Inferring causality from passive observations

Dominik Janzing

Max Planck Institute for Intelligent Systems
Tübingen, Germany

18.-22. August 2014
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 4: Causal inference in time series

- time series as test for causal inference methods
- Granger causality and its limits
- conditional-independence based causal inference in time series
- quantifying causal strength
Time series as test for causal inference methods
consider the following binary classification problem:

Given the values \( X_1, X_2, X_3, \ldots, X_n \) of an empirical time series, infer whether the true time direction is

\[
X_1, X_2, X_3, \ldots
\]

or

\[
X_n, \ldots, X_3, X_2, X_1.
\]

- statistical asymmetries between past and future should be the same as the asymmetries between cause and effect
- provides a simple evaluation of causal inference methods since the time direction is known
- no other direct application
Recall: $P(X, Y)$ may have a linear model from $X$ to $Y$ but not vice versa (i.e. noise is dependent)

We will now see that stochastic processes can be linear in one direction but not the other.
A (weakly stationary) stochastic process \((X_t)_{t \in \mathbb{Z}}\) has an autoregressive moving average process (ARMA) if

\[
X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t,
\]

with iid noise variables \(\epsilon_t\).

It is called \textit{causal} if \(\epsilon_t \perp \perp X_s\) for all \(s < t\).

(regression residuals are independent of the past and not only uncorrelated)
Non-Gaussian ARMA-models have a direction

Theorem (Peters et al, 2009)

Let \( (X_t)_{t \in \mathbb{Z}} \) have a causal ARMA model with non-vanishing AR-part. Then \( (X_{-t})_{t \in \mathbb{Z}} \) has a causal ARMA model if and only if the process is Gaussian.

Note: the theorem is only true if the notion of ARMA model implies independent noise terms. An ARMA model with uncorrelated noise terms exists in both directions.
Experiments with empirical time series showed

(from EEGs, finance...)
If a time series admits a causal ARMA model in one direction but not the other then the former is likely to be the true time direction.

i.e.:

Regressing the future on the past yields residuals that are independent of the past, while

regressing the past on the future yields residuals that are dependent of the future.
Recall: arrow of time in physics

- heat flows from the hot to the cold medium
- any kind of energy can be converted to heat, but not vice versa
- a photo contains information about the past, not the future
- ...

one can link the LINGAM-based asymmetry to the above arrow of time:

Janzing: On the entropy production of time series with unidirectional linearity, Journ. Statistical Physics, 2010
Granger causality and its limits
Granger causality

Let \((X_t)_{t \in \mathbb{Z}}\) and \((Y_t)_{t \in \mathbb{Z}}\) be two time series possibly influencing each other.

For some time instance \(t\), define

\[
Y_{\text{present}} := Y_t \quad \text{and} \quad Y_{\text{past}} := (Y_{t-1}, Y_{t-2}, \ldots)
\]

• how much does \(X_{\text{past}}\) help in predicting \(Y_{\text{present}}\) from \(Y_{\text{past}}\)?
Linear Granger:

• write $Y_t$ as a linear combination of its own past:

$$Y_t = \sum_{j=1}^{\infty} \alpha_j Y_{t-j} + U_t$$

• write $Y_t$ as linear combination of its own past and the past of $X$:

$$Y_t = \sum_{j=1}^{\infty} \alpha_j Y_{t-j} + \sum_{j=1}^{\infty} \beta_j X_{t-j} + E_t.$$ 

If $E_t$ has smaller variance than $U_t$ then the past of $X$ helps in predicting $Y$ from its past.
Transfer Entropy

Information theoretic version of Granger’s idea

\[ TE(X \rightarrow Y) := I(Y_{\text{present}} : X_{\text{past}} | Y_{\text{past}}) . \]
When is Granger causality right?

- Assume there are no instantaneous effects, i.e., no edges between $X_t$ to $Y_t$
- Assume $(X_t)_{t \in \mathbb{Z}}$ and $(Y_t)_{t \in \mathbb{Z}}$ to be causally sufficient, i.e., there are no unobserved common causes

\[
\text{if } Y_{\text{present}} \not\perp\!\!\!\perp X_{\text{past}} | Y_{\text{past}} \text{ there must be arrows from } X \text{ to } Y
\] (otherwise $d$-separation)
Note however...

Although $TE(X \rightarrow Y) \neq 0$ shows the existence of arrows from $X$ to $Y$ under the above assumptions, we don’t agree that the size of $TE(X \rightarrow Y)$ correctly quantifies the strength of the influence of $X$ on $Y$.


Some more explanations later.
Confounded Granger

Hidden common cause $Z$ relates $X$ and $Y$

\[
\begin{align*}
X_{t-3} & \rightarrow X_{t-2} & \rightarrow X_{t-1} & \rightarrow X_t \\
Z_{t-3} & \rightarrow Z_{t-2} & \rightarrow Z_{t-1} & \rightarrow Z_t \\
Y_{t-3} & \rightarrow Y_{t-2} & \rightarrow Y_{t-1} & \rightarrow Y_t
\end{align*}
\]

due to different time delays we have

\[
Y_{\text{present}} \perp \perp X_{\text{past}} \mid Y_{\text{past}}
\]

but

\[
X_{\text{present}} \perp \perp Y_{\text{past}} \mid X_{\text{past}}
\]

Granger erroeeously infers $X \rightarrow Y$
Instantaneous effects

Let $X$ only influence the $Y$ at the same time instance:

\[ X_{t-2} \rightarrow X_{t-1} \rightarrow X_t \rightarrow X_{t+1} \]
\[ Y_{t-2} \rightarrow Y_{t-1} \rightarrow Y_t \rightarrow Y_{t+1} \]

(we call such an influence 'purely instantaneous')

- due to d-separation we have
  \[ Y_{\text{present}} \perp \perp X_{\text{past}} \mid Y_{\text{past}} \]

- thus Granger infers ‘no influence from $X$ to $Y$’

- instantaneous effects often occur when the time steps are large compared to the interaction time
Conditional independence based causal inference in time series
Terminology

Definition (full time graph)

The full time graph of a vector-valued time series \((X_t)_{t \in \mathbb{Z}}\) is the infinite graph on the nodes \(X^i_t\) with an arrow from \(X^i_t\) to \(X^j_{t+s}\) for \(s \geq 0\) whenever there is such an influence. The largest \(s\) is called the order of \((X_t)_{t \in \mathbb{Z}}\), which we assume to be finite.

Note: since there can be arrows from \(X^i_t\) to \(X^j_t\) and vice versa at the same time, it need not be a DAG.

Peters et al: Causal inference on time series using restricted structural equation models, 2013
above we have already used such graphs without definition but the term ‘full time graph’ is helpful to avoid confusion with the following type of graph...
Terminology

Definition (summary graph)

The summary graph has nodes $X^i$ and contains an arrow from $X^i \rightarrow X^j$ whenever the full time graph has an arrow from some $X^i_{t}$ to some $X^j_{s}$.

The summary graph can be cyclic even if the full time graph is acyclic.
Causal Markov condition for infinite DAGs

whenever the full time graph is a DAG (i.e., has no cycles) we postulate:

- **local Markov condition:**
  every node $X^t_i$ is conditionally independent of its non-descendants, given its parents

- **global Markov condition:**
  for any three finite sets of nodes d-separation implies conditional independence
Theorem

If two full time graphs are Markov equivalent DAGs then they coincide up to instantaneous effects

Proof: Markov equivalent DAGs have the same skeleton. The direction of all non-instantaneous arrows follows from the time order.

Hence: the presence or absence of any of the below arrows can be determined.
Theorem

If the summary graph $G$ is known to be acyclic and there is no pair $(X^i, X^j)$ such that the influence is purely instantaneous then $G$ can be uniquely identified using Markov condition and faithfulness.

Proof: the summary graph contains an arrow $X^i \rightarrow X^j$ if and only if the skeleton of the full time graph contains a link $X^i_t \rightarrow X^j_{s+t}$ for some $s > 0$. 
Example for identifying the summary graph

• given two time series \((X_t)_{t \in \mathbb{Z}}\) and \((Y_t)_{t \in \mathbb{Z}}\).
• assume that the summary graph is

\[
\begin{align*}
X & \rightarrow Y \\
Y & \rightarrow X
\end{align*}
\]

or

\[
\begin{align*}
X & \rightarrow Y \\
Y & \rightarrow X
\end{align*}
\]

• assume that the influence is not purely instantaneous

then the full-time graphs have different skeletons:
since both contain arrows from $X$ to $Y$ and from $Y$ to $X$ the difference doesn’t seem so important.
Conclusions...

• since the future cannot influence the past the skeleton of the full time graph almost tells us the DAG
• the usual problem of large Markov equivalence classes is of minor importance for time series

...but we have assumed causal sufficiency so far, detection of confounding is a challenging problem, requires new methods:

Quantifying causal strength
Why Transfer Entropy does not quantify Causal Strength


deterministic influence between $X$ and $Y$

- we have $I(Y_{\text{present}}; X_{\text{past}} | Y_{\text{past}}) = 0$, although the influence is strong, because the past of $Y$ already determines its future
- quantitatively still wrong for non-deterministic relation

we now introduce a new measure for causal strength

Quantifying the strength of an arrow

Given:

- causally sufficient set of variables $X_1, \ldots, X_n$
- causal DAG $G$
- all causal conditionals $P(x_j|pa_j)$ even for values $pa_j$ with probability zero (more than just knowing $P(X_1, \ldots, X_n)$

quantify the strength of $X_i \rightarrow X_j$
Motivation:

Maybe, the true causal DAG is always complete if we also account for weak interactions. Which ones are so weak that we can neglect them?
Strength of a set of arrows

Idea:

• strength of an arrow measures its relevance for understanding the behavior of the system under interventions

• strength of a set of arrows measures their relevance for understanding the behavior of the system under interventions

• if each arrow in $S$ is irrelevant then $S$ could still be relevant
this picture is misleading because for a set $S$ of arrows

- each element may have negligible strength
- but jointly they are not negligible

our causal strength will not be subadditive over the edges!
advantages of information theory

• variables may have different domains
• quantities are invariant under rescaling
• related to thermodynamics
• better for non-statistical generalizations

don’t consider approaches that involve expectations, variances, etc. (ANOVA, ACE)
Some related work

- Avin, Sphitser, Pearl: Identifiability of path-specific effects, 2005.
- Pearl: direct and indirect effects, 2001.
- Robins, Greenland: Identifiability and exchangeability of direct and indirect effects, 1992.

Do not achieve our goal because

- measure impact of switching \( X \) from \( x \) to \( x' \) for one particular pair (\( x, x' \)) on \( Y \) when other paths are blocked
- we want an overall score of the strength of \( X \rightarrow Y \) without referring to particular pairs of values \( x, x' \)
Idea: measures influence of $X$ on $Y$, given $Z$

$$I(X \rightarrow Y \mid do(Z)) :=$$

$$p(z)p(x \mid do(z))p(y \mid do(x, z)) \log \frac{p(y \mid do(xz))}{\sum_{x'} p(x' \mid do(z))p(y \mid do(x', z))}.$$ 

Formally, this is a conditional mutual information $I(X : Y \mid Z)$, but not w.r.t. the observed distribution $P(X, Y, Z)$. Instead it uses the post-interventional distribution

$$\tilde{P}(X, Y, Z) := P(Y \mid do(X, Z))P(X \mid do(Z))P(Z)$$
Let $S$ be a set of arrows.

- Let $\mathcal{C}_S$ denote its strength.

- Postulate desired properties of $\mathcal{C}_S$. 

Axiomatic approach
Postulate 0: causal Markov condition

if $\mathcal{C}_S = 0$ then $P$ is also Markov w.r.t. $G_S$ (after removing all arrows in $S$)
Postulate 1: Mutual information

for this simple DAG we postulate $c_{X \rightarrow Y} = I(X; Y)$

(all the dependences are due to the influence of $X$ on $Y$, hence the strength of dependences can be a measure of the strength of the influence)
Postulate 2: Locality

\[ \xi_{X \rightarrow Y} \text{ is determined by } P(Y|PA_Y) \text{ and } P(PA_Y) \]

\[ Z \text{ is irrelevant in both cases} \]
Postulate 3: Quantitative causal Markov condition

$$\mathcal{C}_{X \rightarrow Y} \geq I(X : Y | PA^X_Y)$$

$PA^X_Y$ (parents of $Y$ without $X$)

Idea: removing $X \rightarrow Y$ would imply $I(X : Y | PA^X_Y) = 0$, therefore we attribute $I(X : Y | PA^X_Y)$ to this arrow
Postulate 4: Heredity

If \( T \supset S \) then \( C_T = 0 \Rightarrow C_S = 0 \)

(subsets of irrelevant sets of arrows are irrelevant)
$I(X : Y)$ is an inappropriate measure for general DAGs

ignores that part of the dependences are due to
a) the confounder $Z$
b) the indirect influence via $Z$
First guess: $I(X : Y | Z)$

- qualitatively, it behaves correctly: screens off the path involving $Z$
- quantitatively wrong because...
Why \( I(X : Y | Z) \) is inappropriate

weakening \( Z \to Y \) converts a) into b), where \( \mathcal{C}_{X \to Y} = I(X; Y) \)
Our approach: measure impact of deleting arrows

To define the strength of $S$, cut every edge in $S$ and feed the open end with an independent copy

- defines new distribution
  \[ p_S(x, y, z) := p(x, z) \sum_{x', z'} p(y|x', z') p(x') p(z') \]
- define causal strength as $C_S := D(p \parallel p_S)$
  (impact of edge deletion)
Relative entropy

(also called Kullback-Leibler divergence)

\[ D(p\|q) := \sum_x p(x) \log \frac{p(x)}{q(x)} \geq 0, \]

with equality iff \( p = q \)

- asymmetric distance measure
- broad applications in statistics, inference, learning
Understanding relative entropy

codelength perspective

• optimal code assigns codelength $-\log p(x)$ to event $x$ that occurs with probability $p(x)$
• Shannon entropy measures expected codelength

$$
\mathbb{E}[-\log p(X)] = H(X) = -\sum_x p(x) \log p(x)
$$

• someone who erroneously assumes that $x$ occurs with probability $q(x)$ uses a code with codelength $-\log q(x)$.
• the expected codelength is larger

$$
\mathbb{E}[-\log q(X)] = -\sum_x p(x) \log q(x) > \mathbb{E}[-\log p(X)].
$$

• relative entropy measures the increase of expected codelength

$$
D(p\|q) = \mathbb{E}[-\log q(X)] - \mathbb{E}[-\log p(X)].
$$
Relation to maximum likelihood estimation

- **goal**: distinguish two densities $q(X)$ and $p(X)$ from observations $x_1, \ldots, x_n$

- **common approach**: choose the one with the higher likelihood: compare

  $$\prod_{j=1}^{n} p(x_j) \quad \text{with} \quad \prod_{j=1}^{n} q(x_j)$$

  equivalently, compare the logarithms

  $$\sum_{j=1}^{n} \log p(x_j) \quad \text{with} \quad \sum_{j=1}^{n} \log q(x_j) .$$

- **asymptotics**: if $x_1, \ldots, x_n$ is sampled from $p(X)$, difference of loglikelihoods increases with $n$ according to $nD(p\|q)$
Idea of edge deletion

- edges are electrical wires
- attacker cuts some wires
- feeds the open ends with random input
- distribution of input chosen like observed marginal distribution
- only distribution that is locally accessible
Why product distribution?

joint distribution \( P(X, Z) \) not accessible to local attacker

Postulate 4 fails with joint distribution

'source exclusion' by Ay & Krakauer
Quantifying the impact of a vaccine

$p_S$ corresponds to an experiment where

- vaccine is randomly redistributed regardless of Age (keeping the fraction of treated subjects)
- the random variable vaccinated is reinterpreted as ‘intention to get vaccinated’
• $Y$ is always 0
• $Y$ is uniformly distributed after deleting $X \to Y$
• $Y$ remains independent of $X$
• $I(X; Y) = 0$ and $I(X; Y | Z) = 0$
• $c_{X \to Y} = 1$
• $A_y$ and Krakauer’s definition yields zero strength
Failure of subadditivity

redundancy code: bit $E$ is copied to all $B_j$

- removing less than half of the arrows $B_j \rightarrow D$ has no impact
- each arrow has strength zero
- all arrow together have strength 1
Applying our measure to time series

\( C_S \) quantifies effect of all \( X \) on \( Y_{t+1} \)

(applying this to the example of Ay & Polani yields a reasonable result)
Take home messages

• none of the existing measures appeared to be conceptually right for measuring strength of sets of edges
• our measure satisfies our postulates
• relies on interventions on edges
• clear operational meaning (does not refer to counterfactuals)
• definitions that rely on interventions on nodes failed although they seem more straightforward
• replacing Transfer Entropy (Granger causality) with our measure seems reasonable
**Exercise**

10 **Granger causality:**

Let \((Z_t)_{t \in \mathbb{Z}}\) be an unobserved time series such that \(Z_t\) influences \(X_{t+1}\) and \(Y_{t+2}\) for every \(t\), as in the earlier example for ‘confounded Granger’. However, now we have also arrows from \(Z_t\) to \(Z_{t+1}\) for every \(t\), as visualized here:

Show that, under the faithfulness assumption, we have

\[
TE(X \rightarrow Y) > 0 \\
TE(Y \rightarrow X) > 0,
\]