Non-statistical notions of independence in causal discovery

Dominik Janzing

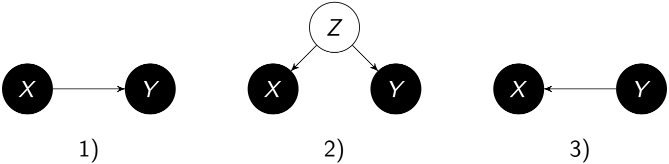
AWS Causality Team
Amazon Research Tübingen, Germany

September 2019

what does statistics tell us about causality?

Reichenbach's principle of common cause (1956)

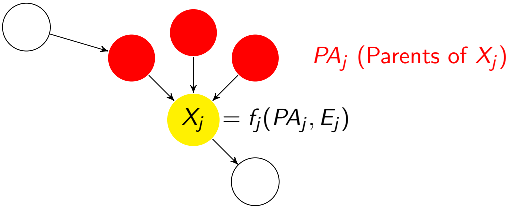
If two variables X and Y are statistically dependent then either



- 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.
- case 2 entails conditional independence $X \perp\!\!\!\perp Y \,|Z$
- cases 1-3 can also occur simultaneously

Functional model of causality Pearl et al

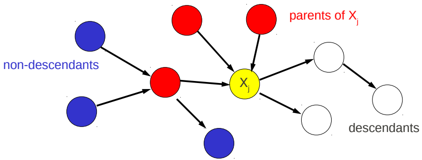
- every node X_j is a function of its parents PA_j and an unobserved noise term E_j
- f_j describes how X_j changes when parents are set to specific values



- 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 relations between single objects

- we don't infer causality only from **statistical** dependences.
- similarities of **single objects** also require a causal explanation

...but only if they are sufficiently complex

Measure complexity via 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

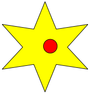
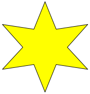
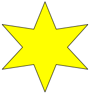
Chaitin, Gacs

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

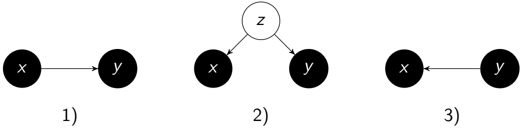
I( : ) = K()

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

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:

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

- 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**

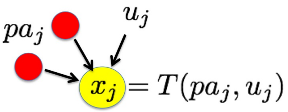
$$K(x_1, ..., x_n) \stackrel{+}{=} \sum_{j=1}^{n} K(x_j | pa_j^*)$$

→ 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

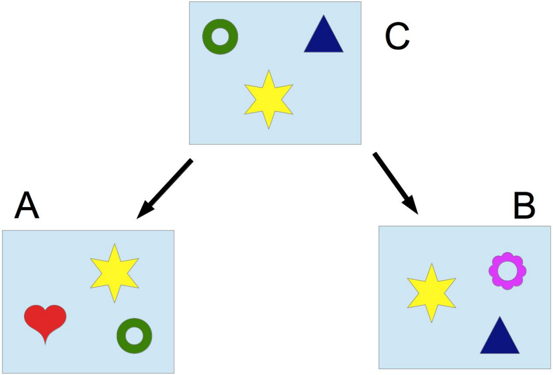
Algorithmic model of causality implies Markov condition

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

We need **computable** information measures instead of K

Ideas:

- compression length w.r.t. existing algorithm
- number of objects of a set
- ...

Questions:

- do they define notion of conditional (in)dependence?
- if yes, should we postulate also a causal Markov condition?

Axiomatic approach: define "information measure"

Given a set $S := \{x_1, \ldots, x_n\}$ of objects, a function $R : 2^S \rightarrow \mathbb{R}_0^+$ is called information measure if

- **normalization:** $R(\emptyset) = 0$
- **monotonicity:** $R(s) \leq R(t)$ for $s \subset t$
- **submodularity:** $R(s) + R(t) \geq R(s \cup t) + R(s \cap t)$

Steudel, DJ, Schölkopf, COLT 2010

Examples of such information measures

- discrete random variables $X_1, \ldots, X_k$

$$R(\{X_1, \ldots, X_k\}) := H(X_1, \ldots, X_k) \quad \text{(Shannon entropy)}$$

- strings $x_1, \ldots, x_k$

$$R(\{x_1, \ldots, x_k\}) := K(x_1, \ldots, x_k) \quad \text{(Kolmogorov complexity)}$$

submodular up to logarithmic terms

- sets $S_1, \ldots, S_k$

$$R(\{S_1, \ldots, S_k\}) := \# \left(\bigcup_j S_j \right) \quad \text{(number of elements)}$$

More examples...

- natural numbers $n_1, \ldots, n_k$

  $R(\{n_1, \ldots, n_k\}) := \log lcm(n_1, \ldots, n_k)$ (least common multiple)

- strings $x_1, \ldots, x_k$

  $R(\{x_1, \ldots, x_k\}) := LZ(x_1, \ldots, x_k)$ (Lempel-Ziv complexity)

  empirical evidence and partial theoretical results suggest that it is approximately submodular

Defining conditional (mutual) information

- **conditional information:**

$$R(s|t) := R(s \cup t) - R(t)$$

(non-negative due to monotonicity)

- **conditional mutual information:**

$$I(s : t|u) := R(s \cup u) + R(t \cup u) - R(s \cup t \cup u) - R(u)$$

(non-negative due to submodularity)

Equivalence of 3 Markov conditions for submodular R

Let $\{x_1, \ldots, x_n\}$ a set of objects, each corresponding to a node of a DAG G. Then the following three conditions are equivalent:

(1) **local Markov condition:** given its parents, every object is conditionally independent of its non-descendants

(2) **global Markov condition:** d-separation of nodes implies conditional independence

(3) the **joint information decomposes** according to the DAG structure

$$R(x_1, \ldots, x_k) = \sum_{j=1}^{k} R(x_j | pa_j)$$

for every causally sufficient subset $\{x_1, \ldots, x_k\}$ of nodes

$\Rightarrow$ mathematically, the Markov condition is well-defined,

but is it also a *reasonable* postulate for general R?

Recall justifications of statistical causal Markov conditions

via a **functional model:**

postulate the existence of unobserved noise variables $N_1, \ldots, N_n$ such that

- noise variables are statistically independent, i.e.,

$$H(N_1, \ldots, N_n) = \sum_j H(N_j)\,.$$

- every variable is a deterministic function of its parents and the noise

$$H(X_j, PA_j, N_j) = H(PA_j, N_j)\,.$$

generalization to arbitrary information measures

Definition: the objects $x_1, \ldots, x_n$ have an R-functional model of causality if there are "noise objects" $n_1, \ldots, n_n$ such that

- **the noise objects are R-independent**

$$R(n_1, \ldots, n_n) = \sum_j R(n_j) \,.$$

- **the causal mechanism is R-deterministic**

$$R(x_j, pa_j, n_j) = R(pa_j, n_j)$$

(the effect only contains information that is already contained in its observed or unobserved causes)

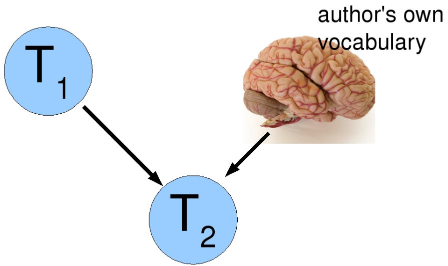
Theorem

the existence of an R-functional model implies the causal Markov condition with respect to R-independence.

this does not really *solve* the problem:

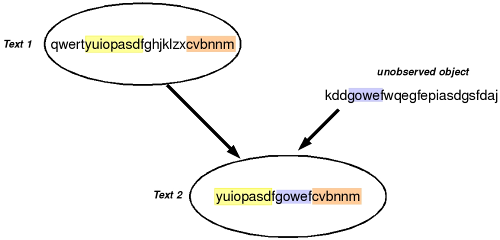
- to decide whether or not an R-functional model is reasonable depends on the domain
- in particular, to decide whether $R(x, y) \ll R(x) + R(y)$ necessarily indicates a causal relation requires domain knowledge

Functional model of plagiarism



- unobserved noise objects: personal vocabulary of every author, assumed to be disjoint
- every author mixes the vocabulary of the templates with his/her own vocabulary

Lempel-Ziv-functional model for texts



- unobserved noise objects $N_1, \ldots, N_n$ (LZ-independent)
- every text T_j is a concatenation of k substrings taken from its parents PA_j and N_j

then the LZ Markov condition holds up to an error term of size k

Non-statistical information on top of statistics

Postulate: Algorithmic Independence of Conditionals

If n random variables $X_1, \ldots, X_n$ are related by a causal DAG G, the conditionals $P(X_j|PA_j)$ in the causal factorization

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

are algorithmically independent.

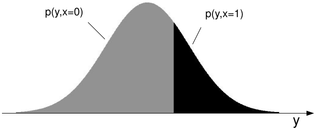
Markov equivalent DAGs may get distinguishable

DJ & Schölkopf, IEEE TIT 2010. Lemeire & DJ, 2012.

Toy example

Let X be binary and Y real-valued.

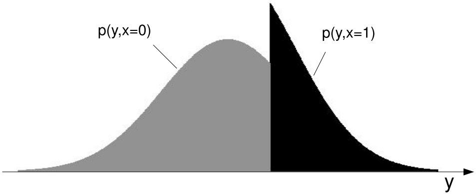
- Let Y be Gaussian and $X = 1$ for all y above some threshold and $X = 0$ otherwise.



- $Y \to X$ is plausible: simple thresholding mechanism
- $X \to Y$ requires a strange mechanism:
  look at $P_{Y|X=0}$ and $P_{Y|X=1}$!

not only $P_{Y|X}$ itself is strange...

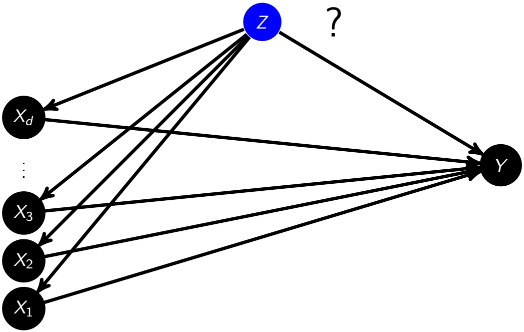
but also what happens if we change P_X:



Hence, reject $X \rightarrow Y$ because it requires *tuning* of P_X relative to $P_{Y|X}$.
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'.

Detect whether a multivariate model is causally sufficient

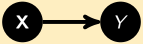
Problem: target Y correlated with potential cause $\mathbf{X} = (X_1, \ldots, X_d)$, but correlation may be due the common cause $\mathbf{Z}$ (e.g.: observed genes may correlate with a phenotype although it is only influenced by unobserved genes)



Goal: infer from $P_{\mathbf{X},Y}$ alone (!) whether hidden common cause $\mathbf{Z}$ exists and whether correlations between $\mathbf{X}$ and Y are dominated by the confounder

Postulate: "Independence of Mechanisms"

For the causal structure



$P_{\mathbf{X}}$ contains no information about $P_{Y|\mathbf{X}}$

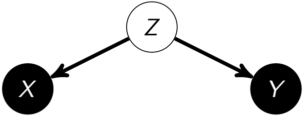
Possible formalizations:

- **algorithmic independence:** knowing $P_{\mathbf{X}}$ does not enable a shorter description of $P_{Y|\mathbf{X}}$ and vice versa (DJ & Schölkopf 2010)
- **no semi-supervised learning in causal direction:** unlabelled **x**-values are useless for learning $P_{Y|\mathbf{X}}$ (Schölkopf, DJ, ... 2012)
- **here: generic orientation of the regression vector:** for

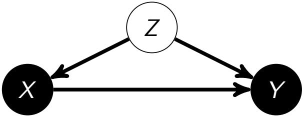
$$Y = \langle \mathbf{a}, \mathbf{X} \rangle + E$$

the vector **a** is not aligned with eigenvectors of $\Sigma_{\mathbf{X},\mathbf{X}}$

Detecting confounding and overfitting



purely confounded



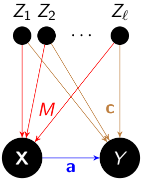
confounded causal relation

- we found different models of confounding for which regression vector is mainly contained in the low eigenvalue subspaces of $\Sigma_{\mathbf{X},\mathbf{X}}$
- same effect also obtained by overfitting small sample sizes

- note: some models of confounding yield concentration in *large* eigenvalue subspaces

DJ & BS, Journal of Causal Inference 2017

Linear model with many independent common causes



$$\mathbf{X} = M\mathbf{Z} \qquad Y = \langle \mathbf{a}, \mathbf{X} \rangle + \langle \mathbf{c}, \mathbf{Z} \rangle$$

($\mathbf{c}$, $\mathbf{a}$ randomly drawn from an isotropic prior)

regression vector:

$$\tilde{\mathbf{a}} := \Sigma_{\mathbf{X},\mathbf{X}}^{-1} \Sigma_{\mathbf{X},Y} = \underbrace{\mathbf{a}}_{\text{causal}} + \underbrace{M^{-T}\mathbf{c}}_{\text{confounding}}$$

results for high dimensions:

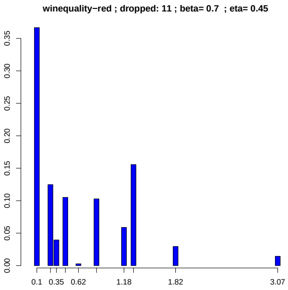
- $M^{-T}\mathbf{c}$ concentrates in low eigenvalue subspace of $\Sigma_{\mathbf{X},\mathbf{X}} = MM^T$
- confounding strength

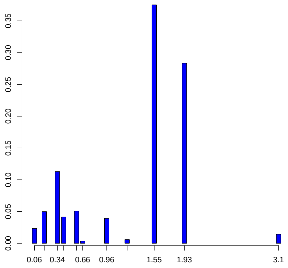
$$\beta := \frac{\|M^{-T}\mathbf{c}\|^2}{\|M^{-T}\mathbf{c}\|^2 + \|\mathbf{a}\|^2}$$

can be estimated from the direction of $\tilde{\mathbf{a}}$

Visualization of the concentration effect

x-**axis:** eigenvalues of $\Sigma_{\mathbf{X},\mathbf{X}}$
y-**axis:** sq.-length of component of $\tilde{\mathbf{a}}$ in the respective eigenspace





confounded case:
strong component for the smallest eigenvalue

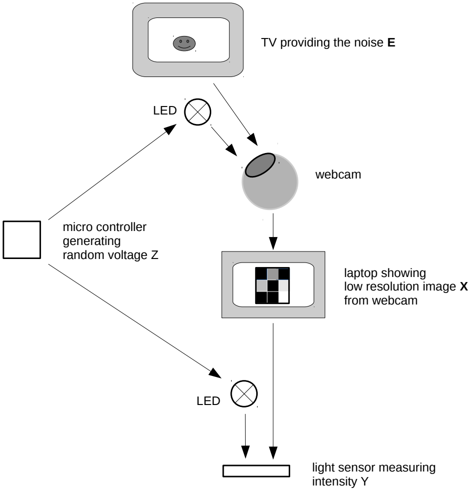
unconfounded case:
strong component at random position

Experiments with real data: taste of wine

- **causes** $X_1, \ldots, X_{11}$**:** ingredients (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol)
- **effect** Y**:** taste between 1 and 10 according to the opinion of human subjects

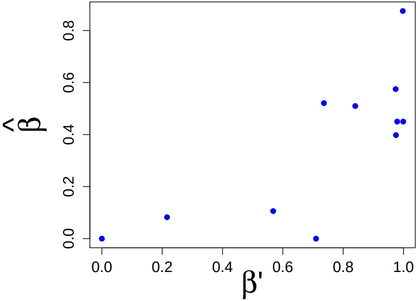
- clearly, $\mathbf{X}$ has some influence on Y (i.e. not purely confounded)
- linear model identifies X_{11} (alcohol) as the strongest influence
- algorithm estimates zero confounding strength ($\beta = 0$)
- algorithm estimates $\beta = 1$ if alcohol is dropped

Optical experiments with known confounding



- **cause X:** pixel vector on Laptop screen
- **target** Y**:** light intensity at the sensor
- **confounder** Z**:** light intensity of LEDs

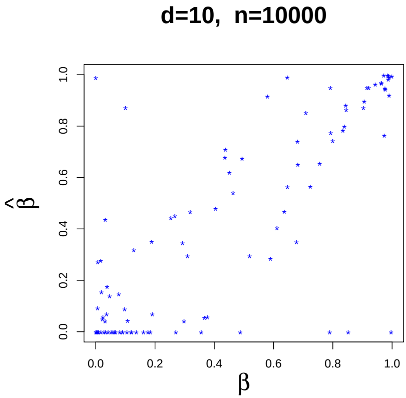
Results: estimated versus true confounding strength



here: systematic underestimation (maybe specific to this particular setup)

Estimated versus true confounding strength in simulations

data sets generated according to the above model
(random choice of **a** and **c**)



Causal regularization

(DJ, NeuriPS 2019)

- **use Ridge and Lasso against confounders:**
  - suppresses part in low eigenvalue space of $\Sigma_{\mathbf{X},\mathbf{X}}$ (employs dependence between $P_{\mathbf{X}}$ and $P_{Y|\mathbf{X}}$)
  - increases prediction error only slightly
  - significantly improves causal model

- **causal learning theory:**
  regression models from small function classes have better chances to be "causal"
  ("generalize" better from observational to interventional distribution)

Take home messages

- **non-statistical dependences** also provide causal information
- they either admit causal inference among **individual objects**
- or they **add a level** to the usual statistical perspective

References

D. Janzing and B. Schölkopf.
Causal inference using the algorithmic Markov condition.
IEEE Transactions on Information Theory, 56(10):5168–5194, 2010.

B. Steudel, D. Janzing, and B. Schölkopf.
Causal Markov condition for submodular information measures.
Proceedings of the 23rd Annual Conference on Learning Theory (COLT), pages 464–476, 2010.

J. Peters, D. Janzing, and B. Schölkopf.
Elements of Causal Inference – Foundations and Learning Algorithms.
MIT Press, 2017.

D. Janzing and B. Schölkopf.
Detecting non-causal artifacts in multivariate linear regression models.
In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

D. Janzing.
Causal regularization.
NeurIPS, 2019.

Thank you for your attention!