Introduction to Probabilistic Graphical Models

Christoph Lampert

IST Austria (Institute of Science and Technology Austria)

Excurse: Causality

Correlation is not Causation – Storks and Babies

Storks Deliver Babies ($p = 0.008$)

KEYWORDS:
Teaching;
Correlation;
Significance;
p-values.

Robert Matthews
Aston University, Birmingham, England.
e-mail: rajm@compuserve.com

Summary
This article shows that a highly statistically significant correlation exists between stork populations and human birth rates across Europe. While storks may not deliver babies, unthinking interpretation of correlation and *p*-values can certainly deliver unreliable conclusions.

◆ INTRODUCTION ◆

Introductory statistics textbooks routinely warn of the dangers of confusing correlation with causation, pointing out that while a high correlation coefficient is indicative of (linear) association, it cannot be taken as a measure of causation. Such

association between storks and the concept of women as bringers of life, and also in the bird's feeding habits, which were once regarded as a search for embryonic life in water (Cooper 1992). The legend lives on to this day, with neonate-bearing storks being a regular feature of greetings cards celebrating births.

[Matthews, Robert. "Storks deliver babies (p= 0.008)." Teaching Statistics 22.2 (2000): 36-38.]

Correlation is not Causation – Storks and Babies

Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries

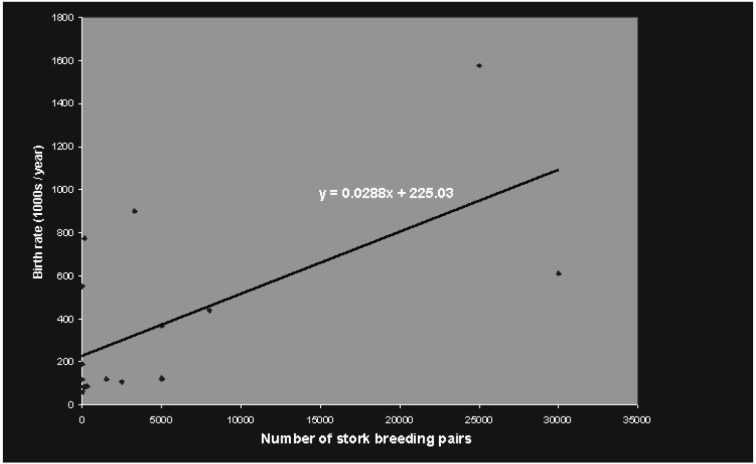


Fig 1. How the number of human births varies with stork populations in 17 European countries.

[Matthews, Robert. "Storks deliver babies (p= 0.008)." Teaching Statistics 22.2 (2000): 36-38.]

Less obvious fallacies (they might not be wrong, just their derivation is)

- Eating red meat causes cancer
- CO_2 deprivation explains near-death experiences
- Women have lower salaries than men
- Immigrants are more often criminals
- Smoking reduces the IQ
- Creative people have more sex
- Happy people are healthier
- Reducing unemployment requires economic growth
- Learning Latin in school helps learning your native language

Less obvious fallacies (they might not be wrong, just their derivation is)

- Eating red meat causes cancer
- CO_2 deprivation explains near-death experiences
- Women have lower salaries than men
- Immigrants are more often criminals
- Smoking reduces the IQ
- Creative people have more sex
- Happy people are healthier
- Reducing unemployment requires economic growth
- Learning Latin in school helps learning your native language

Less obvious fallacies (from the recent news)

JAMA Internal Medicine
Formerly *Archives of Internal Medicine*

April 9, 2012, Vol 172, No. 7 ›

Original Investigation | Apr 9, 2012

Red Meat Consumption and Mortality

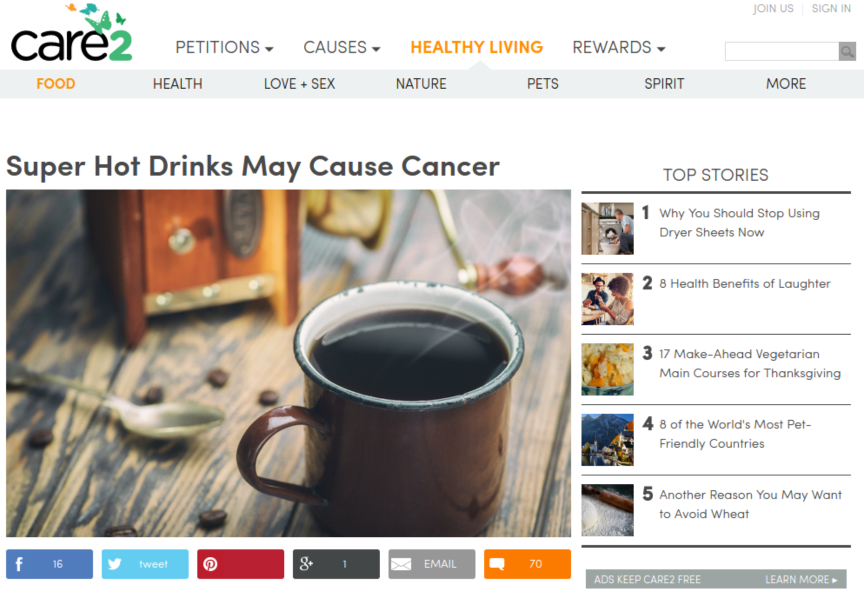
Results From 2 Prospective Cohort Studies FREE

An Pan, PhD; Qi Sun, MD, ScD; Adam M. Bernstein, MD, ScD; Matthias B. Schulze, DrPH; JoAnn E. Manson, MD, DrPH; Meir J. Stampfer, MD, DrPH; Walter C. Willett, MD, DrPH; Frank B. Hu, MD, PhD

[+] Author Affiliations

Arch Intern Med. 2012;172(7):555-563.
doi:10.1001/archinternmed.2011.2287.

Less obvious fallacies (from the recent news)



"Post hoc ergo propter hoc"

Fact: causal effects are time-directed

- I dropped my phone and then the display was cracked.
- I overslept in the morning and then missed the bus.

But: just because something happens after something else doesn't mean one of the cause of the other.

- I dropped my phone and then ran out of data volume.
- I overslept in the morning and then lunch at the restaurant was bad.

"Post hoc ergo propter hoc"

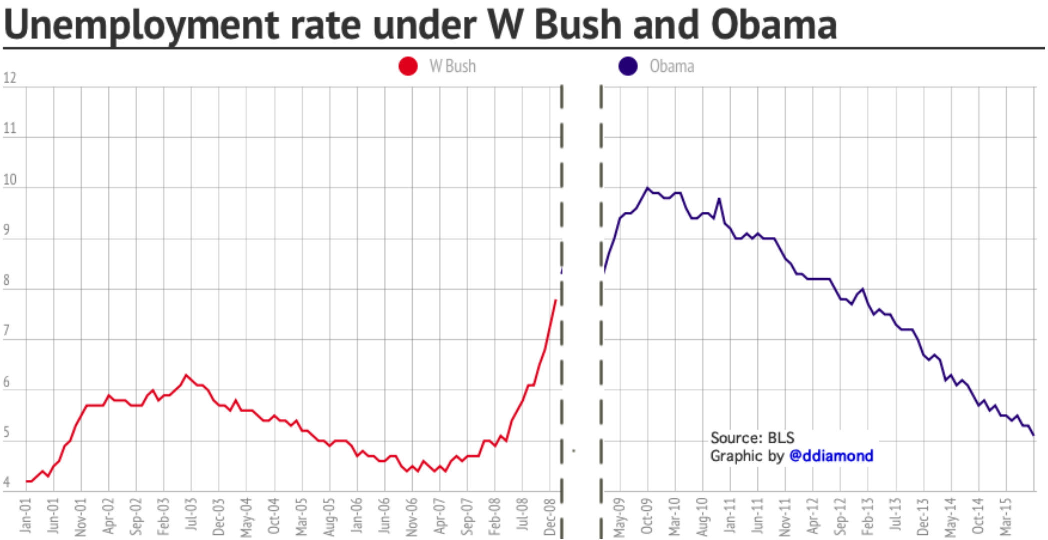
US President Barack Obama Praises Progress On Wages, Employment

BY **AMY NORDRUM** ON 02/05/16 AT 2:05 PM

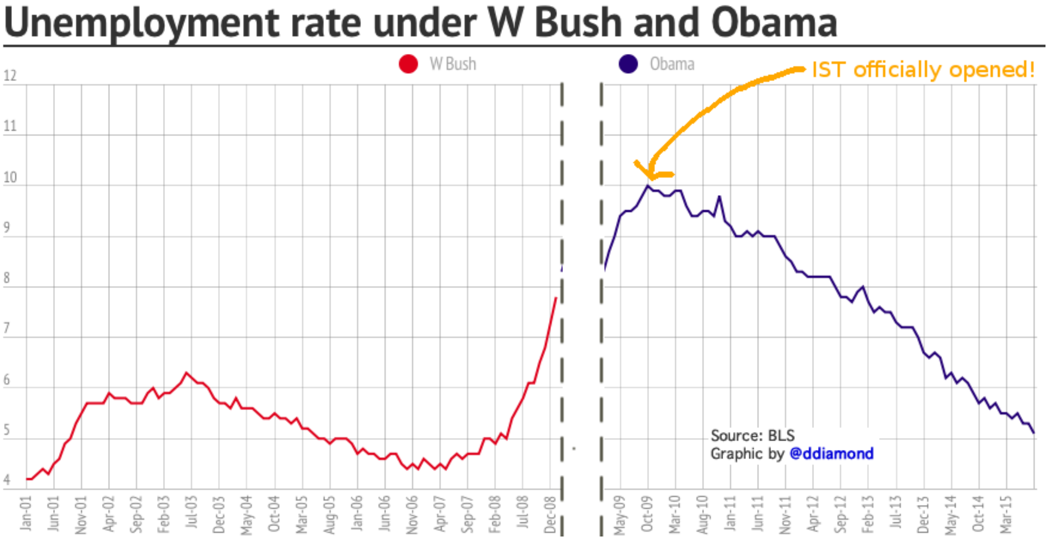
U.S. President Barack Obama discusses the latest unemployment rate within the U.S. economy at the White House, Feb. 5, 2016, in Washington, D.C. Photo: Mark Wilson/Getty Images

U.S. President Barack Obama said he was in a good mood Friday as he praised his administration's progress on the economy after a new jobs report showed wage growth and the lowest rate of unemployment the nation has recorded since 2008.

"Post hoc ergo propter hoc"



"Post hoc ergo propter hoc"



"Post hoc ergo propter hoc" (after this, therefore because of this)

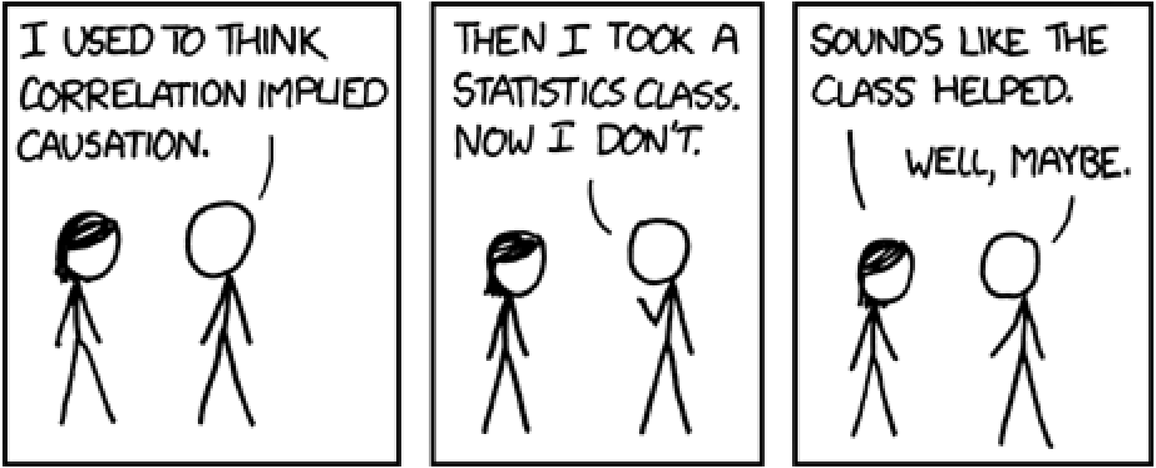


Image: `xkcd.com`

How can we establish a causal relation?

How can we establish a causal relation?

Intervention experiments, e.g. Biology

- observe correlation: bacteria that are resistant against some drug produce protein X, non-resistant bacteria do not.
- hypothesis: X is the cause for the bacteria to be resistant

How can we establish a causal relation?

Intervention experiments, e.g. Biology

- observe correlation: bacteria that are resistant against some drug produce protein X, non-resistant bacteria do not.
- hypothesis: X is the cause for the bacteria to be resistant
- intervention (knock-out): create a mutant without the gene for producing X
- observe: mutant bacteria are not resistant

How can we establish a causal relation?

Intervention experiments, e.g. Biology

- observe correlation: bacteria that are resistant against some drug produce protein X, non-resistant bacteria do not.
- hypothesis: X is the cause for the bacteria to be resistant
- intervention (knock-out): create a mutant without the gene for producing X
- observe: mutant bacteria are not resistant
- intervention (rescue): inject protein X into mutants
- observe: now, mutants are also restistant

Note: we can be pretty sure there is a causal link, even though we don't know if the effect is "direct" or "indirect" (it's not even clear what is meant by that)

Pioneers of Causality Research: Clive W.J. Granger (1934–2009)

Nobelprize.org
The Official Web Site of the Nobel Prize

Video Podcast

Home | Nobel Prizes and Laureates | Nomination | Ceremonies | Alfred Nobel | Educational | Events

Nobel Prizes and Laureates

Prize in Economic S… ‹ 2003 ›

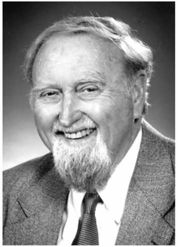
- About the Prize in Economic Sciences 2003
- Robert F. Engle III
- Clive W.J. Granger
  - Facts
  - Biographical
  - Prize Lecture
  - Banquet Speech
  - Prize Presentation
  - Interview
  - Diploma
  - Photo Gallery
  - Other Resources

All Prizes in Economic Sciences
All Nobel Prizes in 2003

The Sveriges Riksbank Prize in Economic Sciences in Memory of Alfred Nobel 2003
Robert F. Engle III, Clive W.J. Granger

Share this:

Clive W.J. Granger - Facts



Clive W.J. Granger

Born: 4 September 1934, Swansea, United Kingdom

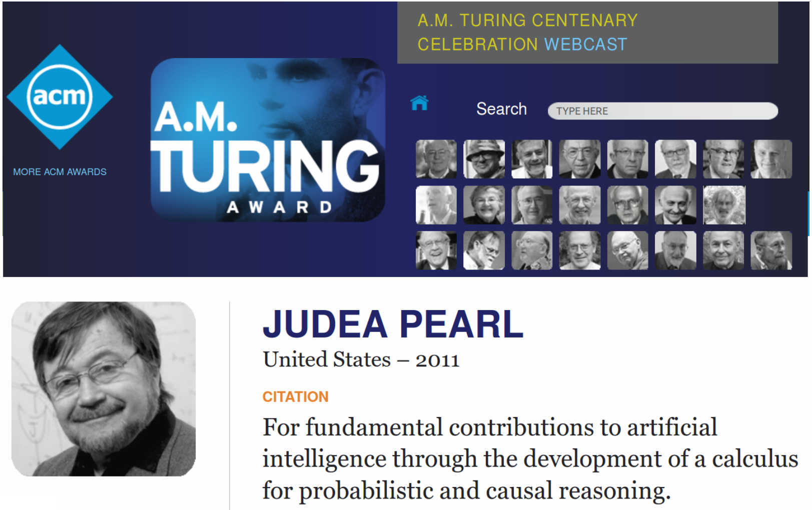
Died: 27 May 2009, San Diego, CA, USA

Affiliation at the time of the award: University of California, San Diego, CA, USA

Prize motivation: "for methods of analyzing economic time series with common trends (cointegration)"

Field: econometrics

Pioneers of Causality Research: Judea Pearl (1936–)



Causal Graph / Causal Bayesian Network

A **causal graph** is a Bayesian network in which arrows indicate **causal relations**

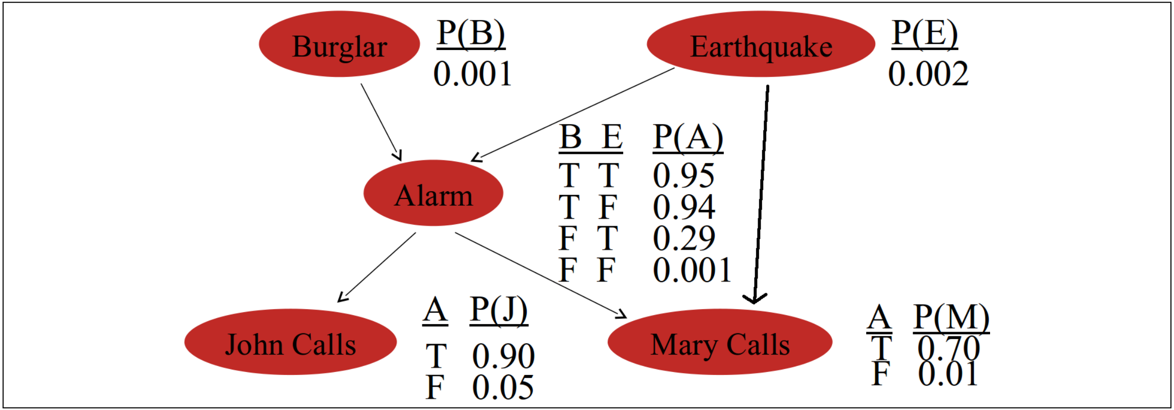


Illustration: adapted from Markus Holzemer "Probabilistic Reasoning"

Causal Graph / Causal Bayesian Network

Equivalent underlying **Bayesian network**, but (some) arrows are not causal

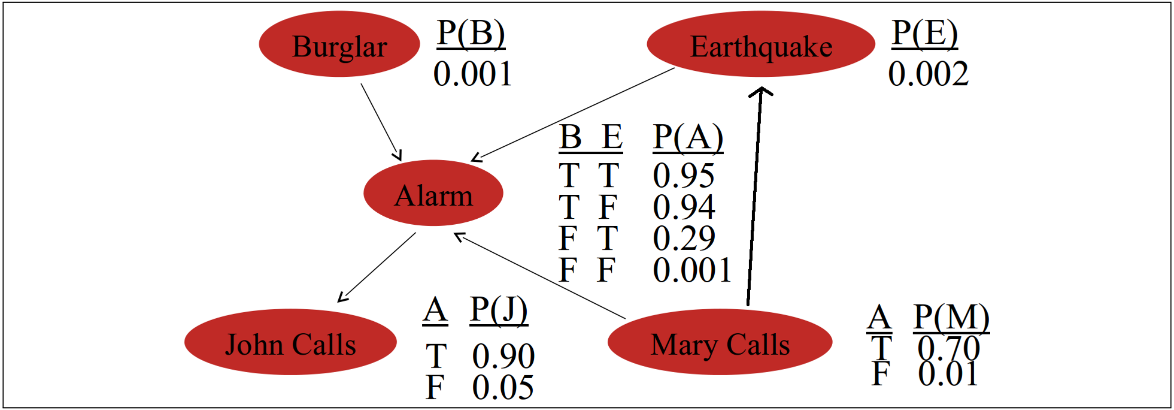


Illustration: adapted from Markus Holzemer "Probabilistic Reasoning"

Causal Inference – do calculus

Probabilistic Inference (purely observational):

- We hear the alarm, what's the probability that Mary calls?

$$Pr(M = \texttt{true} | A = \texttt{true})$$

Causal Inference – do calculus

Probabilistic Inference (purely observational):

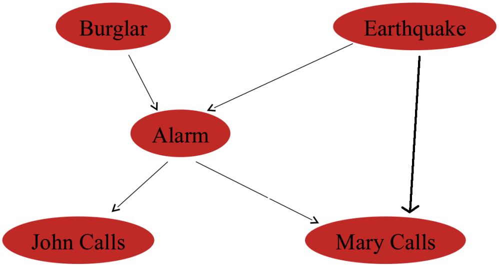
- We hear the alarm, what's the probability that Mary calls?

$$Pr(M = \texttt{true} | A = \texttt{true})$$

Causal Inference:

- We trigger the alarm, what's the probability that Mary calls?

$$Pr(M = \texttt{true} | \ \mathbf{do}(A = \texttt{true})\ )$$



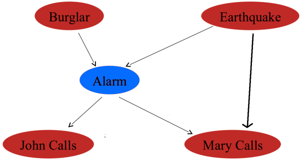
Causal Inference

Mary calls because:

1. she hears the alarm, or
2. she feels the earthquake.

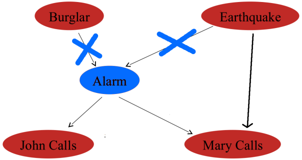
Probabilistic Inference:

- because the alarm rings, the chances of an earthquake higher than normal.

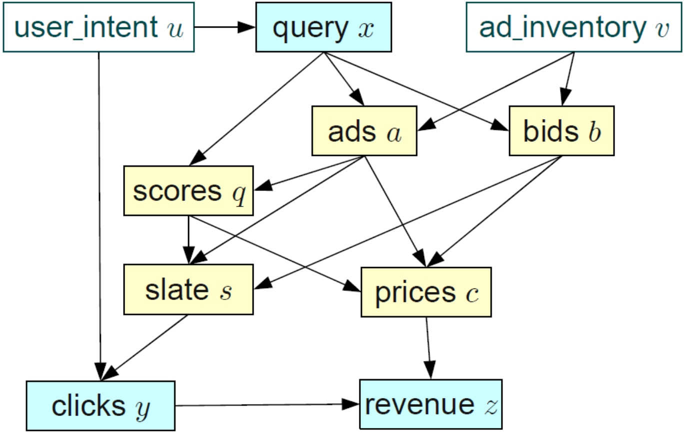


Causal Inference:

- we trigger the alarm ourselves
- the chances of an earthquake are the regular ones.

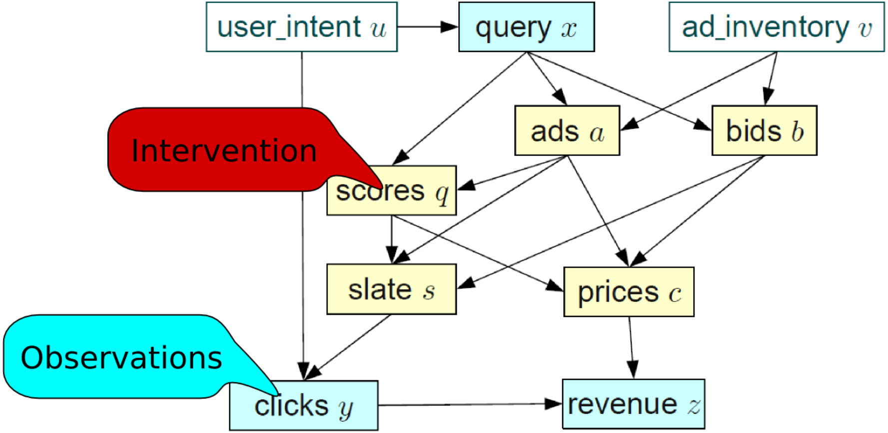


Real world example: ad placement



[Bottou et al, "Counterfactual Reasoning and Learning Systems", JMLR 2013]

Real world example: ad placement

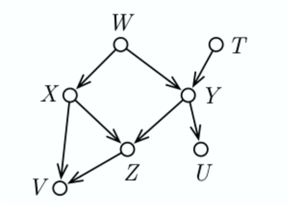


[Bottou et al, "Counterfactual Reasoning and Learning Systems", JMLR 2013]

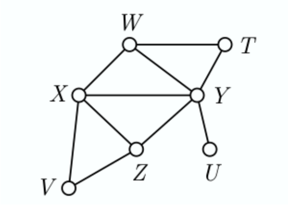
Causality without interventions?

Inductive Causation (IC) algorithm (Verma, Pearl 1990) partially solves the task:

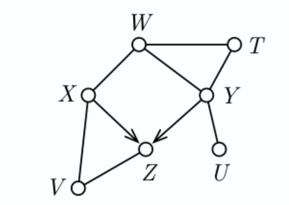
- for any pair of variables, X and Y, identify the smallest set $S_{X,Y}$ such that $X \perp\!\!\!\perp Y | S_{X,Y}$, if any such set exists
- if no such set exists, add a direct connection $X - Y$
- for any substructure $X - Z - Y$, orient it as $X \rightarrow Z \leftarrow Y$ if $Z \notin S_{X,Y}$ ("V-structure")



true causal graph



recovered undirected graph

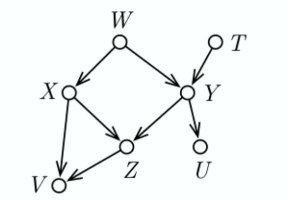


recovered V-structures

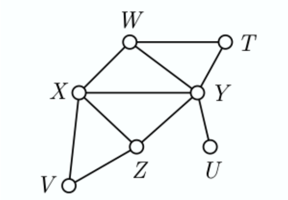
Causality without interventions?

Inductive Causation (IC) algorithm (Verma, Pearl 1990) partially solves the task:

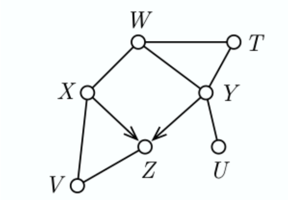
- for any pair of variables, X and Y, identify the smallest set $S_{X,Y}$ such that $X \perp\!\!\!\perp Y | S_{X,Y}$, if any such set exists
- if no such set exists, add a direct connection $X - Y$
- for any substructure $X - Z - Y$, orient it as $X \rightarrow Z \leftarrow Y$ if $Z \notin S_{X,Y}$ ("V-structure")



true causal graph



recovered undirected graph



recovered V-structures

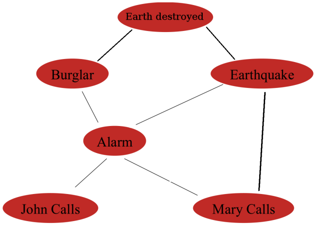
Caveat 1: we assume that some causal structure exists at all

Caveat 2: it's hard to find which nodes are conditional independent given just observations

Causality without interventions?

Build undirected graph based on conditional (in)dependence:

- if we fix a value for 'Earth destroyed' ($\mathtt{D}$), Burglar ($\mathtt{B}$) and Earthquake ($\mathtt{E}$) are **independent**
  $\rightarrow$ no edge,
  $\rightarrow$ memorize $S_{\mathtt{E},\mathtt{B}} = \{\mathtt{D}\}$



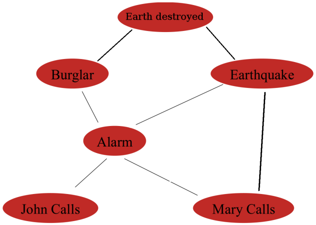
Identify directed V-structures:

- for any unconnected pair X, Y that are both connected to a Z
- if Z is not in $S_{X,Y}$, orient edges $X \rightarrow Z \leftarrow Y$

Causality without interventions?

Build undirected graph based on conditional (in)dependence:

- if we fix a value for 'Earth destroyed' (D), Burglar (B) and Earthquake (E) are **independent**
  $\rightarrow$ no edge,
  $\rightarrow$ memorize $S_{\mathtt{E},\mathtt{B}} = \{\mathtt{D}\}$



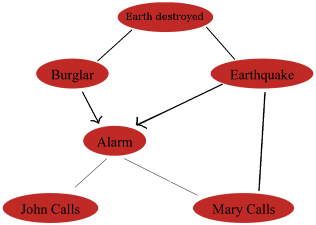
Identify directed V-structures:

- for any unconnected pair X, Y that are both connected to a Z
- if Z is not in $S_{X,Y}$, orient edges $X \rightarrow Z \leftarrow Y$
- *e.g.* Burglar $\rightarrow$ Alarm $\leftarrow$ Earthquake

Causality without interventions?

Build undirected graph based on conditional (in)dependence:

- if we fix a value for 'Earth destroyed' (D), Burglar (B) and Earthquake (E) are **independent**
  $\rightarrow$ no edge,
  $\rightarrow$ memorize $S_{\mathtt{E},\mathtt{B}} = \{\mathtt{D}\}$



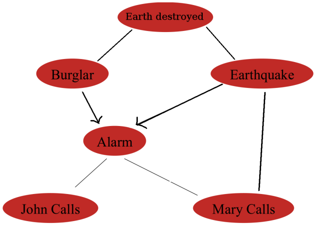
Identify directed V-structures:

- for any unconnected pair X, Y that are both connected to a Z
- if Z is not in $S_{X,Y}$, orient edges $X \rightarrow Z \leftarrow Y$
- *e.g.* Burglar $\rightarrow$ Alarm $\leftarrow$ Earthquake

Causality without interventions?

Build undirected graph based on conditional (in)dependence:

- if we fix a value for 'Earth destroyed' (D), Burglar (B) and Earthquake (E) are **independent**
  $\rightarrow$ no edge,
  $\rightarrow$ memorize $S_{\mathtt{E},\mathtt{B}} = \{\mathtt{D}\}$



Identify directed V-structures:

- for any unconnected pair X, Y that are both connected to a Z
- if Z is not in $S_{X,Y}$, orient edges $X \rightarrow Z \leftarrow Y$
- *e.g.* Burglar $\rightarrow$ Alarm $\leftarrow$ Earthquake
- but not Burglar $\rightarrow$ Earth-destroyed $\leftarrow$ Earthquake

For many edges, we can't decide!

How to Identify Conditional Independence from Observations?

We observe data for three random variables, X, Y and Z. How to tell if $X \perp\!\!\!\perp Y | Z$?

We need to find out if $p(x, y|z) \stackrel{?}{=} p(x|z)p(y|z)$ for every $z \in \mathcal{Z}$

Observation: already without the Z, things are hard: is $p(x, y) \stackrel{?}{=} p(x)p(y)$?

- when we compute an estimate $\hat{p}(x, y)$, from data, this relation will not be fulfilled exactly, e.g.
  - assume $X \perp\!\!\!\perp Y$, with $p(x) = p(y) = 0.5$
  - n Observations: $(0, 0), (0, 1), (0, 0), (1, 1), (0, 0)$
  - $\hat{p}(X = 0) = 0.8 \quad \hat{p}(Y = 0) = 0.6$
  - $\hat{p}(X = 0, Y = 0) = 0.6 \quad \hat{p}(X = 0)\hat{p}(Y = 0) = 0.48$
  - $\hat{p}(X = 1, Y = 0) = 0 \quad \hat{p}(X = 1)\hat{p}(Y = 0) = 0.12$

One would hope that the difference shrinks when $n \to \infty$, but how to measure?

We need of quantitative measure of how much $p(x, y)$ differs from $p(x)p(y)$.

Kullback-Leibler Divergence

Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence between two discrete distribution p and q over t is

$$\mathrm{KL}(p||q) = \sum_i p(t) \log \frac{p(t)}{q(t)}$$

and for continuous distribution with probability densities p and q:

$$\mathrm{KL}(p||q) = \int_t p(t) \log \frac{p(t)}{q(t)}$$

(can be ∞, if $q(t) = 0$ where $p(t) \neq 0$)

One can show that KL divergence is the only measure of difference between probability distributions that satisfies some desirable properties in relation to the entropy (see Wikipedia).

Mutual Information

Observation: both $p(x, y)$ and $p(x)p(y)$ are distributions over (x, y):

Mutual Information

The mutual information between two random variables X, Y is defined as

$$I(X; Y) = \mathrm{KL}(\, p(X, Y) || p(X)p(Y)\,)$$

The mutual information has some nice properties

- $I(X, Y) \geq 0$ positivity
- $I(X, Y) \geq 0$ symmetry
- $I(X, Y) = 0$ if and only if $X \perp\!\!\!\perp Y$

It also has some not so nice properties:

- it's difficult to estimate from finite data

Conditional Mutual Information

For any z, $p(x, y|z)$ and $p(x|z)p(y|z)$ are distributions over (x, y):

Conditional Mutual Information

The mutual information between two random variables X, Y is defined as

$$I(X; Y|Z) = \mathbb{E}_{z\sim Z}\Big[\mathrm{KL}(\, p(X, Y|Z = z)||p(X|Z = z)p(Y|Z = z)\,)\Big]$$

The nice properties of the mutual information still hold

- $I(X, Y|Z) \geq 0$ positivity
- $I(X, Y|Z) \geq 0$ symmetry
- $I(X, Y|Z) = 0$ if and only if $X \perp\!\!\!\perp Y|Z$ almost surely

But the not so nice ones as well:

- it's difficult to estimate from finite data

Estimating (Conditional) Mutual Information in Practice

Finite $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$

- given n observations, compute estimate $\hat{p}(x, y, z)$
- for $n \to \infty$, plug-in estimated mutual information will converge to the true one

Continuous, real-valued $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$

Idea 1: discretize

- quantize $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ into finitely many bins
- use discrete estimate as above
- beware: for $n \to \infty$, bins must shrink (and some assumptions must hold)

Idea 2: hope for some functional relation, e.g.

- use to the observation to learn two functions, $f : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ and $g : \mathcal{X} \to \mathcal{Y}$,
- compare how much f fits the data better than g.

Active Research

The Annals of Statistics
2013, Vol. 41, No. 2, 436–463
DOI: 10.1214/12-AOS1080

GEOMETRY OF THE FAITHFULNESS ASSUMPTION IN CAUSAL INFERENCE[1]

BY CAROLINE UHLER, GARVESH RASKUTTI, PETER BÜHLMANN AND BIN YU

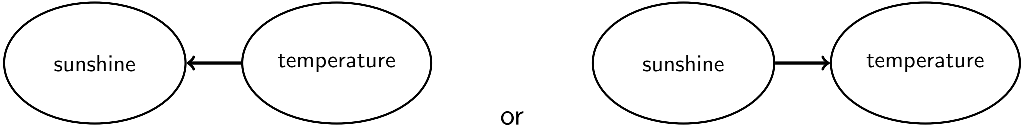
IST Austria, SAMSI, ETH Zürich and University of California, Berkeley

Many algorithms for inferring causality rely heavily on the faithfulness assumption. The main justification for imposing this assumption is that the set of unfaithful distributions has Lebesgue measure zero, since it can be seen as a collection of hypersurfaces in a hypercube. However, due to sampling error the faithfulness condition alone is not sufficient for statistical estimation, and strong-faithfulness has been proposed and assumed to achieve uniform or high-dimensional consistency. In contrast to the plain faithfulness assumption, the set of distributions that is not strong-faithful has nonzero Lebesgue measure and in fact, can be surprisingly large as we show in this paper. We study the strong-faithfulness condition from a geometric and combinatorial point of view and give upper and lower bounds on the Lebesgue measure of strong-faithful distributions for various classes of directed acyclic graphs. Our results imply fundamental limitations for the PC-algorithm and potentially also for other algorithms based on partial correlation testing in the Gaussian case.

Limitations

Problem: many undecidable cases.

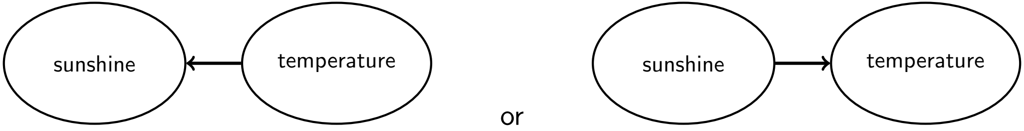
Embarrassing fact: we can't even handle the "easiest possible case": two variables



Limitations

Problem: many undecidable cases.

Embarrassing fact: we can't even handle the "easiest possible case": two variables



Impossible to decide based on just conditional independence.

We need introduce additional assumptions, *e.g.* what is "normal"?

Causality from noise [J. Peters, J. Mooij, B. Schölkopf, 2010s]

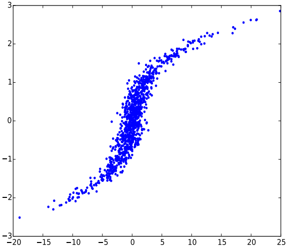
- two random variables, X and Y (*e.g.* sunshine, temperature)
- one causes the other as $Y = f(X) +$ 'noise'
- noise contribution independent of input X
- we observe pairs $(X_1, Y_1), \dots, (X_n, Y_n)$

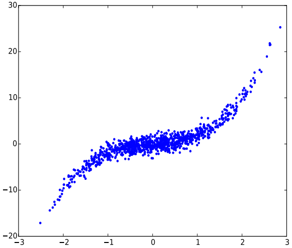
Causality from noise [J. Peters, J. Mooij, B. Schölkopf, 2010s]

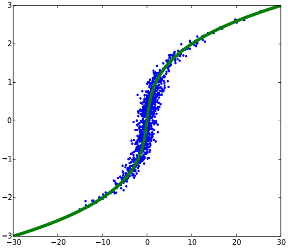
- two random variables, X and Y (*e.g.* sunshine, temperature)
- one causes the other as $Y = f(X) +$ 'noise'
- noise contribution independent of input X
- we observe pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$

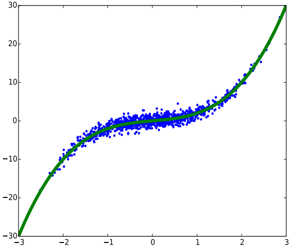
Algorithm:

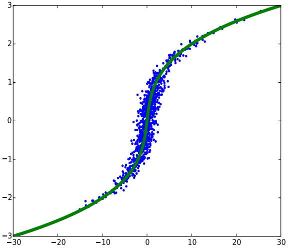
- estimate functions in both directions: g_1, g_2, such that
  $$g_1(X_i) \approx Y_i \qquad \text{and} \qquad g_2(Y_i) \approx X_i$$
- analyze distribution of "noise",
  $$g_1(X_i) - Y_i \qquad \text{and} \qquad g_2(Y_i) - X_i$$
- pick direction in which noise is more independent of input

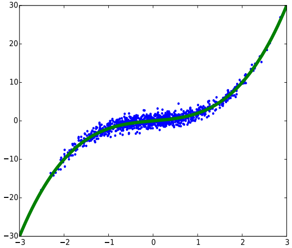

3
2
1
0
−1
−2
−3
−20 −15 −10 −5 0 5 10 15 20 25

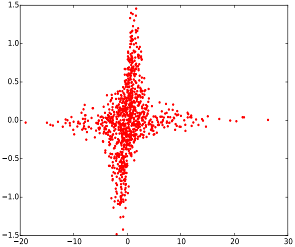

30
20
10
0
−10
−20
−3 −2 −1 0 1 2 3

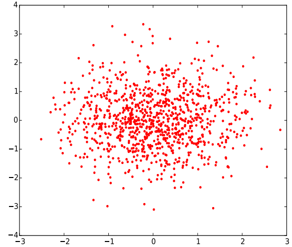

3
2
1
0
−1
−2
−3
−30
−20
−10
0
10
20
30


30
20
10
0
−10
−20
−30
−3
−2
−1
0
1
2
3









unlikely noise distribution ($\mathcal{X} \perp\!\!\!\perp \mathcal{Y}$)

likely noise distribution ($\mathcal{X} \not\perp\!\!\!\perp \mathcal{Y}$)

Summary

- causality is actively researched in machine learning and statistics
- so far, computers are even worse at causal inference than people
- many open challenges, e.g. causality from single examples

Summary

- causality is actively researched in machine learning and statistics
- so far, computers are even worse at causal inference than people
- many open challenges, e.g. causality from single examples

vs.

Images. left: all over the Internet, right: Ritter Sport