But, why?
an introduction to causal inference from observational data

Jilles Vreeken

13 September 2019
Questions of the day

What is *causation*, how can we *measure* it, and how can we *discover* it?
Causation

‘the relationship between something that happens or exists and the thing that causes it’
Correlation vs. Causation

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.

---

<table>
<thead>
<tr>
<th>Country</th>
<th>Area (km²)</th>
<th>Storks (pairs)</th>
<th>Humans ($10^6$)</th>
<th>Birth rate ($10^3$/yr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albania</td>
<td>28,750</td>
<td>100</td>
<td>3.2</td>
<td>83</td>
</tr>
<tr>
<td>Austria</td>
<td>83,860</td>
<td>300</td>
<td>7.6</td>
<td>87</td>
</tr>
<tr>
<td>Belgium</td>
<td>30,520</td>
<td>1</td>
<td>9.9</td>
<td>118</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>111,000</td>
<td>5000</td>
<td>9.0</td>
<td>117</td>
</tr>
<tr>
<td>Denmark</td>
<td>43,100</td>
<td>9</td>
<td>5.1</td>
<td>59</td>
</tr>
<tr>
<td>France</td>
<td>544,000</td>
<td>140</td>
<td>56</td>
<td>774</td>
</tr>
<tr>
<td>Germany</td>
<td>357,000</td>
<td>3300</td>
<td>78</td>
<td>901</td>
</tr>
<tr>
<td>Greece</td>
<td>132,000</td>
<td>2500</td>
<td>10</td>
<td>106</td>
</tr>
<tr>
<td>Holland</td>
<td>41,900</td>
<td>4</td>
<td>15</td>
<td>188</td>
</tr>
<tr>
<td>Hungary</td>
<td>93,000</td>
<td>5000</td>
<td>11</td>
<td>124</td>
</tr>
<tr>
<td>Italy</td>
<td>301,280</td>
<td>5</td>
<td>57</td>
<td>551</td>
</tr>
<tr>
<td>Poland</td>
<td>312,680</td>
<td>30,000</td>
<td>38</td>
<td>610</td>
</tr>
<tr>
<td>Portugal</td>
<td>92,390</td>
<td>1500</td>
<td>10</td>
<td>120</td>
</tr>
<tr>
<td>Romania</td>
<td>237,500</td>
<td>5000</td>
<td>23</td>
<td>367</td>
</tr>
<tr>
<td>Spain</td>
<td>504,750</td>
<td>8000</td>
<td>39</td>
<td>439</td>
</tr>
<tr>
<td>Switzerland</td>
<td>41,290</td>
<td>150</td>
<td>6.7</td>
<td>82</td>
</tr>
<tr>
<td>Turkey</td>
<td>779,450</td>
<td>25,000</td>
<td>56</td>
<td>1576</td>
</tr>
</tbody>
</table>

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

**Correlation does not tell us anything about causality.**  
Instead, we should talk about dependence.
Dependence vs. Causation
What is causal inference?

‘reasoning to the conclusion that something is, or is likely to be, the cause of something else’

Godzillian different definitions of ‘cause’ and ‘effect’
- equally many inference frameworks
- all require (strong) assumptions
- many highly specific
Causal Inference
Naïve approach

If
\[ p(cause)p(\text{effect} \mid cause) > p(\text{effect})p(\text{cause} \mid \text{effect}) \]
then \text{cause} \rightarrow \text{effect}
Naïve approach fails

If

$$p(cause)p(effect \mid cause) = p(effect)p(cause \mid effect)$$

then cause → effect

Both are equal as they are simply factorizations of $p(cause, effect)$
Naïve approach

If

\[ p(cause)p(effect | cause) > p(effect)p(cause | effect) \]

then \( cause \rightarrow effect \)
Naïve approach fails

If

\[ p(cause)p(effect | cause) \not\leftrightarrow p(effect)p(cause | effect) \]

then \( cause \rightarrow effect \)

Depends on distribution and domain size of data, not on causal effect.
Naïve approach

If

\[ p(cause)p(\text{effect} | \text{cause}) > p(\text{effect})p(\text{cause} | \text{effect}) \]

then \textit{cause} \rightarrow \textit{effect}
Naïve approach fails

\[
\frac{p(\text{effect} \mid \text{cause})}{p(\text{effect})} > \frac{p(\text{cause} \mid \text{effect})}{p(\text{cause})} \\
\text{then } \text{cause} \rightarrow \text{effect}
\]

But do we know for sure that the lhs is higher when when \text{cause} \rightarrow \text{effect} ?
What about differences in domain sizes, complexities of distributions, etc
Randomized controlled trials are the de-facto standard for determining whether $X$ causes $Y$

- treatment $X \in \{0,1, \ldots \}$, potential effect $Y$ and co-variates $Z$

Simply put, we

1. gather a large population of test subjects
2. randomly split the population into two equally sized groups $A$ and $B$, making sure that $Z$ is equally distributed between $A$ and $B$
3. apply treatment $X = 0$ to group $A$, and treatment $X = 1$ to group $B$
4. determine whether $Y$ and $X$ are dependent

If $Y \not\perp X$, we conclude that $X$ causes $Y$
The Ultimate Test

Randomized controlled trials are the de-facto standard for determining whether \( X \) causes \( Y \).

- treatment \( X \in \{0, 1, \ldots \} \), potential effect \( Y \) and co-variates \( Z \)

Simply put, we

1. gather a large population of test subjects
2. randomly split the population into two equally sized groups \( A \) and \( B \), making sure that \( Z \) is equally distributed between \( A \) and \( B \)
3. apply treatment \( X = 0 \) to group \( A \), and treatment \( X = 1 \) to group \( B \)
4. determine whether \( Y \) and \( X \) are dependent

If \( Y \perp X \), we conclude that \( X \) causes \( Y \)

Ultimate, but not ideal

- Often impossible or unethical
- Large populations needed
- Difficult to control for \( Z \)
Do, or do not

**Observational** $p(y \mid x)$
- distribution of $Y$ given that we **observe** variable $X$ takes value $x$
- what we usually estimate, e.g. in regression or classification
- **can be inferred from data** using Bayes’ rule $p(y \mid x) = \frac{p(x,y)}{p(x)}$

**Interventional** $p(y \mid do(x))$
- distribution of $Y$ given that we **set** the value of variable $X$ to $x$
- describes the distribution of $Y$ we would observe if we would **intervene** by artificially forcing $X$ to take value $x$, but otherwise use the original **data generating process** ($\neq p(x, y, \ldots)$!)
- the conditional distribution of $Y$ we would get through a randomized control trial!

(Pearl, 1982) 16
Same old, same old?

In general, \( p(y \mid do(x)) \) and \( p(y \mid x) \) are not the same

Let’s consider my espresso machine

- \( y \) actual pressure in the boiler
- \( x \) pressure measured by front gauge

Now,

- if the barometer works well, \( p(y \mid x) \) will be unimodal around \( x \)
- intervening on the barometer, e.g. moving its needle up or down, however, has no effect on the actual pressure and hence, \( p(\ y \mid do(x)) = p(y) \)
What do you want?

Before we go into a lot more detail, what do we want?

If you just want to predict, $p(y \mid x)$ is great
- e.g. when ‘interpolating’ $Y$ between its cause and its effects is fine
- also, boring, because lots of cool methods exist

If you want to act on $x$, you really want $p(y \mid do(x))$
- for example, for drug administration, or discovery
- also, exciting, not so many methods exist
Observational Data

Even if we cannot directly access $p(y \mid do(x))$ e.g. through randomized trials, it does exist.

The main point of causal inference and do-calculus is:

*If we cannot measure $p(y \mid do(x))$ directly in a randomized trial, can we estimate it based on data we observed outside of a controlled experiment?*
Standard learning setup

Training data

Trained model

Observable joint

Observable conditional

\[ q(y \mid x; \theta) \]

\[ p(y \mid x) \]
Causal learning setup

Observable joint

\[ p(y \mid x) \]

Observational conditional

Intervention joint

\[ p(y \mid \text{do}(x)) \]

Intervention conditional

Intervention joint

\[ q(y \mid x; \theta) \]

Intervention model

Training data

\[ ? \]
Causal learning goal

\[ p(y \mid x) \approx \hat{p}(y \mid do(x)) \approx p(y \mid do(x)) \]
If through do-calculus we can derive an equivalent of \( \tilde{p}(y \mid do(x)) \) without any do's, we can estimate it from observational data alone and call \( \tilde{p}(y \mid do(x)) \) identifiable.
Causal Discovery
Causal Discovery
Choices...

For these three, $X \nparallel Z$, and $X \parallel Z \mid Y$ holds

For this one, $X \parallel Z$, and $X \nparallel Z \mid Y$ holds
Reichenbach’s common cause principle links causality and probability

If $X$ and $Y$ are statistically dependent then either

When $Z$ screens $X$ and $Y$ from each other, given $Z$, $X$ and $Y$ become independent.
Causal Markov Condition

Any distribution generated by a Markovian model $M$ can be factorized as

$$p(X_1, X_2, ..., X_n) = \prod_i p(X_i \mid pa_i)$$

where $X_1, X_2, ..., X_n$ are the endogenous variables in $M$, and $pa_i$ are (values of) the endogenous “parents” of $X_i$ in the causal diagram associated with $M$

(Spirtes, Glymour, Scheines 1982; Pearl 2009)
Types of Nodes

Non-descendants of $W$

Parents of $T$

Descendants of $W$
Causal Discovery

Exogenous variables

Endogenous variables

X

Y

Z

T

Q

W

S

V
Causal Discovery

Exogenous variable: A factor in a causal model that is not determined by other variables in the system.

Endogenous variable: A factor in a causal model that is determined by other variables in the system.

Endogenous variables

Exogenous variables
Causal Markov Condition

Any distribution generated by a Markovian model $M$ can be factorized as

$$p(X_1, X_2, ..., X_n) = \prod_i p(X_i | pa_i)$$

where $X_1, X_2, ..., X_n$ are the endogenous variables in $M$, and $pa_i$ are (values of) the endogenous “parents” of $X_i$ in the causal diagram associated with $M$.

(Spirtes, Glymour, Scheines 1982; Pearl 2009)
In other words...

For all distinct variables $X$ and $Y$ in the variable set $V$, if $X$ does not cause $Y$, then $P(X \mid Y, pa_X) = P(X \mid pa_X)$

That is, we can **weed out** edges from a causal graph – we can identify DAGs **up to** Markov equivalence classes.

Which is great, although we are **unable** to choose among these
The PC algorithm is one of the most well-known, and most relied upon causal discovery algorithms
  - proposed by Peter Spirtes and Clark Glymour

Assumes the following
1) data-generating distribution has the causal Markov property on graph $G$
2) data generating distribution is faithful to $G$
3) every member of the population has the same distribution
4) all relevant variables are in $G$
5) there is only one graph $G$ to which the distribution is faithful
The PC algorithm is one of the most well-known, and most relied upon causal discovery algorithms

- proposed by Peter Spirtes and Clark Glymour

Two main steps

1) use conditional independence tests to determine the undirected causal graph (aka the skeleton)
2) apply constraint-based rules to direct (some) edges
Step 1: Discover the Skeleton

\[ \text{for } k = 0 \text{ to } n \]
\[ \text{for all } X, Y \in V \text{ with } (x, y) \in E \]
\[ \text{for all } A \subseteq V \text{ of } k \text{ nodes with } (x, a), (y, a) \in E \]
\[ \text{if } X \cap Y \neq A \]
\[ \text{remove } (x, y) \text{ from } E \]
Step 1: Discover the Skeleton

\[ X \perp Z \mid Y \implies \text{no causal edge} \]

**Algorithm:**

\[
\text{for } k = 0 \text{ to } n \\
\text{for all } X, Y \in V \text{ with } (x, y) \in E \\
\text{for all } A \subseteq V \text{ of } k \text{ nodes with } (x, a), (y, a) \in E \\
\text{if } X \perp Y \mid A \\
\text{remove } (x, y) \text{ from } E
\]
Step 1: Discover the Skeleton

\textbf{for } k = 0 \textit{ to } n \\
\textbf{for } all \ X, Y \in V \textit{ with } (x, y) \in E \\
\textbf{for } all \ A \subseteq V \textit{ of } k \textit{ nodes with } (x, a), (y, a) \in E \\
\textbf{if } X \parallel Y \mid A \\
\textbf{remove } (x, y) \textbf{ from } E
Step 1: Discover the Skeleton

We now have the **causal skeleton**
Step 2: Orientation

We now identify all **colliders** $X \rightarrow Y \leftarrow Z$ considering all relevant pairs once.
Step 2: Orientation

We now identify all *colliders* $X \rightarrow Y \leftarrow Z$ considering all relevant pairs *once*
We then iteratively apply Rules 2—3 until we cannot orient any more edges.
Causal Inference
Causal Inference

We can find the causal skeleton using conditional independence tests.
Causal Inference

We can find the causal skeleton using conditional independence tests, but only few edge directions.
Causal Inference

We can find the causal skeleton using conditional independence tests, but only few edge directions.
Three is a crowd

Traditional causal inference methods rely on **conditional independence tests** and hence require *at least three* observed variables.

That is, they *cannot* distinguish between $X \to Y$ and $Y \to X$ as $p(x)p(y \mid x) = p(y)p(x \mid y)$ are just factorisations of $p(x, y)$.

Can we infer the causal direction between pairs?
Let’s take another look at the definition of causality.

‘the relationship between something that happens or exists and the thing that causes it’

From the do-calculus it follows that if $X$ cause $Y$, we can wiggle $Y$ by wiggling $X$, while when we cannot wiggle $X$ by wiggling $Y$.

But… we only have observational data jointly over $X, Y$, and cannot do any wiggling ourselves...
May The Noise Be With You

\[ y \text{-value with large } H(X \mid y) \text{ and large density } p(y) \]
May The Noise Be With You

“If the **structure** of density of $p(x)$ is not correlated with the slope of $f$, then the flat regions of $f$ induce peaks in $p(y)$.

The causal hypothesis $Y \rightarrow X$ is thus implausible because the causal mechanism $f^{-1}$ appears to be adjusted to the “input” distribution $p(y)$.”
Independence of Input and Mechanism

If $X$ causes $Y$,
the marginal distribution of the cause, $p(X)$
and the conditional distribution of
the effect given the cause, $p(Y|X)$
are independent

That is, if $X \rightarrow Y$
$p(X)$ contains no information about $p(Y|X)$

(Sgouritsa et al 2015)
Additive Noise Models

Whenever the joint distribution $p(X, Y)$ admits a model in one direction, i.e. there exists an $f$ and $N$ such that

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

but does not admit the reversed model, i.e. for all $g$ and $\tilde{N}$ we have

$$X = g(Y) + \tilde{N} \text{ with } \tilde{N} \perp\!\!\!\!\perp Y$$

We can infer $X \rightarrow Y$

ANMs and Identifiability

When are ANMs identifiable?

- what do we need to assume about the data generating process for ANM-based inference to make sense?
- for which functions $f$ and what noise distributions $\mathcal{N}$ are ANMs identifiable from observational data?

- Linear functions and Gaussian noise
- Linear functions and non-Gaussian noise
- For most cases of non-linear functions and any noise

Additive Noise Models

Whenever the joint distribution $p(X, Y)$ admits a model in one direction, i.e. there exists an $f$ and $N$ such that

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

but does not admit the reversed model, i.e. for all $g$ and $\tilde{N}$ we have

$$X = g(Y) + \tilde{N} \text{ with } \tilde{N} \perp\!\!\!\!\!\!\perp Y$$

How do we determine or use this in practice?

Independence of Input and Mechanism

If $X$ causes $Y$, the \textbf{marginal distribution} of the \textit{cause}, $p(X)$ and the \textbf{conditional distribution} of the \textit{effect given the cause}, $p(Y|X)$ are \textit{independent}.

That is, if $X \rightarrow Y$ \textit{$p(X)$ contains no information} about $p(Y|X)$.
Plausible Markov Kernels

In other words, if we observe that

\[ p(cause)p(effect \mid cause) \]

is simpler than

\[ p(effect)p(cause \mid effect) \]

then it is likely that \textit{cause} → \textit{effect}

How to robustly measure ‘simpler’?

(Sun et al. 2006, Janzing et al. 2012)
Kolmogorov Complexity

\[ K(s) \]

The Kolmogorov complexity of a binary string \( s \) is the length of the shortest program \( p^* \) for a universal Turing Machine \( U \) that generates \( s \) and halts.
Algorithmic Markov Condition

If $X \rightarrow Y$, we have, up to an additive constant,

$$K(p(X)) + K(p(Y|X)) \leq K(p(Y)) + K(p(X|Y))$$

That is, we can do **causal inference** by identifying the factorization of the joint with the **lowest Kolmogorov complexity**

(Janzing & Schölkopf, IEEE TIT 2012)
Univariate and Numeric
Two-Part MDL

The Minimum Description Length (MDL) principle

given a model class $\mathcal{M}$, the best model $M \in \mathcal{M}$ is the $M$ that minimises

$$L(M) + L(D \mid M)$$

in which

$L(M)$ is the length, in bits, of the description of $M$

$L(D \mid M)$ is the length, in bits, of the description of the data when encoded using $M$

(see, e.g., Rissanen 1978, 1983, Grünwald, 2007)
MDL and Regression

\[ L(M) + L(D|M) \]

\[ a_1 x + a_0 \]  
errors

\[ a_{10} x^{10} + a_9 x^9 + \ldots + a_0 \]  
\{
\}

VS.
We model $Y$ as

$$Y = f(X) + \mathcal{N}$$

As $f$ we consider linear, quadratic, cubic, exponential, and reciprocal functions, and model the noise using a 0-mean Gaussian. We choose the $f$ that minimizes

$$L(Y \mid X) = L(f) + L(\mathcal{N})$$
SLOPE — computing $L(Y \mid X)$

1. $F = \emptyset$;
2. $f_g \leftarrow$ fit global function and add $f_g$ to $F$;
3. for each function type $t$ do
   4. $F_t \leftarrow F$;
   5. for $x \in X$, $\text{count}(x) > \delta$ do
      6. $f_l \leftarrow$ fit local function on $\tilde{x}$ of $x$;
      7. if adding $f_l$ to $F_t$ reduces overall costs then
         8. $F_t = F_t \cup f_l$;
      end
   end
   10. $F \leftarrow \min(F, F_t)$;
11. end
12. return costs of $Y$ given $F$ and $X$;
Confidence and Significance

How certain are we?

$$\mathbb{C} = |L(X) + L(Y | X) - L(Y) + L(X | Y)|$$

- the higher the more certain

$L(X \rightarrow Y)$  $L(Y \rightarrow X)$
Confidence Robustness

- **RESIT (HSIC idep.)**
- **IGCI (Entropy)**
- **SLOPE (Compression)**
Putting SLOPE to the test

We first evaluate using an ANM, with linear, cubic, or reciprocal functions, sampling $X$ and noise as indicated.

(Resit by Peters et al, 2014; IGCI by Janzing et al 2012)
Performance on Benchmark Data
(Tübingen 97 univariate numeric cause-effect pairs, weighted)
Performance on Benchmark Data
(Tübingen 97 univariate numeric cause-effect pairs, weighted)

SLOPE is 85% accurate with $\alpha = 0.001$

Inferences of state of the art algorithms ordered by confidence values.
Detecting Confounding

(Kaltenpoth & V. Telling Causal From Confounded, SDM’19)
Does Chocolate Consumption cause Nobel Prizes?

$r = 0.791$

$P < 0.0001$
If $X$ and $Y$ are statistically dependent then either

How can we distinguish these cases?
Conditional Independence Tests

If we have measured everything relevant then testing $X \perp Y | Z$ for all possible $Z$ lets us decide whether

Problem: It’s impossible to measure everything relevant
Why not just find a confounder?

We would like to be able to infer a $\hat{Z}$ such that

$$X \perp Y | \hat{Z}$$

if and only if $X$ and $Y$ are actually confounded

**Problem:** Finding such a $\hat{Z}$ is too easy. $\hat{Z} = X$ always works.
Kolmogorov Complexity

\( K(P) \) is the length of the shortest program computing \( P \)

\[
K(P) = \min_p \left\{ |p|: p \in \{0,1\}^*, |U(p, x, q) - P(x)| < \frac{1}{q} \right\}
\]

This shortest program \( p^* \) is the best compression of \( P \)
From the Markov Condition...

An admissible causal network for $X_1, ..., X_m$ is $G$ satisfying

$$P(X_1, ..., X_m) = \prod_{i=1}^{m} P(X_i \mid pa_i)$$

**Problem:** How do we find a simple factorization?
The simplest causal network for $X_1, \ldots, X_m$ is $G^*$ satisfying

$$K(P(X_1, \ldots, X_m)) = \sum_{i=1}^{m} K(P(X_i \mid pa_i^*))$$

**Postulate:** $G^*$ corresponds to the true generating process

(Janzing & Schölkopf, 2010)
AMC with Confounding

We can also include latent variables

$$K(P(X, Z)) = \sum_{i=1}^{m} K(P(X_i \mid pa'_i)) + \sum_{j=1}^{l} K(P(Z_j))$$
We don’t know $P(\cdot)$

$P(X, Z) = P(Z) \prod_{i=1}^{m} P(X_i | Z)$

In particular, we will use probabilistic PCA
Kolmogorov is not computable

For data $X$, the Minimum Description Length principle identifies the best model $M \in \mathcal{M}$ by minimizing

$$L(X, M) = L(M) + L(X \mid M)$$

gives a statistically sound approximation to $K$
Decisions, decisions

If

\[ L(X, Y, | \mathcal{M}_{co}) < L(X, Y | \mathcal{M}_{ca}) \]

then we consider \( X, Y \) to be confounded
Decisions, decisions

If

\[ L(X, Y, \mid M_{co}) > L(X, Y \mid M_{ca}) \]

then we consider \( X, Y \) to be causal.

The difference can be interpreted as confidence.
Confounding in Synthetic Data
There are only two other works directly related to ours

**SA:** Confounding strength in linear models using spectral analysis

**ICA:** Confounding strength using independent component analysis
Confounding in Genetic Networks

More realistically, we consider gene regulation data.
Optical Data
Optical Data
Wait! What about...
Conclusions

Causal inference from observational data
- necessary when making decisions, and to evaluate what-if scenarios
- impossible without assumptions about the causal model

Constraint-based causal discovery
- traditional approach based on conditional independence testing
- PC-algorithm discovers causal skeleton and orients (some) edges

Algorithmic Markov condition works very well in practice
- prefer simple explanations over complex ones
- consider complexity of both the model and the data

There is no causality without assumptions
- early work on relaxing e.g. causal sufficiency, determining confounding
“No causal claim can be established by a purely statistical method, be it propensity scores, regression, stratification, or any other distribution-based design”