Symbolic Causation Entropy

Carlo Cafaro[1], Dane Taylor[2], Jie Sun[1], and Erik Bollt[1]

Clarkson University[1], Potsdam NY, USA
University of North Carolina[2], Chapel Hill NC, USA

cidnet14, Max-Planck Institute, Dresden, Germany

Preliminary thanks

1. Claudia Poenisch (…organization/accomodation…)
2. CIDNET14 Coordinators (…present…)
3. Max-Planck Institute, Dresden (…host…)

$\boxed{2.16 \rightarrow 1.23 \rightarrow 1.14 \rightarrow \text{out!}}$

…my first ten months dealing with **causality** and **networks**…

CIDNET14: Causality, Information Transfer and Dynamical Networks

Problem, relevance, and our approach

- What is the problem?
- Why is it relevant?
- Why is it challenging?
- What is our approach? …applications…

(…and, linking dynamical systems theory to information theory via **symbolic dynamics**)

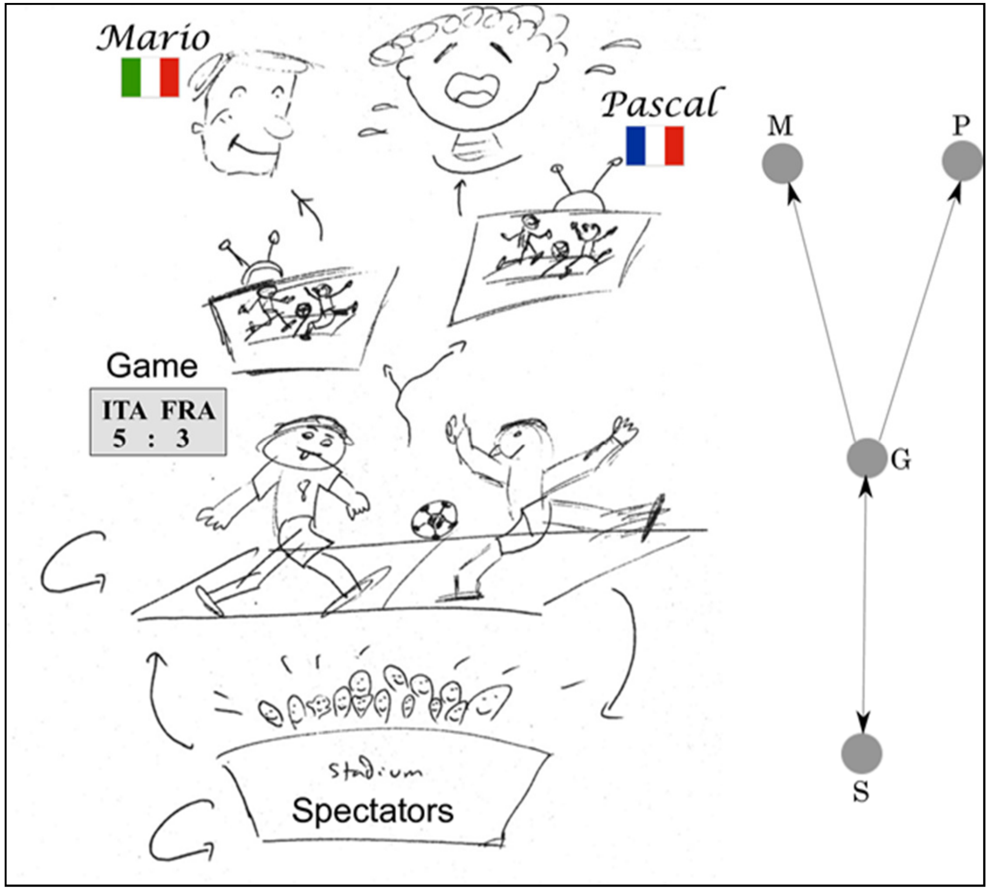
Part 1: Problem, relevance, challenges

The problem in a cartoon

Problem: Infer the coupling structure (cause-effect relationships) of complex systems from time-series data

What is causality?

Causality is the relation between two events (the **cause** and the **effect**), where the second event is understood as a consequence of the first.



Cause-effect relationships in a soccer game

Causality Quiz

Figure 1. Causality quiz.

1. What affects the state of mind of Mario?

☐ *Is Mario happy because Pascal is sad?* No. Mario has no idea who Pascal is.

☐ *Is Mario happy because the spectators are cheering?* No. If anything, Mario is only jealous at those attending the game.

☐ *Is Mario happy because of the game?* Yes. Check the scoreboard.

2. What affects the behavior of the spectators?

☐ *Are the spectators cheering because Mario is happy?* No. Why would they care about someone they don't even know?

☐ *Are the spectators cheering because Pascal is sad?* No. Why would they care about someone they don't even know?

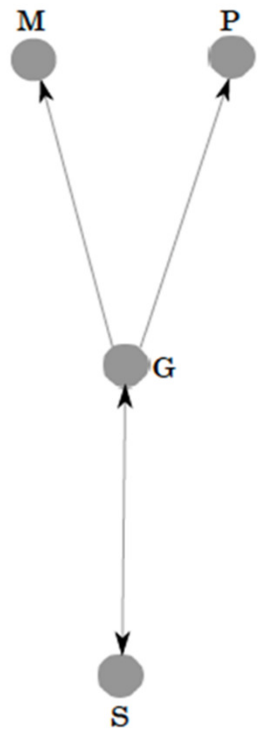
☐ *Are the spectators cheering because of the game?* Yes. They are restless soccer lovers, just like the players.

3. What affects the state of the game?

☐ *Is Mario helping his team to win?* No. Although Mario probably thinks so after too much wine and cheese.

☐ *Is Pascal causing his team to lose?* No. Pascal is only causing his TV to break after kicking a ball against it.

☐ *Do the spectators influence the game?* Yes. This is even scientifically proven.



- Mario and Pascal are *causally disconnected*
- Mario and Pascal are *negatively correlated*

Remark: $\boxed{\textbf{CORRELATION} \neq \textbf{CAUSATION}}$

Example:

$$\text{Summer} \overset{\text{causal}}{\Rightarrow} \text{Drowning}$$

$$\text{Summer} \overset{\text{causal}}{\Rightarrow} \text{Ice-cream consumption}$$

$$\text{Ice-cream consumption} \overset{\text{correlated}}{\leftrightarrow} \text{Drowning}$$

$$\text{Ice-cream consumption} \overset{\text{causal?}}{\Rightarrow} \text{Drowning}$$

correlation **does not** imply causality

Relevance of the problem

Major goals in *theoretical physics* (science, in general):

- **Describe** natural phenomena
- **Understand**, to a certain extent, natural phenomena

…predict the future…
…reconstruct the past…
…control phenomena….

…understand **what causes what** is naturally important…

…there are several practical reasons…

medical diagnosis: identify the causes of a disease in order to suggest effective treatments

Heart dynamics

chaoticity & cardiac rhythm

Angelo Mosso

…*inspired* by D. Chialvo (June19, cidnet14)

Brain dynamics

global synchronization & brain activity

Challenge: a good solution

- Gather a *sufficient* amount of relevant data (experimental work)
- Uncover a *good* causal inference measure (theoretical work)
- Provide an *accurate* (and, possibly, *fast*) estimate of such a measure (numerical and computational work)

…*quoting* K. Lehnertz (June 16, cidnet14): …this is a challenge for the next decades to come…

A good causal network inference measure

1. …general applicability and neat interpretation...

2. …immune to false positives and false negatives…

3. …accurate and fast numerical estimation…

Condition (1) →…linear and nonlinear interactions…

Condition (2) →…correct identification of direct couplings in complex systems with more than two components…

Condition (3) →…appropriate statistical estimation techniques…

A. Kraskov, H. Stogbauer, and P. Grassberger, Phys. Rev. **E69**, 066138 (2004)
J. Runge, J. Heitzig, V. Petoukhov, and J. Kurths, Phys. Rev. Lett. **108**, 2587701 (2012)

Some preliminaries

(information-theoretic approach to cauality inference)

X is a discrete random variable with probability distribution p(x)

$$\boxed{H(X) \stackrel{\text{def}}{=} -\sum_x p(x)\log p(x) \geq 0}$$

Shannon Entropy

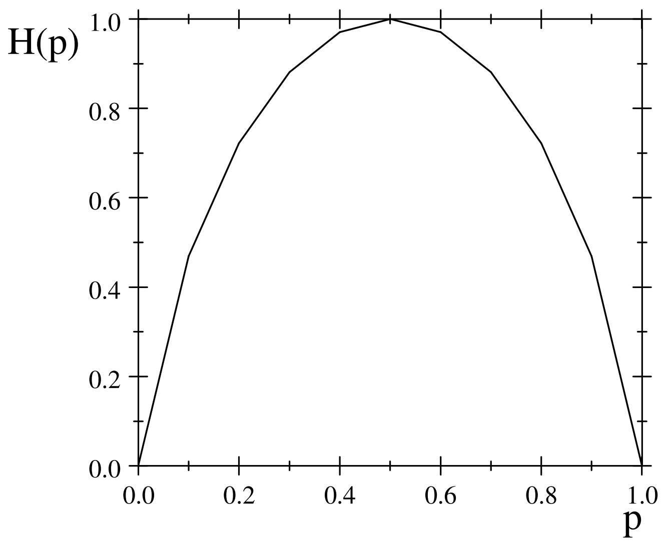
Interpretation: H(X) is a measure of uncertainty associated with X

Example

$$X \stackrel{\text{def}}{=} \begin{cases} 1, \text{ with probability } p \\ 0, \text{ with probability } (1-p) \end{cases}$$

$$H(p) = -p\log_2 p - (1-p)\log_2(1-p)$$

$$[H(p)] = \text{bits}$$



$$\boxed{H(X, Y) \stackrel{\text{def}}{=} -\sum_{x, y} p(x, y) \log p(x, y)}$$

joint entropy

$$\boxed{H(X|Y) \stackrel{\text{def}}{=} -\sum_{x, y} p(x, y) \log p(x|y)}$$

conditional entropy

$$\boxed{M(X, Y) \stackrel{\text{def}}{=} H(X) - H(X|Y) = M(Y, X)}$$

mutual information

Interpretation: M(X, Y) is the reduction in uncertainty of X due to the knowledge of Y

$$\boxed{M(X, Y|Z) \stackrel{\text{def}}{=} H(X|Z) - H(X|Y, Z)}$$

conditional mutual information

Interpretation: M(X, Y|Z) is the reduction in uncertainty of X due to the knowledge of Y when Z is given

Transfer Entropy (TE)

Question: Is there a measure of the magnitude and direction of **information flow** between jointly distributed stochastic processes?

TE:

$$T_{X^{(j)}\to X^{(i)}} = H\left(X_{t+1}^{(i)} \middle| \mathbf{X}_t^{(i)}\right) - H\left(X_{t+1}^{(i)} \middle| \mathbf{X}_t^{(i)}, \mathbf{X}_t^{(j)}\right)$$

$$= M\left(X_{t+1}^{(i)}, \mathbf{X}_t^{(j)} \middle| \mathbf{X}_t^{(i)}\right)$$

$$= \sum p\left(x_{t+1}^{(i)}, \mathbf{x}_t^{(i)}, \mathbf{x}_t^{(j)}\right) \log\left[\frac{p\left(x_{t+1}^{(i)} \middle| \mathbf{x}_t^{(i)}, \mathbf{x}_t^{(j)}\right)}{p\left(x_{t+1}^{(i)} \middle| \mathbf{x}_t^{(i)}\right)}\right]$$

H= Shannon entropy

M= mutual information

$$\mathbf{X}_t^{(i)} = \left(X_t^{(i)}, X_{t-\tau_i}^{(i)}, \ldots, X_{t-(m_i-1)\tau_i}^{(i)}\right)$$

τ_i = time delay parameter

m_i = embedding dimension

T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000)
M. Palus, V. Komarek, Z. Hrncir, K. Sterbova, Phys. Rev. **E63**, 046211 (2001)

Interpretation: Uncertainty reduction of the future states of $X^{(i)}$ as a result of knowing the past states of $X^{(j)}$ given that the past of $X^{(i)}$ is already known.

Transfer Entropy is a pairwise (asymmetric) measure of information flow

Example

$$\boxed{T_{1\to 2} \neq T_{2\to 1}}$$

$T_{1\to 2}$ measures the degree of dependence of 2 on 1 (and NOT viceversa)...

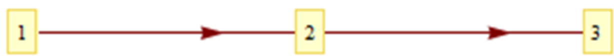
Fact: Pairwise inference methods (bivariate analysis) to identify coupling in networks with more than two nodes warrants caution…

$$\boxed{1 \to 2,\ 2 \to 3,\ 1 \overset{?}{\to} 3}$$

Indirect and direct influences

Fact: If a causal interaction is given by 1→2→3, a bivariate analysis would give a significant link between 1 and 3 that is detected as being only indirect in a multivariate analysis including 2.

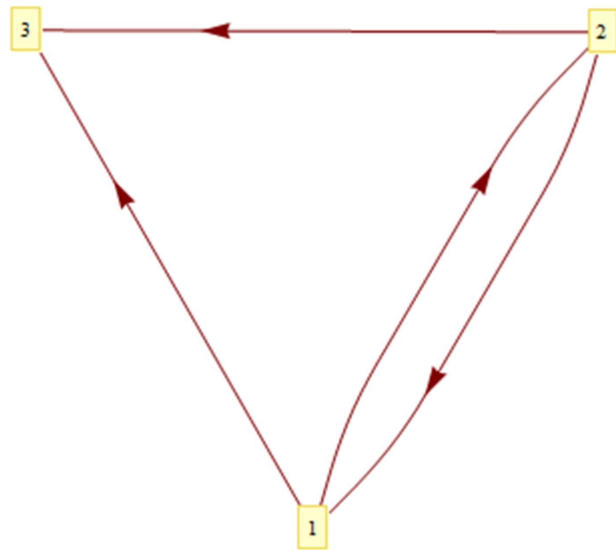
Example



true coupling structure

bivariate analysis: $M_{12} \neq 0$, $M_{13} \neq 0$, $M_{23} \neq 0$

multivariate analysis: $M_{12|3} \neq 0$, $M_{13|2} = 0$, $M_{23|1} \neq 0$



wrong coupling structure

bivariate analysis: $M_{12} \neq 0$, $M_{13} \neq 0$, $M_{23} \neq 0$

multivariate analysis: $M_{12|3} \neq 0$, $M_{13|2} = 0$, $M_{23|1} \neq 0$

- Both coupling structures are consistent with the bivariate analysis (**WITHOUT proper conditioning**);
- Only the true coupling structure is consistent with the multivariate analysis (**WITH proper conditioning)**.

…*quoting* M. Eichler (June 25, cidnet14):…the more…the better…

spurious coupling between 1 and 3

S. Frenzel and B. Pompe, Phys. Rev. Lett. **99**, 204101 (2007)

Some (incomplete) history

- …quantifying information transfer…

T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000) → **(MPI, Dresden)**
M. Palus et *al*., Phys. Rev. **E63**, 046211 (2001)

- …bivariate Gaussian processes and TE…

A. Kaiser and T. Schreiber, Physica **D166**, 43 (2002) → (**MPI, Dresden**)

- …multivariate Gaussian processes and conditional TE…

L. Barnett et *al*., Phys. Rev. Lett. **103**, 238701 (2009)

- …existence and strength of causal relationsips…

J. Runge et *al*., Phys. Rev. **E86**, 061121 (2012)

Summary: Part 1 (problem, relevance, challenges)

Interplay among:

- **Experimentalists** (…experimental design…)
- **Computational Scientists**
(…numerical/computational estimation methods…)
- **Theorists** (…*universal* causal inference measure…)

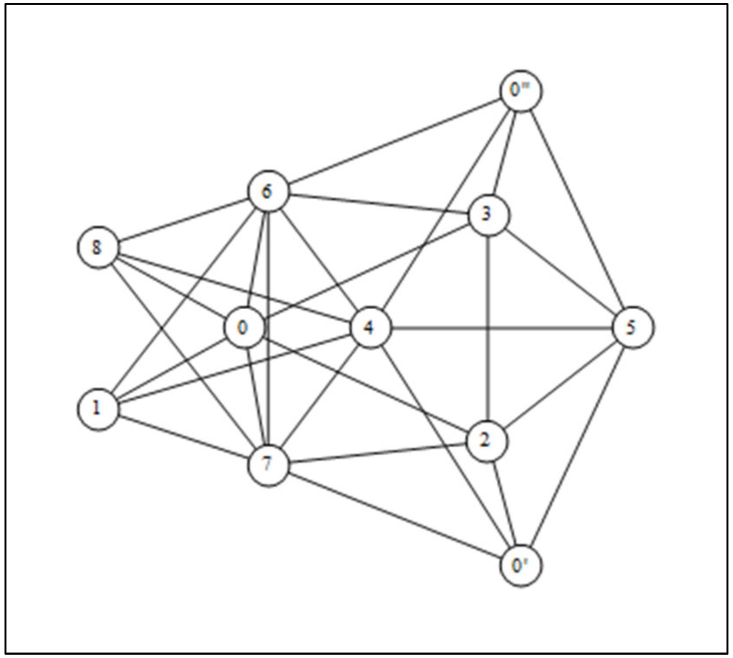
Question: …in the meantime, inspired also by TE, what are we actually doing…?

Part 2: Our approach

(the oCSE approach = the optimal Causation Entropy approach)

The goal

Identify the coupling structure in complex systems described by dynamical networks



- causal network topology (existence)
- link weights (strength)
- functional dependence between nodes (…hard problem…)

Some notations

Remark:

complex systems ↔ networks
networks ↔ (nodes, links)
nodes ↔ dynamical systems
links ↔ interactions

Probabilistic approach: dynamical systems modeled in terms of stationary stochastic processes

n = total number of nodes

$$X_t = \left[X_t^{(1)}, \ldots, X_t^{(n)}\right]$$

$|K|$ = cardinality of a subset of nodes

$$X_t^{(K)} = \left[X_t^{(1)}, \ldots, X_t^{(|K|)}\right], \text{ with } |K| \le n$$

Markovian causal inference framework

Working hypotheses: (*stationary*) stochastic processes satisfying the following (*Markov*) conditions

$$(i)\ :\ p(X_t|X_{t-1}, X_{t-2},...) = p(X_t|X_{t-1}) = p(X_{t'}|X_{t'-1}), \forall t, t';$$

$$(ii)\ :\ p(X_t^{(j)}|X_{t-1}) = p(X_t^{(j)}|X_{t-1}^{(N_j)}), \forall j;$$

$$(iii)\ :\ p(X_t^{(j)}|X_{t-1}^{(K)}) \neq p(X_t^{(j)}|X_{t-1}^{(L)}), \text{ whenever } (K \cap N_j) \neq (L \cap N_j).$$

Main point: for each component j there is a *unique* (and minimal) set of components N_j that renders the rest of the system irrelevant in making inferences about $X^{(j)}$

J. Sun, D. Taylor, and E. M. Bollt, arXiv:cs.IT/1401.7574 (2014)

Causation Entropy (CSE)

CSE:

$$
\begin{aligned}
C_{X^{(J)}\to X^{(I)}|X^{(K)}} &= H\left(X_{t+1}^{(I)}\,\middle|\,\mathbf{X}_t^{(K)}\right) - H\left(X_{t+1}^{(I)}\,\middle|\,\mathbf{X}_t^{(K)},\, \mathbf{X}_t^{(J)}\right) \\
&= M\left(X_{t+1}^{(I)},\, \mathbf{X}_t^{(J)}\,\middle|\,\mathbf{X}_t^{(K)}\right) \\
&= \sum p\left(x_{t+1}^{(I)},\, \mathbf{x}_t^{(J)},\, \mathbf{x}_t^{(K)}\right) \log\left[\frac{p\left(x_{t+1}^{(I)}\,\middle|\,\mathbf{x}_t^{(J)},\, \mathbf{x}_t^{(K)}\right)}{p\left(x_{t+1}^{(I)}\,\middle|\,\mathbf{x}_t^{(K)}\right)}\right]
\end{aligned}
$$

$$\mathbf{X}_t^{(K)} = \left(\mathbf{X}_t^{(k_1)}, \ldots, \mathbf{X}_t^{(k_{|K|})}\right)$$

$$1 \le r \le |K| \,:\, \mathbf{X}_t^{(k_r)} = \left(X_t^{(k_r)},\, X_{t-1}^{(k_r)}, \ldots, X_{t-(\tau_{k_r}-1)m_{k_r}}^{(k_r)}\right)$$

τ_{k_r} = time delay parameter

m_{k_r} = embedding dimension

J. Sun and E. M. Bollt, Physica **D267**, 49 (2014)

Interpretation: Uncertainty reduction of the future states of $X^{(I)}$ as a result of knowing the past states of $X^{(J)}$ given that the past of $X^{(K)}$ is already known.

CSE as a generalization of TE:

$$\boxed{T_{X^{(j)} \to X^{(i)}} \to C_{X^{(J)} \to X^{(I)} | X^{(K)}}}$$

For $J = \{j\}$ and $K = I = \{i\}$:

$$\boxed{C_{X^{(j)} \to X^{(i)} | X^{(i)}} = T_{X^{(j)} \to X^{(i)}}}$$

CSE, uncoditional TE, conditional TE

CSE and unconditional TE

$$T_{Y_-\to X_+} \equiv C_{Y_-\to X_+|X_-} \overset{\text{def}}{=} H(X_+|X_-) - H\left(X_+\middle|Y_-, X_-\right)$$

$\boxed{\text{self-causality cannot be investigated with TE}}$

if $Y_- = X_-$: $T_{X_-\to X_+} = 0$

$$C_{Y_-\to X_+|Z_-} \overset{\text{def}}{=} H(X_+|Z_-) - H\left(X_+\middle|Y_-, Z_-\right)$$

$\boxed{\text{self-causality can be investigated with CSE}}$

if $Z_- = \{\varnothing\}$, and $Y_- = X_-$: $C_{X_-\to X_+} = H(X_+) - H(X_+|X_-)$

N. J. Cowan et al., PlosOne **7**, 1 (2012)

CSE and conditional TE

Notation: $X_t^{(k)} \stackrel{\text{def}}{=} [X_t, X_{t-1},..., X_{t-k+1}]$

$$T_{Y \to X|Z} \stackrel{\text{def}}{=} H\left(X_{t+1} \middle| X_t^{(k)}, Z_t^{(m)}\right) - H\left(X_{t+1} \middle| X_t^{(k)}, Z_t^{(m)}, Y_t^{(l)}\right)$$

conditional TE

$$C_{Y \to X|W} \stackrel{\text{def}}{=} H(X_{t+1}|W_t) - H\left(X_{t+1} \middle| W_t, Y_t^{(l)}\right)$$

CSE

if $W_t = [X_t^{(k)}, Z_t^{(m)}],\ C_{Y \to X|W} = T_{Y \to X|Z}$

if $W_t \neq [X_t^{(k)}, Z_t^{(m)}],\ C_{Y \to X|W} \neq T_{Y \to X|Z}$

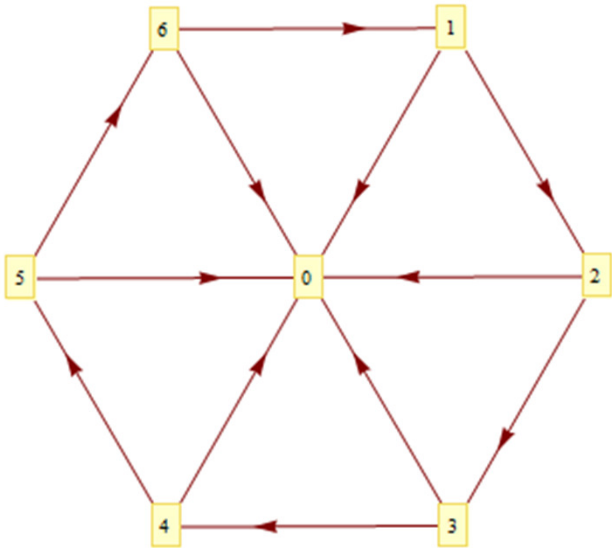
Remark: conditional TE is a special case of CSE

The optimal Causation Entropy approach

i = node of the graph

N_i = set of nodes that are *directly* causally connected to node-i

Goal: For any node-i, find N_i (maximization of CSE)



Algorithm 1: Aggregative Discovery
For any node-i, it finds the set of nodes that are causally connected to node-i (including indirect and spurious causal connections)

Algorithm 2: **Progressive Removal**
For any node-i, it removes indirect and spurious causal connections

J. Sun, D. Taylor, and E. M. Bollt, arXiv:cs.IT/1401.7574 (2014)

Algorithm 1: Aggregative Discovery

Output: For any node-i, it finds $M_i \overset{\text{def}}{=} \{k_1, \ldots, k_j\}$

M_i contains all types of causal links: direct, indirect, spurious

Example:

$$
\boxed{
\begin{array}{l}
k_1 : 0 < C_{k_1 \to i} \text{ is max} \\
k_2 : 0 < C_{k_2 \to i | \{k_1\}} \text{ is max} \\
\vdots \\
k_j : 0 < C_{k_j \to i | \{k_1, \ldots, k_{j-1}\}} \text{ is max} \\
k_{j+1} : C_{k_{j+1} \to i | \{k_1, \ldots, k_j\}} \text{ equals zero} \Rightarrow \text{stop!}
\end{array}
}
$$

Algorithm 2: Progressive Removal

Output: For any node-i, it finds $N_i \stackrel{\text{def}}{=} \{\text{direct causal links to node-}i\}$

k_r with $1 \le r \le j$ is removed from M_i when $C_{k_r \to i | M_i \setminus \{k_r\}} = 0$

Example:

$$k_1 : \begin{cases} C_{k_1 \to i | M_i \setminus \{k_1\}} > 0 \Rightarrow \text{keep } k_1 \\ C_{k_1 \to i | M_i \setminus \{k_1\}} = 0 \Rightarrow \text{remove } k_1 \end{cases}$$

Suppose we remove k_1. Then, $M_i \to M_i' \stackrel{\text{def}}{=} M_i \setminus \{k_1\}$

$$k_2 : \begin{cases} C_{k_2 \to i | M_i' \setminus \{k_2\}} > 0 \Rightarrow \text{keep } k_2 \\ C_{k_2 \to i | M_i' \setminus \{k_2\}} = 0 \Rightarrow \text{remove } k_2 \end{cases}$$

$\vdots$

Finally, after considering all k_r with $1 \le r \le j$, the set M_i becomes N_i

Estimation of CSE

- Gaussian estimator…
- k-nearest neighbor estimator…
- symbolic CSE…

$$C_{J \to I \| K} \to \hat{C}_{J \to I \| K}$$

1)…parametric statistics…

2) non-parametric statistics, multi-dimensional random variables…

3) computational speed, robustness against observational noise, limited data demand…

…*quoting* J. Runge (June 18, cidnet14):…knn is good but some bias appears in the presence of short samples and large dimensions….

A. Kraskov, H. Stogbauer, and P. Grassberger, Phys. Rev. **E69**, 066138 (2004)
M. Vejmelka and M. Palus, Phys. Rev. **E77**, 026214 (2008)

Statistical (permutation) test

Question:

$$C_{J\to I|K} \to \hat{C}_{J\to I|K} : \begin{cases} \hat{C}_{J\to I|K} = 0 \ ? \\ \hat{C}_{J\to I|K} > 0 \ ? \end{cases}$$

$1-\theta$: **significance level**, $0 < (1-\theta) < 1$ $(\theta \approx 99\%)$
$\mathcal{F}$: **empirical cumulative distribution**

$$\boxed{\hat{C}_{J\to I|K} > 0 \Leftrightarrow \mathcal{F}(\hat{C}_{J\to I|K}) > \theta}$$

$$\boxed{\mathcal{F}(\hat{C}_{J\to I|K}) \overset{\text{def}}{=} \frac{\#\left\{\hat{C}^{(s)}_{J\to I|K} : \hat{C}^{(s)}_{J\to I|K} \leq \hat{C}_{J\to I|K}\right\}}{r}}$$

with $1 \leq s \leq r$

r-estimates: $\left\{\hat{C}^{(1)}_{J\to I|K}, \ldots, \hat{C}^{(r)}_{J\to I|K}\right\}$, with $10^3 \lesssim r \lesssim 10^4$

permutation $\pi : \{1,\ldots, T\} \to \{1,\ldots, T\}$

For $j = 1,\ldots,|J|$: $\left\{x_t^{(j)}\right\}_{t=1}^{T} \to \left\{y_t^{(j)}\right\}_{t=1}^{T} \overset{\text{def}}{=} \left\{x_{\pi(t)}^{(j)}\right\}_{t=1}^{T}$

Example: The oCSE approach & the repressilator

$$\begin{cases} \frac{dm_i}{dt} = -m_i + \frac{\alpha}{1+p_j^n} + \alpha_0 \\ \\ \frac{dp_i}{dt} = -\beta(p_i - m_i) \end{cases}$$

$i = lacl, tetR, cl$, and $j = cl, lacl, tetR$

synthetic biological oscillator network

dynamical variables:

p_i= concentration of the protein

m_i= concentration of mRNA

parameters:

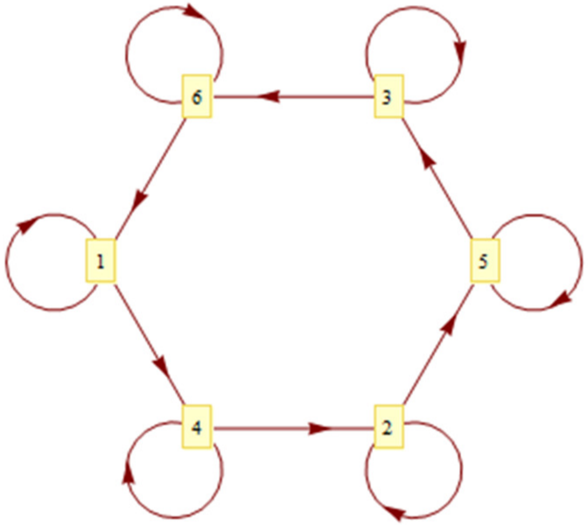
β=ratio of the protein decay rate to the mRNA decay rate

n=Hill coefficient

α_0=leakiness of the promotor

$\alpha+\alpha_0$=additional rate of transcription of the mRNA in the absence of inhibitor

J. Sun, C. Cafaro, and E. M. Bollt, *Entropy* **16**, 3416 (2014)



$$1 \equiv m_{lacI},\ 2 \equiv m_{tetR},\ 3 \equiv m_{cl}$$
$$4 \equiv p_{lacI},\ 5 \equiv p_{tetR},\ 6 \equiv p_{cl}$$

$$1 \to 1,\ 2 \to 2,\ 3 \to 3,\ 4 \to 4,\ 5 \to 5,\ 6 \to 6$$
$$1 \to 4,\ 2 \to 5,\ 3 \to 6,\ 4 \to 2,\ 5 \to 3,\ 6 \to 1$$

hypothesis: $n = 2,\ \alpha_0 = 0,\ \alpha = 10,\ \beta = 10^2$ $\Rightarrow$**equilibrium**: $\vec{x}_{\text{eq.}} = (2, 2, 2, 2, 2, 2)$

Theoretical Jacobian matrix at equilibrium: $J_{ij}^{(\text{theor.})} \overset{\text{def}}{=} \begin{pmatrix} * & 0 & 0 & 0 & 0 & * \\ 0 & * & 0 & * & 0 & 0 \\ 0 & 0 & * & 0 & * & 0 \\ * & 0 & 0 & * & 0 & 0 \\ 0 & * & 0 & 0 & * & 0 \\ 0 & 0 & * & 0 & 0 & * \end{pmatrix}$

$$\boxed{\begin{aligned} J_{ij} &\overset{\text{def}}{=} \partial_j f_i \\ \dot{x} &= f(x) \end{aligned}}$$

Problem statement: The objective of coupling inference is to identify the location of the nonzero entries of the Jacobian matrix through time series generated by the system near equlibrium

Given $\mathcal{I}_6^{\times 2} \equiv \mathcal{I}_6 \times \mathcal{I}_6 \overset{\text{def}}{=} \{1,\ldots,6\} \times \{1,\ldots,6\}$:

- **False negative**: infer nothing when there is something...

$$\boxed{\varepsilon_- \overset{\text{def}}{=} \frac{\text{card}\left\{(i,j) \in \mathcal{I}_6^{\times 2} \;:\; J_{ij}^{(\text{theor.})} \neq 0 \wedge J_{ij}^{(\text{numer.})} = 0\right\}}{\text{card}\left\{(i,j) \in \mathcal{I}_6^{\times 2} \;:\; J_{ij}^{(\text{theor.})} \neq 0\right\}}}$$

- **False positive**: infer something when there is nothing...

$$\boxed{\varepsilon_+ \overset{\text{def}}{=} \frac{\text{card}\left\{(i,j) \in \mathcal{I}_6^{\times 2} \;:\; J_{ij}^{(\text{theor.})} = 0 \wedge J_{ij}^{(\text{numer.})} \neq 0\right\}}{\text{card}\left\{(i,j) \in \mathcal{I}_6^{\times 2} \;:\; J_{ij}^{(\text{theor.})} = 0\right\}}}$$

- Aggregative discovery
- Progressive removal

Some details

- **Preliminary steps**:

i) The system starts at equilibrium $x^* = x_{\text{eq.}}$: $\dot{x}\big|_{x=x_{\text{eq.}}} = 0$;

ii) At time t, apply perturbation ξ to the system: $x^* \to x(t) = x^* + \xi(t)$;

iii) At time $t + \Delta t$, measure the rate of response η: $\eta = \frac{x(t+\Delta t)-x(t)}{\Delta t} = \frac{x(t+\Delta t)-x^*-\xi(t)}{\Delta t}$;

Repeat these steps L-times$\Rightarrow$ $\{\xi_l\}_{l=1}^{L}$ and $\{\eta_l\}_{l=1}^{L}$;

- **Parameters**:

i) $L =$ number of times the perturbation is applied;

ii) $\Delta t^{-1} =$ sample frequency;

iii) $\sigma^2 =$ variance of the Gaussian-distributed variable ξ.

- **Hypotheses**:

i) $\Delta t^{-1} \gg 1$;

ii) $\sigma \ll 1$;

- **Linearized dynamical system**:

For $x = x^* + \delta x$ with $\delta x \ll x^*$, we have:

$$\dot{x} = f(x) \rightarrow \frac{d(\delta x)}{dt} = Df(x^*)\delta x$$

For $\Delta t \ll 1$, we finally have a drive-response type of **linear Gaussian process**:

$$\frac{d\eta_l}{dt} = Df(x^*)\xi_l$$

To write the linear Gaussian process in a more convenient manner, we introduce the following definitions:

$$X_t^{(i)} = \begin{cases} \xi_t^{(i)}, \text{ if } 1 \leq i \leq 6 \\ \\ \eta_{t-1}^{(i-6)}, \text{ if } 7 \leq i \leq 12 \end{cases}.$$

Given this definition, the linear Gaussian process can be finally written as,

$$X_{t+1}^{(I)} = AX_t^{(J)}$$

where $A \stackrel{\text{def}}{=} Df(x^*)$, $I \stackrel{\text{def}}{=} \{7,..., 12\}$, and $J \stackrel{\text{def}}{=} \{1,..., 6\}$.

To be explicit, observe that

$$X_{t+1}^{(I)} = \left[X_{t+1}^{(7)},..., X_{t+1}^{(12)}\right] = \left[\eta_{t+1}^{(7)},..., \eta_{t+1}^{(12)}\right],$$

and,

$$X_t^{(J)} = \left[X_t^{(1)},..., X_t^{(6)}\right] = \left[\xi_t^{(1)},..., \xi_t^{(6)}\right],$$

with $A \stackrel{\text{def}}{=} Df(x^*) \rightarrow A_{ij} \stackrel{\text{def}}{=} \partial_j f_i(x^*)$,

$$A = \begin{pmatrix} A_{11} & \cdot & \cdot & \cdot & \cdot & A_{16} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{61} & \cdot & \cdot & \cdot & \cdot & A_{66} \end{pmatrix}.$$

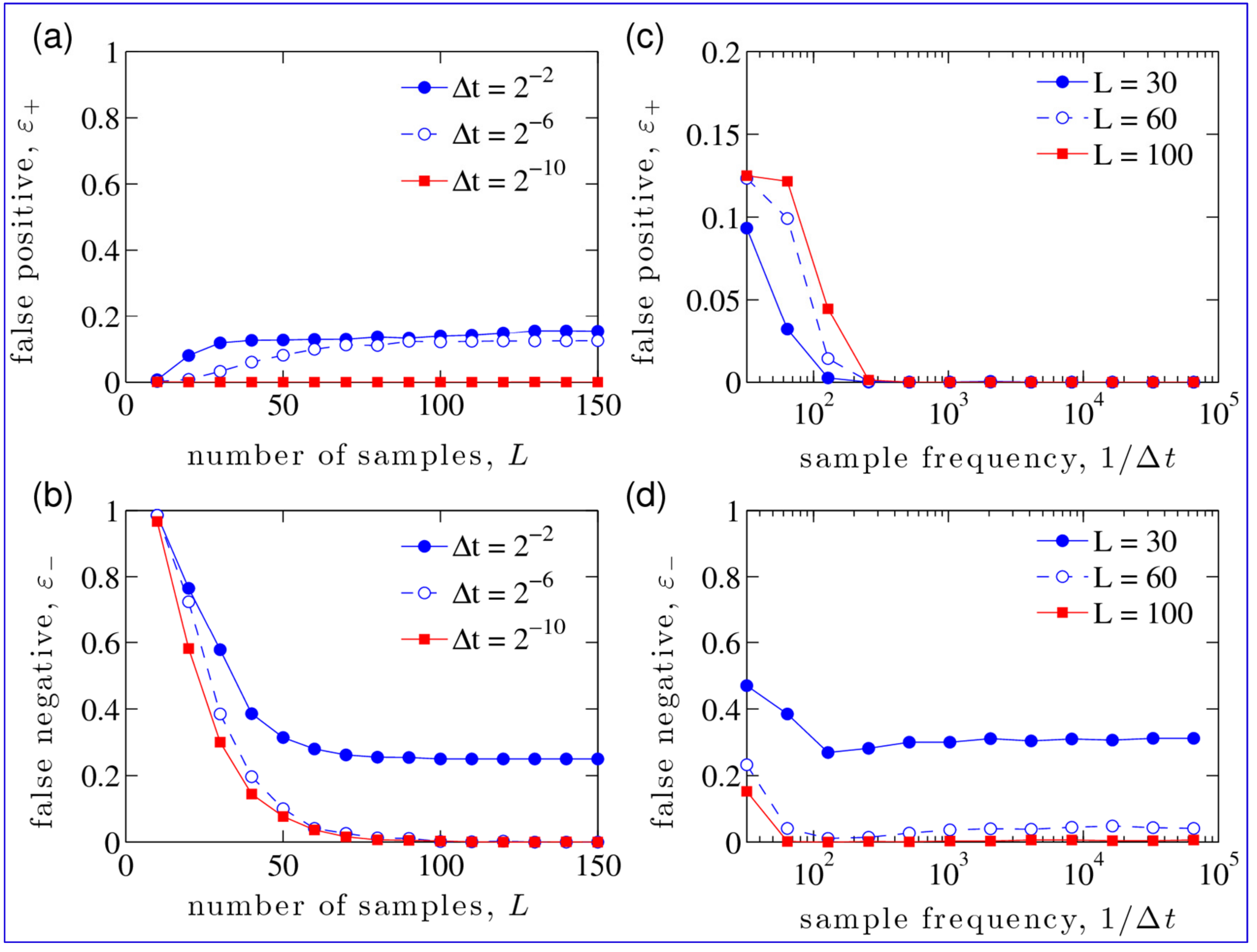
Therefore, the equation $X_{t+1}^{(I)} = AX_t^{(J)}$ becomes

$$\begin{cases} \eta_{t+1}^{(7)} = A_{11}\xi_t^{(1)} + \ldots + A_{16}\xi_t^{(6)} \\ \cdot \\ \cdot \\ \eta_{t+1}^{(12)} = A_{61}\xi_t^{(1)} + \ldots + A_{66}\xi_t^{(6)} \end{cases} .$$

…after some numerical work, we get…

$\varepsilon_{\pm} = \varepsilon_{\pm}(L, \Delta t, \sigma)$

σ^2 = variance of perturbation = 10^{-4}
L = number of samples
Δt^{-1} = sample frequency



Example: The oCSE principle & large scale networks

Network model

signed Erdos-Renyi random network with N≈200 nodes and Gaussian processes

$$X_t = AX_{t-1} + \xi_t$$

Numerical experiments parameters

1. p= connection probability
2. N=n=network size
3. T=sample size
4. ρ(A)= spectral radius (A= adjecency matrix)

Comparisons

- oCSE vs. conditional Granger: $\varepsilon_{\pm}(N)$ vs. N
- oCSE vs. TE: $\varepsilon_{\pm}(\rho(A))$ vs. ρ(A)

- **oCSE vs. Conditional Granger**: $\varepsilon\pm(N)$ vs. N

Np=10 (average degree, density of links)

$\rho(A)=0.8$ (spectral radius, information diffusion rate on networks)

T=200

- **oCSE vs. TE**: $\varepsilon\pm(\rho(A))$ vs. $\rho(A)$

N=200

Np=10

T=2000

Working hypotheses

1. permutation test with r=100 and θ=99%
2. Each data point is the average over 20 independent simulations of the network dynamics

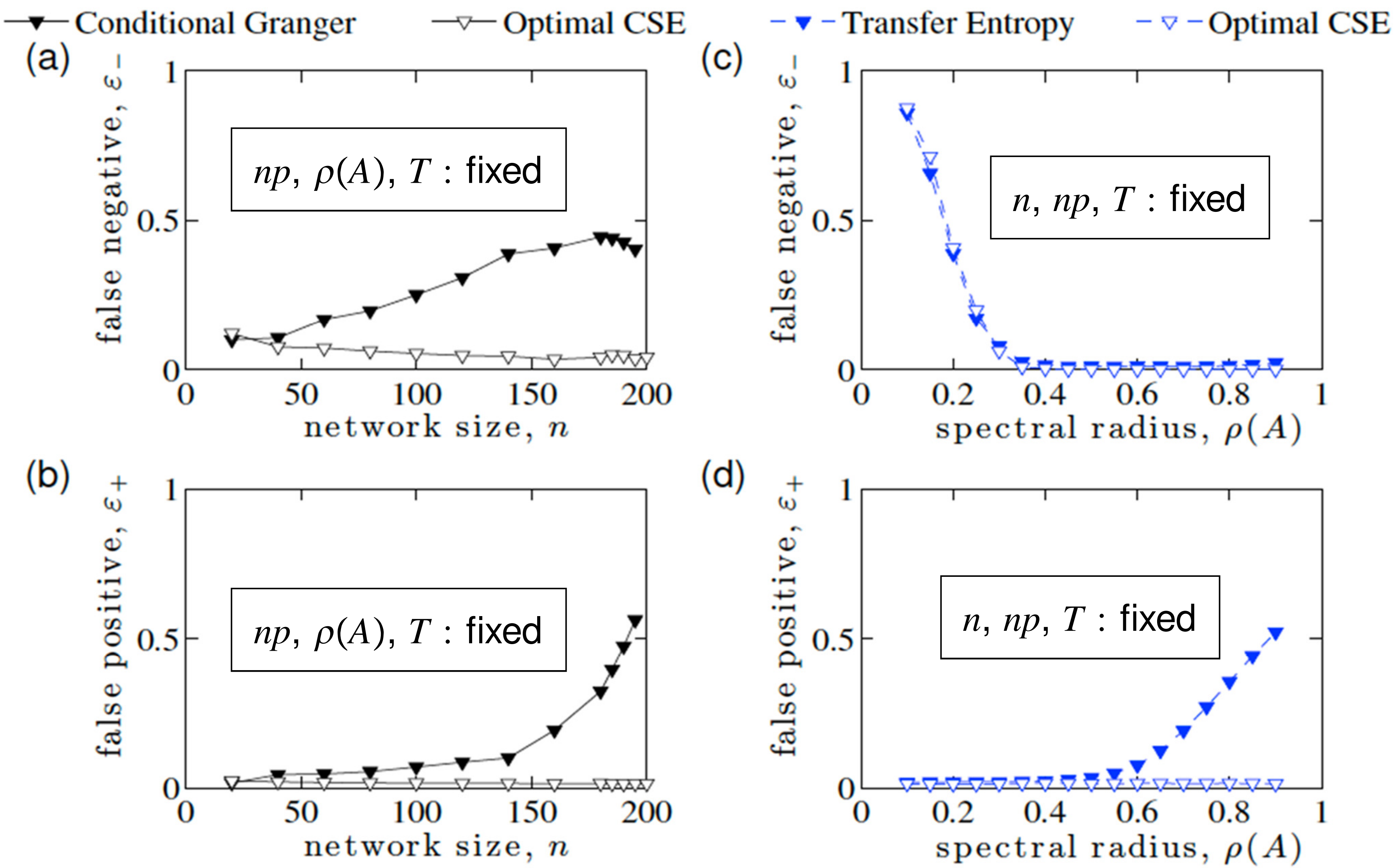


FIG. *Comparison of causal network inference approaches: Conditional Granger, Transfer Entropy, and Optimal CSE. The time series are generated Gaussian process*

On comparing causality inference measures…

Fair comparison:

- Compare measures estimated by means of the same estimation technique…
- Compare measures equally normalized…
- Compare measures that are constructed to capture the same features…

…(also) *inspired* by Xiaogeng Wan's Talk, June 23, cidnet14…

Summary: Part 2 (our approach)

…good news…

- CSE and the oCSE principle seem to be good tools for causal network inference
- Causal network inference based on the the oCSE principle seem to be especially immune to false positives
- The oCSE principle can be extended to arbitrary finite-order Markov processes

…selected challenges…

- loss of Markovianity (infinite memory)
- loss of stationarity
- accurate numerical estimates of CSE in large-scale dynamical systems (non-parametric methods, k-nearest neighbor estimator)
- distinguish anticipatory elements from causal ones (anticipatory dynamics in complex networks)

Part 3: Causal network inference and symbolic dynamics

…on-going research…

Conceptual and computational motivations

Concepts: Can we describe and understand the link between dynamical systems theory and information theory?

Computations/Numerics: What are desirable features of a good estimation method?

1. High computational speed
2. Robustness against observational noise
3. Limited data demands

…apply symbolic computational methods to the theory of dynamical systems on complex networks…

Why symbolic dynamics?

Fact: *the time-evolution of a physical system obtained by means of a classical measurement can be only approximately represented …*

- This approximate representation can be characterized in terms of a sequence of symbols, where *each symbol is the output of a measuring intrument at discrete times*…
- The range of possible symbols is finite since *any measuring instrument has limited resolution*…

Intuition: the symbolic framework is not as demanding on precision and amount of data

Dynamical trajectories, partitioning, symbol sequences

Main idea: …represent trajectories of dynamical systems by infinite length sequences using a finite number of symbols after partitioning the phase space in a convenient manner…

dynamical trajectory→**phase-space partition**→symbol sequence

…the phase-space partition is a key-point…

E. M. Bollt, T. Stanford, Y.-C. Lai, and K. Zyczkowski, Phys. Rev. Lett. **85**, 3524 (2000)
E. M. Bollt, T. Stanford, Y.-C. Lai, and K. Zyczkowski, Physica **D154**, 259 (2001)

Partitioning the phase space of a dynamical system: some facts

- A generating partition is necessary for a faithful symbolic representation of a dynamical system
- The partition is generating if every infinitely long symbol sequence created by the partition corresponds to a single point in phase-space (dynamical trajecteries uniquely defined by symbolic itineraries)

Remark: Any Markov partition is generating but the converse is generally false (generating partitions can be non-Markovian)

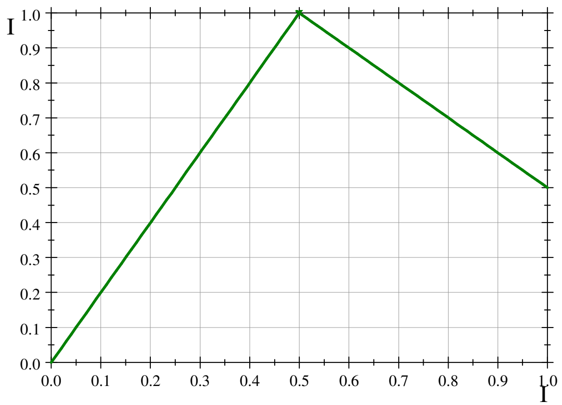
E. M. Bollt and N. Santitissadeekorn*, Applied and Computational Measurable Dynamics*, SIAM (2013)

What is a Markov partition?

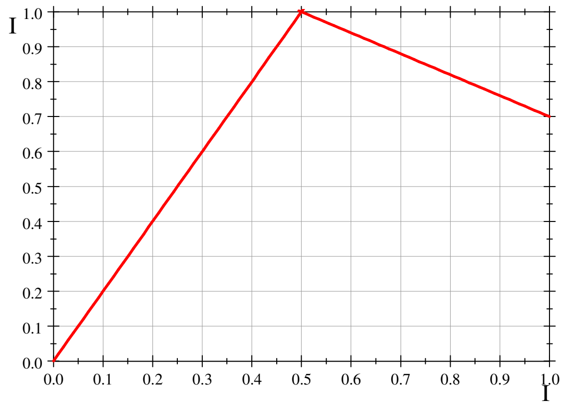
$$\tau\ :\ I \overset{\text{def}}{=} \left[a,b\right] \subset \mathbb{R}^1 \to I$$
$\mathcal{P} =$ partition of I given by points $a \equiv a_0 < \ldots < a_p \equiv b$, with $p \in \mathbb{N}$
$\tau_i = \tau_{|I_i} \to$ (union of intervals of $\mathcal{P}$)
$I_i = \left(a_{i-1}, a_i\right)$ with $i = 1,\ldots,p$

τ is a **Markov transformation** if τ_i is a homeomorphism from I_i onto a union of intervals of $\mathcal{P}$

$\mathcal{P}$ is said to be a **Markov partition** with respect to the function τ



…good…



…bad…

$I \overset{\text{def}}{=} [0, 0.5) \cup [0.5, 1]$

Symbolic description of a dynamical system

Consider a one-humped interval dynamical map with single critical point x_c and two-symbol partition {0,1}

$$(\textbf{dynamical map})\ f : [a,b] \to [a,b]$$

$$(\textbf{initial condition})\ [a,b] \ni x_0 \to \{x_0, f(x_0) = x_1, f^2(x_0) = x_2, \ldots\}\ (\textbf{orbit})$$

$$(\textbf{initial condition})\ x_0 \to \sigma_0(x_0).\sigma_1(x_0)\sigma_2(x_0)\ldots\ (\textbf{symbol sequence})$$

$$\boxed{\sigma_i(x_0) \overset{\text{def}}{=} \begin{cases} 0, \text{ if } f^i(x_0) < x_c \\ 1, \text{ if } f^i(x_0) > x_c \end{cases}}$$

$$(\textbf{Fullshift})\ \Sigma_2 \overset{\text{def}}{=} \{\sigma | \sigma = \sigma_0.\sigma_1\sigma_2\ldots, \text{ with } \sigma_i = 0 \text{ or } 1\}$$

Fact: The correspondence between the orbit of each initial condition x_0 of the map f and the infinite itinerary of 0s and 1s in the shift space Σ_2 can be regarded as a **homemomorphic change of coordinates**.

(**dynamical map**) $f : [a,b] \to [a,b]$

(**subshift**) $\Sigma_2' \subset \Sigma_2$

(**Bernoulli shift map**) $s_B : \Sigma_2' \to \Sigma_2'$, with $s_B(\Sigma_2') = \Sigma_2'$

$$(s_B(\sigma))_i \overset{\text{def}}{=} \sigma_{i+1}$$

(**homeomorphism**) $h : [a,b] - \bigcup_{i=0}^{\infty} f^{-i}(x_0) \to \Sigma_2'$

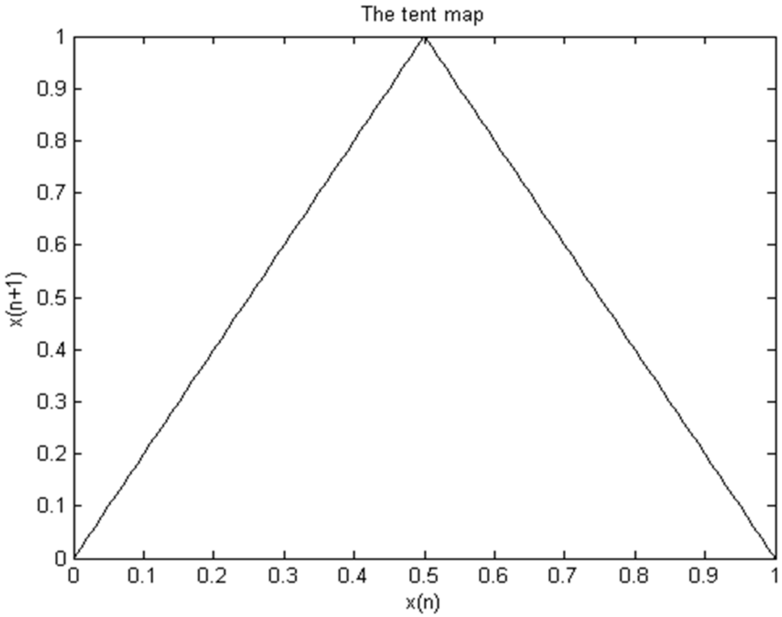
(**conjugacy**) $h \circ f = s_B \circ h$

Remark: conjugacy is the gold standard of equivalence used in dynamical systems theory when comparing two dynamical systems

Example: the tent map

$$f : [0, 1] \ni x \mapsto \begin{cases} 2(1-x), \text{ if } x \geq 0.5 \\ 2x, \text{ if } x < 0.5 \end{cases} \in [0, 1]$$

$$x_{n+1} \stackrel{\text{def}}{=} \begin{cases} 2(1-x_n), \text{ if } x_n \geq 0.5 \\ 2x_n, \text{ if } x_n < 0.5 \end{cases}$$



The symbolic dynamics indicated by the **generating partition** at $x_c = 0.5$ and by the equation,

$$\sigma_i(x_0) \stackrel{\text{def}}{=} \begin{cases} 0, \text{ if } f^i(x_0) < x_c \\ 1, \text{ if } f^i(x_0) > x_c \end{cases},$$

gives the full 2-shift Σ_2 on symbols $\{0, 1\}$.

$$x_c = 0.5$$

$$x_0 = 0.2,\ x_1 = f(x_0) = 0.4,\ x_2 = f^2(x_0) = 0.8,\ x_3 = f^3(x_0) = 0.4,\ldots$$

$$\sigma_0 = 0,\ \sigma_1 = 0,\ \sigma_2 = 1,\ \sigma_3 = 0,\ldots$$

$x_0 \mapsto \left[x_0, f(x_0), f^2(x_0),\ldots\right]$ (**dynamical trajectory**)

$x_0 \mapsto \sigma = \sigma_0.\sigma_1\sigma_2\ldots$ (**sequence of symbols**)

From dynamical systems to stochastic processes via symbolic dynamics

Formal steps

- Consider a **dynamical system**, $f:M\rightarrow M$
- Construct a **symbolic description** of such a system
- Introduce a **sequence of random varibales** defined on the symbol space for any randomly chosen initial condition x in M
- Define a **discrete-time stochastic process** in terms of the introduced sequence of random variables

More explicit steps

dynamical system: $f : M \to M$

measure space: $\left(M, \Sigma, \mu\right)$

partition: $M = \bigcup_{i=0}^{n} \bar{A}_i$, with $A_j \cap A_k = \varnothing$

symbol space (**alphabet**): $\mathcal{A} = \left\{0, 1, \ldots, n-1, n\right\}$

symbolic description: $\{x_t\}_{x_t \in M} \mapsto \{s_t\}$, with $s_t = i \in \mathcal{A}$ if $x_t \in A_i \subset M$

$$\mathfrak{s} : M \ni x \mapsto \mathfrak{s}(x) = \sum_{i=0}^{n} \chi_{A_i}(x) \in \mathcal{A}$$

$$\chi_{A_i}(x) \stackrel{\text{def}}{=} \begin{cases} i, \text{ if } x \in A_i \\ 0, \text{ if } x \notin A_i \end{cases}$$ (indicator function)

random variable: $X : \mathcal{A} \rightarrow \mathbb{R}$

measurable function: $(\mathcal{A}, \mathcal{F}, \mu) =$ measure space $\xrightarrow{\text{measurable function}}$ $(\mathbb{R}, \mathcal{B}(\mathbb{R})) =$ measurable space

$$\forall A \subset \mathcal{B}(\mathbb{R}) : X^{-1}(A) \overset{\text{def}}{=} \{\omega \in \mathcal{A} : X(\omega) \in A\}$$

Remark: If $\mu : \mathcal{F} \rightarrow [0, 1]$ with $\mu(\mathcal{A}) = 1$, then $(\mathcal{A}, \mathcal{F}, \mu) =$ probability space

- A dynamical system describes a discrete-time stochastic process defined by the sequence of random variables
- The support of a stochastic process is represented by a shift space regarded as the set of possible measurement outcomes of the process itself

$X(\mathfrak{s}(T^k(x))) = X_k(\omega)$, with $k = 0, \ldots, \infty$

$\mu(A_\sigma) = P(X_k = \sigma)$, with $\sigma \in \mathcal{A}$

SUMMARY

$\dot{x} = T(x)$

dynamics

$$x(t) = \begin{bmatrix} x_0 \\ \vdots \\ T^N(x_0) \end{bmatrix} \equiv \begin{bmatrix} x_0 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{partition}} \begin{bmatrix} \sigma_0 \\ \vdots \\ \sigma_N \end{bmatrix} \to \begin{bmatrix} X_0 \\ \vdots \\ X_N \end{bmatrix} \to X \stackrel{\text{def}}{=} \{X_k\}_{k=0}^{N}$$

sequence of symbols

sequence of random variables

stochastic process

$X(\mathfrak{s}(T^k(x))) = X_k(\omega)$, with $k = 0, \ldots, \infty$

$\mu(A_\sigma) = P(X_k = \sigma)$, with $\sigma \in \mathcal{A}$

Symbolic Construction of Causation Entropy

Time-series of a dynamical system: sequence of observations

Preliminary notations

graph: $\mathcal{G} = \mathcal{G}(\mathcal{V}, \mathcal{E})$

nodes: $|\mathcal{V}| = \bar{n}$-nodes

observable of single-node dynamics: $X^{(i)} \in \mathbb{R}^N$, with $1 \leq i \leq \bar{n}$

set of observables: $X^{(I)} = \left(X^{(i_1)}, \ldots, X^{(i_{|I|})}\right)$

time-series: $\left\{x_t^{(i_r)}\right\}_{t=1}^{T}$, with $1 \leq r \leq |I|$

$$\boxed{C_{X^{(J)}\to X^{(I)}|X^{(K)}} \overset{\text{def}}{=} \sum p\left(x_{t+1}^{(I)},\, \mathbf{x}_t^{(J)},\, \mathbf{x}_t^{(K)}\right) \log\left[\frac{p\left(x_{t+1}^{(I)} \,\middle|\, \mathbf{x}_t^{(J)},\, \mathbf{x}_t^{(K)}\right)}{p\left(x_{t+1}^{(I)} \,\middle|\, \mathbf{x}_t^{(K)}\right)} \right]}$$

CAUSATION ENTROPY

Explanatory notations

$x_{t+1}^{(I)} \overset{\text{def}}{=} \left[x_{t+1}^{(i_1)}, \ldots, x_{t+1}^{(i_{|I|})} \right]$: set of (scalar) time-series

$\mathbf{x}_t^{(K)} \overset{\text{def}}{=} \left[\mathbf{x}_t^{(k_1)}, \mathbf{x}_t^{(k_2)}, \ldots, \mathbf{x}_t^{(k_{|K|})} \right]$: set of reconstructed (vector) points

$\mathbf{x}_t^{(k_l)} \overset{\text{def}}{=} \left[x_t^{(k_l)}, x_{t-\tau_l}^{(k_l)}, \ldots, x_{t-(m_l-1)\tau_l}^{(k_l)} \right]$: reconstructed points

$\tau_l \equiv \tau_{x^{(k_l)}}$: time delay parameters

$m_l \equiv m_{x^{(k_l)}}$: embedding dimension parameters

Time-series of a stochastic process: sequence of symbols

Preliminary notations

nodes: $|\mathcal{V}| = \bar{n}$-nodes

N-dimensional stochastic components: $X^{(i)} \in \mathbb{R}^N$, with $1 \leq i \leq \bar{n}$

$X^{(i)} = \left(X_1^{(i)},..., X_N^{(i)}\right)$, with $X_l^{(i)} \in \mathbb{R}$, and $1 \leq l \leq N$

1-dimensional stochastic components $X_l^{(i)}$ with values in $\mathcal{A}_m \stackrel{\text{def}}{=} \{1,..., m\}$

N-dimensional stochastic components $X^{(i)}$ with values in $\mathcal{A}_m^N$

$(N|I|)$-dimensional stochastic components $X^{(I)}$ with values in $\mathcal{A}_m^{N|I|}$

An element of $\mathcal{A}_m^{N|I|}$ is a word of length $N|I|$ made of symbols in A_m

$$\omega \in \mathcal{A}_m^{N|I|} \Rightarrow \omega \stackrel{\text{def}}{=} \left(s_1, s_2,...., s_{N|I|-1}, s_{N|I|}\right) \equiv s_1 s_2 ... s_{N|I|-1} s_{N|I|}$$

$$Y_L^{(i)} \stackrel{\text{def}}{=} \left(X_1^{(i)},..., X_L^{(i)}\right) = L\text{-dimensional stochastic process, } L \leq N$$

$$\boxed{\hat{C}_{X^{(J)} \to X^{(I)} | X^{(K)}} \overset{\text{def}}{=} \sum p\left(\hat{x}_{t+1}^{(I)}, \hat{\mathbf{x}}_t^{(J)}, \hat{\mathbf{x}}_t^{(K)}\right) \log\left[\frac{p\left(\hat{x}_{t+1}^{(I)} \middle| \hat{\mathbf{x}}_t^{(J)}, \hat{\mathbf{x}}_t^{(K)}\right)}{p\left(\hat{x}_{t+1}^{(I)} \middle| \hat{\mathbf{x}}_t^{(K)}\right)} \right]}$$

CAUSATION ENTROPY

Explanatory notations

$\left\{\hat{x}_t^{(i)}\right\}$, sequence of symbols $\rightarrow$ $\left\{x_t^{(i)}\right\}$, sequence of observations

$\left\{\hat{x}_{t+1}^{(I)}\right\}$ = sequence of symbols formed by the set of scalar time-series $\left\{x_{t+1}^{(I)}\right\}$

$\left\{\hat{\mathbf{x}}_t^{(K)}\right\}$ = sequence of symbols formed by the set of reconstructed vector points $\left\{\mathbf{x}_t^{(K)}\right\}$

Summary: Part 3 (causal network inference and symbolic dynamics)

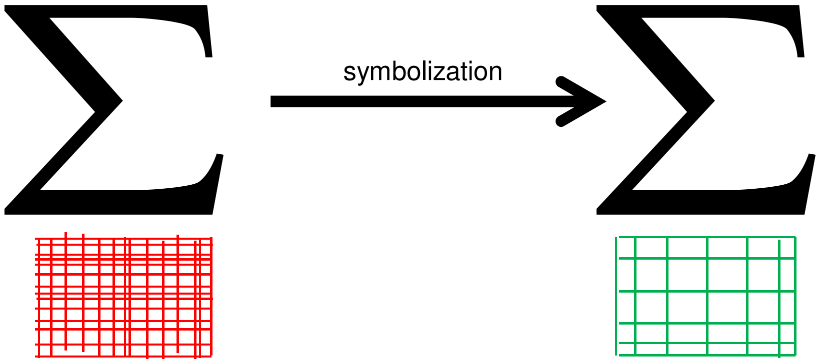
What did we do?

Initial step: Formal construction of the symbolic Causation Entropy

Relevance: Understand the link between dynamical systems and information-theoretic concepts

What is next?

Next steps: Practical estimation of symbolic Causation Entropy for both synthetic and real-world data



(Expected) Relevance:

- Less-demanding computational cost for numerical estimations
- Decrease of the negative effect of observational noise in masking the details of the data-structure

Summary of summaries (parts 1, 2, 3)

- Increase the *interaction* among theorists, computational scientists, and experimentalists

(from synthetic data to real-world data)

- Propose *good* information-theoretic causality measures

(universality and computability)

- Propose *reliable* estimation techniques

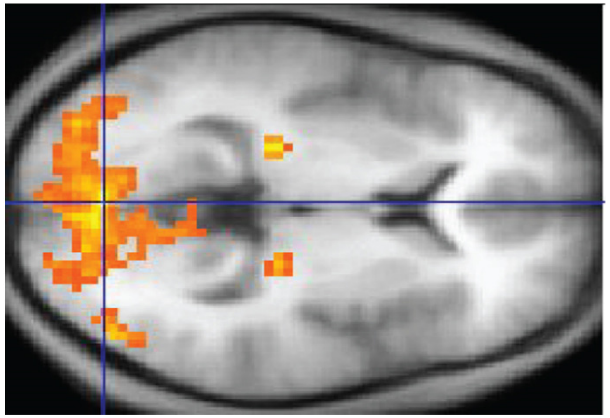
(speed and accuracy)

Causal network inference is important and challenging…

How? Hard work…no *escape*…

…On-going real-world applications of the CSE approach…

1. **Swarm-data**: information flow in a swarm of bugs…
2. **Neuroimaging data** (fMRI-functional magnetic resonance imaging): coupling structure between cerebral blood flow and neural activation…



References

The optimal Causation Entropy principle (**oCSE principle**)

- J. Sun and E. M. Bollt, Physica **D267**, 49 (2014)
- J. Sun, D. Taylor, and E. M. Bollt, arXiv:cs.IT/1401.7574 (2014)
- J. Sun, C. Cafaro, and E. M. Bollt, *Entropy* **16**, 3416 (2014)

Acknowledgements

This work is funded by Army Research Office of United States of America Grant No. 61386-EG

Thanks!