Universal Estimation of Directed Information

Jiantao Jiao, Haim Permuter, Lei Zhao, Young-Han Kim and Tsachy Weissman

Information Systems Laboratory, Stanford University

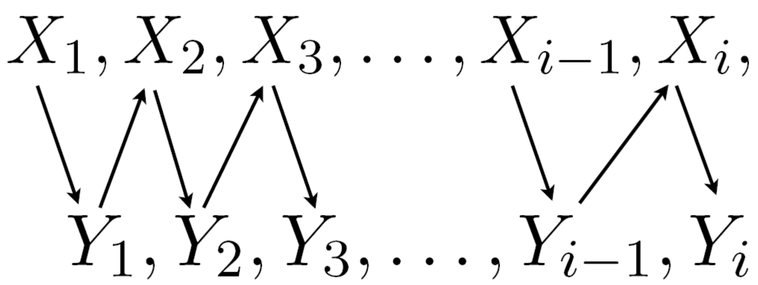
1 Introduction

Directed Information[2]

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1})$$

is an information-theoretic quantity defined for a pair of jointly distributed sequences, which is often a natural measure of the extent to which one sequence is relevant for causal inference on the other. It was found useful in measuring causal relevance in neurological, biological and financial data. The well-known causality measure---Granger Causality---is one of its manifestations under Gaussian and linear assumptions.

2 Mathematical Model



Causally conditional entropy:

$$H(Y^n \| X^n) = \sum_{i=1}^{n} H(Y_i | Y^{i-1}, X^i)$$

Mutual Information:

$$I(X^n; Y^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i | Y^{i-1}, X^n)$$

Directed Information:

$$I(X^n \to Y^n) = H(Y^n) - \sum_{i=1}^{n} H(Y_i | Y^{i-1}, X^i)$$

Causally conditional probability:

$$p(x^n \| y^n) = \prod_{i=1}^{n} p(x_i | x^{i-1}, y^i)$$

Conservation Law:

$$I(X^n; Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n)$$

3 Universal Sequential Probability Assignments

Definition: Probability assignment Q is universal if

$$\lim_{n\to\infty} \frac{1}{n} D(P_{X^n} || Q_{X^n}) = 0$$

for every stationary (true) probability measure P.

Remarks:

1. Universal Data Compression Algorithms can be used to derive Universal Sequential Probability Assignments;
2. In this work we use CTW (Context Tree Weighting) Algorithm[3].

4 Our goal: Good estimators for $\lim_{n\to\infty} \frac{1}{n} I(X^n \to Y^n)$

Four Estimators:

$$\hat{I}_1(X^n \to Y^n) \triangleq \hat{H}_1(Y^n) - \hat{H}_1(Y^n \| X^n),$$

$$\hat{I}_2(X^n \to Y^n) \triangleq \hat{H}_2(Y^n) - \hat{H}_2(Y^n \| X^n),$$

$$\hat{I}_3(X^n \to Y^n) \triangleq \frac{1}{n} \sum_{i=1}^{n} D(Q(y_i | X^i, Y^{i-1}) \| Q(y_i | Y^{i-1})),$$

$$\hat{I}_4(X^n \to Y^n) \triangleq \frac{1}{n} \sum_{i=1}^{n} D(Q(x_{i+1}, y_{i+1} | X^i, Y^i) \| Q(y_{i+1} | Y^i) Q(x_{i+1} | X^i, Y^i)),$$

$$\hat{H}_1(Y^n \| X^n) \triangleq -\frac{1}{n} \log Q(Y^n \| X^n),$$

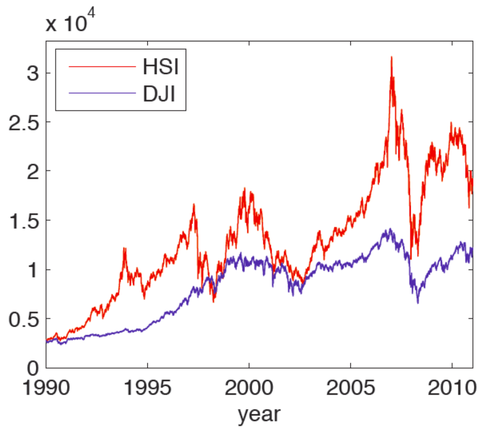
$$\hat{H}_2(Y^n \| X^n) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(Q(x_{i+1}, y_{i+1} | X^i, Y^i)),$$

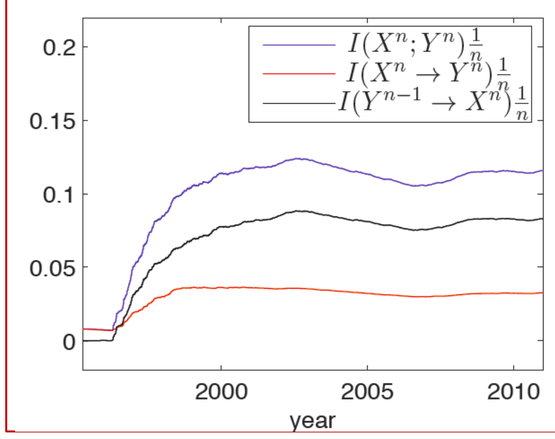
Different estimators have different merits: e.g. $\hat{I}_2$ is bounded, $\hat{I}_3$ and $\hat{I}_4$ are non-negative

Note: $f(\cdot)$ is a functional mapping joint distribution $p(x, y)$ to its conditional entropy $H(Y|X)$

5 Stock Market Experiments

Hang Seng Index (HSI) and Dow Jones Industrial Average (DJIA) 1990-2011





Conclusion:

There are significant causal influence from the DJIA to the HSI, but relatively low causal influence in the reverse direction from 1990 to 2011.

X: HSI
Y: DJIA

6 Main Results

Theorem

Let Q be a universal probability assignment and $(\mathbf{X}, \mathbf{Y})$ be jointly stationary ergodic. Then

$$\lim_{n\to\infty} \hat{I}_1(X^n \to Y^n) = \bar{I}(\mathbf{X} \to \mathbf{Y}) \quad \text{in } L_1$$

If Q is also pointwise universal then the limit holds almost surely as well.

Theorem

Let Q be the CTW sequential probability assignment. If $(\mathbf{X}, \mathbf{Y})$ is a jointly stationary ergodic aperiodic Markov process (of arbitrary order), then there exists a constant C_1 such that

$$\mathbb{E}\left|\hat{I}_1(X^n \to Y^n) - \bar{I}(\mathbf{X} \to \mathbf{Y})\right| \le C_1 n^{-1/2} \log n,$$

and $\forall \epsilon > 0$,

$$\left|\hat{I}_1(X^n \to Y^n) - \bar{I}(\mathbf{X} \to \mathbf{Y})\right| = o(n^{-1/2} (\log n)^{5/2+\epsilon}) \quad P\text{-a.s.}$$

"Theorem" This is essentially the best you can do.

References

[1] J. Jiao, H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, 'Universal Estimation of Directed Information', IEEE Trans. Inf. Theory, vol. 59, no. 10, Oct. 2013.
[2] H. Marko, "The bidirectional communication theory–a generalization of information theory," IEEE Trans. Commum., vol. COM-21, pp. 1345–1351, 1973
[3] F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context-tree weighting method: Basic properties," IEEE Trans. Inf. Theory, vol. 41, no. 3, 1995.