Complex Systems Methods — 6. Interdependence between time series: Granger causality and transfer entropy

Eckehard Olbrich

e.olbrich@gmx.de
http://personal-homepages.mis.mpg.de/olbrich/complex_systems.html

Potsdam WS 2007/08
Overview

1. Interdependence measures for time-series
   - Causality
   - Granger causality

2. Operationalisation
   - Vector autoregressive models
   - Transfer entropy and Granger causality

3. Problems
   - General Problems of observational causality concepts
   - Specific problem: State Dependence
   - Specific problem: Deterministic Dynamics

4. Estimating causal relationships

5. Summary
Why measuring interdependencies?

Consider two processes $X(t)$ and $Y(t)$. Possible questions:

- Is there any connection between the processes? Are they correlated?
- Is there a causal connection between the two processes? Are they coupled?
- Is one process driving the other?
- In particular interesting if no detailed model is available, e.g. in Neurosciences (EEG data) or Econometrics
Causality

There is a difference between correlation and causation.

Reichenbachs principle: Two processes $A$ and $B$ are statistically dependent (correlated) if either $A$ causes $B$, $B$ causes $A$, or both $A$ and $B$ have a common cause $C$.

Causality can be formalized using the concept of an intervention (Pearl): $A$ causes $B$, if we can change $B$ by intervening at (manipulating) $A$.

In models from physics: $B$ is coupled to $A$. 
Some Notation

- “World”: a set $V$ of $1 \leq N < \infty$ elements (agents, nodes) with state sets $\mathcal{X}_v$, $v \in V$.
- Given a probability vector $p$ on $\mathcal{X}_V$ we get random variables $X_V$ on $V$, $X_A$ on $A \subseteq V$ and $X_v$ on $v \in V$.
- World dynamics described as stationary stochastic process $X_V(t) = \{X_v(t)\}$, $v \in V$.
- discrete time
Definition — Wiener 1958, Granger 1964, Granger 1969

- past \( \overline{X}_V(t - 1) = (X_V(t - 1), \ldots, X_V(t - \infty)) \)
- subprocess \( X_{-j} = X_V \setminus \{j\} \)
- \( \sigma(X_A(t)|\overline{X}_A(t - 1)) \) denotes the standard deviation of the error predicting \( X_A(t) \) using \( \overline{X}_A(t - 1) \).

**Definition (Causality)**

\( X_j \) causes \( X_i \), if \( \sigma(X_i(t)|\overline{X}_V(t - 1)) < \sigma(X_i(t)|\overline{X}_{-j}(t - 1)) \), i.e. if the knowledge of the past values of \( X_j \) will improve the prediction of \( X_i \).

**Definition (Instantaneous Causality)**

\( X_j \) instantaneously causes \( X_i \), if

\[
\sigma(X_i(t)|\overline{X}_V(t - 1), X_j(t)) < \sigma(X_i(t)|\overline{X}(t - 1)),
\]

i.e. if the knowledge of the actual value of \( X_j \) will improve the prediction of \( X_i \).
**Axiom A:** The past and the present may cause the future, but the future cannot cause the past

**Axiom B:** \( \mathbf{X}(t) \) contains no redundant information, so that if some variable \( X_k(t') \) is functionally related to one or more other variables, in a deterministic fashion, then \( X_k(t') \) should be excluded from \( \mathbf{X}(t) \).

E.g. \( x_j(t) = f(x_k(t - m)) \), but also \( x_j(t) = f(x_j(t - 1), x_j(t - 2), \ldots, x_j(t - m)) \), i.e. Granger excludes deterministic systems.

---

**Definition**

\( X_j \) causes \( X_i \) if \( p(x_i(t)|\mathbf{X}_V(t - 1)) \neq p(x_i(t)|\mathbf{X}_{-j}(t - 1)) \), i.e. \( X_j \) non-causes \( X_i \) if \( X_i(t) \) is conditionally independent on \( X_j \) given \( \mathbf{X}_{-j}(t - 1) \).
Operationalisation: Vector autoregressive models (VAR)

- A weakly stationary zero mean stochastic process has an autoregressive representation

\[ x_V(t) = \sum_{u=1}^{\infty} a(u)x(t-u) + \epsilon(t) \]

- \(X_j\) is Granger non-causal to \(X_i\) with respect to \(X_V\) if \(a_{ij}(u) = 0\) \(\forall u\).

- \(X_j\) instantaneously non-causes \(X_i\), if \(\Sigma_{ij} = \langle \epsilon_i(t)\epsilon_j(t) \rangle = 0\).

- In the context of Graphical models: structural equations

- Problem: Only linear dependencies!
Transfer entropy — Information theoretic version of Granger causality

- Schreiber 2000: Transfer entropy measures “directed information flow”; originally only bivariate

\[ T_{j \to i} = MI(X_i(t) : \overline{X}_j(t-1)|\overline{X}_i(t-1)) \]
\[ = H(X_i(t)|\overline{X}_i(t-1)) - H(X_i(t)|\overline{X}_i(t-1), \overline{X}_j(t-1)) \]

- Palus 2001: Measuring conditional independence using conditional mutual information \( \Rightarrow \) information theoretic formulation of the Granger causality — \( X_j \) Granger causes \( X_i \) if \( T_{j \to i,V} > 0 \).

\[ T_{j \to i,V} = MI(X_i(t) : \overline{X}_j(t-1)|\overline{X}_{-j}(t-1)) \]
\[ = H(X_i(t)|\overline{X}_{-j}(t-1)) - H(X_i(t)|\overline{X}_V(t-1)) \]
General Problems of observational causality concepts

- World description has to be causally complete in order to exclude common causes.
- Granger causality defined via conditional independence is purely observational, no interventions.
- ⇒ if $X_i$ and $X_j$ are synchronized no causal interaction is detected
- But, this case is excluded by Granger's Axiom B!
Specific problem: State Dependence

Whether e.g. $X_1$ Granger causes $X_2$ depends on the representation of the rest of the world!

\[
\begin{align*}
    x_1(t) &= a_{11} x_1(t-1) + a_{12} x_2(t-1) + a_{13} x_3(t-1) + \epsilon_1(t) \\
    x_2(t) &= a_{21} x_1(t-1) + a_{22} x_2(t-1) + a_{23} x_3(t-1) + \epsilon_2(t) \\
    x_3(t) &= a_{31} x_1(t-1) + a_{32} x_2(t-1) + a_{33} x_3(t-1) + \epsilon_3(t)
\end{align*}
\]

can be transformed into

\[
\begin{align*}
    x_1(t) &= (a_{11} - a_{13} \alpha) x_1(t-1) + (a_{12} - a_{13} \beta) x_2(t-1) + a_{13} \tilde{x}_3(t-1) + \epsilon_1(t) \\
    x_2(t) &= (a_{21} - a_{23} \alpha) x_1(t-1) + (a_{22} - a_{23} \beta) x_2(t-1) + a_{23} \tilde{x}_3(t-1) + \epsilon_2(t) \\
    \tilde{x}_3(t) &= (a_{31} - (a_{33} + a_{11}) \alpha) - a_{13} \alpha^2) x_1(t-1) + \\
    &\quad (a_{32} - (a_{33} + a_{12}) \beta - a_{13} \beta^2) x_2(t-1) + \\
    &\quad (a_{33} - a_{13} \alpha - a_{23} \beta) \tilde{x}_3(t-1) + \epsilon_3(t)
\end{align*}
\]

using $\tilde{x}_3 = x_3 + \alpha x_1 + \beta x_2$ with $\alpha = a_{21} / a_{23} \Rightarrow X_2$ becomes independent on $X_1$ conditioned on $\tilde{x}_3$. 
Specific problem: Deterministic Dynamics

- Deterministic dynamical system:

\[ x(t) = F(x(t - 1)) \]

- Embedding theorem: The map
  \[ x(t) \mapsto s(t) = h(x(t)) \mapsto (s(t), s(t - 1), \ldots, s(t - m + 1)) \]
  is an immersion with nowhere vanishing Jacobian, if \( m > 2D_0 \) with \( D_0 \) the box-counting dimension of the attractor

\[ \Rightarrow \text{state space can be reconstructed from any } X_i \]

- KS-entropy

\[ h_{KS} = \lim_{\epsilon \to 0} h(X(t)|\overline{X}(t - 1), \epsilon) \]

\[ = \lim_{\epsilon \to 0} h(X_i(t)|\overline{X}_i(t - 1), \epsilon) \]

\[ \Rightarrow MI(X_i(t) : \overline{X}_j(t - 1)|\overline{X}_{-j}(t - 1)) = 0 \text{ if } h_{KS} = 0 \]

\[ \Rightarrow \text{No Granger causality in non-chaotic deterministic systems.} \]

- But again, this situation is excluded by Axiom B!
Example: Granger causality in a VAR(2) process

\[ x_1(t) = a_{11} x_1(t - 1) + a_{12} x_2(t - 1) + \epsilon_1(t) \]
\[ x_2(t) = a_{21} x_1(t - 1) + a_{22} x_2(t - 1) + \epsilon_2(t) \]

In which way implies \( a_{12} > 0 \) better predictability of \( X_1 \) knowing \( X_2 \)?
Predicting \( X_1(t) \) using only \( X_1(t - 1) \)

\[ x_1(t) = a_{11} x_1(t - 1) + a_{12} a_{21} x_1(t - 2) \]
\[ + a_{12} a_{22} x_2(t - 2) + a_{12} \epsilon_2(t - 1) + \epsilon_1(t) \]
\[ = a_{11} x_1(t - 1) + a_{12} a_{21} x_1(t - 2) + a_{12} a_{22} a_{21} x_1(t - 3) \]
\[ + a_{12} a_{22} x_2(t - 3) + a_{12} a_{22} \epsilon_2(t - 2) + a_{12} \epsilon_2(t - 1) + \epsilon_1(t) \]

Special case \( a_{22} = 0 \)

\[ x_1(t) = a_{11} x_1(t - 1) + a_{12} a_{21} x_1(t - 2) + a_{12} \epsilon_2(t - 1) + \epsilon_1(t) \]
Transfer entropy and effective noise level

- Granger causality: Improving predictability $\equiv$ Reducing noise level
- Stochastic dynamics for $X_i(t)$:

$$x_i(t) = f(\overline{x}_i(t - 1), \xi_i(t)) \quad \langle \xi_i(t)^2 \rangle = 1$$

- Differential entropy $H(X) = -\int dx \ p(x) \log p(x)$ transforms for invertible function $y = f(x)$ according to

$$H(Y) = H(X) + \int dx \ p(x) \log |f'(x)|$$

because

$$p(x)dx = q(y)dy \quad \Rightarrow \quad q(y) = \left. \frac{p(x)}{df/dx} \right|_{x = f^{-1}(y)}$$

Applying this we get

$$h(x_i(t)|\overline{x}_i(t - 1)) = H(\xi_i) + \langle \ln \left| \frac{\partial f}{\partial \xi_i} \right| \rangle.$$
Transfer entropy and effective noise level

- Using only the dynamics for $X_i(t)$ we got

$$h(x_i(t)|\bar{x}_i(t - 1)) = H(\xi_i) + \langle \ln \left| \frac{\partial f}{\partial \xi_i} \right| \rangle.$$ 

- Stochastic dynamics for $X_i(t)$ and $X_j(t)$:

$$x_i(t) = g(\bar{x}_i(t - 1), \bar{x}_j(t - 1), \xi_{ij}(t)) \quad \langle \xi_{ij}(t)^2 \rangle = 1$$

- Same reasoning gives

$$h(x_i(t)|\bar{x}_i(t - 1), \bar{x}_j(t - 1)) = H(\xi_{ij}) + \langle \ln \left| \frac{\partial g}{\partial \xi_{ij}} \right| \rangle.$$ 

- Therefore

$$T_{j\rightarrow i} = H(\xi_i) - H(\xi_{ij}) + \langle \ln \left| \frac{\partial f}{\partial \xi_i} \right| \rangle - \langle \ln \left| \frac{\partial g}{\partial \xi_{ij}} \right| \rangle.$$
Estimating “causal” relationships

- **Linear:** Fitting a VAR(m) model to the data, e.g. using least square estimation (e.g. ar-model in TISEAN) and then testing the coefficients $a_{ij}$ against zero.

- **Non-linear:** Estimating the conditional mutual informations (transfer entropy) — Partitioning the data (if continuous variables) and estimating the entropies $H(X_i(t) | \mathbf{X}_{-j}(t-1), \epsilon)$ and $H(X_i(t) | \mathbf{X}_V(t-1), \epsilon)$.

  Note that the result depends on the state space, e.g. on the embedding dimensions $m_j, m_i$ in the Transfer entropy

  $$T_{j \rightarrow i}(m_j, m_i, \epsilon) = MI(X_i(t) : X_j(t-1), \ldots, X_j(t-m_j+1) | X_i(t-1), \ldots, X_i(t-m_i+1); \epsilon)$$

  The result might depend on $\epsilon$. But, for stochastic systems the conditional mutual information should converge for $\epsilon \to 0$ to the value for differential entropies!

  You have to correct for finite sample effects. Finite sample effects lead to overestimation.
Dependence on the resolution $\epsilon$


A. Kaiser and T. Schreiber, Information transfer in continuous processes,

$\epsilon$ values with $T_{\text{heart} \rightarrow \text{breath}} > T_{\text{breath} \rightarrow \text{heart}}$ are marked by black squares.

FIG. 3. Bivariate time series of the breath rate (upper) and instantaneous heart rate (lower) of a sleeping human. The data is sampled at 2 Hz. Both traces have been normalized to zero mean and unit variance.

FIG. 4. Transfer entropies $T(\text{heart} \rightarrow \text{breath})$ (solid line), $T(\text{breath} \rightarrow \text{heart})$ (dotted line), and time delayed mutual information $M(\tau = 0.5 \text{ s})$ (directions indistinguishable, dashed line) for the physiological time series shown in Fig. 3.
Correcting for finite sample effects - effective transfer entropy


- Effective transfer entropy: Difference between the usual transfer entropy and the transfer entropy between \( X_i(t) \) and a shuffled version of \( X_j(t) \).

\[
ET_{j\rightarrow i}(m_i, m_j) := T_{j\rightarrow i}(m_i, m_j) - T_{j, shuffled\rightarrow i}(m_i, m_j)
\]

Fig. 6. Comparison of the behaviour of transfer entropy and effective transfer entropy for a varying sample size \( N \). The information flow \( y(t) \) to \( x(t) \) (Eq. (11)), with \( \epsilon = 0.15 \), \( S = 3 \) and \( m = 4 \) was measured for ten different realizations of the process, then average and standard deviation were calculated.
Fig. 2. Transfer entropy measuring the information flow from Dow Jones to DAX series, using various partitions of $S = 2, 3, 4, 5$ symbols (bottom to top). Upper lines have been calculated on the log-returns of DJ and DAX, for the lower ones (triangles) the log-returns of the DJ series have previously been shuffled.

Fig. 3. Effective transfer entropy measuring the information flow between Dow Jones and DAX series, and vice versa, using four different partitions of $S = 2, 3, 4, 5$ symbols (bottom to top).
Summary

- Granger causality asks for interdependencies between stochastic processes.
- It can be expressed using conditional mutual information (Transfer entropy).
- If we consider only linear interdependencies it can be studied with vector autoregressive (VAR) models.
- One has to be careful with causal interpretations because it is a purely observational measure.