Computing Geometric Integrated Information

Jordan Snyder

Department of Mathematics
University of California, Davis

Student-run Math/Applied Seminar, 1 June 2016

Outline

Motivation

The whole is greater than the sum of the parts

Deconstruct causation between individuals in a composite system

Example systems/phenomena:

- Crowd/flocking behavior
- Pattern formation on networks
- Boolean networks
- The brain
- **Multi-layer networks**

Setting

Model class: time-sequence of random vectors

$$X = (X_i^t)_{i \in \mathcal{I}}^{t \in \mathbb{Z}}$$

- Information theory is the state of the art for quantifying complexity, and IT needs probability distributions
- Still reasonably general - free choice of $\mathcal{I}$ (should be finite) and meaning of RV's (should also be drawn from a finite alphabet $\mathcal{A}$)

Setting

Model class: time-sequence of random vectors

$$X = (X_i^t)_{i \in \mathcal{I}}^{t \in \mathbb{Z}}$$

Why this model class?

- Information theory is the state of the art for quantifying complexity, and IT needs probability distributions
- Still reasonably general - free choice of $\mathcal{I}$ (should be finite) and meaning of RV's (should also be drawn from a finite alphabet $\mathcal{A}$)

⚠ Assumptions ⚠

- Markov: $p(X^{t+1}|X^{\leq t}) = p(X^{t+1}|X^{t})$ - allows to consider only two time steps at once
- Stationary: $p(X^{t_1\,:\,t_2}) = p(X^{(t_1+s)\,:\,(t_2+s)})$ for all $s \in \mathbb{Z}$ - allows to average over time

⚠ Assumptions ⚠

- Markov: $p(X^{t+1}|X^{\leq t}) = p(X^{t+1}|X^{t})$ - allows to consider only two time steps at once
- Stationary: $p(X^{t_1\,:\,t_2}) = p(X^{(t_1+s)\,:\,(t_2+s)})$ for all $s \in \mathbb{Z}$ - allows to average over time

$\implies$ Upshot: $p(X^{t}, X^{t+1})$ encodes everything!

Possible Applications

- X_i^t is the opinion of a randomly chosen agent about topic i at time t
- X_i^t is the state (compromised/recovered) of a randomly chosen location with respect to the i^{th} infrastructure type
- X_i^t is the output of node i of a Boolean network at time t given a randomly chosen initial condition
- (current context) X_i^t is the state of a randomly chosen dyad on the i^{th} layer of a relational network at observation time t

Possible Applications

- X_i^t is the opinion of a randomly chosen agent about topic i at time t
- X_i^t is the state (compromised/recovered) of a randomly chosen location with respect to the i^{th} infrastructure type
- X_i^t is the output of node i of a Boolean network at time t given a randomly chosen initial condition
- (current context) X_i^t is the state of a randomly chosen dyad on the i^{th} layer of a relational network at observation time t

⚠: "randomly chosen" … ?

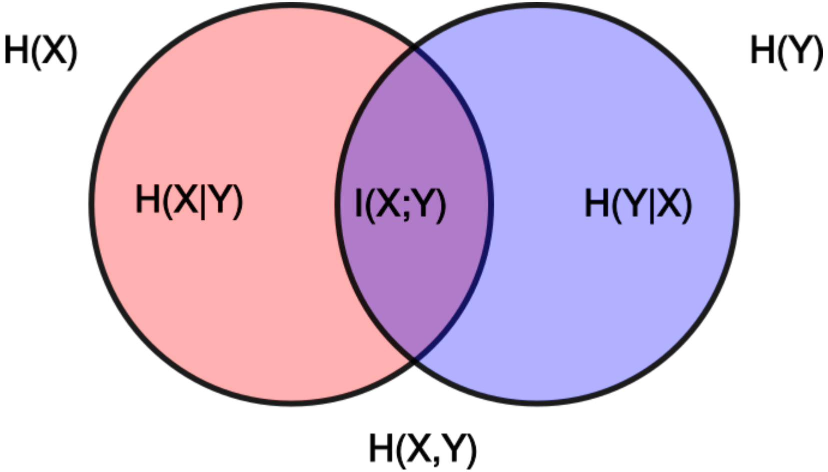
Entropy and Conditional Entropy

The *entropy* of a random variable X, denoted $H(X)$, measures the average uncertainty in the value of X. Mathematically,

$$H(X) = -\sum_{x \in \mathcal{A}} p(X = x) \log_2(p(X = x))$$

Conditional entropy $H(X|Y)$ says: given Y, how much uncertainty remains about X?

Information Venn Diagram



Kullback-Leibler Divergence

A (non-symmetric) information-theoretic measure of difference between two probability distributions. Mathematically,

$$D_{KL}(p\|q) = \sum_{z} p(z) \log_2 \left(\frac{p(z)}{q(z)} \right)$$

Heuristically, if p is reality and q is a model, $D_{KL}(p\|q)$ measures how much complexity is missed by the model.

Properties:

- $D_{KL}(p\|q) \geq 0$, and $D_{KL}(p\|q) = 0$ if and only if $p = q$
- Given p, the function $f(q) = D_{KL}(p\|q)$ is convex (strictly convex if $p(z) > 0$ for all z)

Previous Work

Measures that aim to quantify causal influences using conditional entropy:

- Transfer Entropy

$$TE_{i\to j} = H(X_j^{t+1}|X_{\neg i}^t) - H(X_j^{t+1}|X^t)$$

- Mutual Information

$$I(X^t;X^{t+1}) = H(X^{t+1}) - H(X^{t+1}|X^t)$$

- Stochastic Interaction

$$SI(X^t;X^{t+1}) = \sum_i H(X_i^{t+1}|X_i^t) - H(X^{t+1}|X^t)$$

Unified Perspective

These measures (and more) can all be expressed as

$$\min_{q \in \mathcal{M}} D_{KL}(p \| q)$$

for appropriate choice of null model class $\mathcal{M}$. They are:

- Transfer Entropy

$$\mathcal{M}_{i \to j} = \{q(X^t, X^{t+1}) | q(X_j^{t+1} | X^t) = q(X_j^{t+1} | X_{\neg i}^t)\}$$

- Mutual Information

$$\mathcal{M}_I = \{q(X^t, X^{t+1}) | q(X^t, X^{t+1}) = q(X^t) q(X^{t+1})\}$$

- Stochastic Interaction

$$\mathcal{M}_S = \{q(X^t, X^{t+1}) | q(X^{t+1} | X^t) = \prod_{i=1}^{m} q(X_i^{t+1} | X_i^t)\}$$

Visualizing Minimization

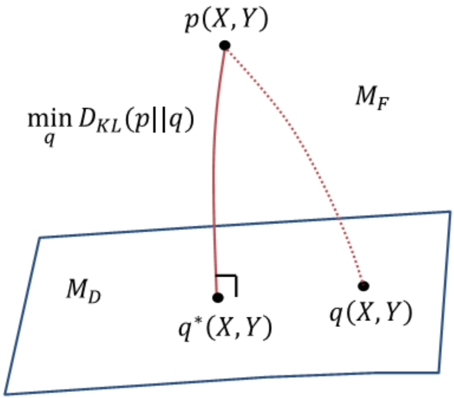


FIG. 1. Information geometric picture for minimizing the KL divergence between the full model $p(X,Y)$ and disconnected model $q(X,Y)$.

Some Minimizers have Closed Form

- Transfer Entropy ($i \to j$)

$$q^*(X^t) = p(X^t), \quad q^*(X_j^{t+1}|X^t) = p(X_j^{t+1}|X_{\neg i}^t)$$

$$q^*(X_{\neg j}^{t+1}|X^t, X_j^{t+1}) = p(X_{\neg j}^{t+1}|X^t, X_j^{t+1})$$

- Mutual Information

$$q^*(X^t, X^{t+1}) = p(X^t)p(X^{t+1})$$

- Stochastic Interaction

$$q^*(X^t, X^{t+1}) = p(X^t)\prod_i p(X_i^{t+1}|X_i^t)$$

... and some do not

We define (geometric) integrated information as

$$\Phi_G = \min_{q \in \mathcal{M}_G} D_{KL}(p \| q)$$

where

$$\mathcal{M}_G = \{q | q(X_i^{t+1} | X^t) = q(X_i^{t+1} | X_i^t), \forall i\}$$

… and some do not

We define (geometric) integrated information as

$$\Phi_G = \min_{q \in \mathcal{M}_G} D_{KL}(p \| q)$$

where

$$\mathcal{M}_G = \{q | q(X_i^{t+1} | X^t) = q(X_i^{t+1} | X_i^t), \forall i\}$$

Important feature:

$$0 \leq \Phi_G \leq I(X^t; X^{t+1}), \text{ since } \mathcal{M}_I \subset \mathcal{M}_G$$

In contrast, it is possible that $SI(X^t; X^{t+1}) > I(X^t; X^{t+1})$

Visualizing Null Model Classes

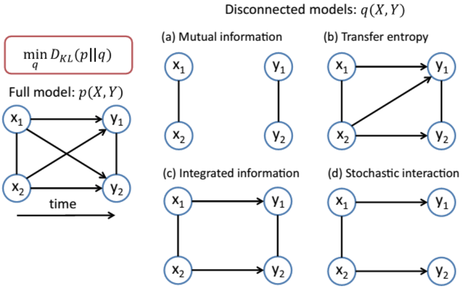
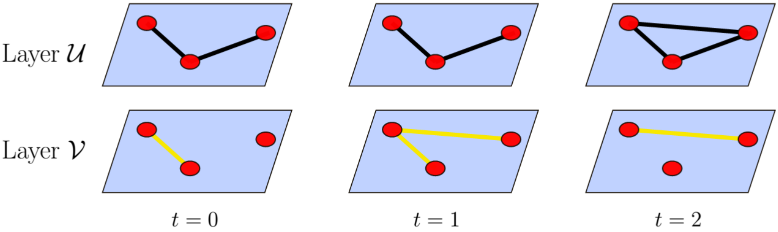


FIG. 2. Minimizing the KL divergence between the full and the disconnected model (a)-(d) lead to various information theoretic quantities; (a) Mutual information, (b) Transfer entropy, (c) Integrated information, and (d) Stochastic interaction. Constraints imposed on the disconnected model are graphically shown.

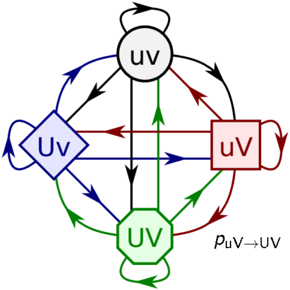
The quantity corresponding to each null model measures the importance of the *missing* arrows/lines

Multiplex Networks



Observing each dyad through time gives a time series of vectors - treat as a sample of $X = (X_i^t)$

The Multiplex Markov Chain



Vijayaraghavan et al. 2015

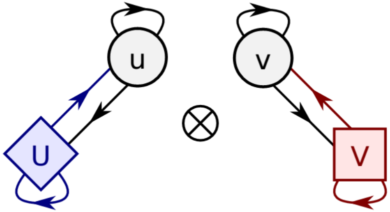
Our assumptions (⚠) mean that everything is captured by *transition probabilities* from one time step to the next

This creates a Markov Chain

Edge weights are of the form

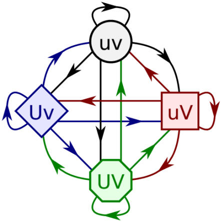
$$p_{\mathrm{uV}\to\mathrm{UV}} = p(X^{t+1} = \mathrm{UV} | X^t = \mathrm{uV})$$

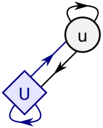
"The" Null Model

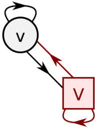


- Layers change state independently
- Joint transition probabilities are product of marginals

Measuring the Difference

$SI(X^t; X^{t+1}) = D_{KL}($ 

$\|$ 

$\otimes$ 

$)$

Integrated Information

No explicit expression for the minimizer.
Solve the constrained optimization problem

$$\min_{q \in \mathcal{M}_G} D_{KL}(p \| q)$$

In our case, $\mathcal{M}_G$ is an 11-dimensional subset of $\mathbb{R}^{16}$
Tools:

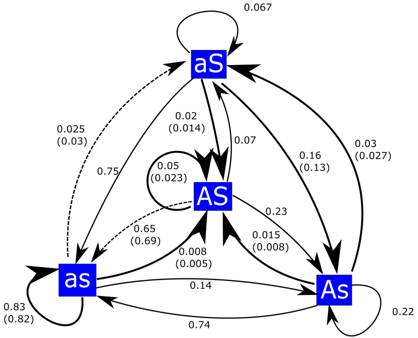
- Julia for Mathematical Programming (JuMP)
- Interior Point Optimization package (Ipopt)

Primate Data

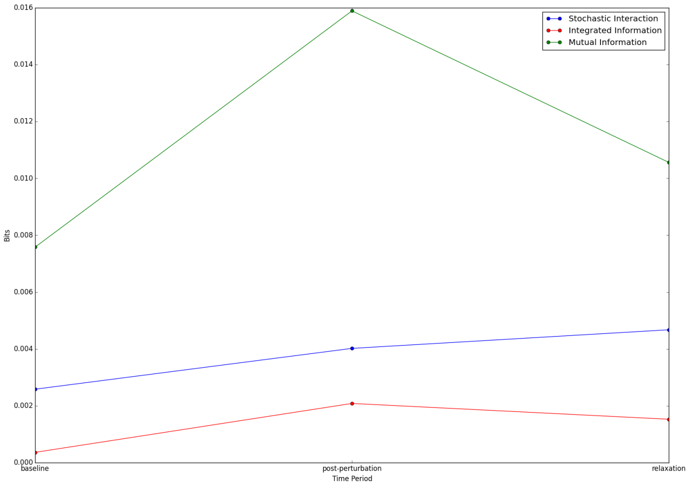
Week-long snapshots of grooming, aggression, and status signaling.

Focus on aggression/status interplay

Look for signatures of social perturbation



Response to Social Perturbation



References

Vikram S Vijayaraghavan, Pierre-André Noël, Zeev Maoz, and Raissa M D'Souza. Quantifying dynamical spillover in co-evolving multiplex networks. *Scientific reports*, 5, 2015.