Machine Learning and Causal Inference

Nataliya Sokolovska

Sorbonne University
Paris, France

Master 2 in Statistics
February, 25, 2020

Outline

My Motivation: Causal Inference by Machine Learning

Causal Graphs and Bayesian Nets

Independence Mechanisms

Noise Models

Information-Geometric Approaches

Algorithmic Information Methods

Causal Discovery in Real Data

Common Hidden Causes

Causal Inference: a View

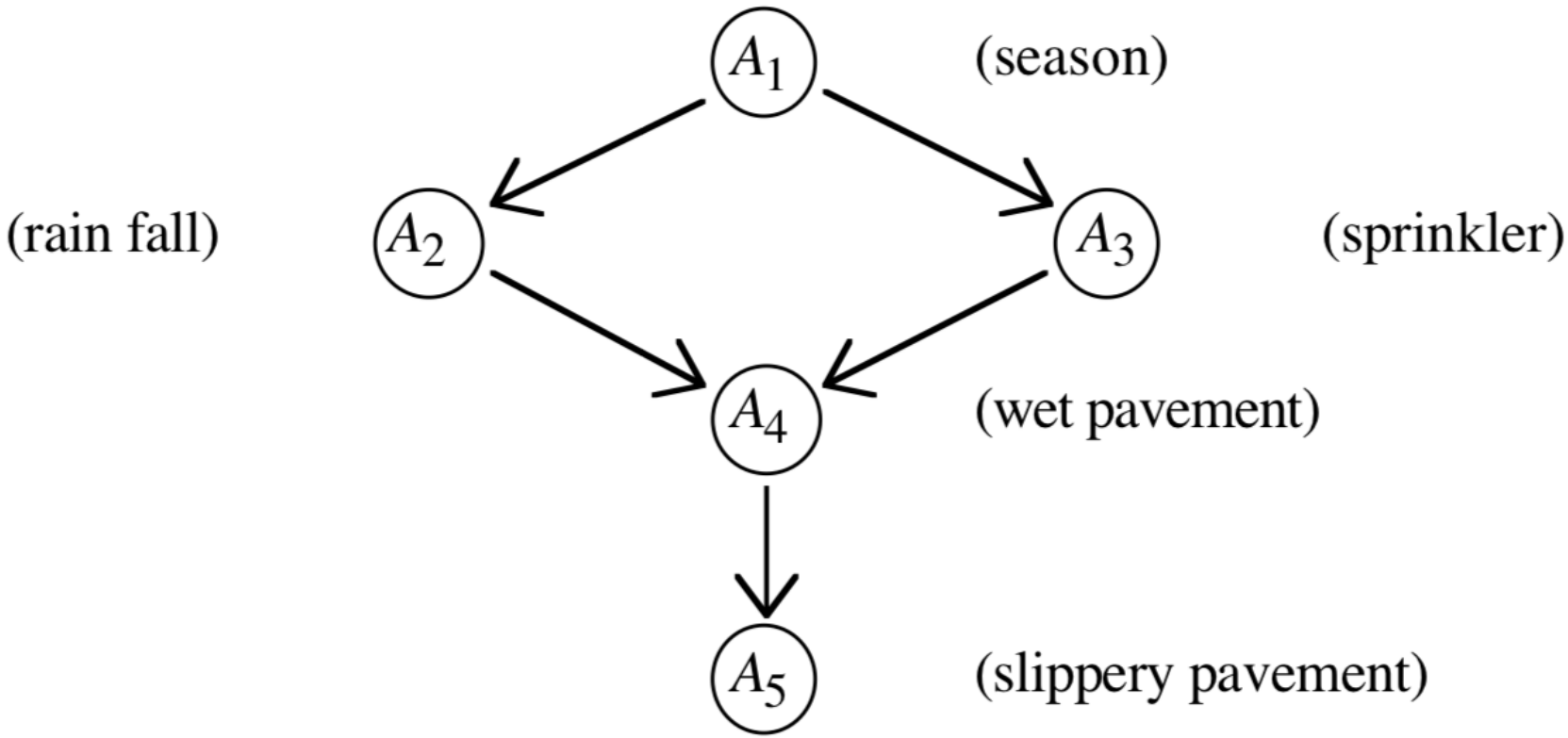
The book *Elements of Causal Inference* by J. Peters, D. Janzing, and B. Schölkopf, 2017

- Computational causality methods are in their infancy
- Bivariate case where the system under analysis contains two observables only
- Machine learning influence
- Absence of time series
- Causal inference is harder than typical ML problems

Causal Graphs and Bayesian Nets

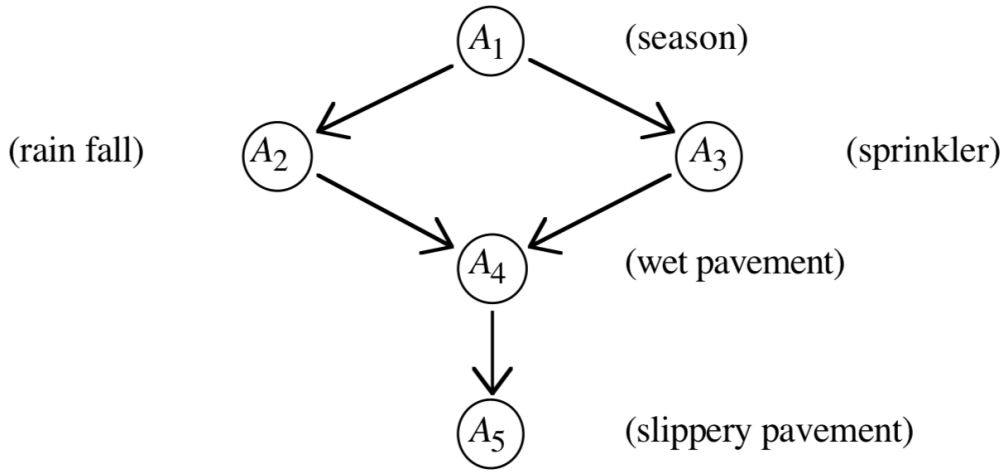
- A DAG $< U, E >$ is a *causal* graph, if the edges in E are *given a causal interpretation*: an edge $A \rightarrow B$ is interpreted as A causes B.
- Graph-theoretic representations:
  - causal dependencies
  - conditional probabilistic dependencies
- There exists a close relation between causal and probabilistic relations:
  - each causal graph is a Bayesian net

Causal Graphs and Bayesian Nets: An Example



from *Pearl, 1998; Spohn, 2009*

Causal Graphs and Bayesian Nets: An Example



- The event $\{A = a\}$ is a direct cause of the event $\{B = b\}$ in the possible world u iff both events occur in u, and if $\{A = a\}$ precedes $\{B = b\}$, and if $\{A = a\}$ is positively related to $\{B = b\}$ (according to P).

from *Pearl, 1998; Spohn, 2009*

Independent Mechanisms

Suppose we have estimated the joint density $p(a, t)$ of the altitude A and the temperature T on a sample of cities in some country.

$$p(a, t) = p(a|t)p(t) \tag{1}$$
$$= p(t|a)p(a) \tag{2}$$

Independent Mechanisms

Suppose we have estimated the joint density $p(a, t)$ of the altitude A and the temperature T on a sample of cities in some country.

$$p(a, t) = p(a|t)p(t) \tag{1}$$
$$= p(t|a)p(a) \tag{2}$$

$$p(a|t)p(t) \text{ (factorization according to } T \to A) \tag{3}$$
$$p(t|a)p(a) \text{ (factorization according to } A \to T) \tag{4}$$

from *Peters, 2017*

Independent Mechanisms

$$p(a|t)p(t) \text{ (factorization according to } T \rightarrow A) \tag{5}$$
$$p(t|a)p(a) \text{ (factorization according to } A \rightarrow T) \tag{6}$$

Consider the **effect of interventions**:

- Intervening on A has changed T
  - We climb higher
  - The temperature is lower
- Intervening on T has not changed A
  - We do not change the altitude
  - We build a massive heating system around the city that raises the temperature

Physical Mechanism

If we change the altitude A, we assume that the physical mechanism

$$p(t|a) \tag{7}$$

is responsible for producing an average temperature, and changes T.

Physical Mechanism

If we change the altitude A, we assume that the physical mechanism

$$p(t|a) \tag{7}$$

is responsible for producing an average temperature, and changes T.

$$\underbrace{p(t|a) \text{ and } p(a)}_{\text{are independent}} \tag{8}$$

- $p(a)$: locations of cities can be different
- $p(t|a)$: would apply for different locations

Even a more detailed example

Let us consider the joint distributions of altitude and temperature in Austria $p^{\ddot{o}}(a, t)$ and Switzerland $p^{s}(a, t)$.

- Austrians and Swiss founded their cities in different places, and $p^{\ddot{o}}(a)$ and $p^{s}(a)$ are different
- However,
  - $p^{\ddot{o}}(a, t) = p(t|a)p^{\ddot{o}}(a)$
  - $p^{s}(a, t) = p(t|a)p^{s}(a)$
- $p(t|a)$ is the same
- no influence of $p^{\ddot{o}}(a)$ and $p^{s}(a)$ on $p(t|a)$

from *Peters, 2017*

Independence Mechanism

If $A \rightarrow T$ is the correct causal structure, then

- it is in principle possible to perform a localised intervention of A, i.e. to change $p(a)$ without changing $p(t|a)$
- $p(a)$ and $p(t|a)$ are autonomous, modular, or invariant mechanisms

Noise Models

Incorporate an independent noise term in the model to compare the evidences of the two directions.
Here are some of the noise models for the hypothesis $Y \rightarrow X$ with the noise E:

- Additive noise: $Y = f(X) + E$
- Linear noise: $Y = pX + qE$
- Post-non-linear: $Y = G(f(X) + E)$
- Heteroskedastic noise: $Y = f(X) + EG(E)$
- Functional noise: $Y = f(X, E)$

The common assumption in these models are:

- There are no other causes of Y
- X and E have no common causes
- Distribution of cause is independent from causal mechanisms

Additive Noise Models

The joint distribution $P_{X,Y}$ is said to admit an Additive Noise Model from X to Y if there is a measurable function f_Y and a noise variable N_Y such that

$$Y = f_Y(X) + N_Y, N_Y \perp\!\!\!\perp X. \tag{9}$$

We say that $P_{Y|X}$ admits an ANM if the equation above holds.

Information-Geometric Causal Inference (IGCI)

Another idea how to formalize independence of $P_{E|C}$ and P_C

- Assumption (strong):
  - Deterministic relation between X and Y
  - $Y = f(X)$ and $X = f^{-1}(Y)$ (the noise variable is constant)
- The independence between P_X and f implies dependence between P_Y and $f^{-1}(Y)$.

from *D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniusis, B. Steudel, B. Schölkopf. Information-geometric approaches to inferring causal directions. 2012.*

IGCI

Definition (IGCI model) $P_{X,Y}$ is said to satisfy an IGCI model from X to Y if the following conditions hold: $Y = f(X)$ for some diffeomorphism f of $[0,1]$ that is strictly monotonic and satisfies $f(0) = 0$ and $f(1) = 1$. Moreover, P_X has the strictly positive continuous density P_X, such that the following "independence condition" holds:

$$cov[\log f', p_X] = 0, \tag{10}$$

where $\log f'$ and p_X are considered as random variables on the probability space $[0,1]$ endowed with the uniform distribution.

log is taken for deterministic monotonically increasing relations

IGCI

Theorem (Identifiability of IGCI models) Assume the distribution $P_{X,Y}$ admits an IGCI model from X to Y. Then the inverse function f^{-1} satisfies:

$$cov[\log f^{-1'}, p_Y] \geq 0, \tag{11}$$

with equality if and only if f is the identity.

In other words, uncorrelatedness of $\log f'$ and p_X implies positive correlation between $\log f^{-1'}$ and p_Y except for the trivial case $f = id$.

Trace Condition

J. Zscheischler, D. Janzing, K. Zhang. Testing whether linear equations are causal: a free probability theory approach, 2012

$$Y = AX + E \text{ or} \tag{12}$$

$$X = \tilde{A}Y + \tilde{E} \text{ ?} \tag{13}$$

Trace Condition

J. Zscheischler, D. Janzing, K. Zhang. Testing whether linear equations are causal: a free probability theory approach, 2012

$$Y = AX + E \text{ or} \tag{12}$$

$$X = \tilde{A}Y + \tilde{E} \text{ ?} \tag{13}$$

- Covariance matrices Σ of the cause and the matrix A (independent mechanism of nature) generating the effect from the cause are independent.
- Matrices B and C are independent if:

$$\tau_n(BC) \approx \tau_n(B)\tau_n(C), \tag{14}$$

$$\tau_n(\cdot) := tr(\cdot)/n. \tag{15}$$

Trace Condition

$$Y = AX + E \text{ or} \tag{16}$$
$$X = \tilde{A}Y + \tilde{E} \text{ ?} \tag{17}$$

Are Σ_X and A independent?

$$\tau_n(AA^T\Sigma_X) \approx \tau_n(A^TA)\tau_n(\Sigma_X) \iff X \to Y \tag{18}$$

- Estime $\Delta_{X\to Y}$ and $\Delta_{Y\to X}$, $\Delta \approx 0$: correct causal direction

$$\Delta_{X\to Y} := \log \tau_n(A\Sigma_X A^T) - \log \tau_n(\Sigma_X) - \log \tau_n(AA^T) \tag{19}$$
$$\tag{20}$$

- requires good estimates of A, $\tilde{A}$, Σ_X, and $\tilde{\Sigma}_X$

Trace Condition

Some details:

- Some theory:

$$\lim_{n\to\infty} \hat{\Delta}_{X_n \to Y_n} = 0 \text{ in the correct direction} \tag{21}$$

$$\lim_{n\to\infty} \hat{\Delta}_{Y_n \to X_n} \leq 0 \text{ violation in the backward direction} \tag{22}$$

- A case with confounders:

$$\lim_{n\to\infty} \hat{\Delta}_{X_n \to Y_n} \leq 0 \text{ and} \tag{23}$$

$$\lim_{n\to\infty} \hat{\Delta}_{Y_n \to X_n} \leq 0 \text{ there exist a confounder } Z \tag{24}$$

Algorithmic Information Models

Compare two programs, both of which output both X and Y.

- Store Y and a compressed form of X in terms of uncompressed Y.
- Store X and a compressed form of Y in terms of uncompressed X.

The shortest such program implies the uncompressed stored variable more-likely causes the computed variable

CURE

- CURE (Causal inference with Unsupervised inverse REgression) infers "X causes Y" if the estimation of the conditional probability $P(X|Y)$ is more accurate than the estimation in the opposite direction.
- How to quantify the accuracy of the conditional probabilities?
- Compare the difference between the unsupervised and supervised log-likelihoods obtained from N pairs $\{X_i, Y_i\}_{i=1}^{N}$ of observations:

$$D_{X|Y} = \mathcal{L}_{X|Y}^{\text{unsup}} - \mathcal{L}_{X|Y}^{\text{sup}} \tag{25}$$

and

$$D_{Y|X} = \mathcal{L}_{Y|X}^{\text{unsup}} - \mathcal{L}_{Y|X}^{\text{sup}} \tag{26}$$

E. Sgouritsa, D. Janzing, P. Henning, B. Schölkopf, *Inference of cause and effect with unsupervised inverse regression, 2015*

CURE

$$D_{X|Y} = \mathcal{L}^{\text{unsup}}_{X|Y} - \mathcal{L}^{\text{sup}}_{X|Y} = \tag{27}$$

$$- \frac{1}{N}\sum_{i=1}^{N} \log p(X_i|Y_i, \mathbf{y}) + \frac{1}{N}\sum_{i=1}^{N} \log p(X_i|Y_i, \mathbf{x}, \mathbf{y}), \tag{28}$$

and

$$D_{Y|X} = \mathcal{L}^{\text{unsup}}_{Y|X} - \mathcal{L}^{\text{sup}}_{Y|X} = \tag{29}$$

$$- \frac{1}{N}\sum_{i=1}^{N} \log p(Y_i|X_i, \mathbf{x}) + \frac{1}{N}\sum_{i=1}^{N} \log p(Y_i|X_i, \mathbf{x}, \mathbf{y}). \tag{30}$$

- The conditional probability $p(X_i|Y_i, \mathbf{y})$ is estimated from Y observed but X are not observed
- $p(Y_i|X_i, \mathbf{x})$ is estimated using observed X only, and $p(X_i|Y_i, \mathbf{x}, \mathbf{y})$ are computed when both X and Y are observed.
- if $D_{X|Y} < D_{Y|X}$, then the inferred causal direction is $X \to Y$, otherwise $Y \to X$.

Causal Discovery with Distance Correlation

- F. Liu and L. Chan. *Causal Inference on Discrete Data via Estimating Distance Correlations*, 2017
- for discrete (or discretized) data

The dependence measures are defined as follows:

$$D_{Y|X} = dCor(P(X), P(Y|X)) \tag{31}$$

$$D_{X|Y} = dCor(P(Y), P(X|Y)), \tag{32}$$

where $dCor(a, b)$ is the distance correlation.

Causal Inference Algorithm with Distance Correlation

Input: Samples X and Y, a threshold ϵ
Output: Causality directions

STEP 1: Compute $P(X)$ and $P(Y|X)$ from data,
Estimate $D_{Y|X} = \mathcal{D}(P(X), P(Y|X))$

STEP 2: Compute $P(Y)$ and $P(X|Y)$ from data,
Estimate $D_{X|Y} = \mathcal{D}(P(Y), P(X|Y))$

STEP 3: Decide the edge direction:
if $D_{Y|X} - D_{X|Y} > \epsilon$ **then**
  Infer $X \rightarrow Y$
end if
if $D_{X|Y} - D_{Y|X} > \epsilon$ **then**
  Infer $Y \rightarrow X$
end if

Note that if $|D_{Y|X} - D_{X|Y}| < \epsilon$, then the approach can not provide any edge orientation.

Comparing Regression Errors

P. Blöbaum, D. Janzing, T. Washio. *Cause-effect inference by comparing regression errors*, 2018.

- Exploit asymmetries in MSE: the prediction error is smaller in causal direction
- cause C, effect E
- $\phi(c) = \mathrm{E}[E|c]$: least squares minimizer to predict E from C
- $\psi(c) = \mathrm{E}[C|e]$: least squares minimizer to predict C from E
- (Intuition: it is easier to predict the effect)

$$\mathrm{E}[(E - \psi(C))^2] \leq \mathrm{E}[(C - \phi(E))^2] \tag{33}$$

$$\mathrm{E}[var(E|C)] \leq \mathrm{E}[var(C|E)] \tag{34}$$

Comparing Regression Errors

$$E_\alpha := \phi(C) + \alpha N, \tag{35}$$

α is a parameter controlling the noise level.

$$\lim_{\alpha \to 0} \frac{\mathrm{E}[var(C|\tilde{E}_\alpha)]}{\mathrm{E}[var(\tilde{E}_\alpha|C)]} \geq 1. \tag{36}$$

- $\tilde{E}_\alpha$ – shifted and rescaled variables
- In practice: fit MSE regression in both directions, and then decide the direction
- Underfitting is good

Causal Discovery in Real Data

The MicroObes corpus contains heterogeneous biomedical data of obese patients:

- NutriOmics team, Pitié-Salpêtrière hospital examined 49 patients.
- Environmental, data, alimentary patterns reflecting nourishing habits of subjects, and also information about their physical activity.
- The host data: measurements of glucose homeostasis markers, blood lipids, inflammatory markers and adipokines, body composition, kidney function, and subcutaneous adipose tissue (AT) markers.
- Abundance matrices of gut flora genes, namely, bacterial quantification (qPCR), and abundance of bacterial clusters (MGS) of individual patients.

How to reveal causal relations between the groups of variables?

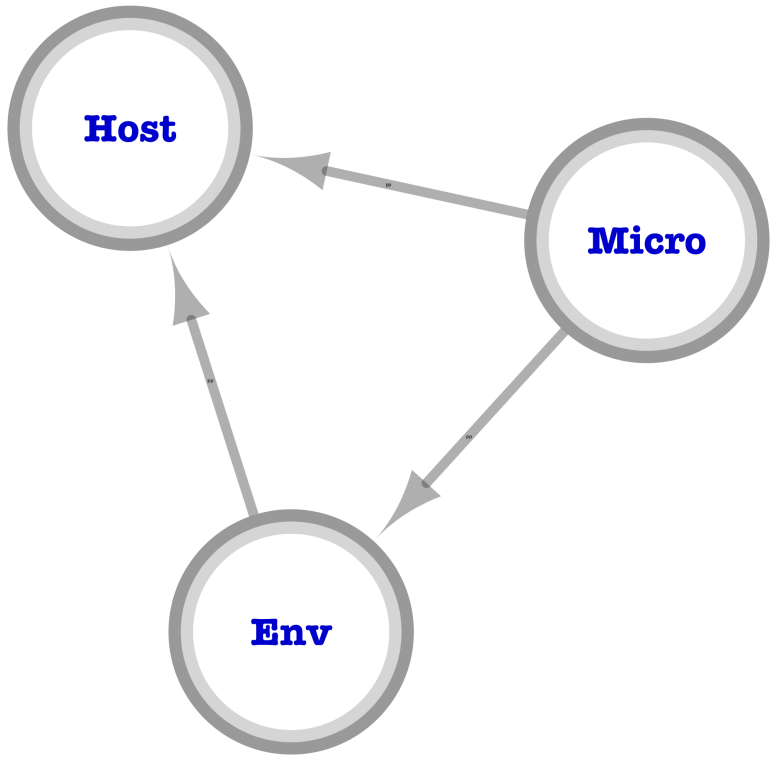
Causality Discovery in Real Data

The parameters of the MicroObes data can be naturally divided into several groups. After discussions with the clinicians of the Pitié-Salpêtrière hospital, we decided to consider two scenarios:

1. Find causal relations between 3 groups (environment, host, and bacteria)
2. Find causal directions between 10 groups (glucose homeostasis markers, blood lipids, inflammatory markers and adipokines, body composition, kidney function, subcutaneous AT markers, food groups, nutrients, physical activity, and gut flora bacteria).

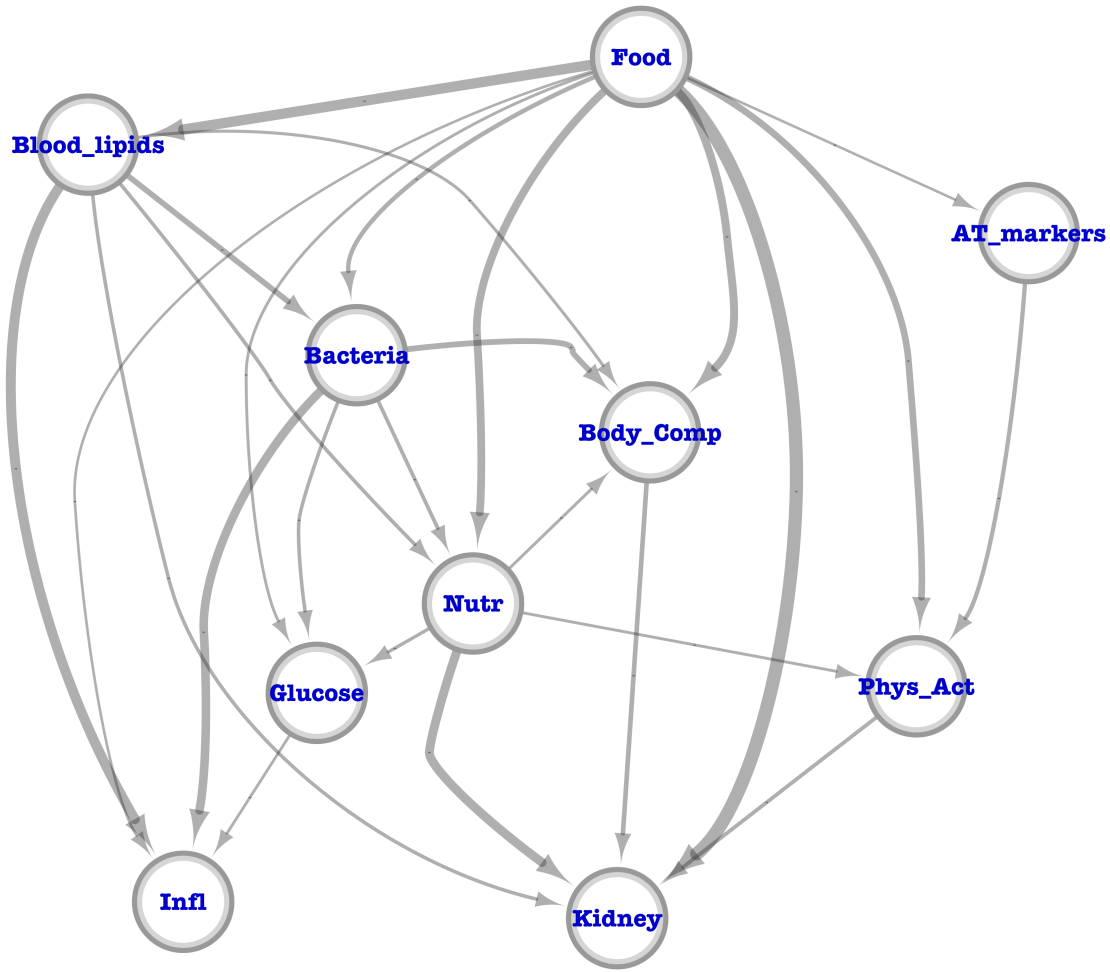
Three groups

Find causal relations between 3 groups (environment, host, and bacteria)



Ten groups

(glucose homeostasis markers, blood lipids, inflammatory markers and adipokines, body composition, kidney function, subcutaneous AT markers, food groups, nutrients, physical activity, and gut flora bacteria).



Ten groups

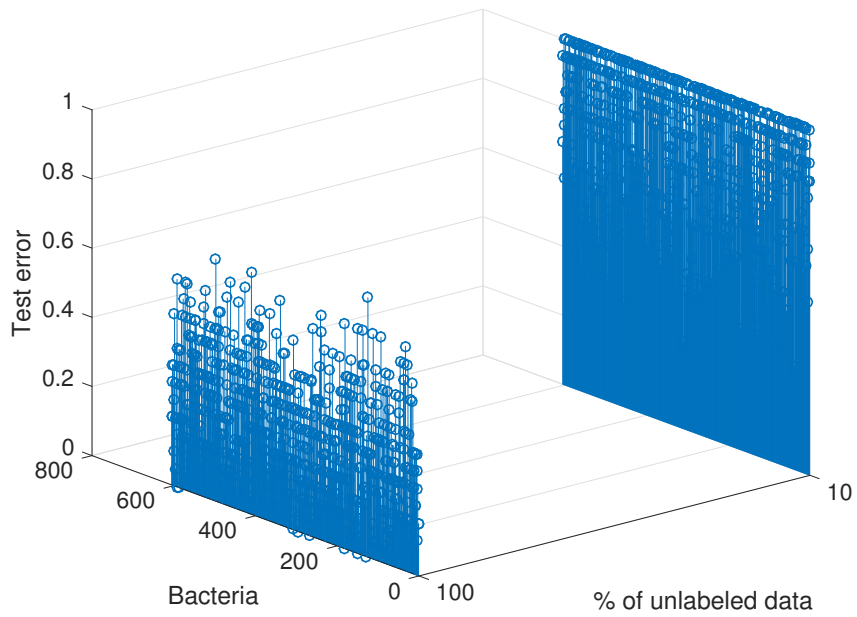
We observe that food has an important impact on the level of blood lipids, on the composition of the human gut, on the markers of the adipose tissue, on body composition, on physical activity, on glucose level, on the inflammation markers, and on the kidney function. The width of an edge is equal to an average taken over 10 runs of the absolute values $|D_{X|Y} - D_{Y|X}|$. Such a heuristic can be a reasonable indicator of significance of a causal direction. We kept all edges whose strength is bigger than 0.1. Note that the node *Nutr* stands for nutrition and includes nutritional values, calculated by clinicians from the food questionnaires, and, therefore, the causal direction $Food \rightarrow Nutr$ was expected. The hypothesis that the gut flora (*Bacteria*) can have an important impact on the inflammatory status is verified in the data.

Impact of Drugs on Human Gut Flora

- Associations between chronic human diseases and alterations in gut microbiome composition[1].
- Treatment causes changes in human gut flora? It affects metabolism?
- The human gut microbiome of type 2 diabetes is confounded by metformin treatment; metformin impacts the composition and richness of the human gut microbiome.
- The data set of *K. Forslund et al., 2015*:
  - a multi-country metagenomic dataset (Danemark, China, and Sweden).
  - The data contains information of 106 patients with type 2 diabetes who take the metformin, and 93 patients with the diabetes who does not take the drug.
  - The features are 785 gut metagenomes or gut bacteria.

[1] K. Forslund et al., Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. 2015

Impact of Drugs on Human Gut Flora



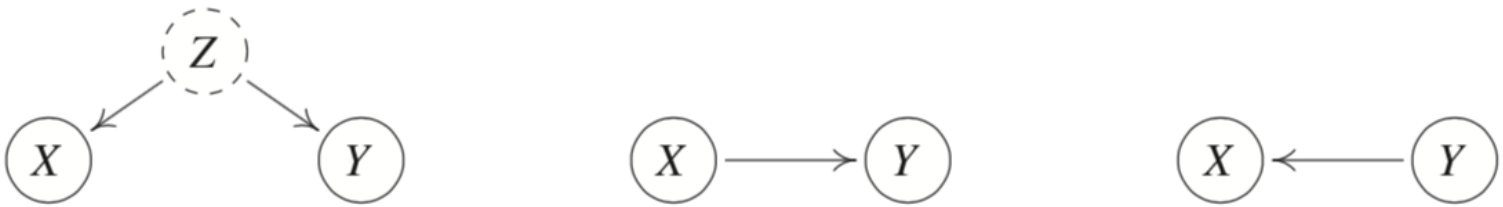
- Make a hypothesis that metformin impacts all bacteria
- For the majority of the bacteria considered in the experiments, this relation is obvious with the error rate equal to 0.
- Current publications focus on a very limited number of bacteria species, and the statement that the metformin impacts the metagenome is not necessarily true for all bacteria of the human gut flora.

Latent variables

According to the Reichenbach's principle of common sense, if there exists a statistical dependency between two observable variables X and Y, it indicates that there exists a variable Z which causes X and Y. If we assume that Z coincides either with X or with Y what leads directly to inferring causality $X \to Y$ or $Y \to X$.

Latent variables

According to the Reichenbach's principle of common sense, if there exists a statistical dependency between two observable variables X and Y, it indicates that there exists a variable Z which causes X and Y. If we assume that Z coincides either with X or with Y what leads directly to inferring causality $X \to Y$ or $Y \to X$.



Detecting Low-Complexity Confounders

Pairwise pure conditionals. The conditional distribution $P(Y|X)$ is said to be pairwise pure if for any two $x_1, x_2 \in \mathcal{X}$ the following condition holds. There is no $\lambda < 0$ or $\lambda > 1$ for which

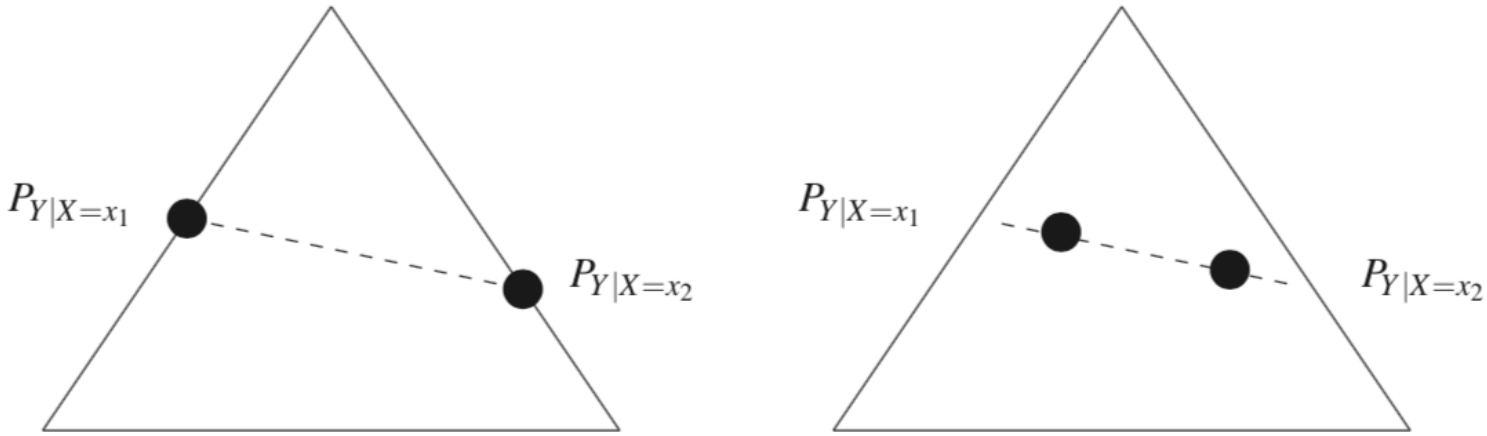
$$\lambda P(Y|X = x_1) + (1 - \lambda)P(Y|X = x_2) \tag{37}$$

is a probability distribution.
The purity is defined by the following condition:

$$\inf_{y \in \mathcal{Y}} \frac{p(y|x_1)}{p(y|x_2)} = 0, \text{ for all } x_1, x_2 \in \mathcal{X}. \tag{38}$$

D. Janzing, E. Sgourisa, O. Stegle, J. Peters, B. Schölkopf. *Detecting low-complexity unobserved causes*, 2011.

Detecting Low-Complexity Confounders



- On the left: a pure conditional, extending the line connecting the two points $P_{Y|X=x_1}$ and $P_{Y|X=x_2}$ would leave the simplex (of probability distributions)
- On the right: a non-pure conditional, the line connecting $P_{Y|X=x_1}$ and $P_{Y|X=x_2}$ can be extended without leaving the simplex

from *Peters et al., 2017*

Detecting Low-Complexity Confounders

In practice, to decide whether the pairwise purity holds, we can estimate

$$\min_{y \in \mathcal{Y}} \frac{\hat{p}(y|x)}{\hat{p}(y|x')} \text{ for all } x, x'. \tag{39}$$

If the conditional distribution is pure, then the path between X and Y is not intermediated by a confounder Z.

D. Janzing, E. Sgourisa, O. Stegle, J. Peters, B. Schölkopf. *Detecting low-complexity unobserved causes*, 2011.