Inferring causality from passive observations

Dominik Janzing

Max Planck Institute for Intelligent Systems
Tübingen, Germany

18.-22. August 2014

MAX-PLANCK-GESELLSCHAFT

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 1, continued: why the relation between statistics and causality is tricky

(remaining part on equivalence of Markov conditions)

Factorization $\Rightarrow$ functional model

Generate each $p(X_j|PA_j)$ in

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

by a deterministic function and a noise variable.

- Idea: "encode" $X_j|pa_j$ (for all values pa_j) into the noise U_j, and pick out the right one depending on pa_j.
- formally, U_j is a map satisfying

$$pa_j \mapsto X_j|pa_j$$

- define structural equation

$$f_j(pa_j, U_j) := U_j(pa_j) = X_j|pa_j$$

- special case easier to understand: if PA_j only takes d values, U_j is an d-dimensional random vector, and the structural equation picks out a component of the vector (see next slide)

Simple case

Goal: generate $P(Y|X)$ via structural equation

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

- let X attain values in $\mathcal{X} = \{x_1, \ldots, x_k\}$
- let Y attain values in $\mathcal{Y}$
- let $E = (E_1, \ldots, E_k)$ be vector valued, each E_j attaining values in $\mathcal{Y}$
- let E_j have distribution $P(Y|x_j)$
- define $f(x_j, E) := E_j$

- note: joint distribution of $E_1, \ldots, E_k$ not relevant (ambiguity of noise distribution), only the marginals $P(E_j)$ matter

Quantitative causal statements

Pearl's do calculus

- **Motivation:** goal of causality is to infer the effect of interventions
- distribution of Y given that X is set to x:

$$p(Y|do(X = x)) \text{ or } p(Y|do(x))$$

- don't confuse with $p(Y|x)$
- can be computed from p and G, but not from p alone

Computing $p(X_1, \ldots, X_n | do(x_i))$

from $p(X_1, \ldots, X_n)$ and G

- Start with causal factorization

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

- Replace $p(X_i | PA_i)$ with $\delta_{X_i x_i}$

$$p(X_1, \ldots, X_n) = \prod_{j \neq i} p(X_j | PA_j) \delta_{X_i, x_i} \,.$$

Computing $p(X_k|do(x_i))$

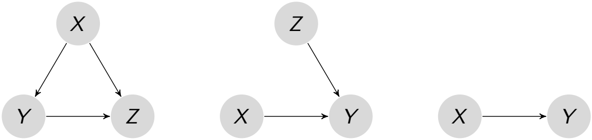
- summation/integration over x_i yields

$$p(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n|do(x_i)) = \prod_{j\neq i} p(X_j|PA_j(x_i)),$$

where $PA_j(x_i)$ is obtained by substituting x_i into PA_j.

- obtain $p(X_k|do(x_i))$ by marginalization

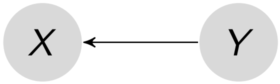
Examples for $p(y|do(x)) = p(y|x)$



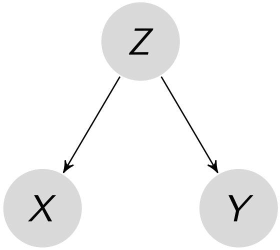
X causes Y and there is no common cause of X and Y

Examples for $p(y|do(x)) \neq p(y|x)$

- $p(Y|do(x)) = p(Y) \neq p(Y|x)$

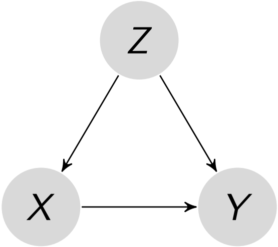


- $p(Y|do(x)) = p(Y) \neq p(Y|x)$



Example: controlling for confounding

$X \not\perp\!\!\!\perp Y$ partly due to the confounder and partly due to $X \rightarrow Y$.



Computing $p(Y|do(x))$ shows the part that is due to X causing Y

Controlling for confounding by deriving $p(Y|do(x))$

- causal factorization

$$p(X, Y, Z) = p(Z)p(X|Z)p(Y|X, Z)$$

- replace $p(X|Z)$ with $\delta_{X,x}$

$$P(X, Y, Z|do(x)) = p(Z)\delta_{X,x}p(Y|X, Z)$$

- marginalize

$$p(Y|do(x)) = \sum_z p(z)p(Y|x, z)$$

Identifiability problem

see e.g. Pearl, Tian

- given the causal DAG G on $X_1, \ldots, X_n$ and two nodes X_k, X_i
- which nodes apart from X_k, X_i need to be observed to compute $p(X_i | do\, x_i)$?

see e.g. results on backdoor criterion and frontdoor criterion in Pearl's book "Causality"

Formal treatment of causal influence

- for **binary** X, Y we define the average causal effect by

$$ACE := p(Y = 1|do(X = 1)) - p(Y = 1|do(X = 0)) \,.$$

(increase of probability of $Y = 1$ by setting X to 1)

- for **real-valued** Y **and binary** X one defines

$$ACE := \mathbb{E}[Y|do(X = 1)] - \mathbb{E}[Y|do(X = 0)]$$

(increase of expectation of Y by setting X to 1

- for **real-valued** X, Y one also uses

$$ACE := \frac{d}{dx}\mathbb{E}[Y|do(X = x)]$$

(increase of expectation of Y by infinitesimal increase of X)

Formal treatment of Simpson's paradox

- $p(y|do(x), z) = p(y|x, z)$ for this DAG.
- the drug helps males and females:

$$\begin{aligned} p(recovery|do(drug), male) &> p(recovery|do(no\ drug), male) \\ p(recovery|do(drug), female) &> p(recovery|do(no\ drug), female) \end{aligned}$$

- the drug helps also on the average:

$$p(recovery|do(drug)) > p(recovery|do(no\ drug)),$$

because

$$p(y|do(x)) = \sum_z p(y|x, z)p(z)$$

Naive solution of the coffee paradox

- observe as many other variables as possible: smoking, nutrition, sports, height, weight, race

- compare life expectancy of people for which all variables are the same except for coffee consumption

Correct solution of the coffee paradox

- X_i (daily coffee consumption) and X_j (length of ones life) are part of a large causal DAG
- there is no arrow from X_j to X_i
- observed $X_i \not\perp\!\!\!\perp X_j$ can be due to directed paths from X_i to X_j or due to paths from common causes to both X_i and X_j
- to asses whether X_i influences X_j we need to block all paths from common causes to X_i

What's wrong with the naive solution

conditioning on all the other variables

- screens off indirect influence
  (e.g. if coffee influences weight and weight influences life expectancy)

- generates dependences via selection bias
  (however, hard to imagine a common effect of coffee drinking and life expectancy)

Part 2: causal inference using conditional independences

- **why the Markov condition is not enough**
  additional postulate: causal faithfulness

- **algorithms for causal inference**

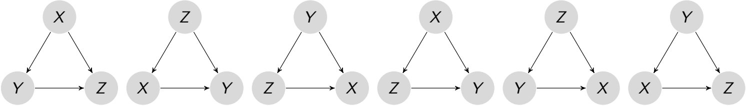
- **causal inference from non-statistical observations**
  defining similarities of single objects

Why the Markov condition is not enough

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



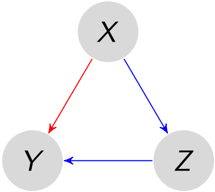
- reasonable rules for prefering simple DAGs required

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

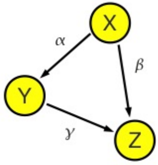
Examples of unfaithful distributions (1)

Cancellation of direct and indirect influence in linear models

$$\begin{aligned} X &= U_X \\ Y &= \alpha X + U_Y \\ Z &= \beta X + \gamma Z + U_Z \end{aligned}$$

with independent noise terms U_X, U_Y, U_Z

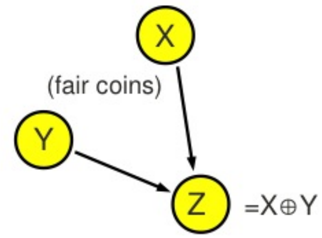
$$\beta + \alpha\gamma = 0 \quad \Rightarrow \quad X \perp\!\!\!\perp Z$$



Examples of unfaithful distributions (2)

binary causes with XOR as effect

- for $p(X), p(Y)$ uniform: $X \perp\!\!\!\perp Z$, $Y \perp\!\!\!\perp Z$.
  i.e., unfaithful (since X, Z and Y, Z are connected in the graph).



- for $p(X), p(Y)$ non-uniform: $X \not\perp\!\!\!\perp Z$, $Y \not\perp\!\!\!\perp Z$.
  i.e., faithful

Conditional-independence based causal inference

Spirtes, Glymour, Scheines and Pearl
Causal Markov condition + Causal faithfulness:

- accept only those DAGs G as causal hypotheses for which

$$(X \perp\!\!\!\perp Y \,|Z)_G \quad \Leftrightarrow \quad (X \perp\!\!\!\perp Y \,|Z)_p \,.$$

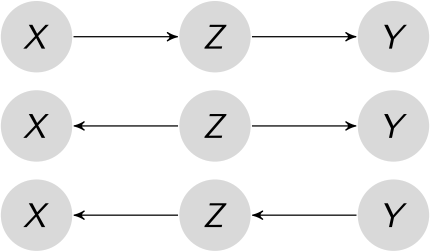
- identifies causal DAG up to Markov equivalence class
(DAGs that imply the same conditional independences)

Markov equivalence class

Theorem (Verma and Pearl, 1990): two DAGs are Markov equivalent iff they have the same skeleton and the same *v*-structures.

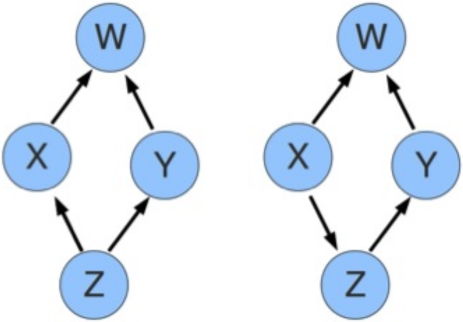
skeleton: corresponding undirected graph
v-structure: substructure $X \to Y \leftarrow Z$ with no edge between X and Z

Markov equivalent DAGs



same skeleton, no *v*-structure

Markov equivalent DAGs



same skeleton, *v* structure at W

Algorithms for causal inference

Algorithmic construction of causal hypotheses

IC algorithm by Verma & Pearl (1990) to reconstruct DAG from p

idea:

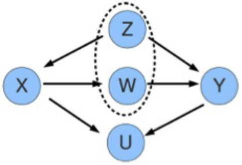
1. Construct skeleton
2. Find v-structures
3. direct further edges that follow from
    - graph is acyclic
    - all v-structures have been found in 2)

Construct skeleton

Theorem: X and Y are linked by an edge iff there is no set S_{XY} such that

$$(X \perp\!\!\!\perp Y \,|\, S_{XY})_G \,.$$

Explanation: dependence that is due to indirect links can be screened off by conditioning



$$X \perp\!\!\!\perp Y \,|\, \{Z, W\} \,.$$

Faithfulness implies: edge $X - Y$ exists iff there is a set $S_{X,Y}$ such that

$$X \perp\!\!\!\perp Y \,|\, S_{XY} \,.$$

(S_{XY} is called a Sepset for X, Y)

Efficient construction of skeleton

PC algorithm by Spirtes & Glymour (1991)

iteration over size of Sepset

1. remove all edges $X - Y$ with $X \perp\!\!\!\perp Y$
2. remove all edges $X - Y$ for which there is a neighbor $Z \neq Y$ of X with $X \perp\!\!\!\perp Y \,|Z$
3. remove all edges $X - Y$ for which there are two neighbors $Z_1, Z_2 \neq Y$ of X with $X \perp\!\!\!\perp Y \,|Z_1, Z_2$
4. ...

Advantages

- many edges can be removed already for small Sepsets
- testing all sets S_{XY} containing the adjacencies of X is sufficient
- depending on sparseness, algorithm only requires independence tests with small conditioning tests
- running time only polynomial in n if DAG has bounded degree

Find v-structures

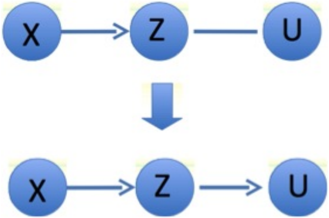
- given $X - Z - Y$ with X and Y non-adjacent
- given S_{XY} with $X \perp\!\!\!\perp Y \,|S_{XY}$
  a priori, there are 4 possible orientations:

$$\left.\begin{array}{l} X \to Z \to Y \\ X \leftarrow Z \to Y \\ X \leftarrow Z \leftarrow Y \end{array}\right\} \qquad Z \in S_{XY}$$

$$X \to Z \leftarrow Y \qquad Z \notin S_{XY}$$

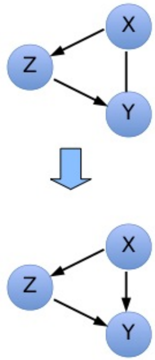
Orientation rule: create v-structure if $Z \notin S_{XY}$

Direct further edges (Rule 1)



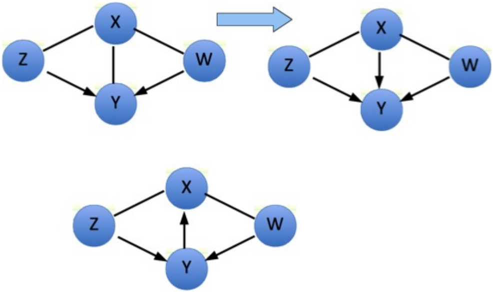
(otherwise we get a new *v*-structure)

Direct further edges (Rule 2)



(otherwise one gets a cycle)

Direct further edges (Rule 3)



(could not be completed without creating a new *v*-structure)

Conditional independence tests

- **discrete case:** contingency tables
  - **idea:** for each $\mathbf{z}$, compare relative frequencies of $(\mathbf{x}, \mathbf{y})$ with product of relative frequency of $\mathbf{x}$ with relative frequency of $\mathbf{y}$
  - **problem:** if $\mathbf{Z}$ attains many different values, many pairs $(\mathbf{x}, \mathbf{y})$ occur only once unless the sample size is huge
- **multi-variate Gaussian case:** covariance matrix contains all information about conditional independences

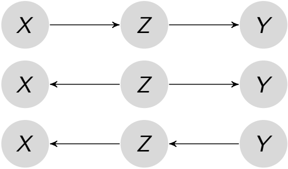
non-Gaussian continuous case: challenging, recent progress via reproducing kernel Hilbert spaces (Fukumizu...Zhang...)

Improvements

- CPC (conservative PC) by Ramsey, Zhang, Spirtes (1995) uses weaker form of faithfulness
- FCI (fast causal inference) by Spirtes, Glymour, Scheines (1993) and Spirtes, Meek, Richardson (1999) infers causal links in the presence of latent common causes
- for implementations of the algorithms see homepage of the TETRAD project at Carnegie Mellon University Pittsburgh

Limitation of independence based approach:

- many DAGs impose the same set of independences

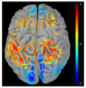


$X \perp\!\!\!\perp Y \,|\, Z$ for all three cases ("Markov equivalent DAGs")

- 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...

Application: Brain Computer Interfaces

- **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)

Take home messages

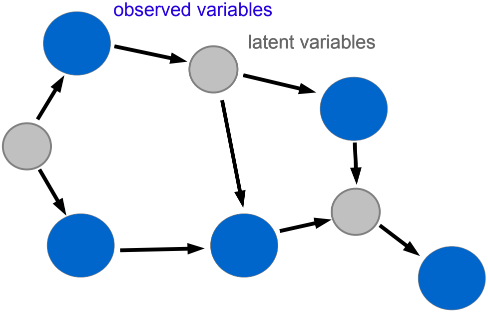
principles of independence based causal inference

- Markov equivalence class is the set of all DAGs that imply the same set of conditional independences
- Markov condition and faithfulness allow for identifying the Markov equivalence class of the true DAG
- algorithms start by first constructing undirected links (uniquely determined by the Markov equivalence class)
- then part of the links are directed and some remain undirected
- the partially directed graph represents the Markov equivalence class

Next step: Dropping causal sufficiency

So far, we have always assumed that no pair of observed variables have unobserved common causes

Let $\mathcal{O} := \{X_1, \ldots, X_n\}$ be observed and $\mathcal{L} := (L_1, \ldots, L_k)$ be latent variables



Then a latent structure is a DAG G on $\mathcal{O} \cup \mathcal{L}$.

We will now see that...

- causal inference is also possible without causal sufficiency

- for given set $\mathcal{O}$, there is an infinity of latent structures
- causal inference then seems to search over an infinity of structures

- nevertheless, inferring causal relations among observed variables requires searching over finitely many structures only

Representing latent structures with MAGs

maximal ancestral graphs:
graph containing 3 types of edges:

$$X \to Y \qquad X \leftarrow Y \qquad X \leftrightarrow Y$$

semantics:

- $X \to Y$ means that there is a directed path from X to Y in G for which each observable non-endpoint node is a collider and an ancestor of X or Y
  (that is, the influence of X and Y cannot screened off by conditioning)
- $X \leftrightarrow Y$ means that there is a confounder, i.e., a node $L \in \mathcal{L}$ having directed paths to X and Y

Idea of latent model reconstruction

- **construction of skeleton:**
  remains the same: X and Y are adjacent iff there is no set S_{XY} with
  $$X \perp\!\!\!\perp Y \,|S_{XY}$$
- **construction of v-structures:**
  orient $X - Z - Y$ to $X- > Z < -Y$ whenever $Z \notin S_{XY}$
  (note: whether there are arrowheads at X and Y is left often)
- **construct definite arrows**
  some independence patters tell us for some edges which of the 3 types of links is present

Example: recognizing Y-structures

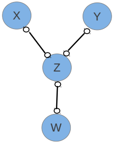
$\mathcal{O} = \{X, Y, Z, W\}$ with

$$X \perp\!\!\!\perp Y$$
$$XY \perp\!\!\!\perp Z \,|\, W$$

as generating set of the independences

Example: recognizing Y-structures

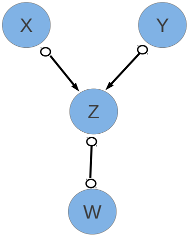
Step 1: construct the skeleton



- remove $X - Y$ because $X \perp\!\!\!\perp Y$
- remove $X - W$ because $X \perp\!\!\!\perp W | Z$
- remove $Y - W$ because $Y \perp\!\!\!\perp W | Z$

Example: recognizing Y-structures

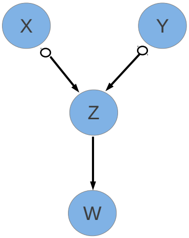
identifying v-structure



- add arrowheads to get a collider at Z because $Z \notin S_{XY}$
- leave it open whether there are arrowheads at the other end of the edges

Example: recognizing Y-structures

orient further edges such that no additional v-structures are created



- there cannot be an arrowhead at Z because $Z \in S_{X,W}$
- there must be an arrowhead at W because every edge has at least one arrowhead

conclusion: Z influences W!

Hence...

- faithfulness and Markov condition on $\mathcal{O} \cup \mathcal{L}$ sometimes imply causal conclusions for $\mathcal{O}$
- one can prove causal influence without assuming causal sufficiency

Causal inference from single observations

- drop the i.i.d. assumption $\rightarrow$ radical approach: causal inference without probabilities
- this approach will help to justify new *statistical* causal inference rules

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 11001001000011111101101010100001...
(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...

$$11.00100100001111110110101010001... = \pi$$

- 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, Solmonoff 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 as backgroun information
- $K(y|x^*)$ 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)

Information of x about y

- $I(x:y) := K(x) + K(y) - K(x,y) \stackrel{+}{=} K(x) - K(x|y^*) \stackrel{+}{=} K(y) - K(y|x^*)$
- Interpretation: number of bits saved when compressing x, y jointly rather than compressing them independently

Analogy to statistics:

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

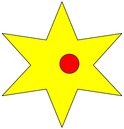
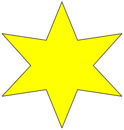
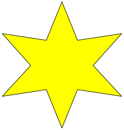
Causal Principle

(Janzing & Schölkopf 2008)

Let x and y be strings that describe two objects/observations in nature. Whenever $I(x : y) \gg 0$, there is some kind of causal relation between x and y.

- causal relation is $x \rightarrow y$, $y \rightarrow x$, or $x \leftarrow z \rightarrow y$
- analogy to Reichenbach's principle of common cause: whenever $I(X : Y) \gg 0$ for two variables X and Y there is a causal relation of the form $X \rightarrow Y$, $Y \rightarrow X$, or $X \leftarrow Z \rightarrow Y$.

algorithmic mutual information: example

I( : ) = K()

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:

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

Postulate: Algorithmic Markov condition

(Janzing & Schölkopf), Causal inference using the algorithmic Markov condition, 2010

Given n observations $x_1, ..., x_n$ (formalized as strings)
Given its direct causes pa_j, every x_j is conditionally algorithmically independent of its non-effects:

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

Equivalence of algorithmic Markov conditions

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

Take home messages

- we have developed a framework for causal inference from individual observations
- it should be applicable to real world data where good compression schemes are available
- we will later use it for justifying novel *statistical* causal inference rules: select among causal DAGs within the same Markov equivalence class

Generalized PC

Steudel et al: Causal Markov condition for submodualr information measures

PC algorithm also works with generalized conditional independence derived from information functions R other than Shannon entropy

Examples:

1. $R :=$ number of different words in a text
2. $R :=$ compression length (e.g. Lempel Ziv is approximately submodular)
3. $R :=$ logarithm of period length of a periodic function

example 2 yielded reasonable results on simple real texts (different versions of a paper abstract)

4. **do-calculus:**
Let X, Y, Z be binary and coupled by the deterministic structural equations

$$\begin{aligned} Z &= U_Z \\ X &= Z \\ Y &= X \oplus Z \oplus U_Y \,, \end{aligned}$$

where $\oplus$ denotes the XOR, U_Y is a binary attaining 1 with probability $\epsilon < 1/2$ and U_Z is a binary attaining 1 with probability $\delta \neq 1/2$

- Argue/show that $X \perp\!\!\!\perp Y$.
- Show formally that X has an influence on Y by proving that

$$p(Y|do(X=1)) \neq p(Y|do(X=0))\,.$$

5. **Faithfulness:**
   Given the causal DAG $X \to Y \to Z$. Let Y be deterministically depend on X, i.e., the structural equation for Y reads

$$Y = f(X)\,.$$

   Show that the joint distribution of X, Y, Z is not faithful.

Exercises

6. **Markov equivalence:**
   give all DAGs with 3 nodes whose Markov equivalence class consists of only one element.