A common principle behind thermodynamics and causal Inference

Dominik Janzing

Max Planck Institute for Intelligent Systems
Tübingen, Germany

29. April 2015

MAX-PLANCK-GESELLSCHAFT

Outline

1. **Causal inference using conditional statistical independences**
(conventional approach since early 90s)

2. **Causal inference using the shape of probability distributions**
(first ideas around 2003, major results since 2008)

3. **Relating these new causal inference methods to the Arrow of Time**

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

⇒ Relation between statistical and causal dependences is tricky

Example for causal problems from our collaborations

- **Brain Research:**
  which brain region influences which one during some task?
  (goal: help paralyzed patients, given: EEG or fMRI data)

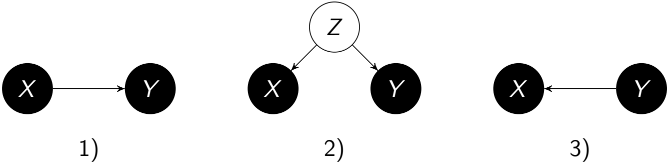
- **Biogenetics:**
  which genes are responsible for certain diseases?

- **Climate research:**
  understand causes of global temperature fluctuations

Part 1: Causal inference using conditional statistical independences

Reichenbach's principle of common cause (1956)

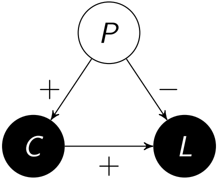
If two variables X and Y are statistically dependent then either



- in case 2) Reichenbach postulated $X \perp\!\!\!\perp Y \,|\, Z$ and linked this to thermodynamics in his book 'The direction of time' (1956)
- 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 and the focus of this talk.

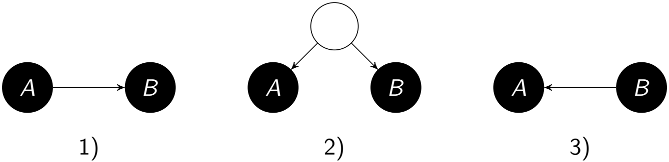
Coffee example

- coffee drinking C increases life expectancy C
- common cause “Personality” P increases coffee drinking C but decreases (via other habits) life expectancy L
- negative correlation by common cause stronger than positive by direct influence



Quantum causality

Ried, Agnew, Vermeyden, Janzing, Spekkens, Resch Nature Physics 2015



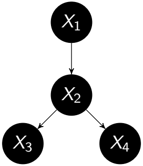
Observe dependences between measurements at system A and system B.

- **acausal state:** in scenario 2) there is a joint density operator on $\mathcal{H}_A \otimes \mathcal{H}_B$
- **causal state:** in scenario 1) and 3) there is an operator on $\mathcal{H}_A \otimes \mathcal{H}_B$ whose partial transpose is a density operator

There are dependences between A and B that can clearly be identified as 2) and those that can be identified as 1) or 3)

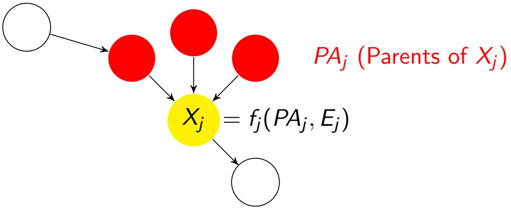
Causal inference problem, general form Spirtes, Glymour, Scheines, Pearl

- 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)



Functional model of causality Pearl et al

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

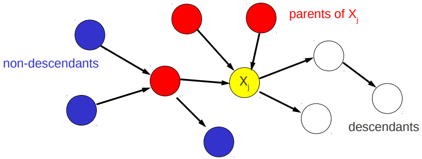


- all noise terms E_j are statistically independent (causal sufficiency)

- which properties of $P(X_1, \ldots, X_n)$ follow?

Causal Markov condition (4 equivalent versions) Lauritzen et al, Pearl

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



(information exchange with non-descendants involves parents)

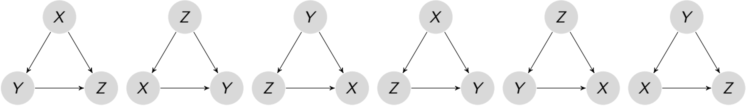
- **global Markov condition:** describes all ind. via d-separation
- **Factorization:** $P(X_1, \ldots, X_n) = \prod_j P(X_j | PA_j)$

(every $P(X_j | PA_j)$ describes a causal mechanism)

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



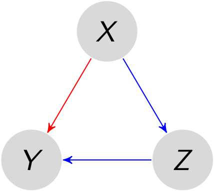
- reasonable rules for prefering simple DAGs required

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

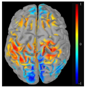


- not faithful, direct and indirect influence compensate
- **Application:** PC and FCI algorithm infer causal structure from conditional statistical independences

Application: Brain Computer Interfaces

Grosse-Wentrup & Schölkopf, 2011

- **Goal:** Paralyzed subjects communicate by activating certain brain regions

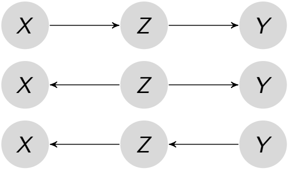


- **Open problem:** Performance of subjects varies strongly
- **Hypothesis:** Attention influenced by oscillations in the γ-frequency band
  - indeed, γ seems to influence the sensorimotor rhythm (SMR) since conditional dependences support the DAG

(Grosse-Wentrup, Schölkopf, Hill *NeuroImage* 2011)

Limitation of independence based approach:

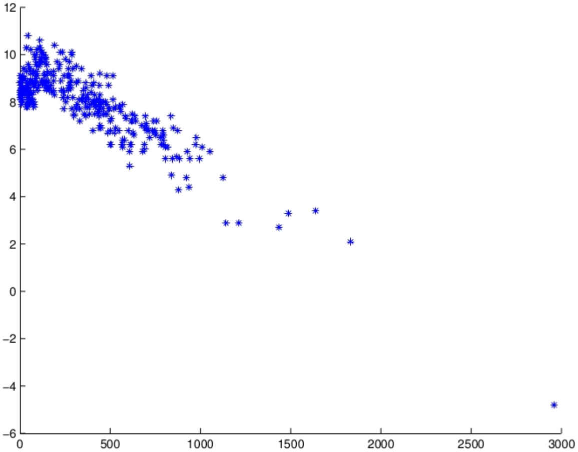
- many DAGs impose the same set of independences



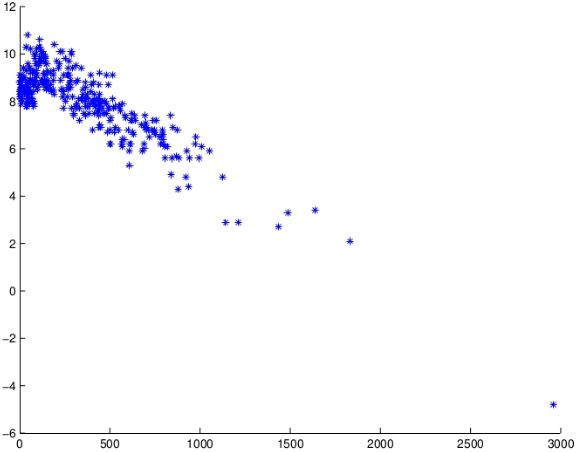
$X \perp\!\!\!\perp Y \,|Z$ for all three cases ("Markov equivalent DAGs")

- 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...

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

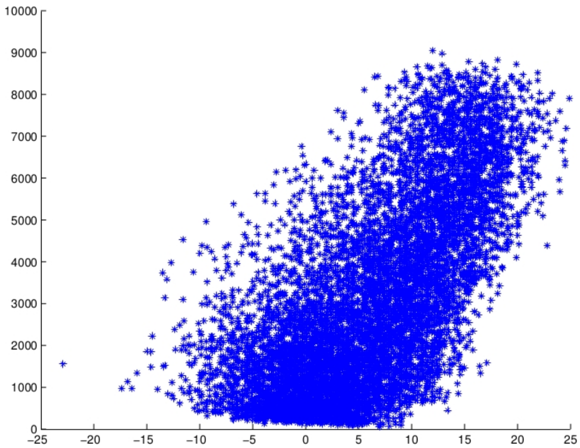


What's the cause and what's the effect?

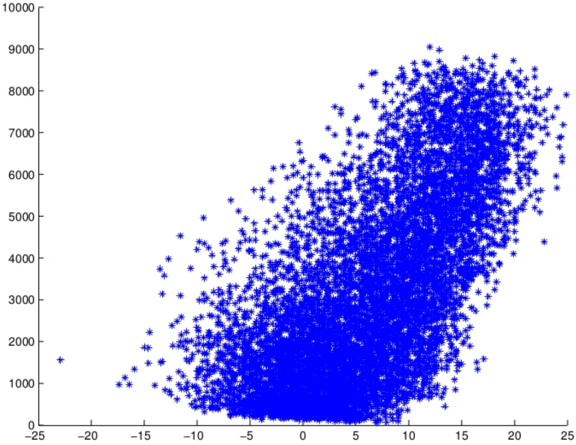


X (Altitude) $\to$ Y (Temperature)

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

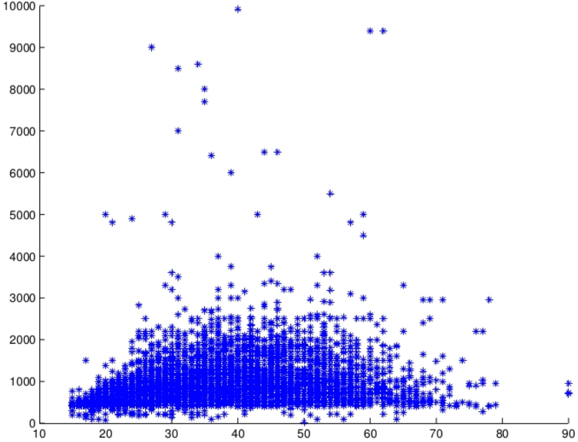


What's the cause and what's the effect?

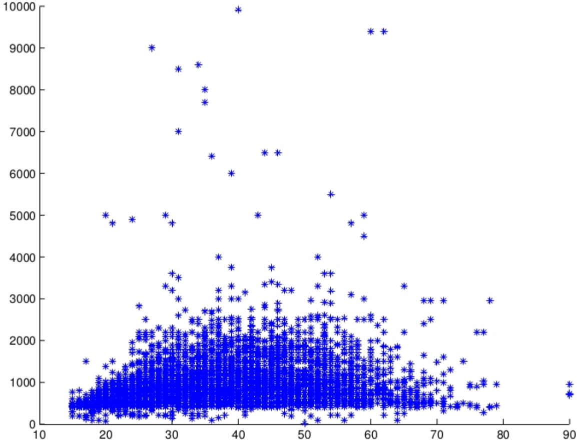


Y (Solar Radiation) $\rightarrow$ X (Temperature)

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



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



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

Hence...

- there are asymmetries between cause and effect apart from those formalized by the causal Markov condition

- new methods that employ these asymmetries need to be developed

Linear non-Gaussian models

Kano & Shimizu 2003

Theorem

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

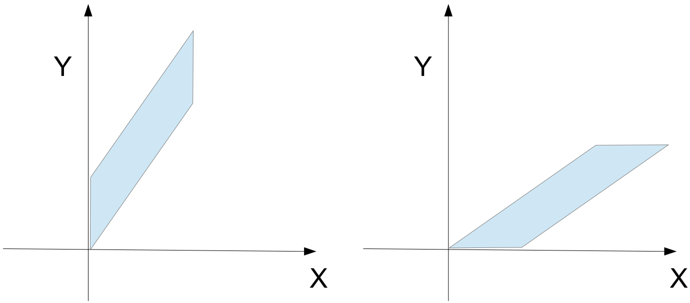
$$
\begin{aligned}
Y &= \alpha X + U_Y \text{ with } U_Y \perp\!\!\!\perp X \\
X &= \beta Y + U_X \text{ with } U_X \perp\!\!\!\perp Y\,,
\end{aligned}
$$

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):

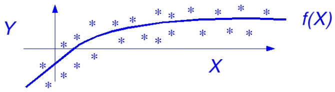


uniformly distributed Y and U_X with $X = \beta Y + U_X$ induces the diamond on the right.

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...

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

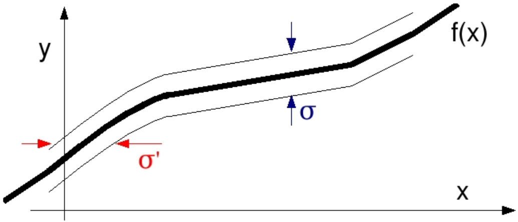
can model any conditional $P(Y|X)$

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

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)
- performed better than chance on real data with known ground truth

Justification of these methods

seems quite ad hoc: one defines a model class and believes that it is related to causal directions...

To avoid arbitrariness when inventing new inference methods we need a deeper foundation...

Tool: Algorithmic Information Theory Kolmogorov, Chaitin, Solomonoff, Gacs

- **Kolmogorov complexity:** $K(x)$: length of the shortest program on a universal Turing machine that outputs x
- **conditional Kolmogorov complexity:** $K(y|x^*)$ length of the shortest program that generates the output y from the shortest compression of x
- **algorithmic mutual information:**

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

measures the number of bits that a joint description of x, y saves compared to separate descriptions

Postulate: 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)

- idea: each $P(X_j|PA_j)$ describes independent mechanism of nature
- special case: shortest desription of $P(\mathrm{effect}, \mathrm{cause})$ is given by separate descriptions of $P(\mathrm{cause})$ and $P(\mathrm{effect}|\mathrm{cause})$.
- implication of a general theory connecting causality with description length

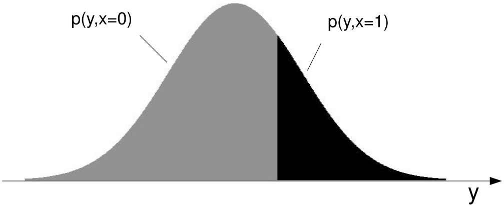
Janzing, Schölkopf: Causal inference using the algorithmic Markov condition, IEEE TIT (2010).

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

Illustrative toy example

Let X be binary and Y real-valued.

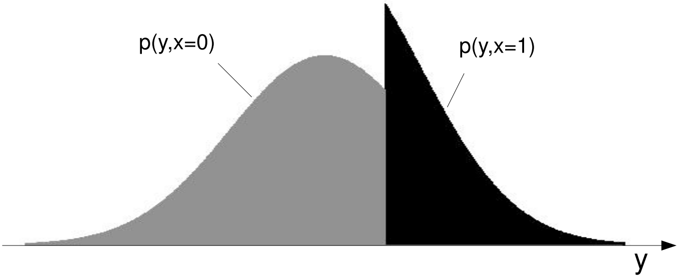
- Let Y be Gaussian and $X = 1$ for all y above some threshold and $X = 0$ otherwise.



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

Strange relation between $P(Y|X)$ and $P(X)$...

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



- $P(X)$ and $P(Y|X)$ seem to be adjusted to each other
- 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'.

Part 1: Relating these methods to the Arrow of Time

Arrow of time in stationary stochastic processes

Peters, DJ, Gretton, Schölkopf ICML 2009

- **Theorem:** If $(X_t)_{t\in\mathbb{Z}}$ has an autoregressive moving average (ARMA) model

$$X_t = \sum_{j=1}^{p} \alpha_j X_{t-j} + \sum_{j=1}^{q} \beta_j E_{t-j} + E_t \quad \text{with independent } E_t$$

  there is no such autoregressive model for (X_{-t}), unless E_t is Gaussian or $\alpha_j = 0$.

- **Experiment:** infer the direction of real-world time series (finance, EEG...)

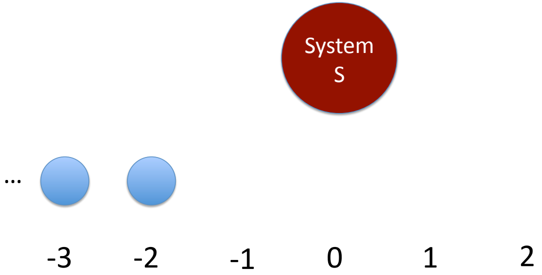
- **Result:** more often linear in forward than in backward direction

smells like an arrow of time, right?

Physical toy model for $X_t = \alpha X_{t-1} + E_t$

DJ, Journ. Stat. Phys. 2010

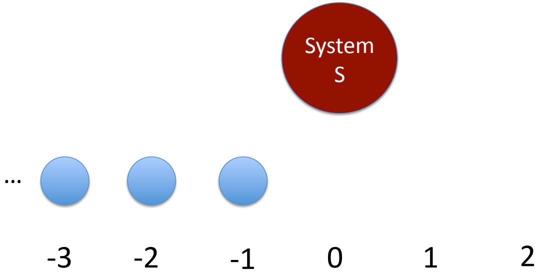
- X_t: physical observable of a fixed system S at time t.
- noise term provided by propagating particle beam (shift on $\mathbb{Z}$)



Physical toy model for $X_t = \alpha X_{t-1} + E_t$

DJ, Journ. Stat. Phys. 2010

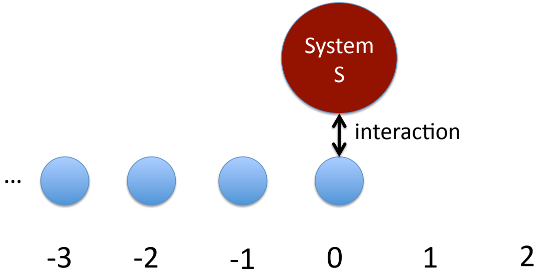
- X_t: physical observable of a fixed system S at time t.
- noise term provided by propagating particle beam (shift on $\mathbb{Z}$)



Physical toy model for $X_t = \alpha X_{t-1} + E_t$

DJ, Journ. Stat. Phys. 2010

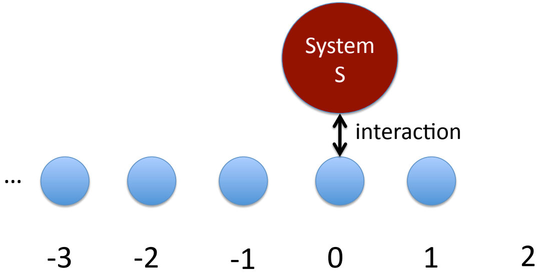
- X_t: physical observable of a fixed system S at time t.
- noise term provided by propagating particle beam (shift on $\mathbb{Z}$)



Physical toy model for $X_t = \alpha X_{t-1} + E_t$

DJ, Journ. Stat. Phys. 2010

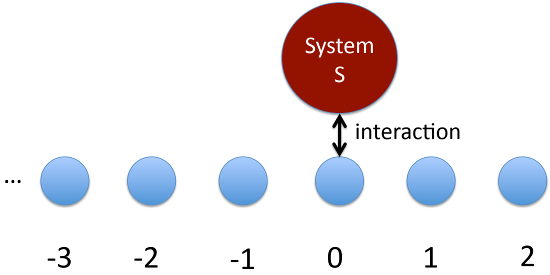
- X_t: physical observable of a fixed system S at time t.
- noise term provided by propagating particle beam (shift on $\mathbb{Z}$)



Physical toy model for $X_t = \alpha X_{t-1} + E_t$

DJ, Journ. Stat. Phys. 2010

- X_t: physical observable of a fixed system S at time t.
- noise term provided by propagating particle beam (shift on $\mathbb{Z}$)



Model and its implications

Assumptions:

- interaction is rotation on phase space of S and particle at position 0
- incoming particles statistically independent

Implications:

- outgoing particles are dependent (except for Gaussian states)
- coarse-grained entropy increased
- $P(X_t|X_{t-1})$ is linear, but not $P(X_{t-1}|X_t)$

Time-reversed process unlikely...

- incoming particles are statistically dependent
- interaction with S removes dependences
- outgoing particles independent
- rotation angle must be adapted to the dependences
- model requires adjustments between incoming state and rotation angle

Note the analogy...

- the input state (of the particles) and the mechanism transforming the state are independently chosen by nature

- $P(\textit{cause})$ and $P(\textit{effect}|\textit{cause})$ are independently chosen by nature

Another view on the Arrow of Time

This seems to be its crucial idea:

The initial state and the dynamical law are algorithmically independent

Arrow of time

- **typical closed system dynamics:**

$$\text{simple state} \to \text{complex state}$$

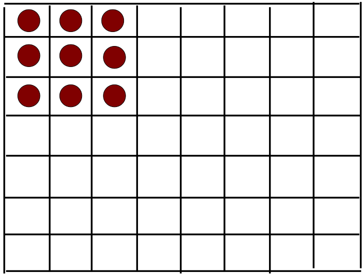
- **unlikely:**

$$\text{complex state} \to \text{simple state}$$

(thermodynamic entropy = Kolmogorov complexity?)

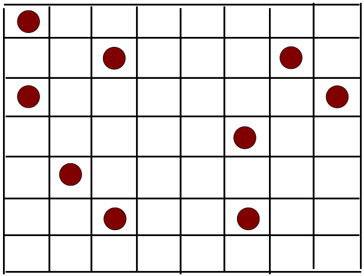
Zurek: Algorithmic randomness and physical entropy, PRA 1989

Discrete dynamical system



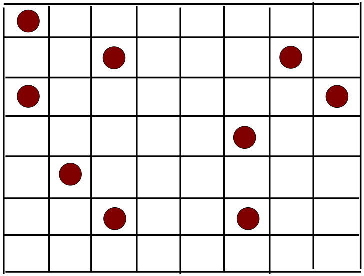
initial state s with low description length

Discrete dynamical system



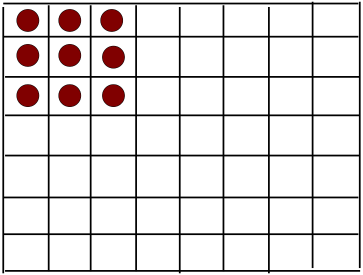
state $D(s)$ with large description length after applying bijective dynamical law D

Time reversed scenario



initial state with large dscription length $K(s)$

Time reversed scenario



final state with low description length $K(D(s))$

Independence principle induces Arrow of Time

initial state s, bijective dynamics D

- assume $K(D(s)) < K(s)$
- then $K(s|D) \stackrel{+}{=} K(D(s)|D) \stackrel{+}{\leq} K(D(s)) < K(s)$
- hence, s contains algorithmic information about D

Independence principle more general than Arrow of Time

Postulate:

$$K(s|D) \stackrel{+}{=} K(s)$$

also for non-bijective D

- implication $K(D(s)) \geq K(s)$ only holds for bijective D
- lower bounds for $K(D(s))$ in terms of non-bijectivity of D
- postulate makes also sense if D is probabilistic

- replace $s \equiv P(\mathrm{cause})$ and $D \equiv P(\mathrm{effect}|\mathrm{cause})$

Wrong approach to distinguish cause and effect

"Variable with lower entropy is the cause"
(motivated by thermodynamics)

- Cause may be continuous, effect binary
- entropy depends on scaling
- application of non-linear functions tends to decrease entropy

Conclusions

- Arrow of Time can be derived from algorithmic independence between initial state and dynamical law

- Algorithmic independence between $P(\text{cause})$ and $P(\text{effect}|\text{cause})$ implies novel causal inference rules

References

1. Spirtes, Glymour, Scheines: Causation, Prediction, and Search, 1993
2. Pearl: Causality. 2000
3. Kano & Shimizu: Causal Inference using non-normality, 2003.
4. Hoyer, Janzing, Mooij, Peters, Schölkopf: Nonlinear causal discovery with additive noise models, NIPS 2008.
5. Janzing & Schölkopf: Causal Inference using the algorithmic Markov condition, IEEE TIT 2010.
6. Peters, Janzing, Gretton, Schölkopf: Detecting the Direction of Causal Time Series, ICML 2009
7. Janzing: On the Entropy Production of Time Series with unidirectional linearity. J. Stat. Phys, 2010.

Thank you for your attention!