acyclic** artificial graphs (DAG) of 20 variables

	PC Gauss	PC HSIC	GES	MMHC	DAGL1	LINGAM	CAM	SAM
Linear	0.36	0.29	0.40	0.36	0.30	0.31	0.29	**0.49**
Sigmoid AM	0.28	0.33	0.18	0.31	0.19	0.19	0.72	**0.73**
Sigmoid Mix	0.22	0.25	0.21	0.22	0.16	0.12	0.15	**0.52**
GP AM	0.21	0.35	0.19	0.21	0.15	0.17	**0.96**	0.74
GP Mix	0.22	0.34	0.18	0.22	0.19	0.14	0.61	**0.66**
Polynomial	0.27	0.31	0.20	0.11	0.26	0.32	0.47	**0.65**
NN	0.40	0.38	0.42	0.11	0.43	0.36	0.22	**0.60**
Execution time	1s	10h	$<$1s	$<$1s	2s	2s	2.5h	1.2h

Quantitative benchmark - artificial DG (with cycles)

Directed **cyclic** artificial graphs of 20 variables

	CCD	PC Gauss	GES	MMHC	DAGL1	LINGAM	CAM	SAM
Linear	0.44	0.44	0.20	0.34	0.26	0.19	0.23	**0.51**
Sigmoid AM	0.31	0.31	0.16	0.32	0.17	0.24	0.37	<u>**0.47**</u>
Sigmoid Mix	0.31	0.35	0.18	0.34	0.19	0.17	0.22	<u>**0.49**</u>
GP AM	0.30	0.32	0.17	0.30	0.15	0.23	0.50	**0.56**
GP Mix	0.24	0.25	0.15	0.24	0.16	0.18	0.26	<u>**0.49**</u>
Polynomial	0.25	0.33	0.20	0.25	0.17	0.22	0.33	<u>**0.42**</u>
NN	0.25	0.18	0.18	0.24	0.18	0.16	0.22	<u>**0.40**</u>
Execution time	1s	1s	$<$1s	$<$1s	2s	2s	2.5h	1.2h

Causal Modeling and Human Resources

Known:

- A Quality of life at work — employee's perspective
- B Economic performance — firm's perspective
- ... are correlated

Question: Are there causal relationships ?
$A \rightarrow B$; or $B \rightarrow A$; or $\exists C \ / \ C \rightarrow A$ and $C \rightarrow B$

Data

- Polls from Ministry of Labor
- Gathered by Group Alpha Secafi (trade union advisor)
- Tax files + social audits for 408 firms

Economic sectors: low tech, medium-low, medium-high and high-tech.

Variables

Economic indicators

- Total number of employees
- Capitalistic intensity, Total payroll, Gini index
- Average salary (of workers, technicians, managers)
- Productivity, Operating profits, Investment rate

People

- Average age, Average seniority, Physical effort,
- Permanent contract rate, Manager rate, Fixed-term contract rate, Temporary job rate, Shift and night work, Turn-over
- Vocational education effort, duration of stints, Average stint rate (for workers, technicians, managers);

Variables, cont'd

Quality of life at work

- Frequency & Gravity of work injuries, Safety expenses, Safety training expenses
- Absenteism (diseases), Occupational-related diseases
- Resignation rate, Termination rate, Participation rate
- Subsidy to the works council

Men/Women

- Percentage of women (employees, managers)
- Wage gap between women and men (average, for workers, technicians, managers)

General Causal Relations

Access to training $\nearrow$

- $\searrow$ Gravity of work injuries
- $\searrow$ Occupational-related diseases

Termination rate $\nearrow$

- $\nearrow$ Absenteeism (diseases)

Percentage of managers $\nearrow$

- $\nearrow$ Access to training
- $\searrow$ Shift or night working hours

Age $\nearrow$

- $\searrow$ Fixed-term contract rate
- $\searrow$ Productivity (weak impact)

?

- Productivity $\nearrow$ $\rightarrow$ Participation rate $\nearrow$

Global relations between QLW and performance ?

Failure

- Nothing conclusive

Interpretation

- Exist confounders (controlling QLW and performance) $C \rightarrow A$ and $C \rightarrow B$
- One such confounder is the activity sector
- In different activity sectors, causal relations are different (hampering their identification)
- $\Rightarrow$ Condition on confounders

Low-tech sector

- Resignation rate $\nearrow$, Productivity $\searrow$
- Average salary $\nearrow$, Productivity $\nearrow$ very significant
- Occupational-related diseases $\nearrow$, Productivity $\searrow$
- Temporary job rate $\nearrow$, Gravity of work injuries $\nearrow$
- Permanent contract rate $\nearrow$, Safety training $\searrow$
- Duration training stints $\nearrow$, Termination rate $\searrow$

Outcomes & Limitations

Causal modeling and exploratory analysis

- Efficient filtering of plausible relations (several orders of magnitude);
- Complementary w.r.t. visual inspection (experts can be fooled and make sense of correlations & hazards);
- Multi-factorial relations ? yes; but even harder to interpret.

Not a ready-made analysis

- Causal relations must be
  - interpreted
  - confirmed by field experiments; polls; interviews.

A data-driven approach to individual dietary recommendations

Context

- Long-term goal: Personalized dietary recommendations
- Requirement: identify risk index associated to food products
- At a coarse-grained level (lipid, protein, glucid), nothing to see
- At a fine-grained level: 300+ types of pizzas, ranging from ok to very bad.

The wealth of Kantar data

- $\sim$22,000 households $\times$ 10 years (this study: 2014)
- 19M total purchases/year (180,000 products)
- Socio-demographic attributes, varying size

Beware: data rarely collected as should be...

Raw description can hardly be used for meaningful analysis

- 170,000 products for 22,000 households
- Data gathered with (among others) marketing goals
  where bought, which conditioning
- Most products are sold by 1 vendor
- Most families are going to one vendor

Manual pre-processing

- Consider 10 categories of interest, e.g. bio/non-bio; alcohol yes/no; fresh/frozen
- Merge products with same categories
- 170,000 $\rightarrow \approx$ 4,000 products

Example: for beer, we only selected as features of interest: colour (blonde, black, etc.); has-alcohol (yes, no); organic (yes, no)

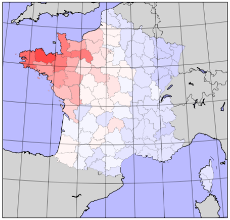
Methodology

Dimensionality reduction

1. Borrowing Natural Language Processing tools, with
   vector of purchase ≈ document
   food product ≈ word
2. Using Latent Dirichlet Association to extract "dietary topics"

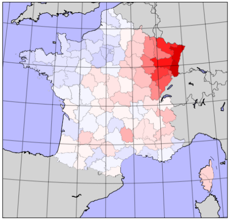
Blei et al. 03

Some topics can be directly interpreted The darker the region, the more present the topic (NB: regions are not used to build topics)



Topic 2

"Brittany"



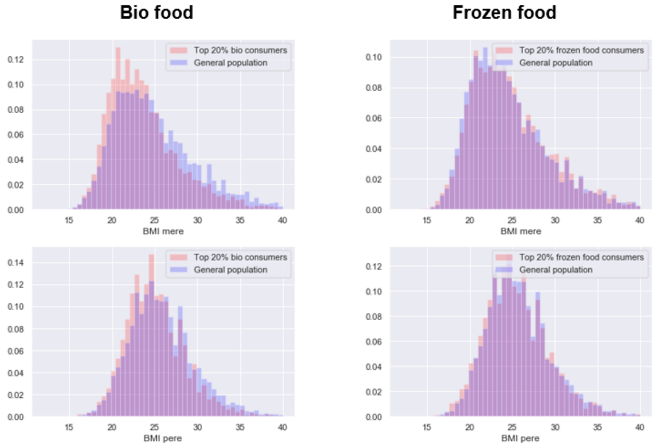
Topic 16

"Sausages++"

Focus: impact of topics on BMI

Left: Bio/organic topic
Top row: Women

Right: Frozen food topic
Bottom row: Men



High weight of Bio topic is correlated with lower BMI ($p < 5\%$) (particularly so for women).

Does A (eat bio) cause B (better BMI) ?

Three cases

- A does cause B (bio food is better)
- Confounder: exists C that causes A and B (rich/young/educated people tend to consume bio products and have lower BMI);
- Backdoor effects: exists C correlated with A which causes B (people eating bio also tend to eat more greens, which causes lower BMI);

Goal: Find out which case holds

Causal models

- Ideally based on randomized controlled trials   Imbens Rubins 15

Proposed Methodology

Taking inspiration from Abadie Imbens 06

Target population: "Bio" people = top quantile coordinate on bio topic.

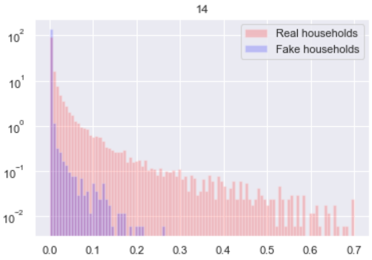
RCT would require a control population

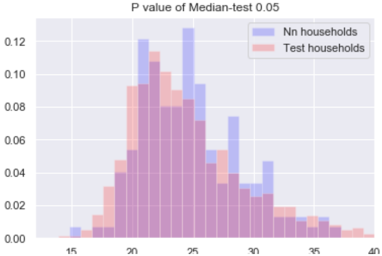
Building a control population finding matches

- For each bio person, take her consumption z (basket of products)
- Create a falsified consumption z' (replacing each bio product with same, but non-bio, product)
- Find true consumption $z``$ nearest to z' (in LDA space)
- Let the true person with consumption $z``$ be called "falsified bio"

Compare bio and "falsified bio" populations wrt BMI

Bio vs Falsified Bio populations





Left

- Projection on the Bio topic (in log scale)
- (Falsified bio population not 0: the bio topic contains e.g. sheep yogurt).

Right

- BMI Histograms of both bio and falsified bio populations
- Statistically significant difference

Next

Chasing confounders

- Discriminating bio from "falsified bio" populations w.r.t. socio-professional features: accuracy $\approx 60\%$
- Candidate confounder: mother education level (on-going study)

Next steps

- Confirm conjectures using longitudinal data (2015-2016)
- Interact with nutritionists / sociologists
- Extend the study to consider the impact of, e.g.
  - Price of the food
  - Amount of trans fats
  - Amount of added sugar

Perspectives: Causality analysis and Big Data

Finding the needle in the haystack

- Redundant variables (e.g. in economics) $\rightarrow$ un-interesting relations
- Variable selection
- Feature construction dimensionality reduction

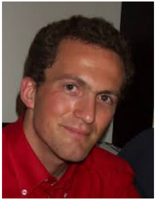
Beyond causal sufficiency

- Confounders are all over the place (and many are plausible, e.g. age and size of firm; company ownership and shareholdings)
- When prior knowledge available, condition on counfounders
- Use causal relationships on latent variables Wang and Blei, 19
  to filter causal relationships on initial variables

Thanks!







Isabelle Guyon, Diviyan Kalainathan, Olivier Goudet, David Lopez-Paz,
Philippe Caillou, Paola Tubaro,
Ksenia Gasnikova