Probality-free causal inference via the Algorithmic Markov Condition

Dominik Janzing

Max Planck Institute for Intelligent Systems
Tübingen, Germany

23. June 2015

Can we infer causal relations from passive observations?

Recent study reports negative correlation between coffee consumption and life expectancy

Paradox conclusion:

- drinking coffee is healthy
- nevertheless, strong coffee drinkers tend to die earlier because they tend to have unhealthy habits

$\Rightarrow$ Relation between statistical and causal dependences is tricky

Statistical and causal statements...

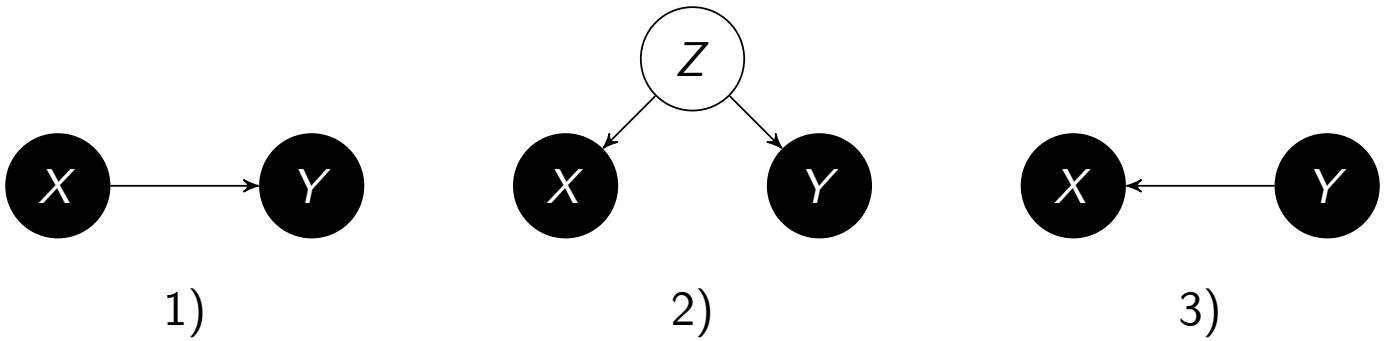
...differ by **slight** rewording:

- **"The life of coffee drinkers is** 3 **years shorter (on the average)."**

- **"Coffee drinking shortens the life by** 3 **years (on the average)."**

Reichenbach's principle of common cause (1956)

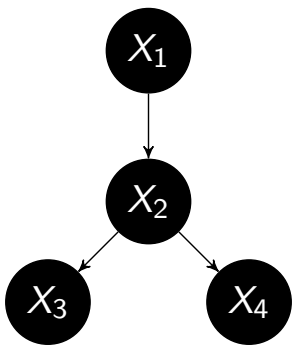
If two variables X and Y are statistically dependent then either



- in case 2) Reichenbach postulated $X \perp\!\!\!\perp Y \,|Z$.
- every statistical dependence is due to a causal relation, we also call 2) "causal".
- distinction between 3 cases is a key problem in scientific reasoning.

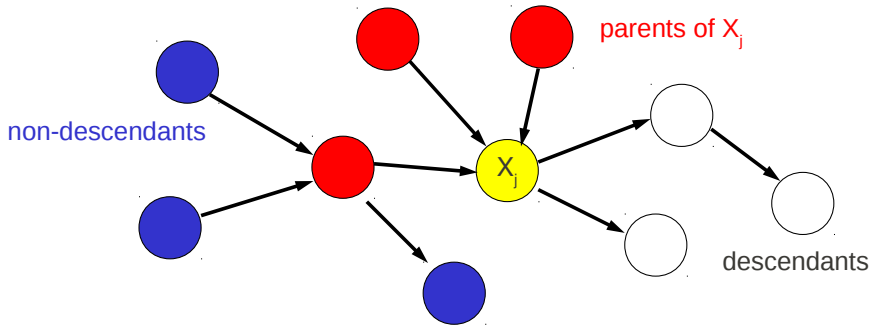
Causal inference problem, general form Spirtes, Glymour, Scheines, Pearl

- Given variables $X_1, \ldots, X_n$
- infer causal structure among them from n-tuples iid drawn from $P(X_1, \ldots, X_n)$
- causal structure = directed acyclic graph (DAG)



Causal Markov condition (3 equivalent versions) Lauritzen et al

- **local Markov condition:** every node is conditionally independent of its non-descendants, given its parents



- **global Markov condition:** If the sets S, T of nodes are d-separated by the set R, then

$$S \perp\!\!\!\perp T \,|R\,.$$

- **factorization of joint density:** $p(x_1, \ldots, x_n) = \prod_j p(x_j|pa_j)$ (subject to a technical condition)

Relevance of Markov conditions

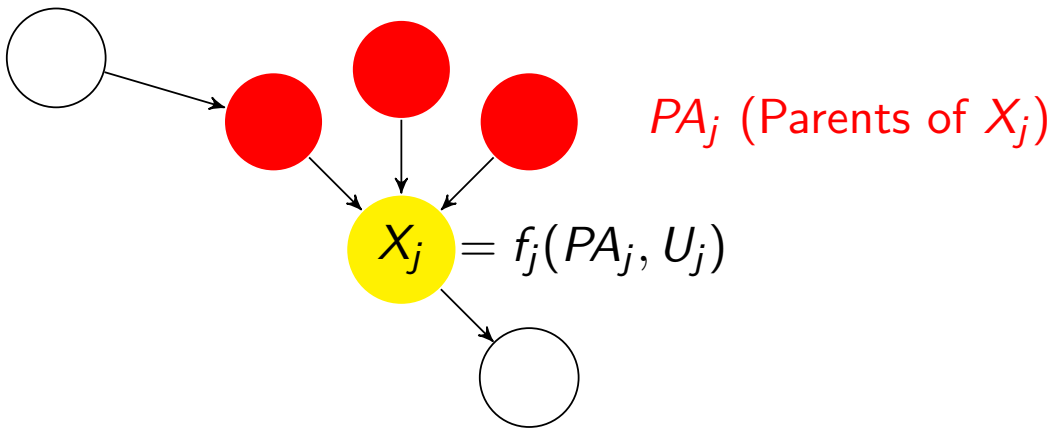
- **local Markov condition:** Most intuitive form, formalizes that every information exchange with non-descendants involves the parents

- **global Markov condition:** graphical criterion describing all independences that follow from the ones postulated by the local Markov condition

- **factorization:** every conditional $p(x_j|pa_j)$ describes a causal mechanism

Justification: Functional model of causality Pearl,...

- every node X_j is a function of its parents and an unobserved noise term U_j



- all noise terms U_j are statistically independent (causal sufficiency)

Functional model implies Markov condition

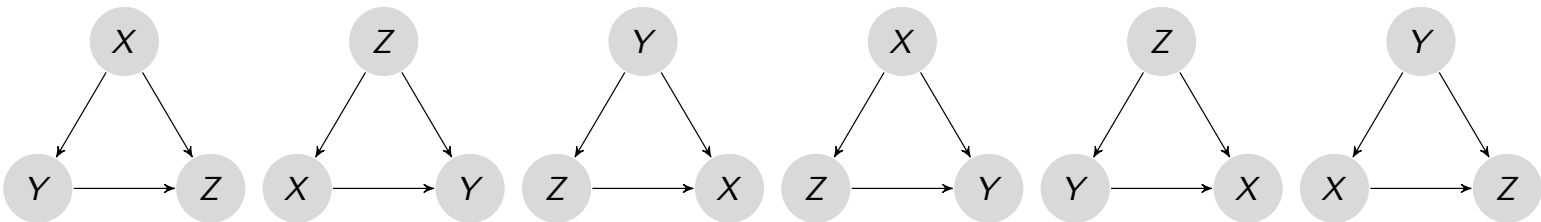
Theorem (Pearl 2000)

If $P(X_1, \ldots, X_n)$ is generated by a functional model according to a DAG G, then it satisfies the 3 equivalent Markov conditions with respect to G.

Causal inference from observational data

Can we infer G from $P(X_1, \ldots, X_n)$?

- MC only describes which sets of DAGs are consistent with P
- $n!$ many DAGs are consistent with any distribution



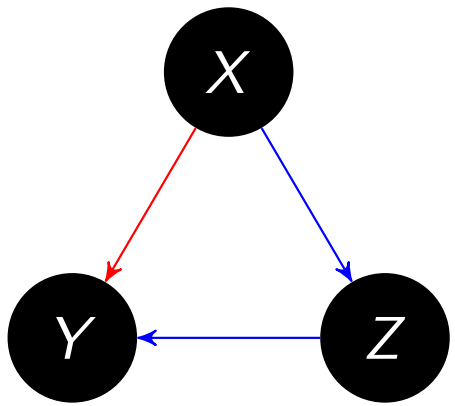
- reasonable rules for preferring simple DAGs required

Causal faithfulness

Spirtes, Glymour, Scheines, 1993

Prefer those DAGs for which all observed conditional independences are implied by the Markov condition

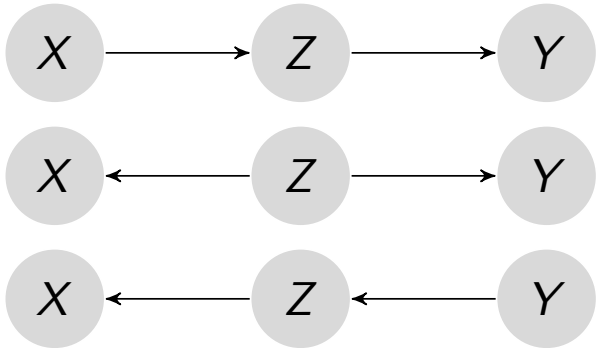
- **Idea:** generic choices of parameters yield faithful distributions
- **Example:** let $X \perp\!\!\!\perp Y$ for the DAG



- not faithful, direct and indirect influence compensate
- **Application:** PC and FCI infer causal structure from conditional statistical independences

Limitation of independence based approach:

- many DAGs impose the same set of independences



  $X \perp\!\!\!\perp Y \,|Z$ for all three cases ("Markov equivalent DAGs")

- method useless if there are no conditional independences
- non-parametric conditional independence testing is hard
- ignores important information:
  only uses yes/no decisions "conditionally dependent or not"
  without accounting for the kind of dependences...

We will see that causal inference should not only look at **statistical** information...

forget about statistics for a moment...

– how do we come to causal conclusions in *every-day* life?

these 2 objects are similar...

– *why* are they so similar?

Conclusion: common history

similarities require an *explanation*

what kind of similarities require an explanation?

here we would *not* assume that anyone has copied the design...

..the pattern is too simple

- similarities require an explanation only if the pattern is sufficiently complex

consider a binary sequence

Experiment:
2 persons are instructed to write down a string with 1000 digits

Result:
Both write 1100100100001111110110101010001...
(all 1000 digits coincide)

the **naive** statistician concludes

"There must be an agreement between the subjects"

correlation coefficient 1 (between digits) is highly significant for sample size 1000 !

- reject statistical independence
- infer the existence of a causal relation

another mathematician recognizes...

$$11.001001000011111101010101001... = \pi$$

- subjects may have come up with this number independently because it follows from a simple law
- superficially strong similarities are not necessarily significant if the pattern is too simple

How do we measure simplicity versus complexity of patterns / objects?

Kolmogorov complexity

(Kolmogorov 1965, Chaitin 1966, Solomonoff 1964)
of a binary string x

- K(x) = length of the shortest program with output x (on a Turing machine)
- interpretation: number of bits required to describe the rule that generates x

  neglect string-independent additive constants; use $\stackrel{+}{=}$ instead of $=$
- strings x, y with low $K(x)$, $K(y)$ cannot have much in common
- $K(x)$ is uncomputable
- probability-free definition of information content

Conditional Kolmogorov complexity

- $K(y|x)$: length of the shortest program that generates y from the input x.
- number of bits required for describing y if x is given
- $K(y|x^*)$ length of the shortest program that generates y from x^*, i.e., the shortest compression x.
- subtle difference: x can be generated from x^* but not vice versa because there is no algorithmic way to find the shortest compression

Algorithmic mutual information

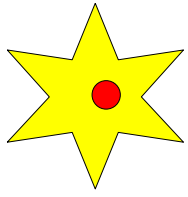
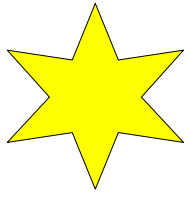
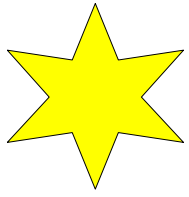
Chaitin, Gacs

Information of x about y (and vice versa)

- $I(x:y) := K(x) + K(y) - K(x,y)$
  $\stackrel{+}{=} K(x) - K(x|y^*) \stackrel{+}{=} K(y) - K(y|x^*)$

- Interpretation: number of bits saved when compressing x, y jointly rather than compressing them independently

Algorithmic mutual information: example

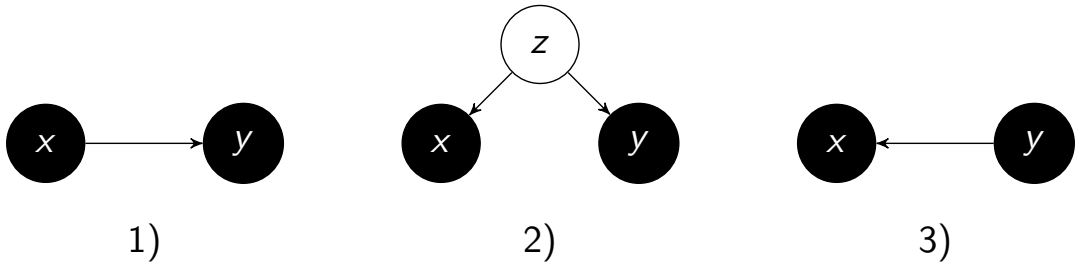
I( : ) = K()

Analogy to statistics:

- replace strings x, y (=objects) with random variables X, Y
- replace Kolmogorov complexity with Shannon entropy
- replace algorithmic mutual information $I(x : y)$ with statistical mutual information $I(X; Y)$

Causal Principle

If two strings x and y are algorithmically dependent then either



- every algorithmic dependence is due to a causal relation
- algorithmic analog to Reichenbach's principle of common cause
- distinction between 3 cases: use conditional independences on more than 2 objects

DJ, Schölkopf IEEE TIT 2010

Relation to Solomonoff's universal prior

- string x occurs with probability $\sim 2^{-K(x)}$
- if generated independently, the pair (x, y) occurs with probability $\sim 2^{-K(x)}2^{-K(y)}$
- if generated jointly, it occurs with probability $\sim 2^{-K(x,y)}$
- hence $K(x, y) \ll K(x) + K(y)$ indicates generation in a joint process
- $I(x : y)$ quantifies the evidence for joint generation

conditional algorithmic mutual information

- $I(x : y|z) = K(x|z) + K(y|z) - K(x, y|z)$
- Information that x and y have in common when z is already given
- Formal analogy to statistical mutual information:

$$I(X : Y|Z) = S(X|Z) + S(Y|Z) - S(X, Y|Z)$$

- Define conditional independence:

$$I(x : y|z) \approx 0 :\Leftrightarrow x \perp\!\!\!\perp y|z$$

Algorithmic Markov condition

Postulate (DJ & Schölkopf IEEE TIT 2010)

Let $x_1, ..., x_n$ be some observations (formalized as strings) and G describe their causal relations.
Then, every x_j is conditionally algorithmically independent of its non-descendants, given its parents, i.e.,

$$x_j \perp\!\!\!\perp nd_j \,|pa_j^*$$

Equivalence of algorithmic Markov conditions

Theorem

For n strings $x_1, ..., x_n$ the following conditions are equivalent

- **Local Markov condition:**

$$I(x_j : nd_j | pa_j^*) \stackrel{+}{=} 0$$

- **Global Markov condition:**
  R d-separates S and T implies $I(S : T|R^*) \stackrel{+}{=} 0$
- **Recursion formula for joint complexity**

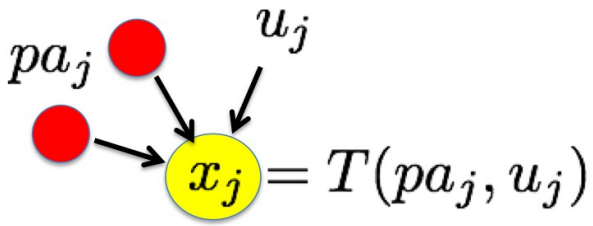
$$K(x_1, ..., x_n) \stackrel{+}{=} \sum_{j=1}^{n} K(x_j | pa_j^*)$$

$\rightarrow$ another analogy to statistical causal inference

Algorithmic model of causality

Given n causality related strings $x_1, \ldots, x_n$

- each x_j is computed from its parents pa_j and an unobserved string u_j by a Turing machine T



- all u_j are algorithmically independent
- each u_j describes the causal mechanism (the program) generating x_j from its parents
- u_j is the analog of the noise term in the statistical functional model

Interpretation

- **Church-Turing-Deutsch Principle:** Every physical process can be simulated on a Turing machine

- **Algorithmic model of causality:** Every physical multipartite process can be simulated by multiple Turing machines influencing each other via the same DAG as the process

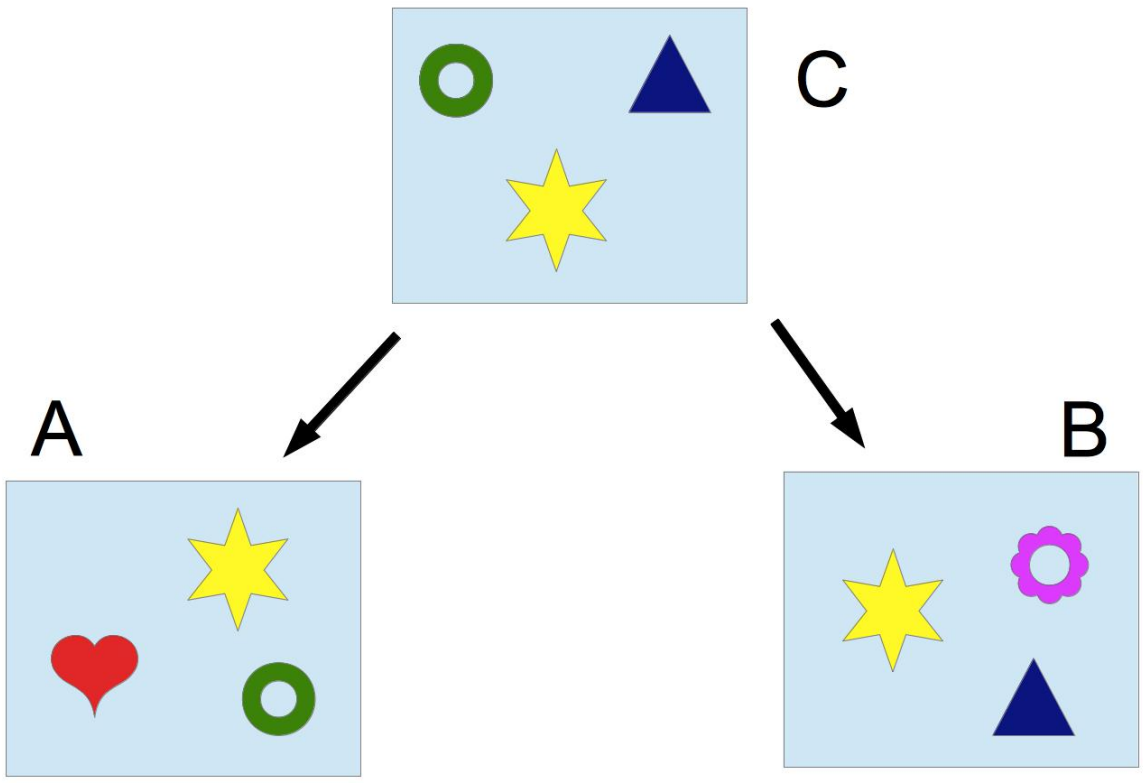
Algorithmic model of causality implies Markov condition

Theorem

If $x_1, \ldots, x_n$ are generated by an algorithmic model of causality according to the DAG G then they satisfy the 3 equivalent algorithmic Markov conditions.

Causal inference for single objects

3 carpets



conditional independence $A \perp\!\!\!\perp B \,||\, C$

Applications

- **Approximate K by existing compression schemes** (e.g. infer causal relations between texts by Lempel-Ziv compression. Steudel, DJ, Schölkopf COLT 2010)

- **Use algorithmic Markov condition as foundation for new statistical inference rules**

Algorithmic Independence of Conditionals

Postulate (DJ & Schölkopf 2010, Lemeire & DJ 2012)

If $P(X_1, \ldots, X_n)$ is generated by the causal DAG G, then the conditionals $P(X_j|PA_j)$ in the decomposition

$$P(X_1, \ldots, X_n) = \prod_{j=1}^{n} P(X_j|PA_j)$$

are algorithmically independent

Relation to algorithmic Markov condition

- If one assumes that nature chooses the mechanisms $P(X_j|PA_j)$ independently, then they should be algorithmically independent due to the causal principle

- Applying the algorithmic Markov condition to the single instances in the statistical sample yields something closely related

Two-variable case

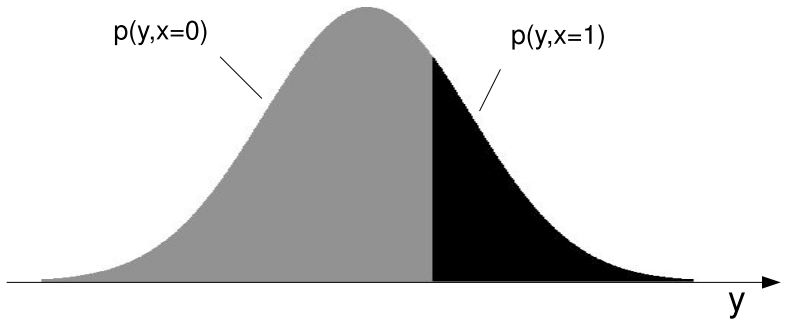
If $X \to Y$ then

- $P(X)$ and $P(Y|X)$ are algorithmically independent while $P(Y)$ and $P(X|Y)$ need not
- shortest description of $P(X, Y)$ is given by separate dscriptions of $P(X)$ and $P(Y|X)$
- defines an asymmetry of cause and effect although the literature often claims that $X \to Y$ and $Y \to X$ cannot be distinguished from observing $P(X, Y)$.

Toy example

Let X be binary and Y real-valued.

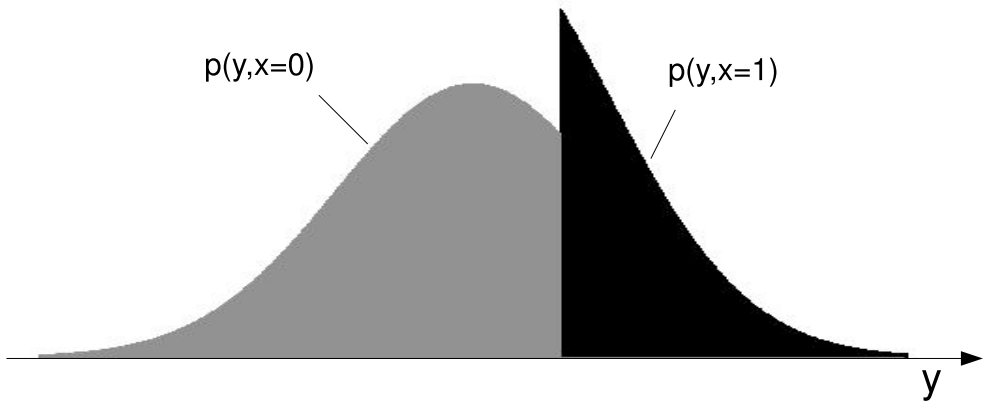
- Let Y be Gaussian and $X = 1$ for all y above some threshold and $X = 0$ otherwise.



- $Y \rightarrow X$ is plausible: simple thresholding mechanism
- $X \rightarrow Y$ requires a strange mechanism:
  look at $P(Y|X = 0)$ and $P(Y|X = 1)$!

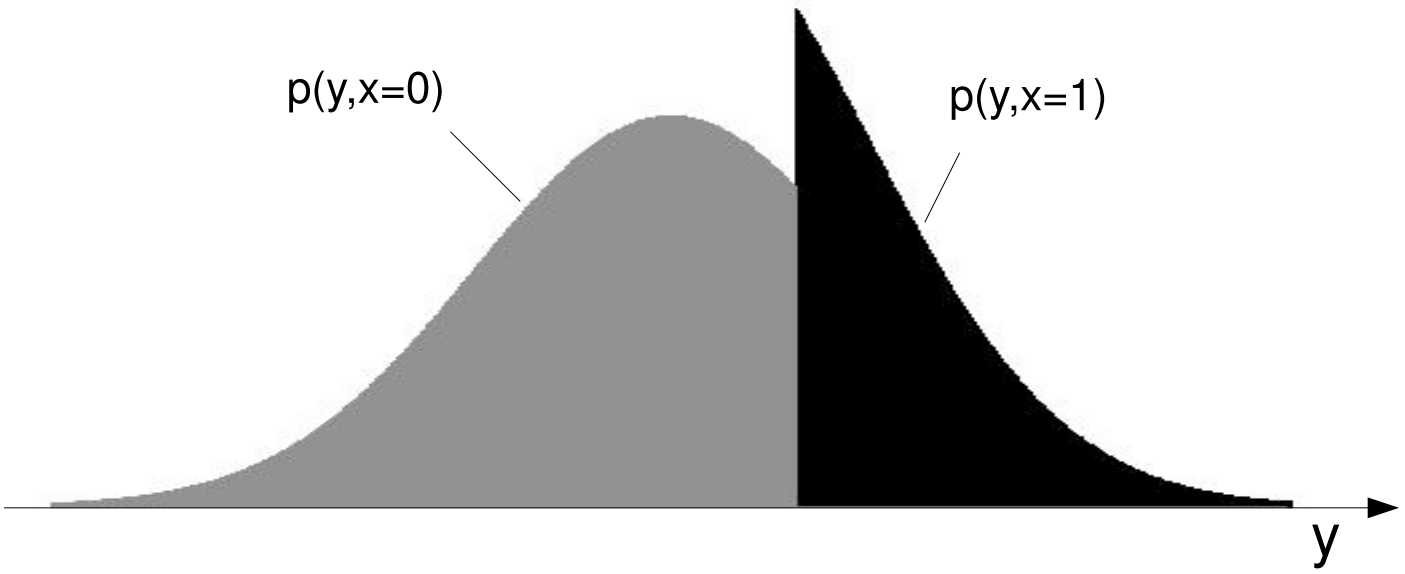
not only $P(Y|X)$ itself is strange...

but also what happens if we change $P(X)$:



Hence, reject $X \rightarrow Y$ because it requires tuning of $P(X)$ relative to $P(Y|X)$.

Violation of independence of conditionals

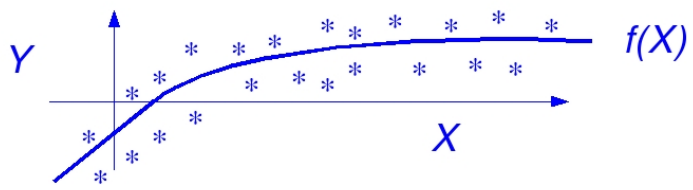


Knowing $P(Y|X)$, there is a short description of $P(X)$, namely 'the unique distribution for which $\sum_x P(Y|x)P(x)$ is Gaussian'.

Non-linear additive noise based inference Hoyer, Janzing, Peters, Schölkopf, 2008

- Assume that the effect is a function of the cause up to an additive noise term that is statistically independent of the cause:

$$Y = f(X) + E \quad \text{with} \quad E \perp\!\!\!\perp X$$



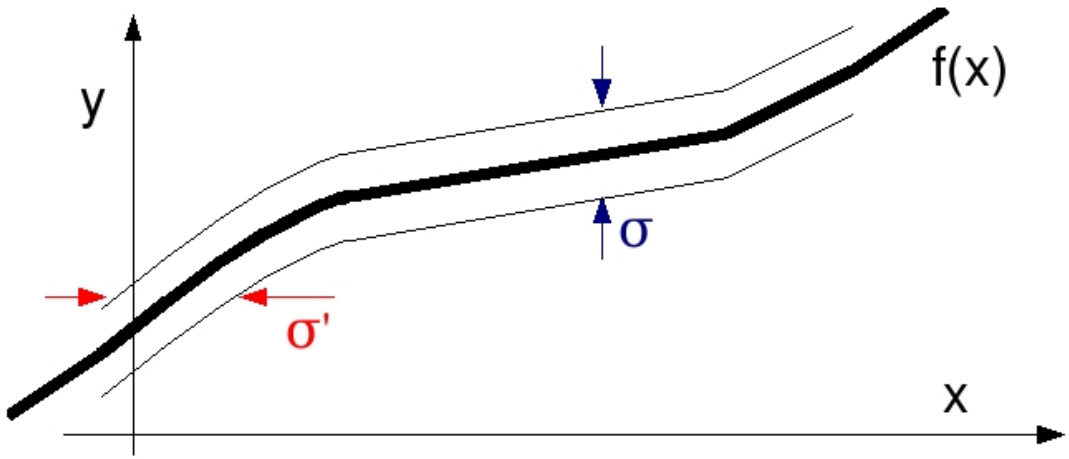
- there will, in the generic case, be no model

$$X = g(Y) + \tilde{E} \quad \text{with} \quad \tilde{E} \perp\!\!\!\perp Y\,,$$

even if f is invertible! (proof is non-trivial)

Intuition

- additive noise model from X to Y imposes that the width of noise is constant in x.
- for non-linear f, the width of noise wont't be constant in y at the same time.



Causal inference method:

Prefer the causal direction that can better be fit with an additive noise model.

Implementation:

- Compute a function f as non-linear regression of Y on X, i.e., $f(x) := \mathbb{E}(Y|x)$.
- Compute the residual

$$E := Y - f(X)$$

- check whether E and X are statistically independent (uncorrelated is not sufficient, method requires tests that are able to detect higher order dependences)

Justifying additive noise based causal inference

Assume $Y = f(X) + E$ with $E \perp\!\!\!\perp X$

- Then $P(Y)$ and $P(X|Y)$ are related:

$$\frac{\partial^2}{\partial y^2} \log p(y) = -\frac{\partial^2}{\partial y^2} \log p(x|y) - \frac{1}{f'(x)} \frac{\partial^2}{\partial x \partial y} \log p(x|y) \,.$$

  $\Rightarrow \frac{\partial^2}{\partial y^2} \log p(y)$ can be computed from $p(x|y)$ knowing $f'(x_0)$ for one specific x_0

- Given $P(X|Y)$, $P(Y)$ has a short description.

- We reject $Y \to X$ provided that $P(Y)$ is complex

Janzing, Steudel, OSID (2010)

Cause-effect pairs

- `http://webdav.tuebingen.mpg.de/cause-effect/`
- contains currently 86 data sets with X, Y where we believe to know whether $X \to Y$ or $Y \to X$, e.g.

day in the year	$\to$	temperature
age of snails	$\to$	length
drinking water access	$\to$	infant mortality rate
open http connections	$\to$	bytes sent
outside room temperature	$\to$	inside room temperature
age of humans	$\to$	wage per hour

- goal: collect more pairs, diverse domains
- ground truth should be obvious to non-experts

Additive noise based inference...

- about 75% correct decisions for 70 cause-effect pairs with known ground truth
- fraction even better if we allow "no decision"
- we do not claim that noise is always additive in real life, but if it is for one direction this is unlikely to be the wrong one
- generalization to n variables outperformed PC

(Peters, Mooij, Janzing, Schölkopf *UAI 2011*)

Conclusions

Conventional causal inference is based on conditional statistical dependences. This is insufficient because...

- not every causal conclusion refers to statistical data, we often infer causal relations between single objects.

- even in statistical data one should not only look at statistical information. Also the description length of the distribution contains information about the causal structure.

The algorithmic Markov condition inspired us in developing new statistical inference methods

Thank you for your attention!

References

- Janzing, Schölkopf: **Causal inference using the algorithmic Markov condition.** IEEE TIT (2010).

- Lemeire, Janzing: **Replacing causal faithfulness with the algorithmic independence of conditionals**, Minds & Machines (2012).

- Janzing, Steudel: **Justifying additive-noise based causal discovery via algorithmic information theory.** Open Systems & Information Dynamics (2011)