Inferring causality from passive observations

Dominik Janzing

Max Planck Institute for Intelligent Systems
Tübingen, Germany

22.-28. August 2014

MAX-PLANCK-GESELLSCHAFT

Outline

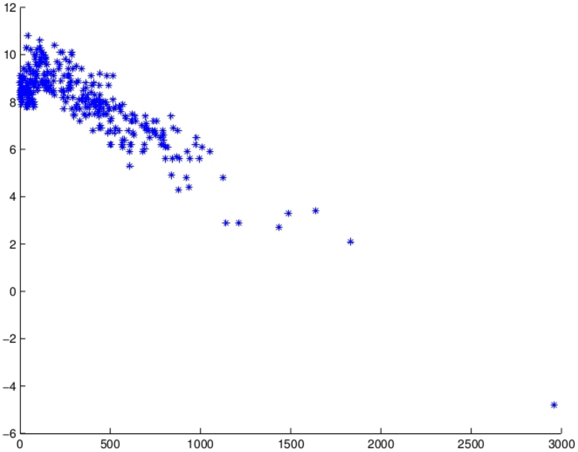
1. **why the relation between statistics and causality is tricky**
2. **causal inference using conditional independences (statistical and general)**
3. **causal inference using other properties of joint distributions**
4. **causal inference in time series, quantifying causal strength**
5. **why causal problems matter for prediction**

Part 3: causal inference using other properties of joint distributions

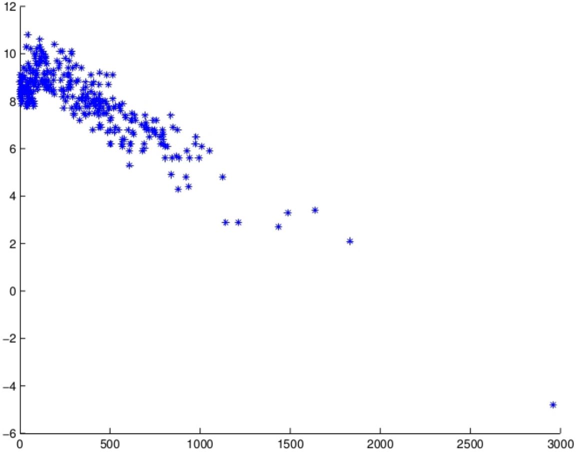
- intuitive approach to distinguishing cause and effect
- new foundations of causal inference
- additive noise based causal inference
- information-geometric causal inference

Intuitive approach to distinguishing cause and effect

What's the cause and what's the effect?

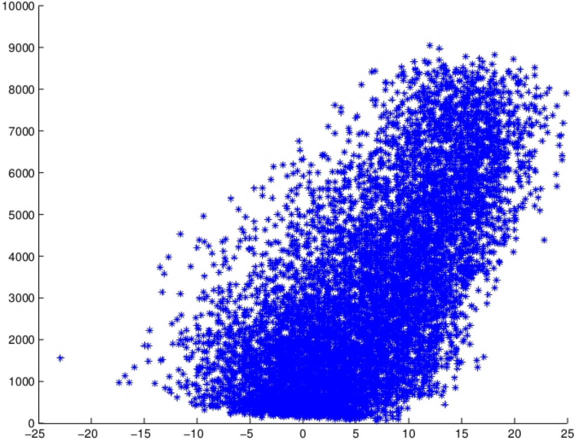


What's the cause and what's the effect?

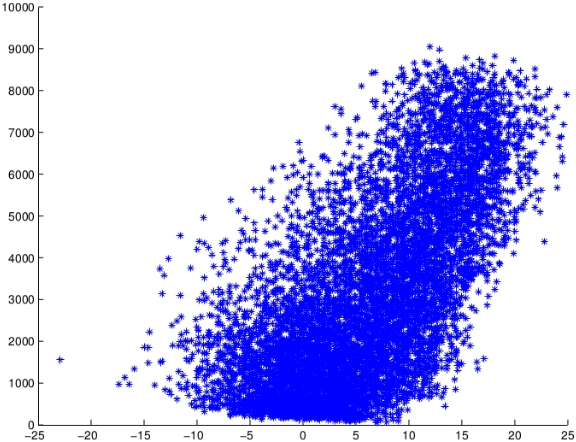


X (Altitude) $\rightarrow$ Y (Temperature)

What's the cause and what's the effect?

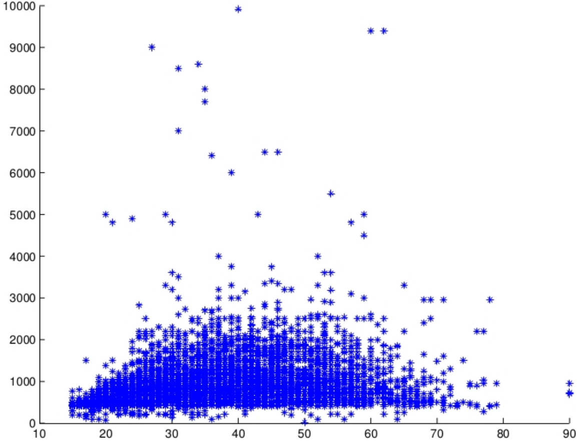


What's the cause and what's the effect?

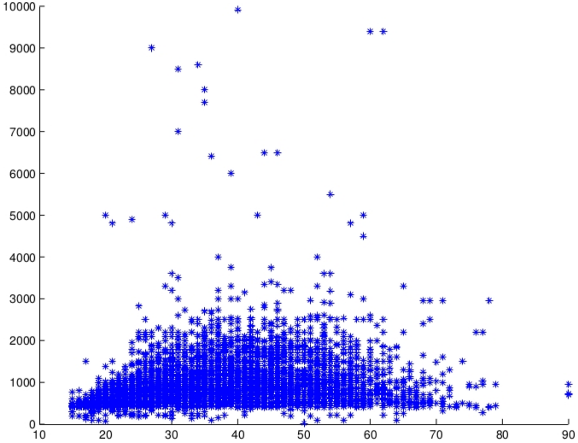


Y (Solar Radiation) $\rightarrow$ X (Temperature)

What's the cause and what's the effect?



What's the cause and what's the effect?



X (Age) $\rightarrow$ Y (Income)

Idea of new inference rules

Consider two decompositions of $P(X, Y)$:

$$P(X)P(Y|X) \quad \text{or} \quad P(Y)P(X|Y),$$

- does one of it look simpler than the other?
- if yes, assume this to be the causal direction

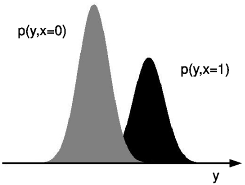
Implementing this idea is a challenging research program:

- defining simplicity/complexity
- estimating it from finite data
- justifying why this is related to causality

Particularly nice toy examples (1)

Janzing & Schölkopf: Causal inference using the algorithmic Markov condition, IEEE TIT 2010

Let X be binary and Y real-valued. Observe that both $P(Y|X=0)$ and $P(Y|X=1)$ are Gaussians with different mean:

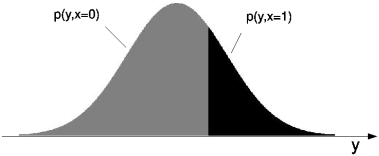


$X \rightarrow Y$ more plausible:

- simple effect of X: shift the mean of Y
- if Y was the cause it would be implausible that conditioning on X separates the two modes of $P(Y)$

Particularly nice toy examples (2)

Let $P(Y)$ be Gaussian and $X = 1$ above a certain threshold y_0:



$Y \to X$ more plausible:

- simple effect of Y: set X via a threshold
- $P(Y|X=0)$ and $P(Y|X=1)$ look strange (truncated Gaussians)

Philosophical basis for new methods

- so far, it seems arbitrary how to defined simplicity/complexity
- we will recall and criticise the justification of faithfulness
- we replace faithfulness with a principle that we consider more fundamental

Justifying faithfulness

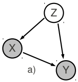
Unfaithful distributions occur with probability zero if

- nature chooses each $P(X_j|PA_j)$ independently
- each $P(X_j|PA_j)$ is chosen from a probability density in parameter space (e.g. uniform distribution)

  here the parameter space of each conditional is a subset of $\mathbb{R}^k$
  with $k := \{x_j\}^{\{pa_j\}}$
  (see next slide)

C. Meek: Strong completeness and faithfulness in Bayesian networks. (UAI 1995)

What we mean by parameter space

Consider the DAG  with binary Z, X, Y.

- $P(Z)$ is described by the value $P(Z = 1)$
  (parameter space: $[0, 1]$)
- $P(X|Z)$ is described by the values
  $P(X = 1|Z = 0), P(X = 1|Z = 1)$
  (parameter space: $[0, 1]^2$)
- $P(Y|X, Z)$ is described by the values $P(Y = 1|X = i, Z = j)$
  with $i, j \in \{0, 1\}$
  (parameter space: $[0, 1]^4$)

in total: 7 free parameters, set of parameters that induce unfaithful distributions is a lower dimensional submanifold in this 7-dimensional space.

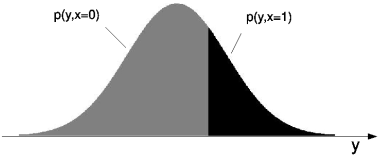
However, we will argue...

- There are cases of obvious parameter tuning that do not generate additional independences
  ($\Rightarrow$ faithfulness is too weak)

- Not every violation of faithfulness is due to parameter tuning since we do not believe in *densities* on the parameter space
  ($\Rightarrow$ faithfulness is too strong)

Why faithfulness is too weak

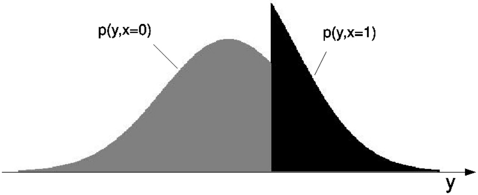
recall the motivating example:



we reject $X \to Y$ not only because $Y \to X$ yields simpler explanations for the shape of the distribution, but

For $X \to Y$ generation of $P(X, Y)$ requires tuning

look what happens if we change $P(X)$:



Hence, reject $X \to Y$ because it requires tuning of $P(X)$ relative to $P(Y|X)$. Faithfulness would accept both causal directions.

Why faithfulness is too strong

Consider deterministic relations

$$Y = f(X)$$

- unfaithful because... (homework for today)
- but there is no adjustment between $P(X)$, $P(Y|X)$, $P(Z|Y)$
- only $P(Y|X)$ is 'non-generic'

We don't want to reject non-generic conditionals, we only want to reject non-generic **relations** between conditionals

New foundations of new inference rules

Algorithmic independence of conditionals

The **shortest** description of $P(X_1, \ldots, X_n)$ is given by **separate** descriptions of $P(X_j|PA_j)$.

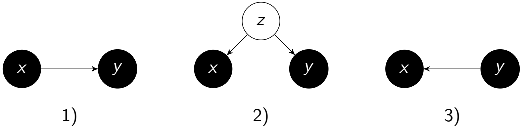
(Here, description length = Kolmogorov complexity)

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).

Recall: 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

Apply the causal principle to conditionals

$K(P(X_j|PA_j))$ denotes the length of the shortest program computing $P(x_j|pa_j)$ from (x_j, pa_j).

- If nature chooses each mechanism $P(X_j|PA_j)$ independently they are algorithmically independent, e.g.,

$$I(P(X_j|PA_j) : P(X_1|PA_1), P(X_2|PA_2), \ldots) \stackrel{+}{=} 0 \quad \forall j \, .$$

- equivalent to

$$K(P(X_1, \ldots, X_n)) \stackrel{+}{=} \sum_{j=1}^{n} K(P(X_j|PA_j))$$

(shortest description of the joint is given by separate descriptions of the causal conditionals)

Bivariate case

If $X \rightarrow Y$ then

$$I(P(X) : P(Y|X)) \stackrel{+}{=} 0$$

$P(X)$ contains no information about $P(Y|X)$ and vice versa.
(note: here we are not talking about information in the sense of Shannon mutual information)

Equivalent formulation

- Describing both $P(X)$ and $P(Y|X)$ describes $P(X, Y)$.
- Moreover, describing $P(X, Y)$ describes $P(X)$ and $P(Y|X)$.
- Therefore,

$$K(P(X), P(Y|X)) \stackrel{+}{=} K(P(X, Y)) \,.$$

- Thus,

$$K(P(X)) + K(P(Y|X)) \stackrel{+}{=} K(P(X), P(Y|X)) \,,$$

  is equivalent to

$$K(P(X)) + K(Y|X)) \stackrel{+}{=} K(P(X, Y)) \,.$$

- Hence, the algorithmic independence of $P(X)$ and $P(Y|X)$ is equivalent to

$$K(P(X, Y)) \stackrel{+}{=} K(P(X)) + K(P(Y|X)) \,.$$

Relation to Occam's Razor

Note:

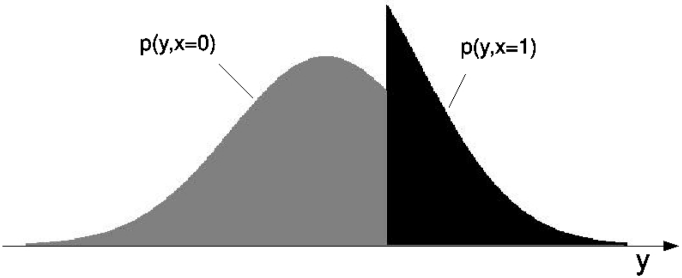
$$K(P(X,Y)) \stackrel{+}{=} K(P(X)) + K(P(Y|X)).$$

implies

$$K(P(X)) + K(P(Y|X)) \leq K(P(Y)) + K(P(X|Y)).$$

but not vice versa.

Revisiting the motivating example



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'.

Although Kolmogorov complexity is uncomputable...

we apply the principle of algorithmically independent conditionals:

- find notions of dependence of conditionals that capture essential aspects
- use it as a foundation/justification of new inference rules

Some new inference rules...

...that can also be justified by our philosophical principle

Additive noise based causal inference

Linear non-Gaussian models

Kano & Shimizu 2003

Theorem

Let $X \not\perp\!\!\!\perp Y$. Then $P(X, Y)$ admits linear models in both directtion, i.e.,

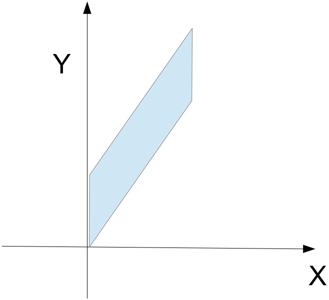
$$\begin{aligned} Y &= \alpha X + U_Y \text{ with } U_Y \perp\!\!\!\perp X \\ X &= \beta Y + U_X \text{ with } U_X \perp\!\!\!\perp Y, \end{aligned}$$

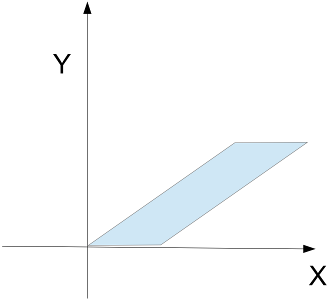
if and only if $P(X, Y)$ is bivariate Gaussian

- if $P(X, Y)$ is non-Gaussian, there can be a linear model in at most one direction.
- LINGAM: causal direction is the one that admits a linear model

Intuitive example:

Let X and U_Y be uniformly distributed. Then $Y = \alpha X + U_Y$ induces uniform distribution on a diamond (left):





uniformly distributed Y and U_X with $X = \beta Y + U_X$ induces the diamond on the right.

Proof via Darmois-Skitovic

Theorem

Let $X_1, \ldots, X_n$ be independent random variables and

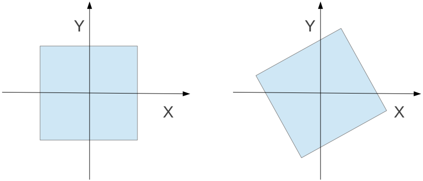
$$
\begin{aligned}
Y_1 &:= a_1 X_1 + \cdots + a_n X_n \\
Y_2 &:= b_1 X_1 + \cdots + b_n X_n
\end{aligned}
$$

be independent. Then each X_i with $a_i b_i \neq 0$ is Gaussian.

proof involves Fourier transforms of probability distributions

Example for Darmois-Skitovic

Let $P(X, Y) = P(X)P(Y)$ be uniform on $[0, 1]^2$



rotating the axis by a generic angle generates dependences between X and Y (although they are still uncorrelated)

Proof of Theorem 1

- Assume independence of $Y - \alpha X$ and X.
- Assume independence of $X - \beta Y$ and Y.
- Set

$$\begin{aligned} X_1 &:= X \\ X_2 &:= Y - \alpha X\,. \end{aligned}$$

- Set

$$\begin{aligned} Y_1 &:= Y \\ Y_2 &:= X - \beta Y\,. \end{aligned}$$

- Then Y_1 and Y_2 can be written as linear combinations of X_1 and X_2. If X_1 or X_2 are non-Gaussian, then Y_1 and Y_2 cannot be independent.

Independent component analysis

Jutten & Hérault 1991

Theorem

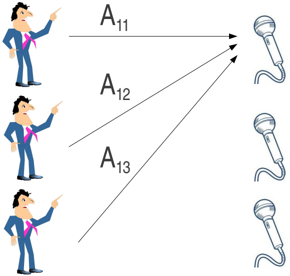
Let $\mathbf{U} := (U_1, \ldots, U_n)$ *be independent non-Gaussian random variables and* $\mathbf{X} := A\mathbf{U}$ *where* A *is an* $n \times n$ *matrix. Then* $\mathbf{U}$ *can be determined from* $\mathbf{X}$ *up to permutation and rescaling of components* U_j.

follows from Darmois-Skitovic

Application: blind source separation

e.g. Hyvärinen 1998

- n microphones record n speakers simultaneously.
- due to the different distance, each speaker j occurs with different weight A_{ij} in microphone i



ICA recovers the signal of each speaker

Applications of LINGAM

- applications in brain research (e.g. fMRI data) are considered promising by several people (see e.g. talks of Hyvärinen)
- supported by positive results on simulated data where LINGAM performed better than traditional methods like Granger causality)
- not easy to find data with known ground truth

LINGAM as a special case of ICA

Shimizu et al. 2006

Let $P(X_1, \ldots, X_n)$ be generated by the linear structural equation

$$X_i = \sum_j b_{ij} X_j + U_i \text{ with independent } U_i \,,$$

where the set of non-zero b_{ij} define a DAG G. Then G can be uniquely identified from $P(X_1, \ldots, X_n)$:

- write structural equation as $\mathbf{X} = B\mathbf{X} + \mathbf{U}$
- hence $(1 - B)\mathbf{X} = \mathbf{U}$
- rewrite as $\mathbf{X} = (1 - B)^{-1}\mathbf{U}$
- define $A := (1 - B)^{-1}$ to obtain usual ICA problem
- no ambiguity regarding permuting and scaling U_j: all diagonal entries of $1 - B$ are 1 (B is lower triangular with respect to an appropriate ordering of nodes, therefore inverse is easy to see)
- compute $B = 1 - A^{-1}$ to recover the structural equation

Distinguishing cause and effect via ICA/LINGAM

- Assume

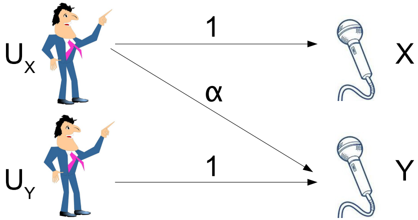
$$\begin{aligned} X &= U_X \\ Y &= \alpha X + U_Y\,, \end{aligned}$$

  where U_X and U_Y are independent 'sources'.
- Hence,

$$\begin{aligned} X &= U_X \\ Y &= \alpha U_X + U_Y \end{aligned}$$

- The cause X contains only one source and the effect Y contains both sources.

Analogy to blind source separation problem



- The cause is like a microphone that receives only the signal from 1 speaker
- The effect receives signal from both speakers

- ICA can easy decide which one is which

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)

Note...

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

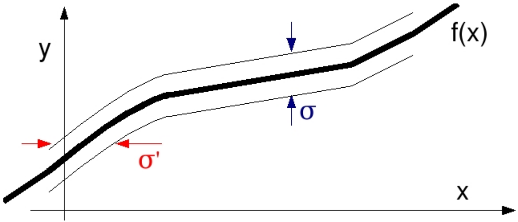
can model any conditional $P(Y|X)$

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

restricts the class of possible $P(Y|X)$

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)

Why f is given by the conditional expectation

- If $Y - f(X)$ should be independent of X, its expectation needs to be independent of X, i.e.,

$$\mathbb{E}[Y - f(x)|x]$$

  needs to be constant in x.
- Assume $\mathbb{E}[Y - f(x)|x] = 0$ without loss of generality because this changes only the offset of the noise
- Hence, $\mathbb{E}[Y|x] = f(x)$.

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 70% 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*)
- generalization to time series

  (Peters, Janzing, Schölkopf *NIPS 2013*)
- note the paradigm shift: a model is good if the noise is independent, not if the noise is small
  (if it's dependent it may not be noise?)

Additive noise based inference with n variables

Peters et al. 2014

- Step 1: find a causal order
- Step 2: drop unnecessary edges in the corresponding complete DAG

Step 1: find the causal order

i.e., find $\pi \in S_n$ such that there is no arrow $X_{\pi(i)} \to X_{\pi(j)}$ for $\pi(i) > \pi(j)$

- compute regression for each X_j:

$$X_j = f(X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n) + E_j$$

- check dependence between E_j and X_j
- let $\pi(n)$ be the node for which the dependence is minimal
- drop $X_{\pi(n)}$ and repeat the procedure with $n-1$ variables

Step 2: remove unnecessary edges

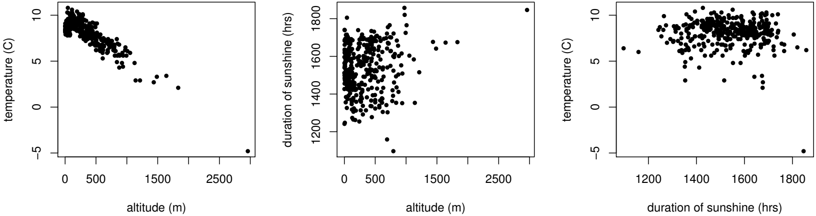
Apply the following procedure to each node $X_{\pi(j)}$:

1. for each $X_{\pi(j)}$ let the parents be all its predecessors w.r.t. order π
2. check each parent whether removing it still yields independent noise
3. repeat 2 until no further parents can be removed

note: step 2 performs a conditional independence test. The additive noise assumption reduces it to testing independence of error term.

Application to real data

- A: altitude of 349 places in Germany
- T average temperature
- D duration of sunshine



the method preferred $T \leftarrow A \rightarrow D$

Inferring confounders with additive noise

$$
\begin{aligned}
X &= f_X(T) + U_X \\
Y &= f_Y(T) + U_Y
\end{aligned}
$$

with jointly independent T, U_X, U_Y.

note: contains $X \to Y$ by setting $f_X = id$ and $U_X = 0$. Similar for $Y \to X$.

conjecture: f_X, f_Y can be inferred up to bijective transformations of T

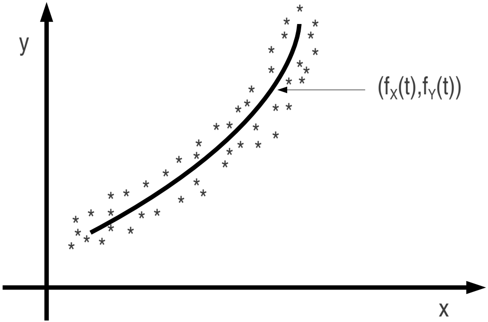
argument: suggested by a theoretical result with small noise

interpretation: constructing f_X, f_Y amount to distinguishing between the three cases

$$
X \to Y \qquad X \leftarrow T \to Y \qquad X \leftarrow Y \,.
$$

Intuition

- without noise, the points describe the line $(f_X(t), f_Y(t))$



- independent noise U_X and U_Y is added in X and Y directions
- original line can be obtained by deconvolution with an appropriate product distribution

Post-nonlinear model

Let $P(X, Y)$ be generated by

$$Y = g(f(X) + U) \quad \text{with } U \perp\!\!\!\perp X\,.$$

Then there is in the generic case no triple $\tilde{g}, \tilde{f}, \tilde{U}$ such that

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

However...

employing properties of the noise

is not the only way

of inferring causal directions

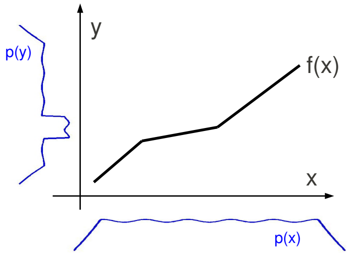
$\rightarrow$ look at the **noiseless** case...

Information-geometric causal inference

Inferring deterministic causality

Daniusis, Janzing,... *UAI 2010*, Janzing et al. *AI 2012*

- Problem: infer whether $Y = f(X)$ or $X = f^{-1}(Y)$ is the right causal model
- Idea: if $X \to Y$ then f and the density p_X are chosen independently "by nature"
- Hence, peaks of p_X do not correlate with the slope of f
- Then, peaks of p_Y correlate with the slope of f^{-1}



Formalization

Let f be a monotonously increasing bijection of $[0, 1]$

- **Postulate:**

$$\int_0^1 \log f'(x) p(x) dx = \int_0^1 \log f'(x) dx \text{ (approximately)}$$

- **Idea:** averaging log of slope of f over p is the same as averaging over uniform distribution

- **Implication:**

$$\int_0^1 \log f^{-1'} p(y) dy \geq \int_0^1 \log f^{-1'}(y) dy$$

Testable implication / inference rule

- If $X \to Y$ then

$$\int \log |f'(x)| p(x) dx \leq \int \log |f^{-1'}(y)| p(y) dy$$

(high density $p(y)$ tends to occur at points with large slope)

- empirical estimator

$$\hat{C}_{X \to Y} := \frac{1}{m} \sum_{j=1}^{m} \log \left| \frac{y_{j+1} - y_j}{x_{j+1} - x_j} \right| \approx \int \log |f'(x)| p(x) dx$$

- infer $X \to Y$ whenever

$$\hat{C}_{X \to Y} < \hat{C}_{Y \to X} \,.$$

"information geometric causal inference (IGCI)"

Experiments

Rhine data:

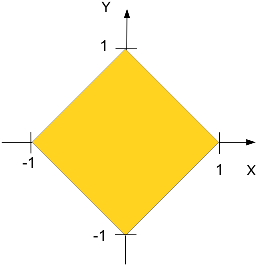
- water levels at 22 cities measured in 15 minutes intervals from 1990 to 2008,
- pick 231 random pairs and decide which one is "upstream"
- 87% correct decisions

Note: IGCI actually not suitable for non-deterministic relations yet although several positive results have been reported

7. **Additive noise models:**
Let X be uniformly distributed on $[-1, 1]$ and $Y = X^2$. Show that there is no function g such that

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

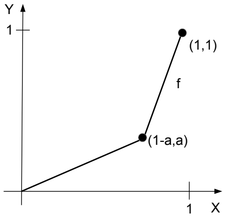
8. **Darmois-Skitovic:**
Let $P(X, Y)$ be the uniform distribution on the below diamond.



- Show that X and Y are uncorrelated.
- Show that $X \not\perp\!\!\!\perp Y$.

Convincing arguments are at least as good as calculations! (no lengthy calculations necessary)

9 Information-geometric causal inference:
Let f be the following bijection of the interval $[0, 1]$.



Let X be uniformly distributed on $[0, 1]$, i.e., $p(X) = 1$ (w.r.t. the Lebesgue measure) and $Y = f(X)$.

- Compute the density $p(Y)$ w.r.t. the Lebesgue measure.
- Argue in what sense $p(Y)$ contains information about f.