Inferring causality from passive observations

Dominik Janzing

Max Planck Institute for Intelligent Systems
Tübingen, Germany

18.-22. August 2014

Preliminarities

- **Interdisciplinary topic:** between computer science, mathematics, philosophy of science, relations to physics, applications in all kind of sciences such that economy, psychology, biology,...
- **Switches between vague and precise:** causality is hard to formalize. Justifying mathematical assumptions about causality involves philosophical issues. However, once we have stated assumptions, we prove precise mathematical theorems.
- **Challenging** both from the conceptual and the mathematical perspective
- **Ask questions on all levels:** during and after the lectures and excercises as much as you like! Gaps that appear to be huge can usually be closed quickly. Don't ask scientific questions by email!
- **Structure:** the slides are carefully structured and contain the main material. My explanations on the blackboard are spontaneous and need not be well-structured.

Schedule

- **morning sessions:** lectures and (at the end) presentation of the questions to be done until the next day exercises session.

- **afternoon sessions:**
  - Monday: Questions and feedback (optional, but highly recommended)
  - Tuesday to Friday: Solution of the homework from the previous day
  - Friday: brainstorming about future directions

Requirements for passing

- **Homework assignments:**
  50 out of 100 credits
- **Presence:** obligatory unless there are good reasons

Literature:

- Peter Spirtes, Clark Glymour, Richard Scheines: **Causation, Prediction, and Search,** 1993
- Judea Pearl: **Causality. Models, Reasoning, and Inference,** 2000.

references to articles are given on the respective slides.

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 1: the tricky relation between statistics and causality

- **what's wrong with common causal conclusions:**
  motivation of the problem
- **mathematics tools:**
  measure theoy, statistical (in)dependences vs. correlations, information theory
- **first basis for correct causal conclusions:**
  Reichenbach's principle of common cause
- **a language for causal relations:**
  directed acyclic graphs (DAGs), structural equations
- **cornerstone of causal inference:**
  causal Markov condition
- **quantitative causal statements:**
  Pearl's do calculus
- **counterfactual causal statements**

What's wrong with common causal conclusions

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 relations 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)."**

Statistical relations and causal statements...

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

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

  statistical statement:
  can be tested by standard statistical tools

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

  causal statement:
  no standard methods available, this week will give partial answers, don't expect simple ones!

Goal of causal inference...

...in the sense of this lecture:
Predict the effect of interventions without doing them

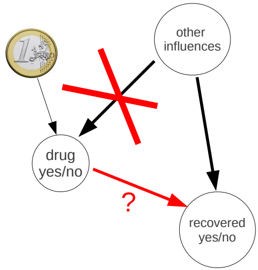
(e.g. what would have happened if someone had changed his/her coffee drinking habits?)

- therefore the lecture is called “Causal inference from *passive* observations”
- statistical evaluation of causal effects of *true* interventions is sometimes also called causal inference, but that’s not what we have in mind

Example for perfect interventions

double-blind randomized medical test

- toss a coin which patient gets a medical drug and which one the placebo
- the decision whether the drug helped is made by a doctor who doesn't know who got the drug



Why interventions may be difficult

- **expensive:**
  test the impact of changing the interest rate
- **unethical:**
  give patients a treatment that is already believed (but not proven) to be bad
- **impossible:**
  move the moon to check whether its really the cause of a solar eclipse

Difficulties in defining interventions

- Assume X is the variable `gross national product`

- what does 'setting X to x' mean?

- changing X is logically impossible without the change of some other variables (e.g., production of companies, consumption of goods)

Is causal inference science at all?

“The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm.”

(Betrand Russell, 1913)

Idea of such a skeptical view

- Interpreting phenomena in nature as causal is just an artefact of our mind

- Physical laws are given by equations that describe relations between observations (e.g. differential equations). Unclear how causal language fits into such concepts.

Our working hypotheses..

- **Causal questions are scientific questions**

  (whether or not a medical drug helps or not is a *scientific* question and definitely an important one)

- **Despite all the difficulties about the philosophical meaning of causality it's possibe to do research on causality**

  (the philosophical interpretation of quantum physics has also caused headache since one century – nevertheless modern technology uses it)

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

Mathematical tools

Measures

A **measure** on the set Ω is a map μ assigning a number to each 'measurable' subset $A \subset \Omega$ such that

- $\mu(A) \in \mathbb{R}_0^+ \cup \infty$
- $\mu(\emptyset) = 0$
- $\mu(\cup_j A_j) = \sum_j \mu(A_j)$ for every countable family of disjoint sets $A_j \subset \Omega$.

(Why 'measurable' instead of general $A \in 2^\Omega$: There are subsets that are so weird that one cannot assign a measure to them. E.g. not all subsets of $[0, 1]$ have a length, see also Banach-Tarski-paradox.)

μ is a **probability measure** if $\mu(\Omega) = 1$

Measure-theoretic integral

There is a precise sense in which every measure μ defines an integral

$$\int f(\omega) d\mu(\omega)\,,$$

for every 'measurable function' f, i.e., function that is sufficiently well-behaved.

Idea: μ defines how much each point is weighted.
(Don't ask: why not weighting each point equally much? This already refers to a measure!)

Examples for two measures on $\mathbb{R}$

- **counting measure on integers:**

$$\nu(A) = \text{ number of integers in } A$$

- **Lebesgue measure:**

$$\lambda(A) := \text{ length of } A$$

Densities

a measure $\tilde{\mu}$ is said to have a density f w.r.t. μ if

$$\tilde{\mu}(A) = \int_A f(\omega) d\mu(\omega)\,,$$

for all measurable A.

Idea: $\tilde{\mu}$ can be obtained from μ by reweighting points via the factor f (not possible if there are sets A with $\mu(A) = 0$ and $\tilde{\mu}(A) \neq 0$).

Examples and counterexamples

- **Gaussian distribution** with expectation μ and standard deviation σ on $\mathbb{R}$ has the density

$$p(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

  w.r.t. the Lebesgue measure

- **counting measure** has no density w.r.t. Lebesgue measure
- **Lebesgue measure** has no density w.r.t. counting measures

Product measure

Let μ_1, μ_2 be measures on Ω_1, Ω_2, respectively. Then

$$(\mu_1 \otimes \mu_2)(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)\,.$$

(Write general $A \subset \Omega_1 \times \Omega_2$ as infinite disjoint union of cartesian products)

Example: Lebegue measure on $\mathbb{R}^2$ (=area) is the product of Lebesgue measure on $\mathbb{R}$ (length)

Notation and terminology

- **Random variables:** denoted by capital letters, e.g., X, Y, Z with ranges $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$
- specific values by $x \in \mathcal{X}, y \in \mathcal{Y}, z \in \mathcal{Z}$

- **vector-valued random variables:** ($=$ sets of scalar random variables) denoted by $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ with values $\mathbf{x}, \mathbf{y}, \mathbf{z}$.

- **functions vs. values of functions:** by $f(X)$ we mean the function $x \mapsto f(x)$

Joint distributions and joint probability densities

- **Probability distribution:** $P(X_1, \ldots, X_n)$ describes probabilities for events like $(X_1, \ldots, X_n) \in A \subset \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$

- **Probability density:** $p(X_1, \ldots, X_n)$ is called the density for $P(X_1, \ldots, X_n)$ if

$$P\{(X_1, \ldots, X_n) \in A\} = \int_A p(x_1, \ldots, x_n) d\mu(x_1, \ldots, x_n),$$

where μ should be clear from the context.

Our two main examples for densities:

- **for continuous variables:**

$$P\{(X_1,\dots,X_n)\in A\} = \int_A p(x_1,\dots,x_n)d^n(x_1,\dots,x_n)\,.$$

(μ is the Lebesgue measure, drop it because this is the usual integral)

- **for discrete variables**

$$P\{(X_1,\dots,X_n)\in A\} = \sum_{(x_1,\dots,x_n)\in A} p(x_1,\dots,x_n)\,.$$

(μ is the counting measure on the discrete set $\mathcal{X}_1 \times \cdots \mathcal{X}_n$. Then p is also called the probability mass function.)

Advantage of the measure theoretic integral

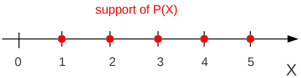
- common framework for discrete and continuous variables
- sums and integrals are both measure theoretic integrals
- part of the variables in $p(x_1, \dots, x_n)$ may be continuous and others discrete. Then we still have

$$P\{(X_1, \dots, X_n) \in A\} = \int_A p(x_1, \dots, x_n) d\mu(x_1, \dots, x_n),$$

  and μ is a tensor product that consists of Lebesgue measures (for the continuous variables) and counting measures (on the discrete ones).

Examples for probability densities: discrete case

Let X attain values in $\{1, \ldots, n\}$ with probability $1/n$ each.



Then

$$p(x) = \begin{cases} 1/n & \text{for } x \in \{1, \ldots, n\} \\ 0 & \text{for } x \in \mathbb{R} \setminus \{1, \ldots, n\} \end{cases}$$

Then,

$$P(A) = \int p(x) d\nu(x) \,,$$

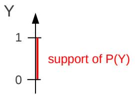
where μ is the counting measure, i.e.,

$$\nu(A) = \text{ number of integers in } A$$

for all measurable subsets A of $\mathbb{R}$.

Examples for probability densities: continuous case

Let Y be uniformly distributed in $[0, 1]$.



Then

$$p(y) = \begin{cases} 1 & \text{for } y \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Then,

$$P(A) = \int p(y) d\lambda(y) \,,$$

where λ is the Lebesgue measure, i.e., $\lambda(A)$ is the length of A. In this case, we often drop λ and write

$$P(A) = \int_A p(y) dy \,.$$

Examples for probability densities: hybrid case

The product density reads

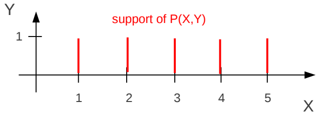
$$p(x,y) = p(x)p(y)\,.$$

Then,

$$P(A) = \int p(x,y)d(\nu \otimes \lambda)(x,y)\,,$$

where $\mu \otimes \lambda$ is the product of counting measure and Lebesgue measure, i.e.,

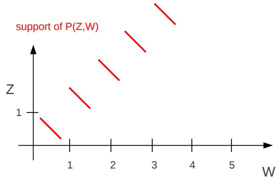
$$\mu(A \times B) = \text{ (number of integers in } A) \;\cdot\; (\text{length of } B)\;.$$



Difficult case

Rotate the distribution $P(X, Y)$:

$$Z \;:=\; \frac{1}{\sqrt{2}}(X+Y), \qquad W := \frac{1}{\sqrt{2}}(X-Y)$$

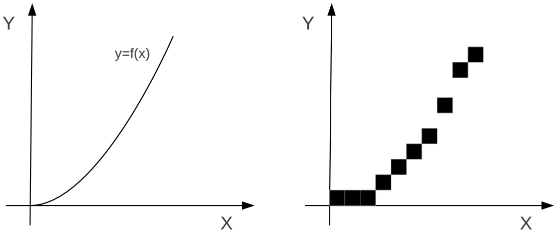


- there is no density w.r.t. any *product* measure
- Z, W are both continuous, but the way they are related is discrete
- for such distributions we avoid using *densities* and describe $P(Z, W)$ in a different way.

Why using continuous variables at all...

...empirical data is always discrete anyway? - Then we don't have all these issues.

Answer: many interesting models contain continuous variables. E.g. discretizations of bijective functions are neither injective not surjective:



$\Rightarrow$ despite all the issues with continuous variables, they are sometimes simpler

Expectation and covariance

- **Expectation:**

$$\mathbb{E}[X] := \int x dP(x) = \int x p(x) d\mu(x)\,.$$

  Note: the probability distribution is also a measure, it therefore also defines an integral!

- **Covariance:**

$$Cov[X, Y] := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- **Variance:**

$$V[X] := Cov[X, X]$$

- **Standard deviation:**

$$\sigma_X := \sqrt{V[X]}$$

  note: σ_X has the same unit as X, while $V[X]$ does not.

Geometric interpretation

- Set of random variables with finite variance is a vector space $\mathcal{V}$
- Variables with zero mean define a subspace $\mathcal{V}_0$
- covariance defines an inner product on $\mathcal{V}_0$
- variance is squared length, standard deviation the length

Covariance matrix

- **Cross covariance matrix:**
  Let $\mathbf{X} = (X_1, \ldots, X_n)$ and $\mathbf{Y} := (Y_1, \ldots, Y_k)$ be vector-valued variables. Then

$$\Sigma_{\mathbf{X},\mathbf{Y}} := (Cov[X_i, Y_j])_{i,j} \,.$$

- **Covariance matrix:**

$$\Sigma_{\mathbf{X}} := \Sigma_{\mathbf{X},\mathbf{X}}$$

Correlation

- **correlation coefficient:**

$$cor[X, Y] := \frac{Cov[X, Y]}{\sigma_X \sigma_Y} \quad \in [-1, 1]$$

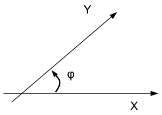
- **interpretation:**
  positive/negative correlation means tha6 large X tend to occur together with large/small Y

$$cor[X, Y] = \pm 1 \quad \Leftrightarrow \quad X = \alpha Y \text{ with } \alpha \neq 0$$

- **geometric picture:**

$$cor[X, Y] = \cos\phi$$

  in the space of centered variables with finite variance



Why the geometric picture helps

Two equivalent formulation of linear regression:

- find $c \in \mathbb{R}$ such that $Y - cX$ has minimal variance
- find $c \in \mathbb{R}$ such that $Y - cX$ and X are uncorrelated

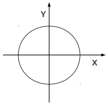
equivalent because orthogonal projection minimizes the distance

Statistical independence

$$X \perp\!\!\!\perp Y \quad :\Leftrightarrow \quad P(X \in A, Y \in B) = P(X \in A)P(X \in B) \quad \forall A, B$$

in terms of densities: $p(X, Y) = p(X)p(Y)$

- **implies uncorrelatedness**, i.e., $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- **uncorrelatedness does not imply independence:**
  Let $P(X, Y)$ be uniform distribution on the circle, i.e., $X^2 + Y^2 = 1$, where $P(X)$ and $P(Y)$ are uniformly distributed on $[-1, 1]$



(uncorrelated because $P(X, Y)$ is symmetric under reflection $X \mapsto -X$)

Statistical independence

- uncorrelated and independent is equivalent for binary variables and for jointly Gaussian variables
- **joint independence:**

$$X_1, \ldots, X_n \text{ jointly ind. } :\Leftrightarrow p(X_1, \ldots, X_n) = p(X_1) \cdots p(X_n)\,.$$

- **conditional independence:** for three sets of variables

$$\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\mathbf{Z} \quad \text{if} \quad p(\mathbf{x}, \mathbf{y}|\mathbf{z}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{y}|\mathbf{z}) \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z}$$

- difficult to test: each $\mathbf{z}$ defines a different distribution

Semi-graphoid axoims

the following rules apply to conditional independence

- **symmetry:**

$$X \perp\!\!\!\perp Y \,|Z \Leftrightarrow Y \perp\!\!\!\perp X \,|Z$$

- **decomposition:**

$$X \perp\!\!\!\perp YW \,|Z \Rightarrow X \perp\!\!\!\perp Y \,|Z$$

- **weak union:**

$$X \perp\!\!\!\perp YW \,|Z \Rightarrow X \perp\!\!\!\perp Y \,|ZW$$

- **contraction:**

$$X \perp\!\!\!\perp Y \,|Z \quad \& \quad X \perp\!\!\!\perp W \,|ZY \Rightarrow X \perp\!\!\!\perp YW \,|Z$$

in distributions with strictly positive density one also has the **intersection property:**

$$X \perp\!\!\!\perp W \,|ZY \quad \& \quad X \perp\!\!\!\perp Y \,|ZW \Rightarrow X \perp\!\!\!\perp YW \,|Z$$

Pearl: Causality, 2000

Notion of a generating set for independences

Given a joint distribution P, a generating set is a list of independences from which all the independences follow that hold for P.

Gaussian variables

- **joint density:** if $\Sigma_{\mathbf{X},\mathbf{X}}$ is invertible, we have

$$p(\mathbf{x}) \sim e^{-\frac{1}{2}(\mathbf{x}-\mu)^t C(\mathbf{x}-\mu)},$$

  where $C := \Sigma_{\mathbf{XX}}^{-1}$ is the concentration matrix and μ is the vector of expectations.

- **conditional distributions:**
  Let $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ and $\mu = (\mu_1, \mu_2)$ and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

  Then $p(\mathbf{X}_1 | \mathbf{x}_2)$ is a Gaussian with mean $\mu_1 + \Sigma_{11}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2)$ and covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

- **conditional indepedence:** can be seen from $\Sigma_{\mathbf{XX}}$ alone

Some information theory

- **joint Shannon entropy of set of random variables:**

$$H(\mathbf{X}) := -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x})$$

(differential entropy for continuous variables $-\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$ has less nice properties)

- **conditional entropy:**

$$H(\mathbf{Y}|\mathbf{X}) = \sum_{x} p(\mathbf{x}) H(\mathbf{Y}|\mathbf{x}) = -\sum_{\mathbf{x}} \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}) \log p(\mathbf{y}|\mathbf{x}) .$$

- **additivity:**

$$H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}|\mathbf{X}) = H(\mathbf{Y}) + H(\mathbf{X}|\mathbf{Y}) .$$

- **mutual information:**

$$I(\mathbf{X} : \mathbf{Y} \,|\mathbf{Z}) := H(\mathbf{X}|\mathbf{Z}) + H(\mathbf{Y}|\mathbf{Z}) - H(\mathbf{X}, \mathbf{Y} \,|\mathbf{Z}) .$$

zero if and only if $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} \,|\mathbf{Z}$.

On the i.i.d. assumption

independently identically distributed
"Let $x_1, \ldots, x_n$ be i.i.d. drawn from $P(X)$" means that every x_j is drawn from the same distribution $P(X)$

- what does this mean?
- when is this justified?
- also applicable to humans although everyone is different?
  E.g., let x_j be the height of the jth person.

When is height of different persons i.i.d.?

Consider two different experiments:

1. On a long hike from Denmark to the South of Italy, measure the height of every person you meet and obtain $x_1, \ldots, x_n$
2. Write all the heights of a small piece of paper, mix all the pieces and draw $x_{\pi(1)}, \ldots, x_{\pi(n)}$.

$x_1, \ldots, x_n$ isn't i.i.d. (people are taller in the North).

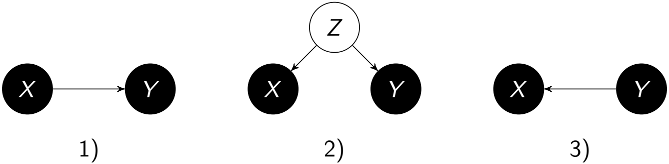
Whether or not some data is i.i.d. is not a property of the world but of the way we acquire the data. Here, the mixing generates the i.i.d. property despite the different races.

de Finetti's theorem: i.i.d. properties come from symmetries of distributions.

First basis for causal conclusions

Reichenbach's principle of common cause (1956)

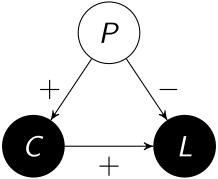
If two variables X and Y are statistically dependent then either



- 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.

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



Simpson's paradox

Fort a certain disease, observe that

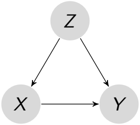
- people taking a certain drug recover less often than the ones that didn't take it (drug seems to hurt instead of helping)

- females taking the drug recover more often than females not taking it (drug seems to help females)
- males taking the drug recover also more often (drug seems to help males)

how can a drug hurt on the average when it helps males and females?

Resolving Simpson's paradox

Z: gender
X: taking the drug or not
Y: recover or not

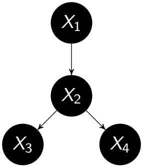


- assume females take the drug more often and recover less often.
- then gender induces a negative correlation between taking and recovery
- negative correlation overcompensates the positive effect of the drug

A Language for causal conclusions

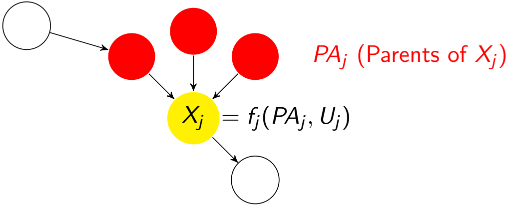
Causal inference problem, general form Spirtes, Glymour, Scheines, Pearl

- 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 U_j



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

The meaning of the DAG and the structural equations

- result of adjusting all parents: setting parents PA_j of X_j to pa_j changes X_j to $x_j = f_j(pa_j, u_j)$.

- result of adjusting a subset of parents: distribution of X_j can be computed from structural equation, details later

- adjusting children of X_j has no effect on X_j

Justification and limits of structural equations

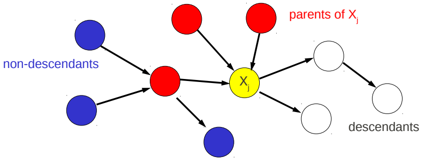
- **independence of noise:**
  if some noise terms $U_1, \ldots, U_k$ were dependent, they had a common cause that needs to occur explicitly in the model

- **determinism:**
  - here we have indeterminism only because we don't know the values of the noise variables
  - inconsistent with modern physics: quantum theory states existence of absolute randomness in microphysics, two identically prepared electrons do not necessarily hit the same point on a screen even if all background conditions are exactly the same

Cornerstone of causal inference: causal Markov condition

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:** If Z d-separates X, Y then $X \perp\!\!\!\perp Y \,|Z$ (definition follows)
- **Factorization:** $p(X_1, \ldots, X_n) = \prod_j p(X_j | PA_j)$ (subject to a technical condition)

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

d-separation

(Pearl 1988)

Path = sequence of pairwise distinct nodes where consecutive ones are adjacent

A path q is said to be **blocked** by the set Z if

- q contains a *chain* $i \rightarrow m \rightarrow j$ or a *fork* $i \leftarrow m \rightarrow j$ such that the middle node is in Z, or
- q contains a *collider* $i \rightarrow m \leftarrow j$ such that the middle node is not in Z and such that no descendant of m is in Z.

Z is said to **d-separate** X and Y in the DAG G, formally

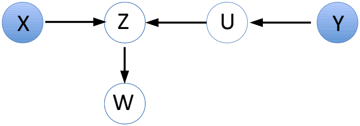
$$(X \perp\!\!\!\perp Y \,|\, Z)_G$$

if Z blocks every path from a node in X to a node in Y.

Example (blocking of paths)

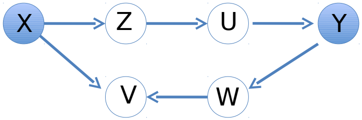
path from X to Y is blocked by conditioning on U or Z or both

Example (unblocking of paths)



- path from X to Y is blocked by $\emptyset$
- unblocked by conditioning on Z or W or both

Example (blocking and unblocking of paths)

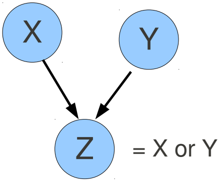


several options for blocking all paths between X and Y:

$$(X \perp\!\!\!\perp Y \,|\, ZW)_G$$
$$(X \perp\!\!\!\perp Y \,|\, ZUW)_G$$
$$(X \perp\!\!\!\perp Y \,|\, VZUW)_G$$

Unblocking by conditioning on common effects

Berkson's paradox (1946), selection bias. Example: X, Y, Z binary



- assume language skils and science skills are independent a priori
- assume pupils go to highschool if they have good skills in science or languages
- then there is a negative correlation between science skills and language skills in high school

Sometimes selection bias cannot be avoided

Hypothetical poll among students in Jyväskylä:

- 'Do you like cultural life in Jyväskylä?' $C =$Yes/No
- 'Do you like the academic programs at the University of Jyväskylä?' $A =$Yes/No

Result: C and A are negatively correlated

Possible explanations

- $C \to A$: Students who enjoy cultural life spend to little time with learning. Then they hate the academic program because they get lost.

- $A \to C$: Students who like the academic program ignore cultural life and therefore underestimate it

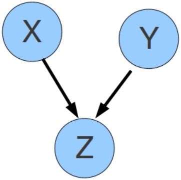
- $A \leftarrow P \to C$: common cause 'Personality' influences both

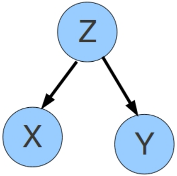
- $A \to S \leftarrow C$: Students who hate both leave Jyväskylä. Therefore our poll describes $P(A, C|S = 1)$ where S labels whether someone stays.

$\Rightarrow$ extend Reichenbach's principle by a fourth alternative: the dataset conditions on a common effect of X and Y without noticing

Asymmetry under inverting arrows

Reichenbach (1956)





$$X \perp\!\!\!\perp Y$$
$$X \not\!\perp\!\!\!\perp Y \,|\, Z$$

$$X \not\!\perp\!\!\!\perp Y$$
$$X \perp\!\!\!\perp Y \,|\, Z$$

Equivalence of Markov cond.: Local $\Rightarrow$ factorization

- Proof by induction. Note the factorization is trivial for $n = 1$.
- Assume that local Markov for $n - 1$ nodes implies

$$p(x_1, \ldots, x_{n-1}) = \prod_{j=1}^{n-1} p(x_j | pa_j) \, .$$

- By local Markov, $X_n \perp\!\!\!\perp ND_n \, | PA_n$. Assume X_n is a terminal node, i.e., it has no descendants, then $ND_n = \{X_1, \ldots, X_{n-1}\}$. Thus

$$X_n \perp\!\!\!\perp \{X_1, \ldots, X_{n-1}\} \, | PA_n$$

and hence the general decomposition

$$p(x_1, \ldots, x_n) = p(x_n | x_1, \ldots, x_{n-1}) p(x_1, \ldots, x_{n-1}).$$

becomes $p(x_1, \ldots, x_n) = p(x_n | pa_n) p(x_1, \ldots, x_{n-1})$.
- By induction, $p(x_1, \ldots, x_n) = \prod_{j=1}^{n} p(x_j | pa_j) \, .$

Equiv: Factorization $\Rightarrow$ global Markov

Need to prove $(X \perp\!\!\!\perp Y \,|\, Z)_G \Rightarrow (X \perp\!\!\!\perp Y \,|\, Z)_p$. Rough idea:

Assume $(X \perp\!\!\!\perp Y \,|\, Z)_G$

- define the smallest subgraph G' containing X, Y, Z and all their ancestors
- consider moral graph G'^m (undirected graph containing the edges of G' and links between all parents)
- use results that relate factorization of probabilities with separation in undirected graphs

Equiv: Global Markov $\Rightarrow$ local Markov

Know that if Z d-separates X, Y, then $X \perp\!\!\!\perp Y \,|Z$.
Need to show that $X_j \perp\!\!\!\perp ND_j \,|PA_j$.

Simply need to show that the parents PA_j d-separate X_j from its non-descendants ND_j:

All paths connecting X_j and ND_j include a $P \in PA_j$, but never as a collider

$$\cdot \to P \leftarrow X_j$$

Hence all paths are chains

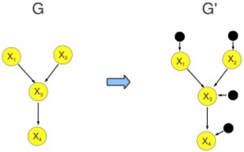
$$\cdot \to P \to X_j$$

or forks

$$\cdot \leftarrow P \to X_j$$

Therefore, the parents block every path between X_j and ND_j.

Functional model $\Rightarrow$ local Markov



- augmented DAG G' contains unobserved noise
- local Markov-condition holds for G':
  - (i): the unexplained noise terms U_j are jointly independent, and thus (unconditionally) independent of their non-descendants
  - (ii): for the X_j, we have

$$X_j \perp\!\!\!\perp ND'_j \,|\, PA'_j$$

    because X_j is a (deterministic) function of PA'_j.
- local Markov in G' implies global Markov in G'
- global Markov in G' implies local Markov in G (proof as last slide)

Exercises

1. **Confounding:** Let X, Y, Z be real-valued variables coupled by the structural equations

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

Find values α, β, γ such that

- X and Y are uncorrelated but X influences Y
- X and Y are positively correlated although X has a negative effect on Y

Prove your claims. 10 credits.

2. **Conditional independences implied by structural equations:**
Let X, Y, Z be related by the structural equations

$$
\begin{aligned}
X &= U_X \\
Y &= f_Y(X) + U_Y \\
Z &= f_Z(Y) + U_Z
\end{aligned}
$$

Show that the joint independence of U_X, U_Y, U_Z implies $X \perp\!\!\!\perp Z \,|\, Y$ without using the equivalence of different Markov conditions. 5 credits.

Exercises

3. Given the causal structure $X \to Y \to Z \to W$. Show that the local Markov condition togther with the semi-graphoid axioms imply

$$X \perp\!\!\!\perp W \,|\, Y\,.$$

5 credits