Causalité observationnelle

Découverte de liens de cause à effet sans expériences randomisées

Diviyan Kalainathan

Doctorant, TAU, INRIA,
Université Paris Sud,
Université Paris Saclay,
France

Plan

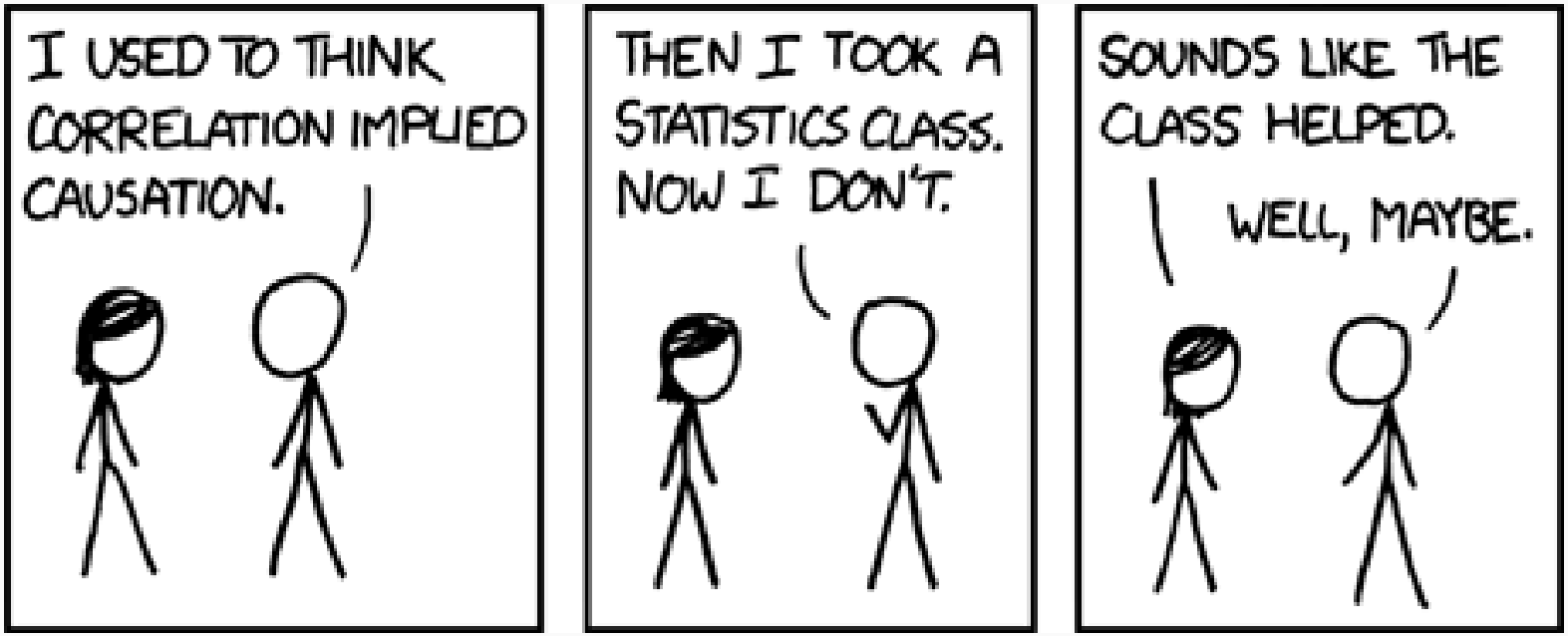
1. Introduction
2. Causal discovery
3. Algorithmes
4. Conclusion

1. Introduction

Causalité ?

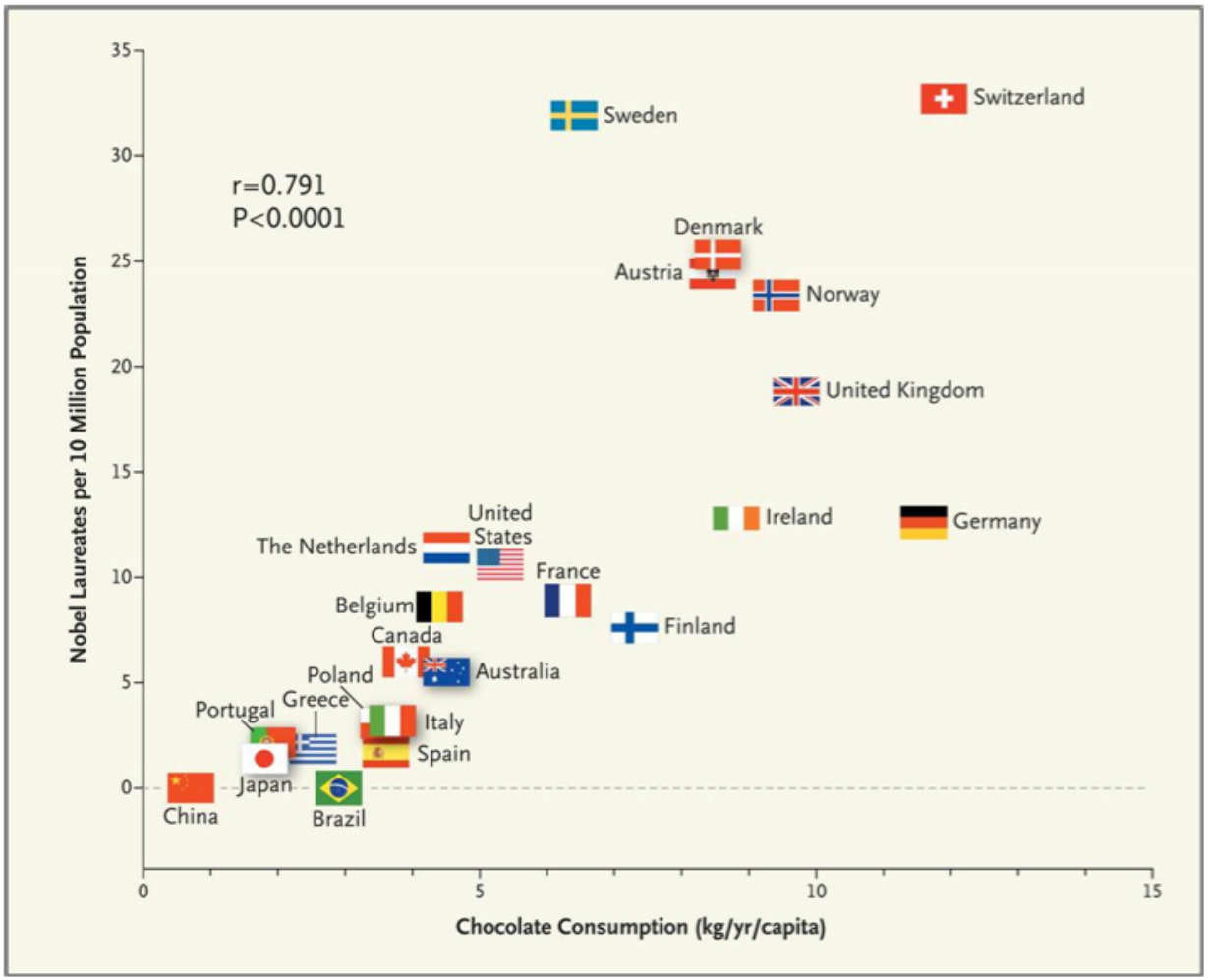
Manque de précautions $\rightarrow$ Douleurs

Correlation $\neq$ Causation



Credit: XKCD

Correlation $\neq$ Causation (2)



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

Credit: Jonas Peters

Une définition de causalité

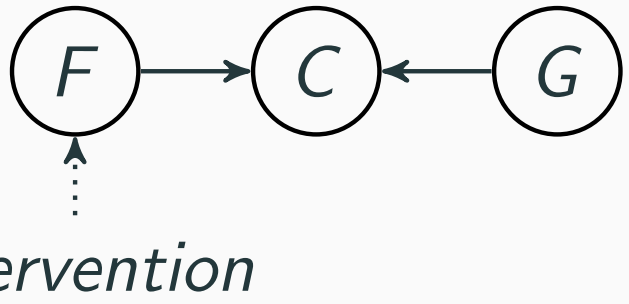
- Intervention $do(X = x_1)$ force la variable X à la valeur x_1 [Pearl, 2009]
- Direct cause $X \to Y$:

$$P_{Y|\mathrm{do}(X=x_1,\mathbf{S}_{\setminus XY}=\mathbf{c})} \neq P_{Y|\mathrm{do}(X=x_2,\mathbf{S}_{\setminus XY}=\mathbf{c})} \quad (1)$$

avec $\mathbf{S}_{\setminus XY}$ l'ensemble de toutes les variables privées de X et Y.

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

C: Cancer, F : Fumer, G : Facteurs génétiques.



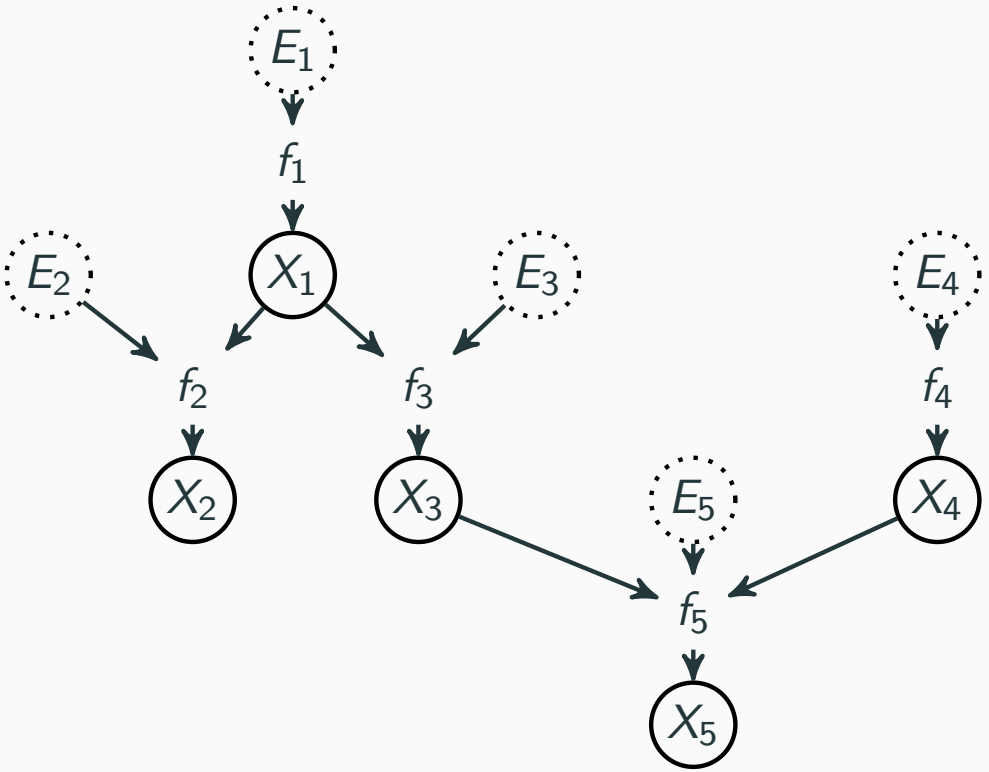
Objectif

- Entrée: Données $\mathbf{X}$: n examples $\times$ d features $(X_1, \ldots, X_d)$
- Contexte: Données observationnelles, pas d'intervention possible
- Sortie: Graphe $\mathcal{G}$ causal

Functional Causal models

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

$X_{\mathrm{Pa}(i;\mathcal{G})}$ est le set de parents de X_i dans $\mathcal{G}$, E_i est une variable de bruit aléatoire, f_i est une fonction déterministique

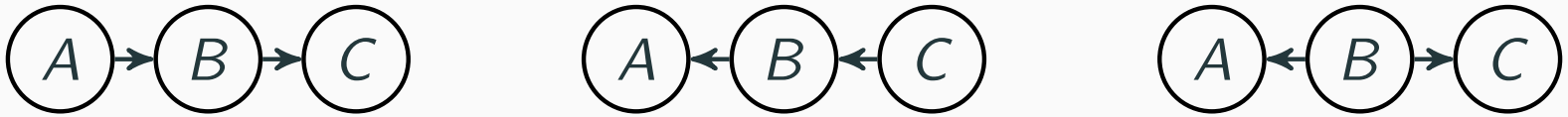


$$\begin{cases} X_1 = f_1(E_1) \\ X_2 = f_2(X_1, E_2) \\ X_3 = f_3(X_1, E_3) \\ X_4 = f_4(E_4) \\ X_5 = f_5(X_3, X_4, E_5) \end{cases}$$

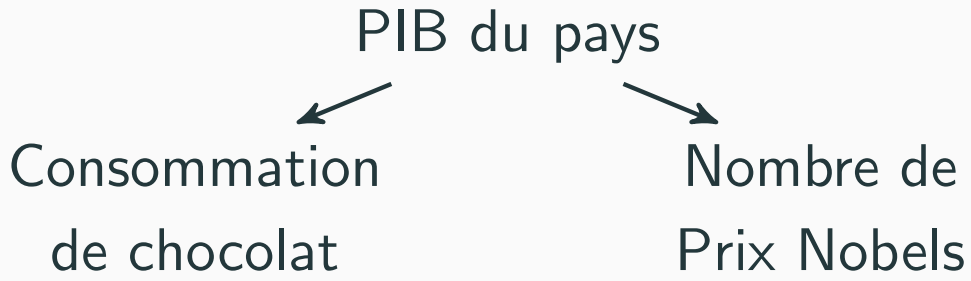
2. Causal discovery

Relations statistiques causales

Markov Equivalent Classes: $A \not\perp\!\!\!\perp C$ **and** $A \perp\!\!\!\perp C | B$

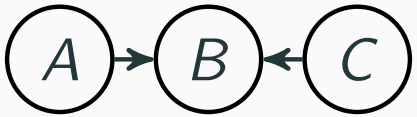


Example

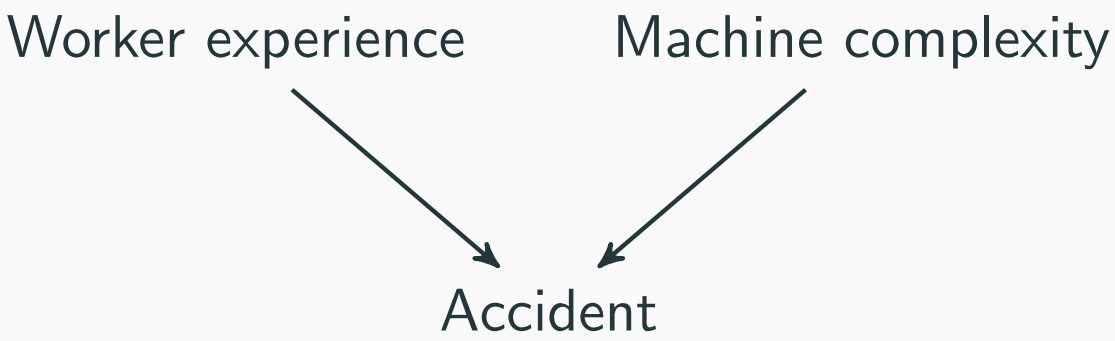


V-structures

V-Structure: $A \perp\!\!\!\perp C$ **and** $A \not\perp\!\!\!\perp C | B$

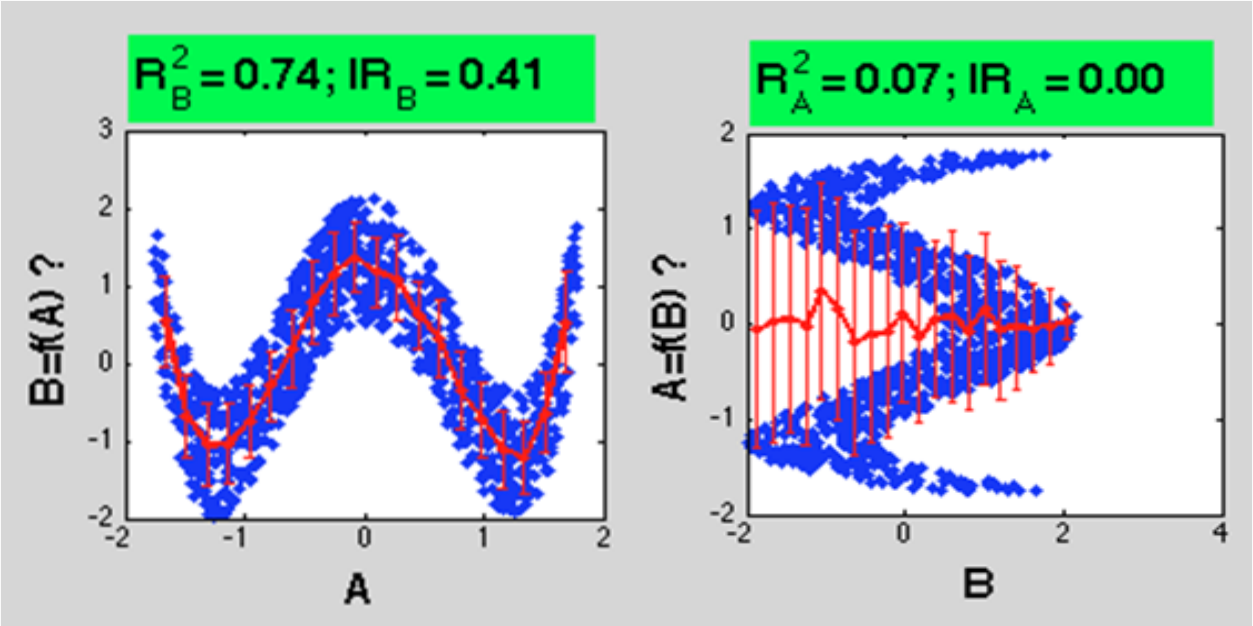


Example



Assymétries dans les distributions

- Utilise les asymmétries dans les distributions de variables
- Argument de simplicité



Paradoxe de Simpson

Traitements de calculs rénaux

Succès en fonction de la taille des calculs

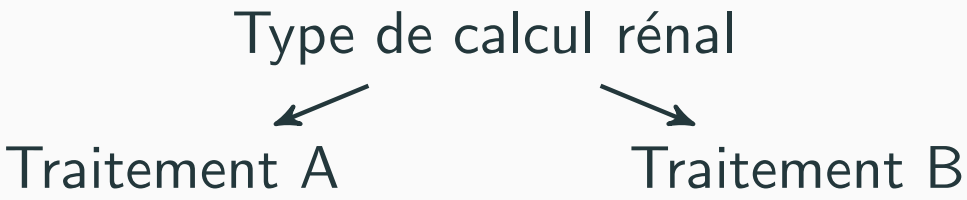
petits calculs		gros calculs	
Traitement A	**Traitement B**	**Traitement A**	**Traitement B**
93 % (81/87)	87 % (234/270)	73 % (192/263)	69 % (55/80)

Résultats agrégés

Taux de succès (succès/total)

Traitement A	Traitement B
78 % (273/350)	83 % (289/350)

Que s'est il passé ?



$\Rightarrow$ **Présence de facteurs confondants + Aggrégation: très dangeureux !**

Hypothèses

- **Causal Sufficiency**: Pas de variable confondantes en dehors du dataset
- **Causal Markov**: Toutes les variables sont indépendantes de leurs non-effets conditionnellement aux parents
- **Causal Faithfulness**: Toute dépendence conditionelle statistique présente dans les données provient du vrai graphe $\mathcal{G}$

3. Algorithmes

Algorithmes à contraintes

Utilisent les dépendences et indépendances conditionnelles sur des variables et ensembles de variables, ainsi que de la propagation de contraintes pour orienter les liens.

Algorithmes en 2 étapes:

1. Trouver le squelette du graphe avec des relations de dépendence.
2. Orienter le squelette trouvé avec les v-structures et la propagation de contraintes.

[Spirtes et al., 2000],[Drton and Maathuis, 2016]

Limitations

- Choix du test statistique ?
- Complexité de l'algorithme $\Rightarrow$ temps de calcul

Algorithmes à score

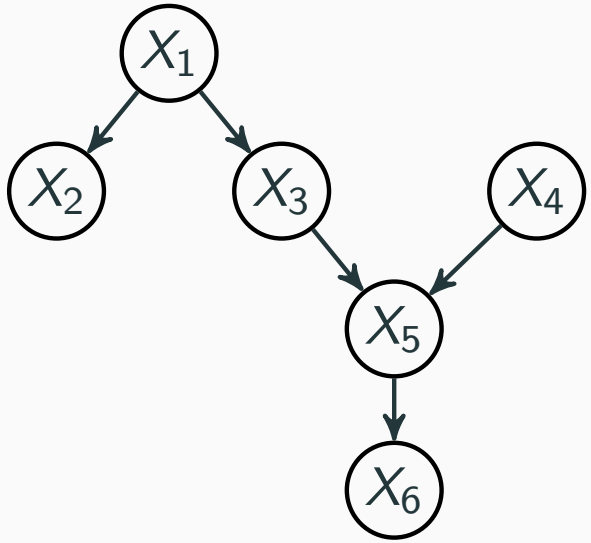
A partir d'un graphe candidat, modéliser l'ensemble du graphe et mesurer l'erreur. Tester heuristiquement différents graphes candidats dans le but de minimiser le score.

1. Modéliser et évaluer le nouveau graphe candidat
2. Si l'erreur est plus faible que le meilleur score obtenu, sauver le nouveau graphe et son score
3. Modifier le graphe candidat en retirant, ajoutant ou inversant des arcs.
4. Retourner à l'étape 1. jusqu'à convergence.

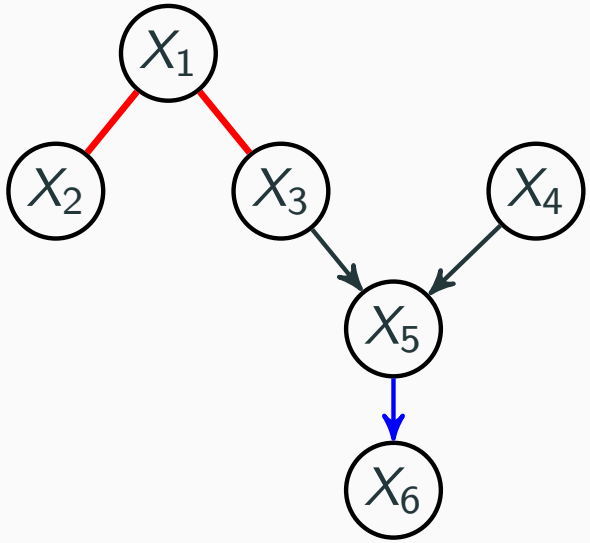
[Tsamardinos et al., 2006], [Drton and Maathuis, 2016], [Chickering, 2002]

CPDAG

Les méthodes à score ou à contraintes produisent un **CPDAG** (Completed Partially Directed Acyclic Graph)



(a) The exact DAG of $\mathcal{G}$.



(b) The CPDAG of $\mathcal{G}$.

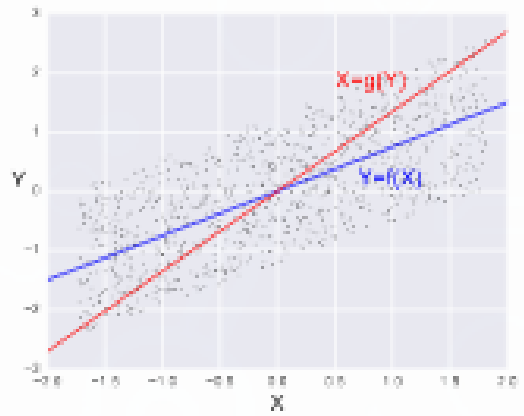
Méthodes bivariées

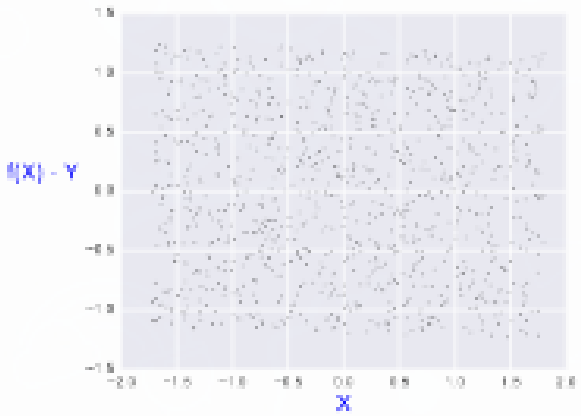
Pose un modèle assez simple, tel que la distribution est reproductible avec le modèle dans un sens causal, mais pas dans l'autre sens.

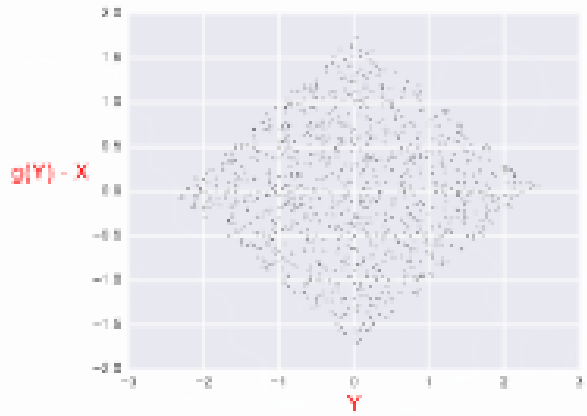
ANM (Additive Noise Model) [Hoyer et al., 2009]:

$Y = f(X) + E$,

f continue et E un bruit indépendant de la cause X







Shameless advertisement

- Learning Functional Causal Models with Generative Neural Networks, Explainable and Interpretable Models in Computer Vision and Machine Learning, Springer, Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, David Lopez-Paz, Isabelle Guyon, Michèle Sebag
- SAM: Strutural Agnostic Model, Causal Discovery and Penalized Adversarial Learning, 2018, ArXiv, Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, Michèle Sebag

4. Conclusion

Causalité observationnelle

- Outil pas universel, plutôt des solutions au cas par cas
- Permet d'avoir une idée de la structure des données
- Beaucoup d'hypothèses
- Utile pour planifier des expériences et en tirer un maximum d'informations

Chickering, D. M. (2002).
Optimal structure identification with greedy search.
Journal of machine learning research, 3(Nov):507–554.

Drton, M. and Maathuis, M. H. (2016).
Structure learning in graphical modeling.
Annual Review of Statistics and Its Application.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2009).
Nonlinear causal discovery with additive noise models.
In *Advances in neural information processing systems*, pages 689–696.

Pearl, J. (2009).
Causality.
Cambridge university press.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000).
Causation, prediction, and search.
MIT press.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006).
The max-min hill-climbing bayesian network structure learning algorithm.
Machine learning, 65(1):31–78.