A la découverte du graph causal !

Exploration sur des cas d'usage concrets

Fabien Faivre
Responsable R&D et innovation data

Julien Capitaine
Data Scientist

Georges Oppenheim
Professor emeritus

Alessandro Leite
Advanced Research Position

Tiphanie Huang
Junior Consultant

Audrey Poinsot
PhD student

Marianne Clausel
Professor

Myriam Tami
Associate Professor in AI

Agenda de l'atelier

1h

Présentation

Théorie & Méthode Eki

3h

Mise en pratique

Benchmark Eki method

Plan de la présentation

1
Récap - Atelier 1
Corrélation n'implique pas effet causal

2
Theorie
Causal discovery : les algos!

3
Méthode Eki
Une méthode hybride - data & expertise

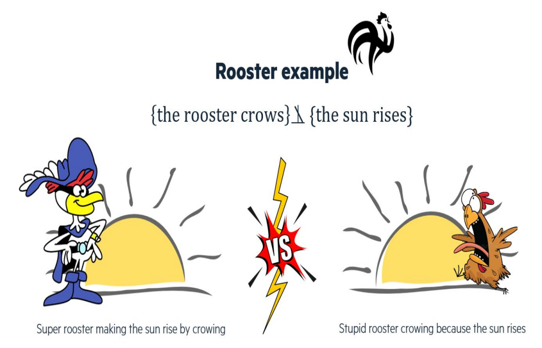
1

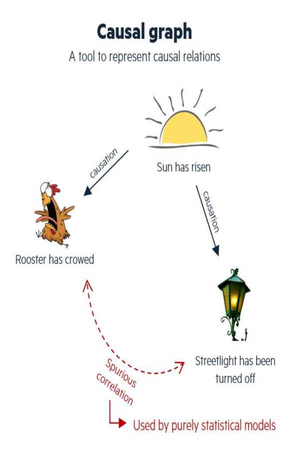
Récap - Atelier 1

Corrélation n'implique pas effet causal

Corrélation n'implique pas effet causal

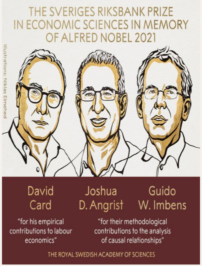
"You are smarter than your data. Data do not understand causes and effects; humans do." *Judea Pearl, The book of Why*



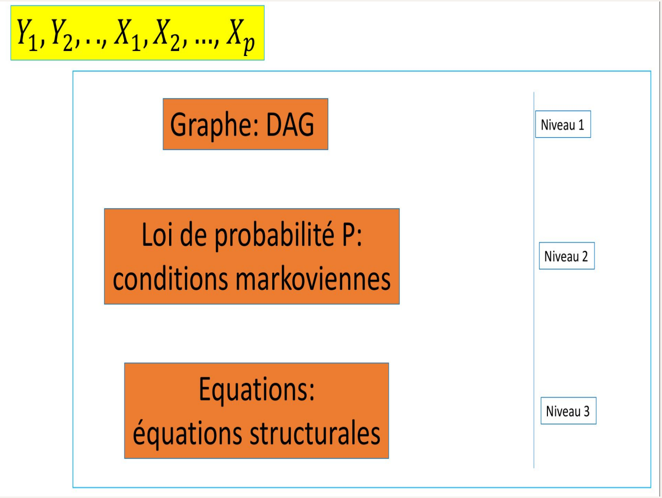


Enjeux : Généralisation & compréhension des phénomènes

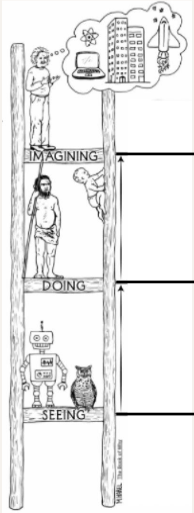
Un intérêt grandissant pour la causalité



Modèle CAUSAL



De nouvelles opportunités avec la causalité



3. Contrefait - Penser l'existant modifié

Si j'avais agi différemment, quel aurait été le résultat?
Si je n'avais pas pris l'aspirine, est ce que j'aurais toujours mal à la tête ?

→ Et si X=x avait été X=x' ?
$P(\mathbf{y'}|do(x'), \mathbf{y})$

2. Intervention - Agir sur le monde

Que serait Y, si je faisais X ?
Si je prends de l'aspirine, est-ce que ma douleur s'arrêtera ?

→ Si je fais X=x ?
$P(y|\mathbf{do}(x))$

1. Observation - Etre passif

Dans quelle mesure, observer X change ma croyance sur Y?
Est ce que le symptôme X est lié / associé à la maladie Y ?

→ Si je vois X=x ?
$P(y|x)$

2

Theory

Causal Discovery : les algos!

Observational causal discovery

- **Similar to machine learning**

Observational causal discovery

- **Similar to machine learning**
  - Given the data, infer the causal models

Observational causal discovery

- **Similar to machine learning**
  - Given the data, infer the causal models
  - Data quality, quantity, and learning criterion may be challenging

Observational causal discovery

- **Similar to machine learning**
  - Given the data, infer the causal models
  - Data quality, quantity, and learning criterion may be challenging
- **Difference**: functional causal models

Observational causal discovery

- **Similar to machine learning**
  - Given the data, infer the causal models
  - Data quality, quantity, and learning criterion may be challenging
- **Difference**: functional causal models
  - **Assumptions**

Observational causal discovery

- **Similar to machine learning**
  - Given the data, infer the causal models
  - Data quality, quantity, and learning criterion may be challenging
- **Difference**: functional causal models
  - **Assumptions**
    - **Causal sufficiency**: no unobserved confounders

Observational causal discovery

- **Similar to machine learning**
  - Given the data, infer the causal models
  - Data quality, quantity, and learning criterion may be challenging
- **Difference**: functional causal models
  - **Assumptions**
    - **Causal sufficiency**: no unobserved confounders
    - **Causal Markov**: all d-separations in the causal graph G imply conditional independence in the observational distribution P

Observational causal discovery

- **Similar to machine learning**
  - Given the data, infer the causal models
  - Data quality, quantity, and learning criterion may be challenging
- **Difference**: functional causal models
  - **Assumptions**
    - **Causal sufficiency**: no unobserved confounders
    - **Causal Markov**: all d-separations in the causal graph G imply conditional independence in the observational distribution P
    - **Causal faithfulness**: all conditional independence in P imply d-separations in G

Observational causal discovery

- **Similar to machine learning**
  - Given the data, infer the causal models
  - Data quality, quantity, and learning criterion may be challenging
- **Difference**: functional causal models
  - **Assumptions**
    - **Causal sufficiency**: no unobserved confounders
    - **Causal Markov**: all d-separations in the causal graph G imply conditional independence in the observational distribution P
    - **Causal faithfulness**: all conditional independence in P imply d-separations in G

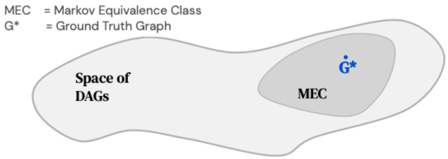


Image credit Rosemary and Bauer, 2021

Challenges and principles

- In general, causal discovery from observational data is impossible.
- But, it is possible under additional assumptions.
- Several approaches in the literature
  - Constraint based methods: run local tests of independence to create constraints on space of possible graphs.
  - Score-based methods: use the fact that each DAG can be scored in relation to the data, by using a penalized likelihood score
  - Noise based methods: find footprints in the noise that imply causal asymmetry.
  - ...

Assumptions and output format of causal discovery methods

	PC	FCI	GES	GIES	MMHC	LINGAM	backShift
Causal sufficiency	✓	✗	✓	✓	✓	✓	✗
Causal faithfulness	✓	✓	✓	✓	✓	✗	✗
Acyclicity	✓	✓	✓	✓	✓	✓	✗
Non-Gaussian errors	✗	✗	✗	✗	✗	✓	✗
Unknown shift interventions	✗	✗	✗	✗	✗	✗	✓
Known do-interventions	✗	✗	✗	✓	✗	✗	✗
Output	CPDAG	PAG	CPDAG	PDAG	DAG	DAG	DG

CPDAG – completed partially directed acyclic graph, **DAG** – directed acyclic graph, **FCI** – fast causal inference, **GES** – greedy equivalence search, **GIES** – greedy interventional equivalence search, **LINGAM** – linear non-Gaussian acyclic models, **MMHC** – max-min hill climbing, **PAG** – partial ancestral graph, **PC** – Peter-Clark, **PDAG** partially directed acyclic graph

Causal discovery: preparing the data I

1. **Handle missing data** appropriately (e.g., by interpolating or excluding them) or choose causal discovery methods that are robust to them
2. Ensure that variables are **"semantically independent"** – (not mathematically interdefinable) and **independently manipulable**
   - remove redundant variables (e.g., HDL cholesterol, LDL cholesterol, and total cholesterol, where total = HDL + LDL)
   - The variables to remove depend on *domain knowledge*, as there is no universal rules for determining which one to remove
   - General guideline is to ensure that there is no collinearity in the data set (e.g., checking if the covariance matrix is invertible)
3. Most causal search algorithms assume that variables are either continuous or categorical
   - **Discretization** should be done very carefully
   - Different discretizations can lead to various independence judgement and consequently different inferred causal structures

Causal discovery: preparing the data II

- Discretization can also make nonlinear causal dependencies difficult to detect
- Ideally, a discretization strategy should result in *causally-appropriate bins* preserving relevant causal relationships

4. There may exist multiple **proxy measurements** for non-observable variables of interest
   - Ensure that proxy measurements are accurate estimates of a single non-observable causal factor
   - Otherwise, choose a search method that can discover proxy relationships
5. Ensure the observations represent measurements of different individuals or of the same individual over time
   - Time series data require additional constraints for causal inference, and thus, demand different causal search algorithms
6. Ensure correct background knowledge about the potential causal relations

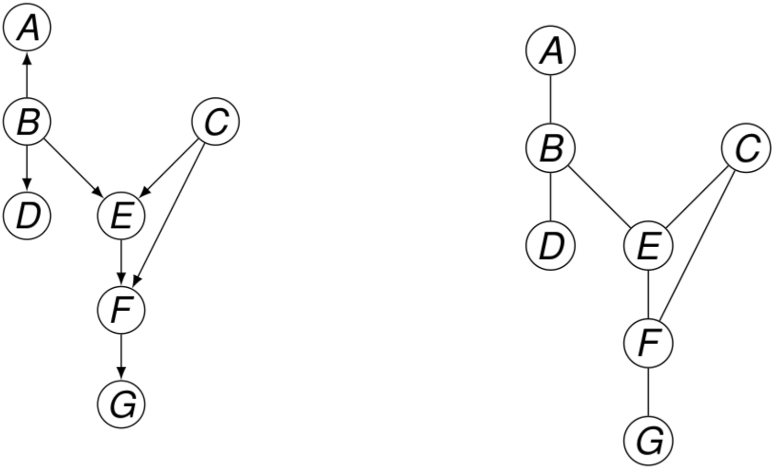
Principles of constraint-based methods

- Focus on discovering the set of causal graphs that imply the conditional independencies found in the data by performing a sequence of hypothesis tests.

Main steps

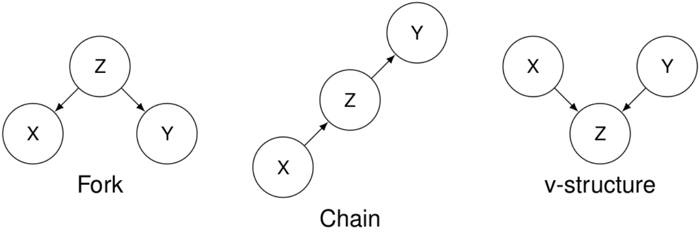
- Find skelton
- Find v-structures
- Orient other edges using basic rules

Principles of constraint-based methods



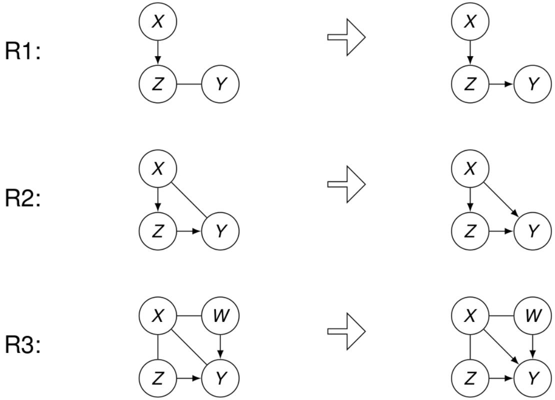
a DAG and its corresponding skeleton

Principles of constraint-based methods



Fork, chains and v-structures

Principles of constraint-based methods



Basic rules

Principles of constraint-based methods

Algorithm 1 SGS

Input: $P(\mathcal{V})$
Output: CPDAG $\mathcal{G}^*$

Form the complete undirected graph $\mathcal{G}^*$ on vertex set $\mathcal{V}$
for all $X - Y$ in $\mathcal{G}^*$
and subsets $\mathcal{S} \subseteq \mathcal{V} \backslash \{X, Y\}$ **do**
  if $\exists \mathcal{S} \subseteq \mathcal{V} \backslash \{X, Y\}$ such that $X \perp\!\!\!\perp_P Y \mid \mathcal{S}$ **then**
    Delete edge $X - Y$ from $\mathcal{G}^*$
  end if
end for
for all $X - Z - Y$ in $\mathcal{G}^*$ such that $X \notin Adj(Y, \mathcal{G})$ **do**
  if $\nexists\, \mathcal{S} \subseteq \mathcal{V} \backslash \{X, Y\}$ such that $Z \in \mathcal{S}$ and $X \perp\!\!\!\perp_P Y \mid \mathcal{S}$ **then**
    Orient $X \rightarrow Z \leftarrow Y$ in $\mathcal{G}^*$
  end if
end for
Recursively apply rules R1-R3 until no more edges can be oriented
Return $\mathcal{G}^*$

Independence tests: some examples

Type of variable	Example of independence test
Discrete	χ^2 test
Gaussian	Test based on the precision matrix
Non Gaussian continuous	Non parametric tests Mutual Information (MI), RKHS

See notebook `CI.ipynb` for more details

The concept of d separation

Blocked paths

A path is said to be blocked by a set of vertices Z if:

- it contains a chain $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ and $B \in Z$, or
- it contains a collider $A \rightarrow B \leftarrow C$ such that no descendant of B is in Z

The concept of d separation

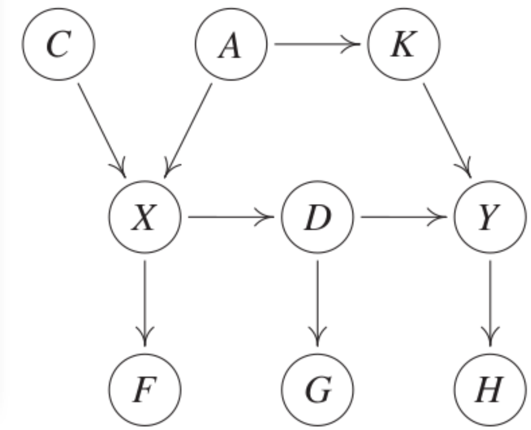
Definition

Two (sets of) nodes X and Y are d-separated by a set of nodes Z if all of the paths between (any node in) X and (any node in) Y are blocked by Z. We denote $X \perp\!\!\!\perp_G Y|Z$

Theorem

Two DAGs G_1 and G_2 have the same d-separations iff they have the same skeleton and the same v-structures.

The concept of *d* separation



For this DAG : $C \perp\!\!\!\perp_G G | \{X\}$ and $C \not\perp\!\!\!\perp_G G | \{X, H\}$

The PC algorithm

- PC algorithm : optimized version of SGS
- Infer causal structure with the PC algorithm?
  - Infer mutual dependencies between variables : skeleton of the causal graph
  - Distinguish between causes and effects : orientation of the v-structures of the causal graph

The PC algorithm

An example

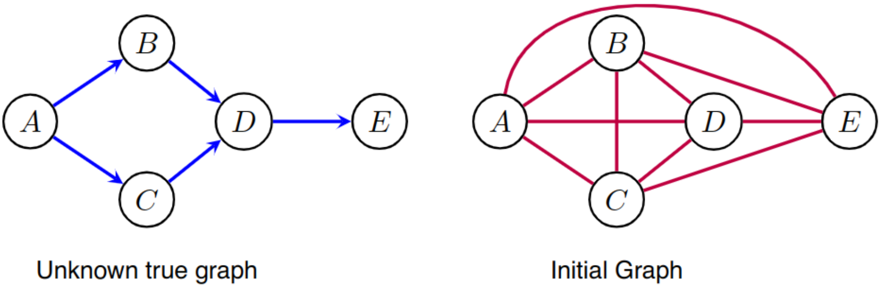


Figure: The initial graph is complete

The PC algorithm

An example

Example of PC Algorithm : k=0

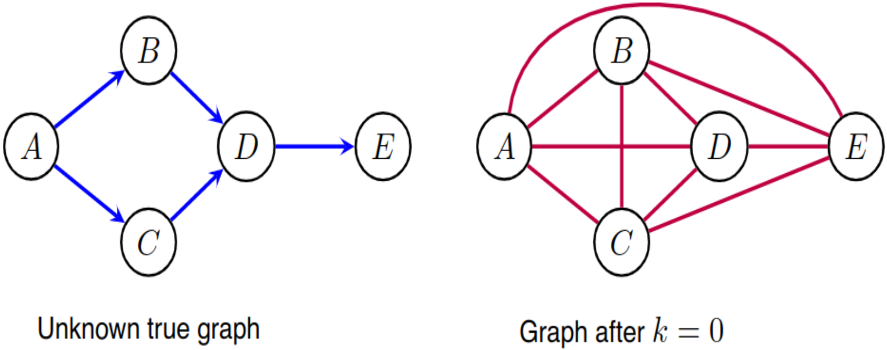


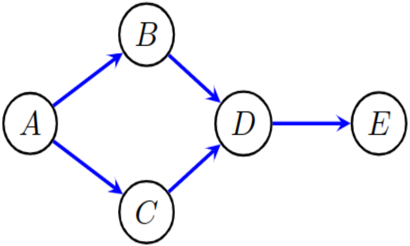
Figure: The initial graph is complete

There is no pair of variables d-separated given $\emptyset$, so the graph is

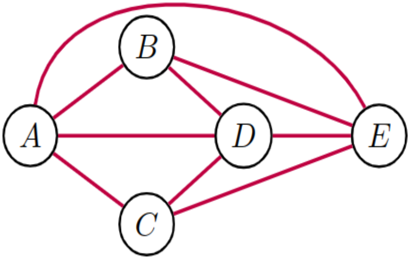
The PC algorithm

An example

Example of PC Algorithm $k = 1$ with (B, C)



Unknown true graph



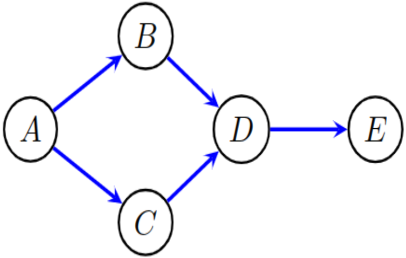
Graph

Since B and C are d-separated given $\{A\}$ we remove the $B - C$ edge and record $S_{BC} = \{A\}$.

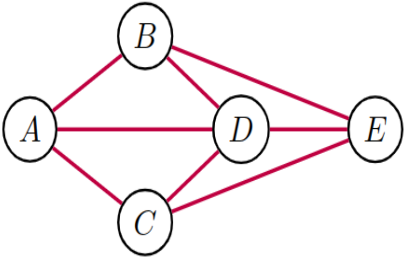
The PC algorithm

An example

Example of PC Algorithm : $k = 1$ with (A, E)



Unknown true graph



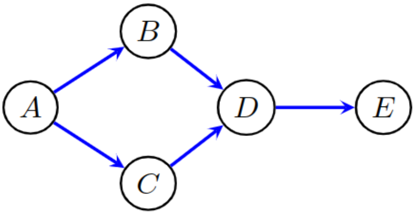
Graph

Since A and E are d separated given $\{D\}$ we remove the $A - E$ edge and record $S_{AE} = \{D\}$.

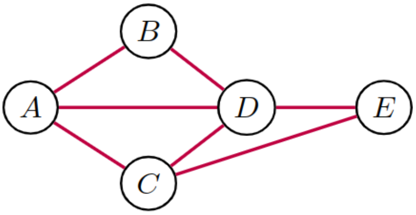
The PC algorithm

An example

Example of PC Algorithm : $k = 1$ with (B, E)



Unknown true graph



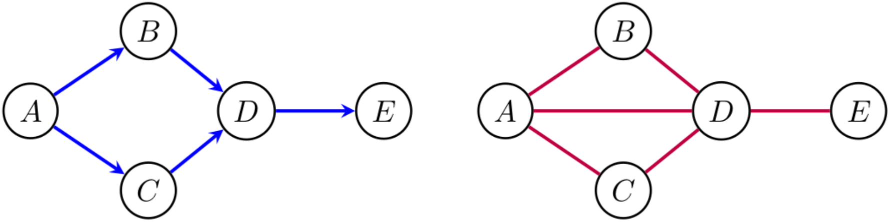
Graph

Since B and E are d separated given $\{D\}$ we remove the $B - E$ edge and record $S_{BE} = \{D\}$.

The PC algorithm

An example

Example of PC Algorithm : $k = 1$ with (C, E)



Unknown true graph

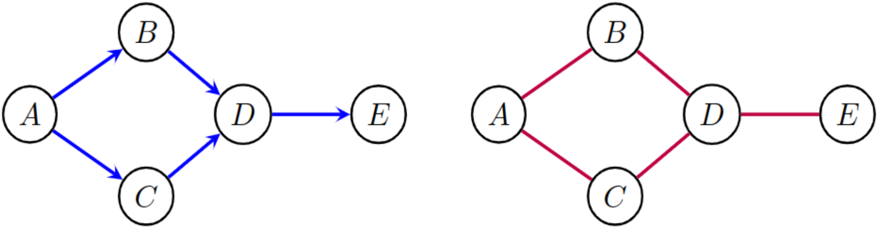
Graph after $k = 1$

Since C and E are d separated given $\{D\}$ we remove the $C - E$ edge and record $S_{CE} = \{D\}$. This completes this stage

The PC algorithm

An example

Example of PC Algorithm : $k = 2$ with (A, D)



Unknown true graph

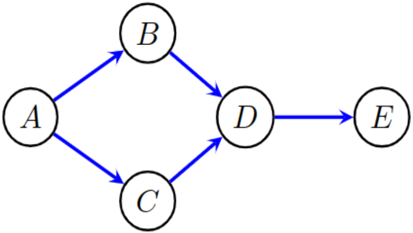
Graph after $k = 2$

Since A and D are d separated given $\{B, C\}$ we remove the $A - D$ edge and record $S_{AD} = \{B, C\}$.This completes this stage.

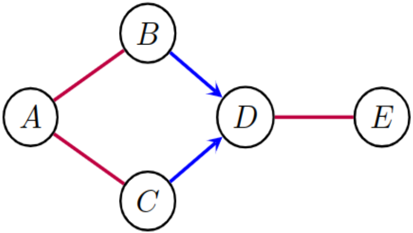
The PC algorithm

An example

Orienting unshielded colliders from separating sets



Unknown true graph



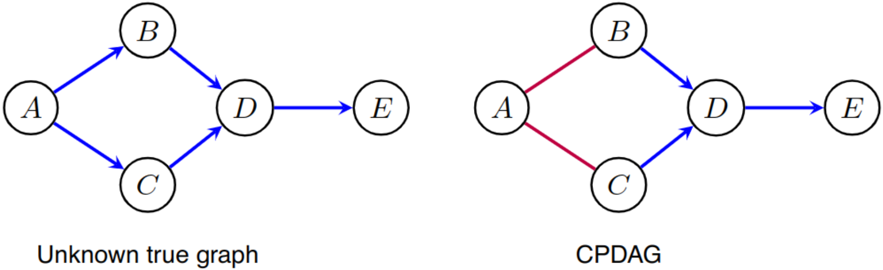
Graph after $k = 1$

Since $D \notin S_{BC} = \{A\}$, we orient $B \to D \leftarrow C$.The other triples (B, A, C),(A, B, D),(A, C, D),(B, D, E) and (C, D, E) do not lead to further orientation; since the middle vertex is in each separating set

The PC algorithm

An example

Additional orientations to form CPDAG



Since (B, D, E) is not a collider, but $B \rightarrow D$ we can orient $D - E$ as $D \rightarrow E$

Cause or consequence?

- Can we distinguish cause from effect?
- That is distinguish between these two causal graphs

$$X \to Y$$

or

$$Y \to X$$

using observational data.

`Not always possible!`

Cause or consequence?

The example of linear structural equation [f linear]

X cause Y if there exists $a \in \mathbb{R}$, ε^Y s.t.

$$Y = aX + \varepsilon^Y, X \perp\!\!\!\perp \varepsilon^Y.$$

Distinguish cause from consequence? [Shimizu et al., 2006]

Assume that $Y = aX + \epsilon^Y, X \perp\!\!\!\perp \epsilon^Y$ where all r.v. are continuous. Then

$$\exists b \in \mathbb{R}, \varepsilon^X \text{ s.t. } X = bY + \varepsilon^X, Y \perp\!\!\!\perp \varepsilon^X$$

iff (X, ε^X) are Gaussian random variables.

Existence of a non-linear extension of this result.

Noise based algorithm

Theorem (LINGAM)

Assume a linear SCM with graph $G = (V, E)$ and a compatible distribution $P(V)$ such that or all $Y \in V$

$$Y = \sum_{X \in Pa(Y)} a_{xy} X + \xi_Y$$

where all ξ_Y are jointly independent and non-Gaussian distributed. Additionally, we require that for all $Y \in V$, $X \in Pa(Y)$, $a_{xy} \neq 0$. Then, the graph G is identifiable from $P(V)$.

Noise based algorithm

LINGAM

Algorithm 1 LiNGAM

Input: $P(\mathcal{V})$
Output: $\mathcal{G}$

Form an empty graph $\mathcal{G}$ on vertex set $\mathcal{V} = \{X_1, \cdots, X_p\}$
Let $S = \{1, \cdots, p\}$ and $\mathcal{T} = []$
repeat
  $H = []$
  for $i \in S$ **do**
    for $j \in S \setminus \{i\}$ **do**
      $\hat{\xi}_{ij} = X_j - \frac{cov(X_i, X_j)}{var(X_i)} X_i$
    end for
    $h = \sum_{j \in S\setminus\{i\}} \mathcal{I}(X_i, \hat{\xi}_{ij})$
    $H = [H, h]$
  end for
  $i^* = arg \min_{i \in S} H$
  $S = S \setminus \{i^*\}$
  $\mathcal{T} = [\mathcal{T}, i^*]$
  $\forall j \in S, X_j = \hat{\xi}_{i^* j}$
until $|S| = 0$
Append($\mathcal{T}$, S_0)
Construct a strictly lower triangular matrix by following the order in $\mathcal{T}$, and estimate the connection strengths $a_{i,j}$ by using some conventional covariance-based regression.
if $a_{i,j} > 0$ **then**
  Add $X_i \rightarrow X_j$ to $\mathcal{G}$
end if
Return $\mathcal{G}$

The PC algorithm

An example

Algorithm 1 The PC_{pop}-algorithm

INPUT: Vertex Set V, Conditional Independence Information
OUTPUT: Estimated skeleton C, separation sets S (only needed when directing the skeleton afterwards)
Form the complete undirected graph $\tilde{C}$ on the vertex set V.
$\ell = -1; \quad C = \tilde{C}$
repeat
    $\ell = \ell + 1$
    repeat
        Select a (new) ordered pair of nodes i, j that are adjacent in C such that $|adj(C,i) \setminus \{j\}| \geq \ell$
        repeat
            Choose (new) $\mathbf{k} \subseteq adj(C,i) \setminus \{j\}$ with $|\mathbf{k}| = \ell$.
            if i and j are conditionally independent given $\mathbf{k}$ **then**
                Delete edge i, j
                Denote this new graph by C
                Save $\mathbf{k}$ in $S(i,j)$ and $S(j,i)$
            end if
        until edge i, j is deleted or all $\mathbf{k} \subseteq adj(C,i) \setminus \{j\}$ with $|\mathbf{k}| = \ell$ have been chosen
    until all ordered pairs of adjacent variables i and j such that $|adj(C,i) \setminus \{j\}| \geq \ell$ and $\mathbf{k} \subseteq adj(C,i) \setminus \{j\}$ with $|\mathbf{k}| = \ell$ have been tested for conditional independence
until for each ordered pair of adjacent nodes i, j: $|adj(C,i) \setminus \{j\}| < \ell$.

The PC algorithm

An example

Algorithm 2 Extending the skeleton to a CPDAG

INPUT: Skeleton G_{skel}, separation sets S
OUTPUT: CPDAG G
for all pairs of nonadjacent variables i, j with common neighbour k **do**
  if $k \notin S(i, j)$ **then**
    Replace $i - k - j$ in G_{skel} by $i \rightarrow k \leftarrow j$
  end if
end for

In the resulting PDAG, try to orient as many undirected edges as possible by repeated application of the following three rules:

R1 Orient $j - k$ into $j \rightarrow k$ whenever there is an arrow $i \rightarrow j$ such that i and k are nonadjacent.

R2 Orient $i - j$ into $i \rightarrow j$ whenever there is a chain $i \rightarrow k \rightarrow j$.

R3 Orient $i - j$ into $i \rightarrow j$ whenever there are two chains $i - k \rightarrow j$ and $i - l \rightarrow j$ such that k and l are nonadjacent.

R4 Orient $i - j$ into $i \rightarrow j$ whenever there are two chains $i - k \rightarrow l$ and $k \rightarrow l \rightarrow j$ such that k and l are nonadjacent.

Noise based algorithm

Theorem (ANM)

Assume that an SCM with graph $G = (V, E)$ is given and a compatible distribution P(V) such that for all $Y \in V$

$$Y = f((X \in Pa(Y)) + \xi_Y$$

where all ξ_Y are jointly independent.
Then, the graph G is identifiable from $P(V)$.

Noise based algorithm

ANM

Algorithm 2 ANM

Input: $P(\mathcal{V})$

Output: $\mathcal{G}$

Form an empty graph $\mathcal{G}$ on vertex set $\mathcal{V} = \{X_1, \cdots, X_p\}$

Let $S = \{1, \cdots, p\}$ and $\mathcal{T} = []$

repeat

  $H = []$

  for $j \in S$ **do**

    $\hat{f}_{.j}$: Regress X^j on $\{X_i\}_{i \in S \backslash \{j\}}$

    $\hat{\xi}_{.j} = X_j - \hat{f}_{.j}(X_i)$

    $h = \hat{I}(\{X_i\}_{i \in S \backslash \{j\}}, \hat{\xi}_{.j})$

    $H = [H, h]$

  end for

  $i^* = \arg\min_{i \in S} H$

  $S = S \backslash \{i^*\}$

  $\mathcal{T} = [i^*, \mathcal{T}]$

until $|S| = 0$

for $j \in \{2, \cdots, p\}$ **do**

  for $i \in \{\mathcal{T}_1, \cdots, \mathcal{T}_{j-1}\}$ **do**

    $\hat{f}_{.j}$: Regress X^j on $\{X_k\}_{k \in \{\mathcal{T}_1, \cdots, \mathcal{T}_{j-1}\} \backslash \{i\}}$

    $\hat{\xi}_{.j} = X_j - \hat{f}_{.j}(X_i)$

    if $\{X_k\}_{k \in \{\mathcal{T}_1, \cdots, \mathcal{T}_{j-1}\} \backslash \{i\}} \not\perp\!\!\!\perp_P \hat{\xi}_{.j}$ **then**

      Add $X_i \to X_j$ to $\mathcal{G}$

    end if

  end for

end for

Return $\mathcal{G}$

3

Méthode Eki

Une méthode hybride - data & expertise

Notre constat: Les données et les humains sont “biaisés”

Fully data-based Causal Discovery

- Measurement error
- Unobserved variables
- Observational bias
- Small data regimes
- ... and many more

Fully expert-based Causal Discovery

- Wrong knowledge
- non-instantaneous reasoning
- Human biases
- ... and many more

Hybrid method

Hoping for :

- humans to give hints on potential data biases
- data to reveal human biases

Notre but: Faire coïncider les indépendances cond.

Data

- **Indep. 2 à 2**
  - $X \not\perp\!\!\!\perp Y$
- **Indep. Cond.**
  - $X \perp\!\!\!\perp Y | W$

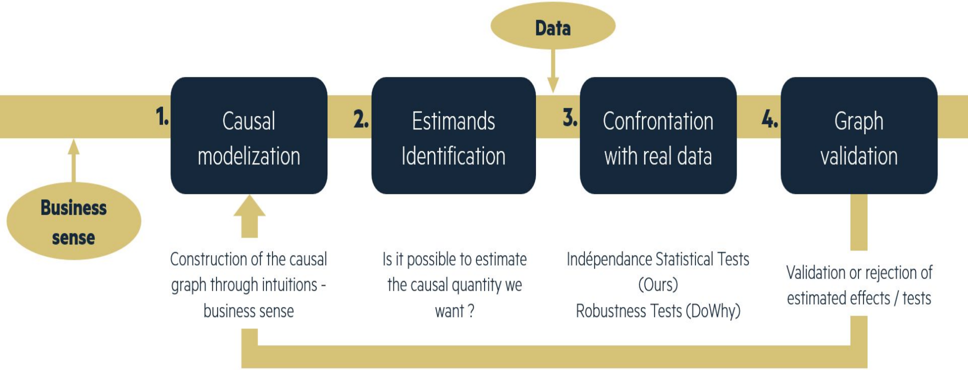
Graph

- **Existence d'un chemin →**
  - *X vers Y ou*
  - *Y vers X ou*
  - *X et Y ont un parent en commun*
- **D-separation**
  - *X et Y sont d-séparés par W*

Les résultats des tests d'indépendances sont toujours acceptés/rejetés par un humain

Notre méthode

Inputs : Data + expertise/apriori about the input variables causal relation



Output : *(A non-unique)* Causal Graph in line with the data

A vous de jouer !

Le but: Faire échouer notre méthode / identifier ses limites

A dispo:

- Code méthode hybride Eki
- Code pour générer des SCMs
- Un dataset “marketing”
- D’autres méthodes de Causal Discovery

Possibles axes d’exploration:

- Faible régime de données, grande dimension, niveaux de bruit, ...
- Connaissance experte erronée
- Benchmark face à d’autres méthodes de causal discovery
- ...

Annexes