Causality Through Directed Information

Young-Han Kim

University of California, San Diego

SNU Institute for Research in Finance and Economics
April 30, 2013

Joint work with Jiantao Jiao (Stanford), Haim Permuter (Ben Gurion),
Tsachy Weissman (Stanford), and Lei Zhao (Jump Operations)

Supported in part by National Science Foundation (NSF),
US–Israel Binational Science Foundation (BSF), and BSF Bergmann Memorial Award

Related publications

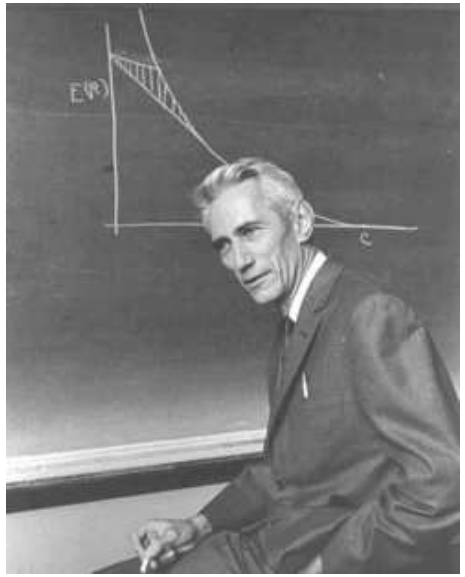
- Haim H. Permuter, Young-Han Kim, and Tsachy Weissman, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *IEEE Transactions on Information Theory,* vol. 57, no. 6, pp. 3248–3259, June 2011.

- Tsachy Weissman, Young-Han Kim, and Haim H. Permuter, "Directed information, causal estimation, and communication in continuous time," *IEEE Transactions on Information Theory,* vol. 59, no. 3, pp. 1271–1287, March 2013.

- Jiantao Jiao, Lei Zhao, Haim H. Permuter, Young-Han Kim, and Tsachy Weissman, "Universal estimation of directed information," to appear in *IEEE Transactions on Information Theory*, 2013.

- For more information, visit http://circuit.ucsd.edu/~yhk

Shannon's information measures (1948)



- Entropy: "uncertainty in a random variable X"

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)}$$

- Mutual information: "information about X provided by Y"

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

- Relative entropy (Kullback–Leibler 1951): "distinction between p and q"

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Where do they come from?

- Mathematical communication theory (Shannon 1948)
  - Fundamental limits on communication and compression
  - Probability theory and statistics

- Axiomatic definitions (Aczél–Daróczy 1975)
  - "Reasonable" properties for information measures
  - Functional equations: $f(p \times q) = f(p) + f(q) \;\Rightarrow\; f \cong H$

- How about finance and economics?

Gambling in horse races

- Horses: $1, 2, \ldots, m$
- Odds: $o(1), o(2), \ldots, o(m)$ (say, $o(x) \equiv m$)
- Win probabilities: $p(1), p(2), \ldots, p(m)$

Optimal gambling

- Bets: $b(1), b(2), \ldots, b(m)$
  - No short: $b(x) \geq 0, x = 1, 2, \ldots, m$
  - No margin: $\sum_x b(x) = 1$
  - In other words, $b(x)$ lies in the probability simplex

- Payoff: If horse x wins (with probability $p(x)$), then \$1 turns into $\$b(x)o(x)$

Question

How should we choose our portfolio $b(x)$?

Kelly gambling and log-optimal portfolio

- Kelly (1956), "A new interpretation of information rate":

$$b^*(x) = p(x)$$

- Maximize $\mathsf{E}[\log(b(X)o(X))]$
  - Logarithmic utility
  - Growth rate optimality
  - Competitive optimality (Bell–Cover 1980)
  - Other properties (MacLean–Thorp–Ziemba 2011)

- Optimal growth rate:

$$W^*(X) = \max_{b(x)} \mathsf{E}[\log(b(X)o(X))] = \mathsf{E}(\log o(X)) - H(X)$$

With $o(x) \equiv m$,

$$W^*(X) = \log m - H(X)$$

Entropy

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)} = \mathsf{E}\left[\log \frac{1}{p(X)}\right]$$

- Amount of randomness (information, uncertainty) in X
- Fundamental limit on lossless compression (Shannon 1948)
- Can be generalized to measures other than the counting measure

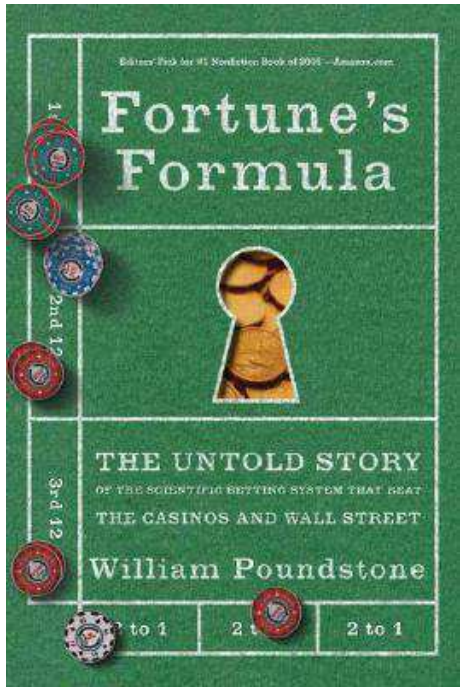
- Conditional entropy:

$$H(X|Y) = \sum_{x,y} p(x,y) \log \frac{1}{p(x|y)} = \mathsf{E}\left[\log \frac{1}{p(X|Y)}\right]$$

Gambling with side information

- Side information Y about the horse race outcome X
- Bets: $b(x|y), x = 1, 2, \ldots, m$
- Kelly gambling: $b^*(x|y) = p(x|y)$
- Optimal growth rate:

$$W^*(X|Y) = \max_{b(x|y)} \mathsf{E}[\log(b(X|Y)o(X))]$$
$$= \mathsf{E}(\log o(X)) - H(X|Y)$$



Value of side information (Kelly 1956)

$$\Delta W = W^*(X|Y) - W^*(X) = I(X;Y)$$

Mutual information

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- Amount of information about X provided by Y (and vice versa)
  - For a general stock market (Barron–Cover 1988): $\Delta W \leq I(X;Y)$
- Fundamental limit on communication (Shannon 1948)
- Fundamental limit on lossy compression/quantization (Shannon 1959)
- Can be generalized to any pair of random objects
- Conditional mutual information:

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

Repeated gambling in horse races with memory

- Win probabilities: $p(x_1), p(x_2|x_1), p(x_3|x_1, x_2), \ldots, p(x_n|x^{n-1})$
- Odds: $o(x_i) \equiv m$
- Bets: $b(x_1), b(x_2|x_1), b(x_3|x_1, x_2), \ldots, b(x_n|x^{n-1})$
- Kelly gambling: $b^*(x_i|x^{i-1}) = p(x_i|x^{i-1}), i = 1, 2, \ldots$
- Optimal growth rate:

$$W^*(X^n) = \log m - \frac{1}{n} H(X^n) = \log m - \frac{1}{n} \sum_{i=1}^{n} H(X_i|X^{i-1})$$

- If the horse race process $\{X_n\}$ is stationary ergodic, then
  - $(1/n)H(X^n) \to H^*(X)$
  - $W^*(X^n) \to W^*(X)$
  - wealth $\doteq 2^{nW^*}$ almost surely (Shannon 1948, McMillan 1953, Breiman 1957)

Gambling with causal side information

- Side information $Y_1, Y_2, \ldots$
- Bets: $b(x_i|x^{i-1}, y^i)$
- Kelly gambling: $b^*(x_i|x^{i-1}, y^i) = p(x_i|x^{i-1}, y^i), i = 1, 2, \ldots$
- Optimal growth rate:

$$W^*(X^n\|Y^n) = \log m - \frac{1}{n}\sum_{i=1}^{n} H(X_i|X^{i-1}, Y^i) = \log m - \frac{1}{n}H(X^n\|Y^n)$$

- If the $\{(X_n, Y_n)\}$ is stationary ergodic, then $(1/n)H(X^n\|Y^n) \to H^*(X\|Y)$

Value of causal side information (Permuter–K–Weissman 2011)

$$\Delta W = W^*(X^n\|Y^n) - W^*(X^n) = \frac{1}{n}(H(X^n) - H(X^n\|Y^n)) = \frac{1}{n}I(Y^n \to X^n)$$

Directed information

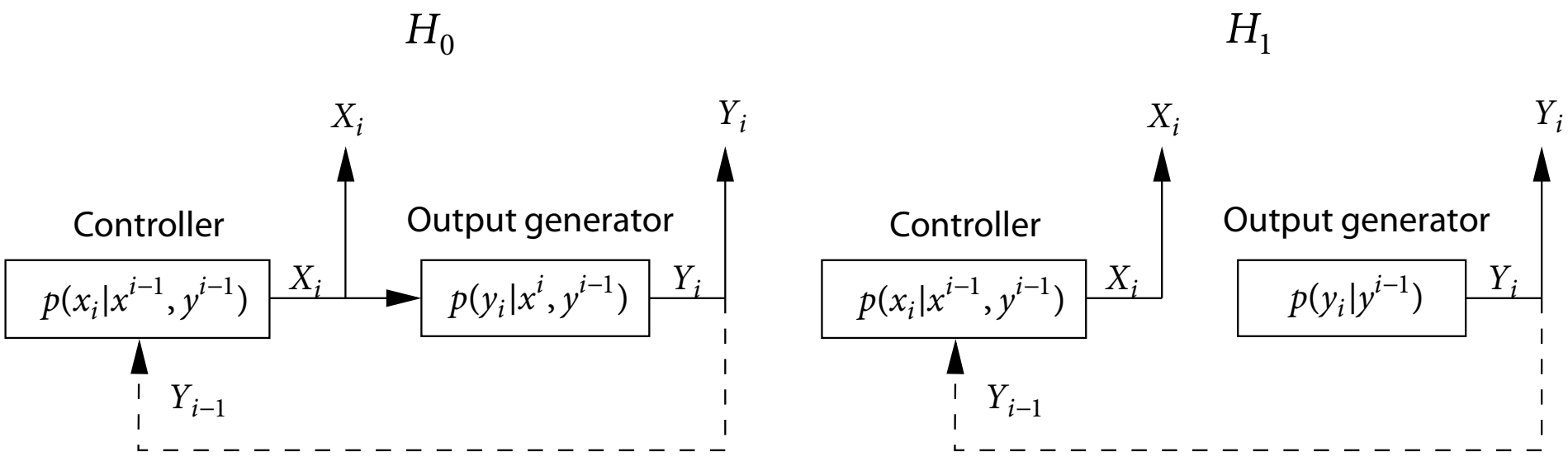
$$I(Y^n \to X^n) = H(X^n) - H(X^n \| Y^n) = \sum_{i=1}^{n} I(X_i; Y^i | X^{i-1})$$

- Amount of information about X causally provided by Y
  - For a general stock market: $\Delta W \leq (1/n) I(Y^n \to X^n)$
- Arrow of time: directed and asymmetric

$$I(Y^n \to X^n) \neq I(X^n \to Y^n)$$

- Fundamental limit on feedback communication (Tatikonda–Mitter 2009, K 2008, Permuter–Weissman–Goldsmith 2009)
- Can be generalized to continuous time (Weissman–K–Permuter 2013)

Test for causal dependence



- Type-I and type-II error probabilities: $\alpha = \mathrm{P}(\mathcal{A}^c|H_0)$, $\beta = \mathrm{P}(\mathcal{A}|H_1)$

Chernoff–Stein lemma for the causal dependence test

$$\beta^* = \min_{\mathcal{A} \subseteq \mathcal{X}^n \times \mathcal{Y}^n : \alpha < \epsilon} \beta \doteq 2^{-I(X^n \to Y^n)}$$

Brief history

- Marko (1973), "The bidirectional communication theory: A generalization of information theory"
  - Direction of information flow for mutually coupled statistical systems
  - Cybernetics: Group behavior with monkeys
- Massey (1990), "Causality, feedback, and directed information"

Relationship to other notions for causality

- Granger causality (Granger 1969, Geweke 1982):

$$G(X^n \to Y^n) = \sum_{i=1}^{n} \log \frac{\mathrm{LMMSE}(Y_i | Y_{i-p}^{i-1})}{\mathrm{LMMSE}(Y_i | Y_{i-p}^{i-1}, X_{i-p}^{i})}$$

  - The higher $G(X^n \to Y^n)$ is, the more X influences Y
  - If $\{(X_n, Y_n)\}$ is Gauss–Markov of order p, then

$$I(X^n \to Y^n) \equiv G(X^n \to Y^n)$$

- Transfer entropy (Schreiber 2000):

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

  - The higher $T_i(X \to Y)$ is, the more X influences Y (with one step delay)
  - If $\{(X_n, Y_n)\}$ is stationary, then

$$\frac{1}{n} I(X^{n-1} \to Y^n) \to T(X \to Y)$$

Causal conditioning

- Causally conditional probability (Kramer 1998):

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

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

- Causally conditional entropy:

$$H(Y^n \| X^n) = -\mathsf{E}[\log p(Y^n \| X^n)],$$

$$H(Y^n \| X^{n-1}) = -\mathsf{E}[\log p(Y^n \| X^{n-1})]$$

Chain rules

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

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

Properties of directed information

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

- $I(X^n \to Y^n) \leq I(X^n; Y^n)$
- $I(X^n \to Y^n) = I(X^n; Y^n)$ if $p(x^n \| y^{n-1}) = p(x^n)$
- $I(X^n \to Y^n) = I(X^n; Y^n) = nI(X; Y)$ if $\{(X_n, Y_n)\}$ is IID

Conservation law

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

- Measure of causal influence

Universal estimation of directed information

- In reality, the probability distribution may not be known or may not even exist

Something out of nothing

- Can we perform as if the distribution were known?
- Can we perform as well as the best estimator in a given class?

- Answer: Yes! (Jiao–Zhao–Permuter–K–Weissman 2013)

Universal probability assignments

- Probability assignment: $q(x^n)$
- Sequential probability assignment: $q(x_1), q(x_2|x_1), q(x_3|x_1, x_2), \ldots, q(x_n|x^{n-1})$
- Probability assignment q is universal if

$$\lim_{n\to\infty} \frac{1}{n} D(p(x^n)\|q(x^n)) = 0$$

for every stationary distribution p

- Probability assignment q is pointwise universal if

$$\limsup_{n\to\infty} \frac{1}{n} \log \frac{p(X^n)}{q(X^n)} \leq 0 \quad p\text{–a.s.}$$

for every stationary ergodic distribution p

- (Pointwise) universal probability assignments
  - Compression-based approaches: Ziv–Lempel (1978), Willems–Shtarkov–Tjalkens (1995)
  - Ergodic theoretic approaches: Ornstein (1978), Morvai–Yakowitz–Algoet (1997)

Algorithm 1

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

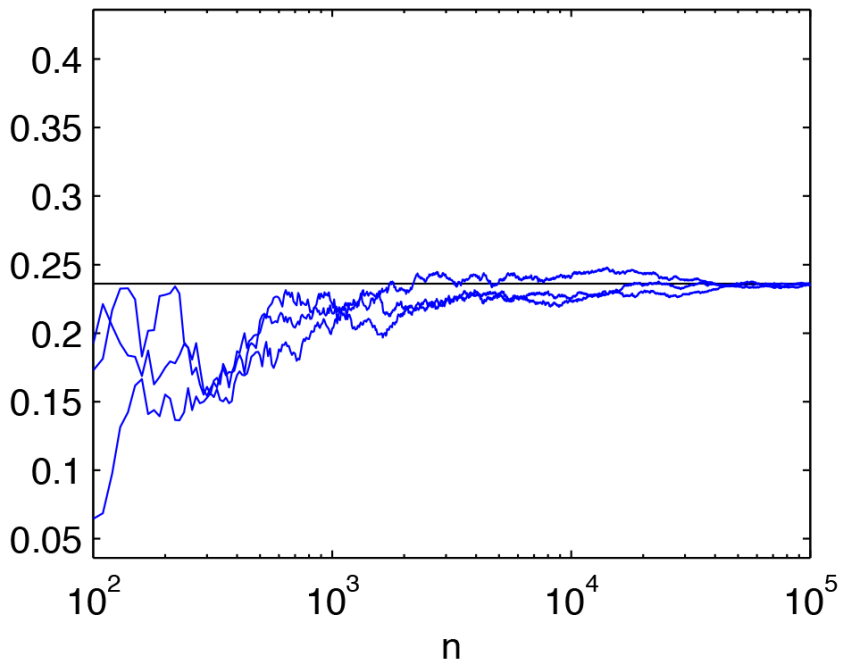
- $\hat{H}_1(Y^n) = -\frac{1}{n} \log q(Y^n)$ and $\hat{H}_1(Y^n \| X^n) = -\frac{1}{n} \log q(Y^n \| X^n)$

☺ Consistency (almost sure and L_1 convergence)

☺ Essentially optimal convergence rate

☹ Erratic for small n

☹ Unbounded support



Algorithm 2

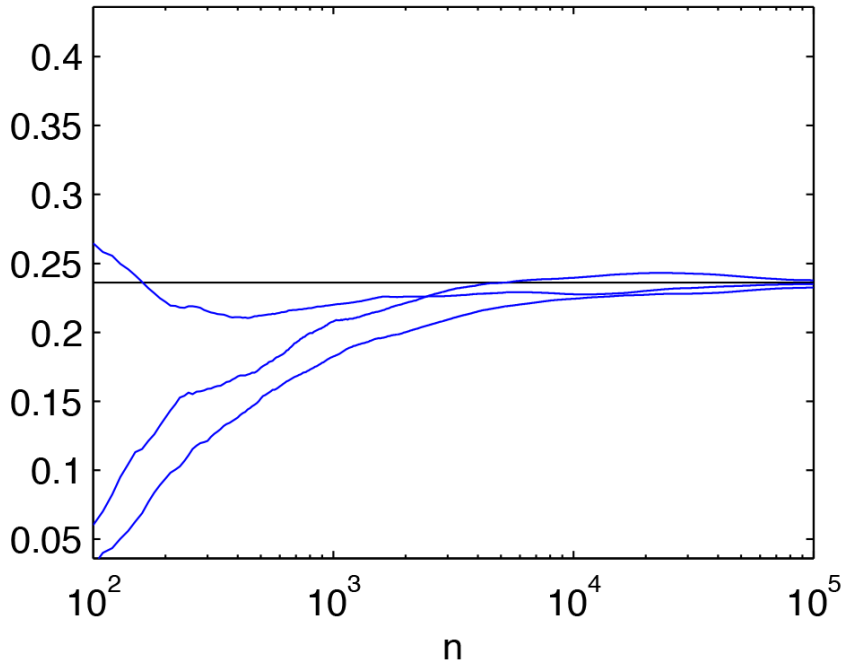
$$\hat{I}_2(X^n \to Y^n) = \hat{H}_2(Y^n) - \hat{H}_2(Y^n \| X^n)$$

- $\hat{H}_2(Y^n) = \frac{1}{n}\sum_{i=1}^{n} H(q(y_i|Y^{i-1}))$ and $\hat{H}_2(Y^n\|X^n) = \frac{1}{n}\sum_{i=1}^{n} H(q(y_i|x_i, X^{i-1}, Y^{i-1}))$

☺ Similar convergence rate as $\hat{I}_1$

☺ Smooth and bounded support

☹ Can be negative

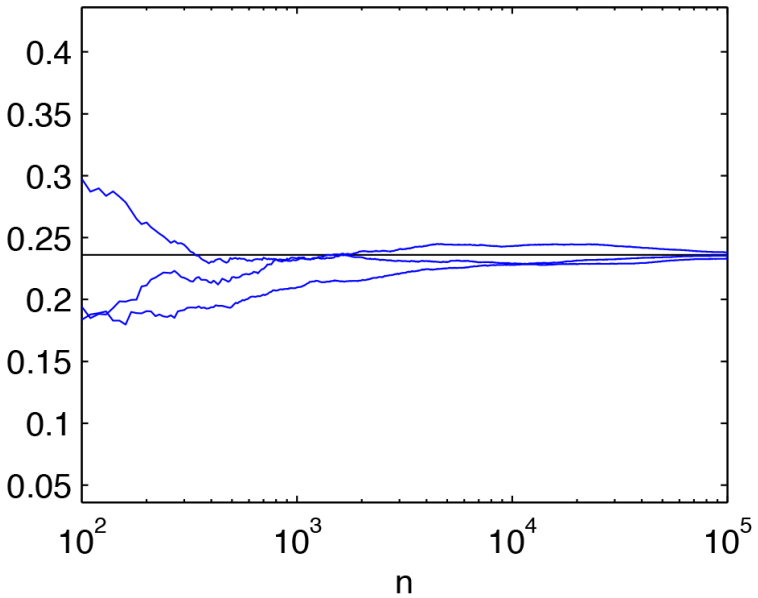


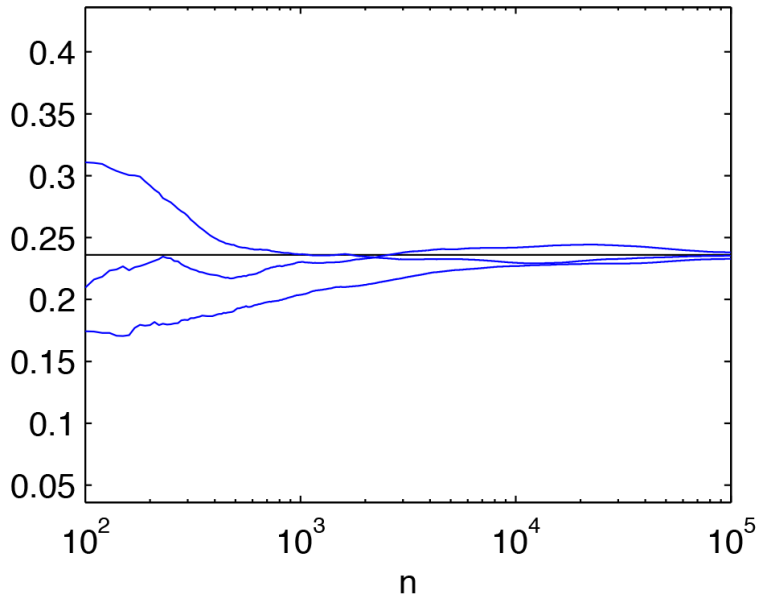
Algorithms 3 and 4

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

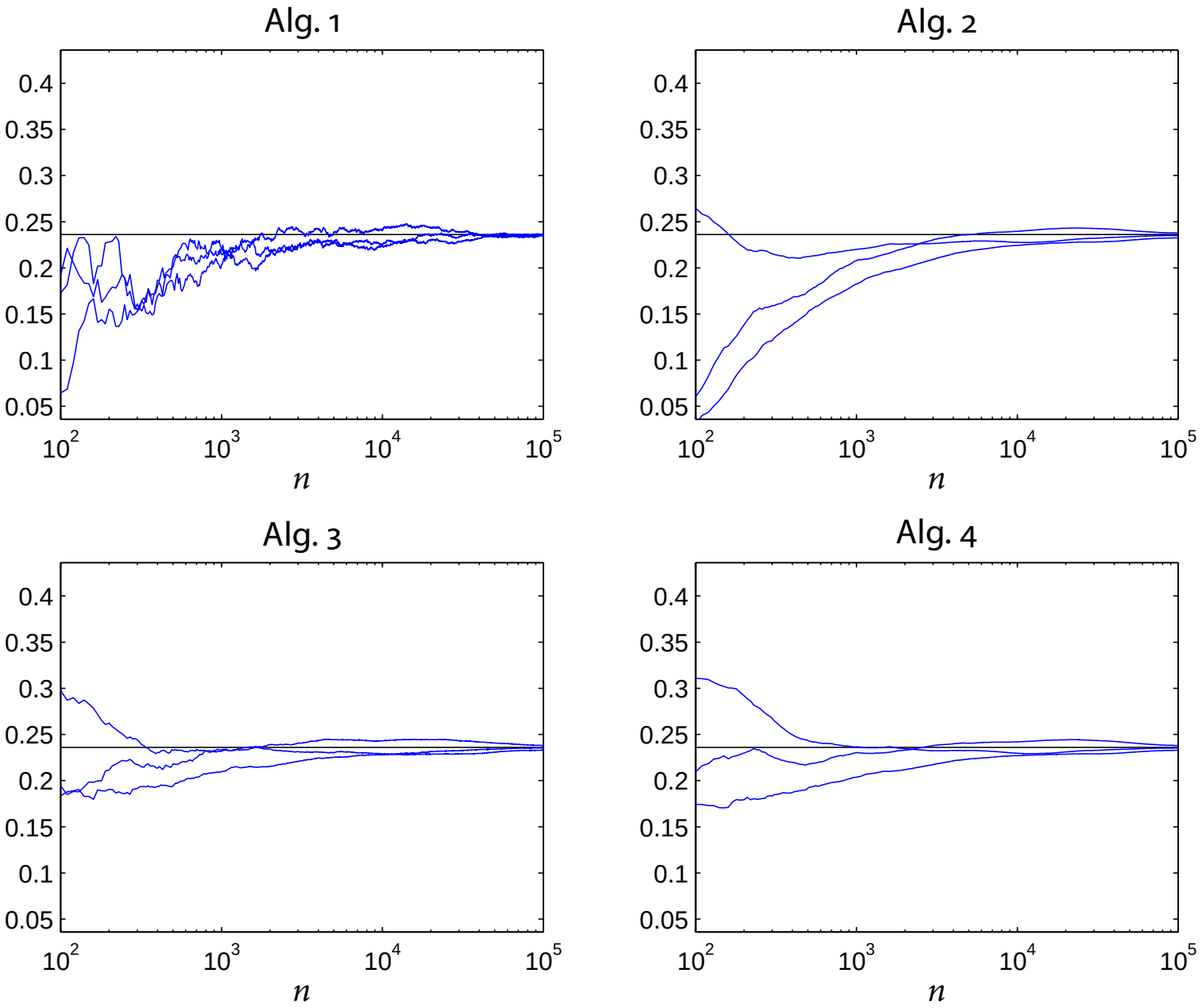
$$\hat{I}_4(X^n \to Y^n) = \frac{1}{n}\sum_{i=1}^{n} D(q(x_i, y_i|X^i, Y^{i-1})\|q(y_i|Y^{i-1})q(x_i|X^i, Y^i))$$

- ☺ Smooth, nonnegative, and bounded support
- ☹ Weaker performance guarantee than $\hat{I}_1$ and $\hat{I}_2$

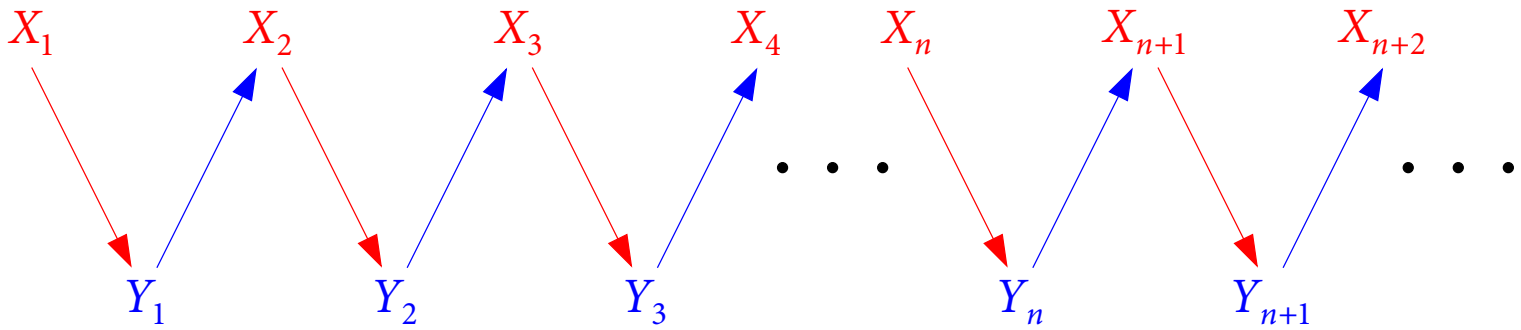




Performance comparison



Causal influence

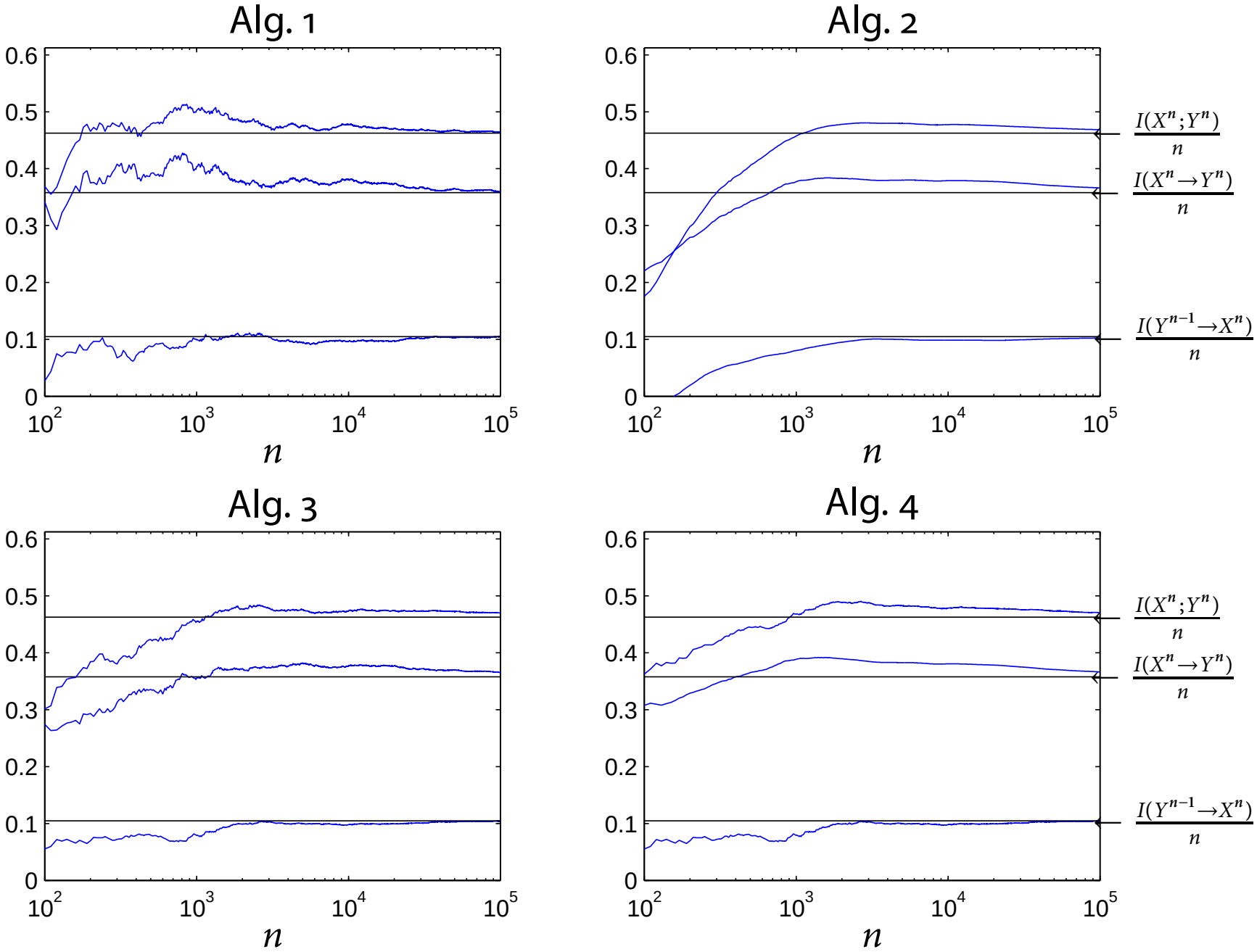


Question

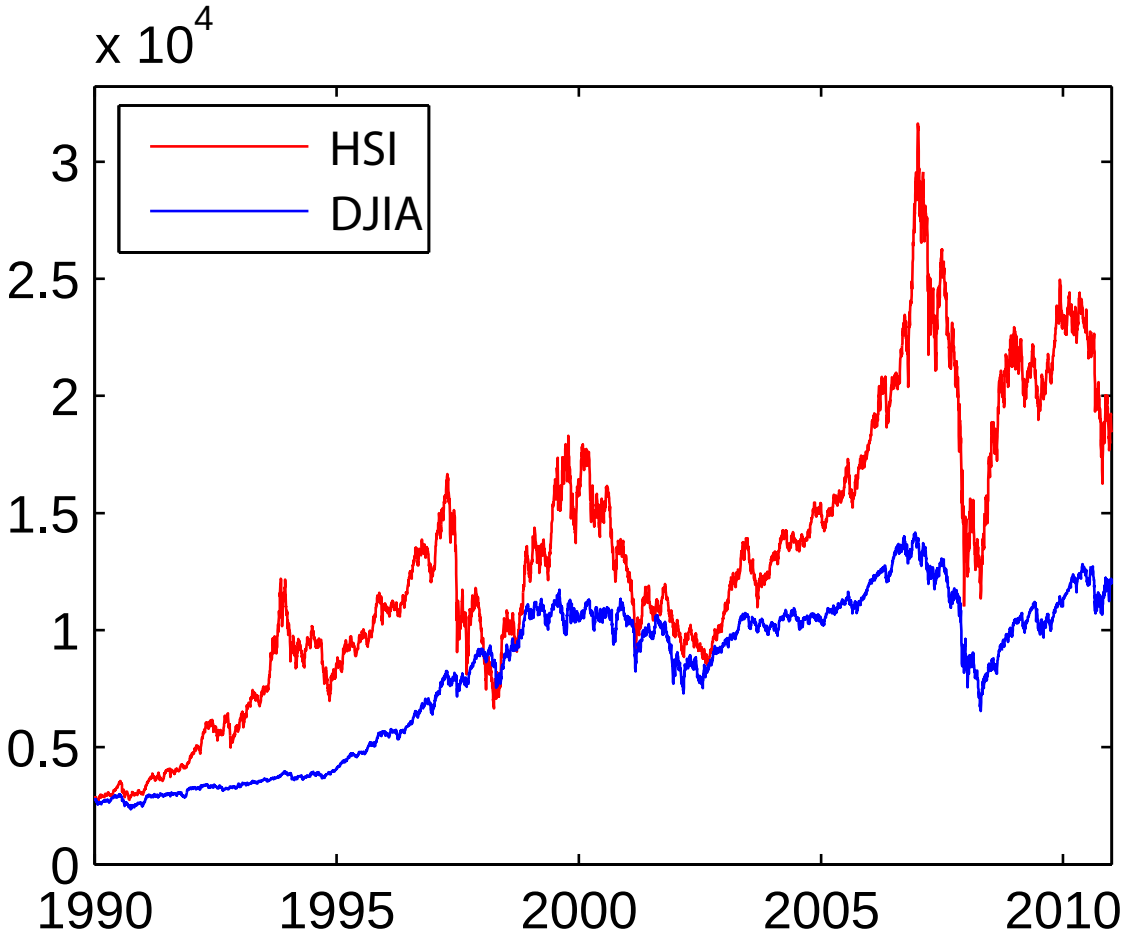
Which process influences the other?

- Conversation law: $I(X^n; Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n)$
- If $I(X^n \to Y^n) \gg I(Y^{n-1} \to X^n)$, then X causes Y
- If $I(X^n \to Y^n) \ll I(Y^{n-1} \to X^n)$, then Y causes X

Causal influence



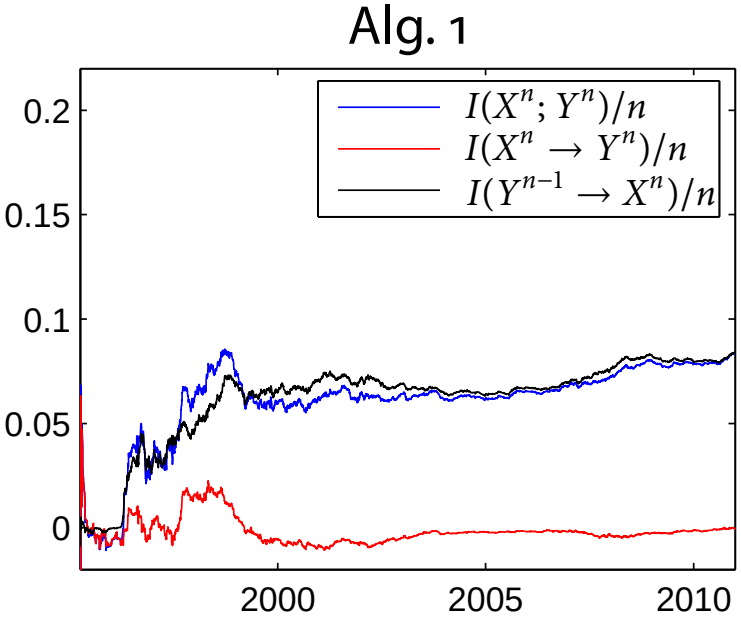
HSI versus DJIA

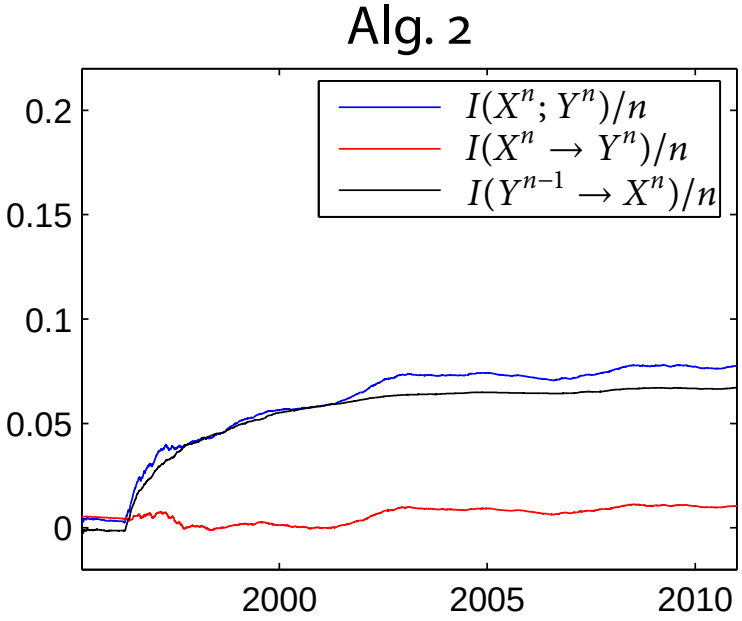


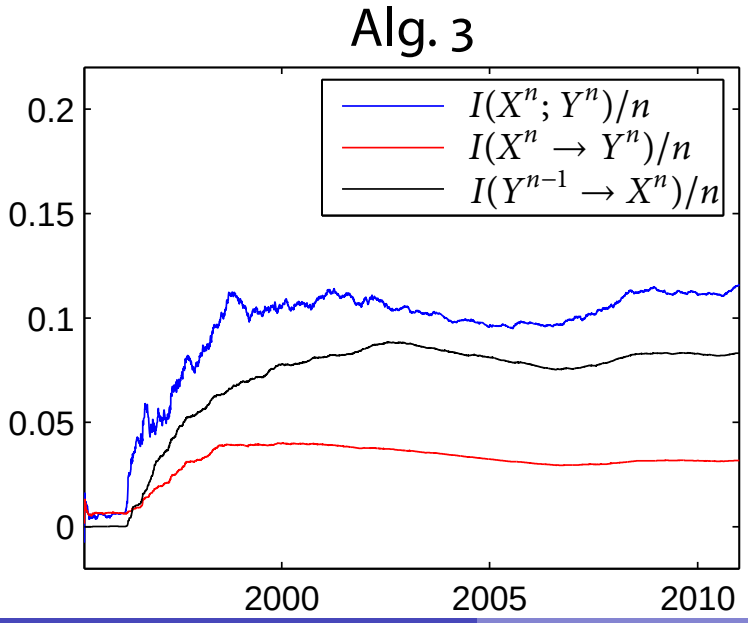
Questions

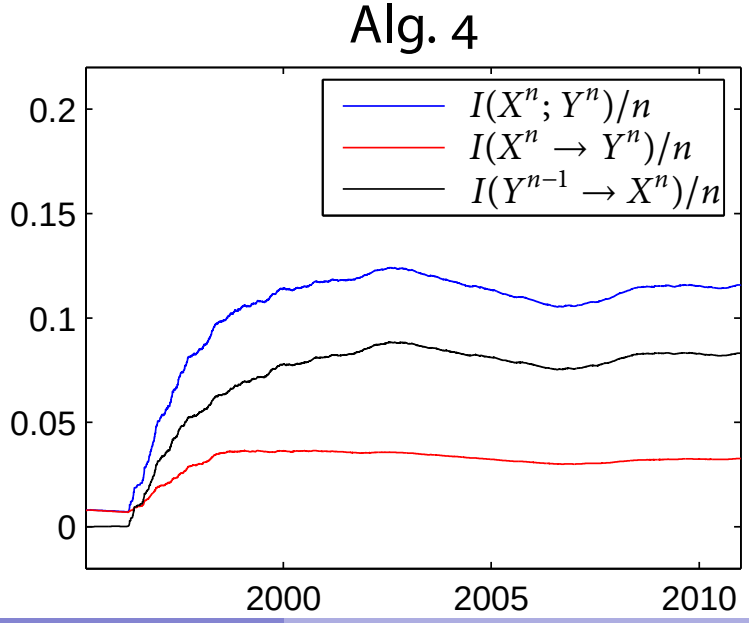
- Are these markets correlated?
- Which index leads the other?

HSI versus DJIA

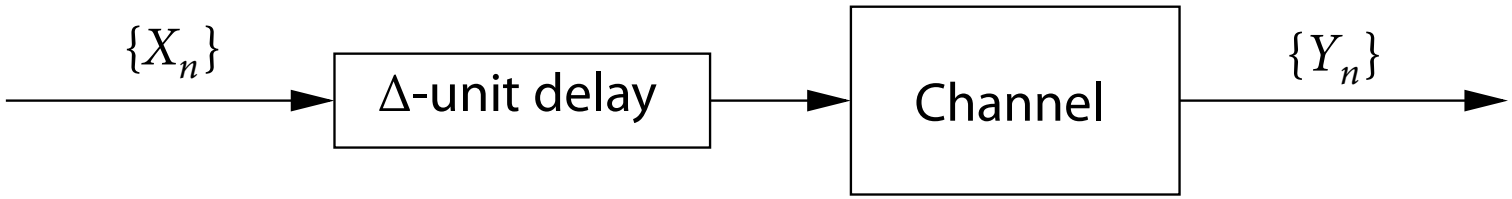








Delay estimation



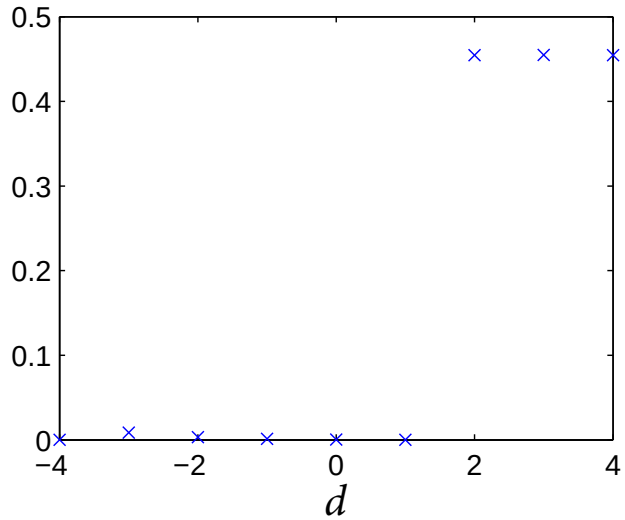
Question

Can we estimate Δ efficiently?

- Shifted directed information

$$I(Y_{d+1}^n \to X^{n-d}) = \sum_{i=1}^{n-d} H(X_i|X^{i-1}) - H(X_i|X^{i-1}, Y_{d+1}^{d+i})$$

- If $d < \Delta$, then $I(Y_{d+1}^n \to X^{n-d}) = 0$
- If $d \geq \Delta$, then $I(Y_{d+1}^n \to X^{n-d}) \gg 0$



Concluding remarks

- Directed information $I(X^n \to Y^n)$
  - Arrow of time + Shannon's mutual information
  - A natural generalization of Granger causality
  - Causation beyond correlation
  - Universal estimation algorithms (MATLAB codes available)

- Looking forward
  - Large alphabets
  - More than a pair of random processes
  - Piecewise stationary processes
  - More applications (economics, biology, climate change, …)

References

Aczél, J. and Daróczy, Z. (1975). *On Measures of Information and Their Characterizations*. Academic Press, New York.

Barron, A. R. and Cover, T. M. (1988). A bound on the financial value of information. *IEEE Trans. Inf. Theory*, IT-34, 1097–1100.

Bell, R. M. and Cover, T. M. (1980). Competitive optimality of logarithmic investment. *Math. Oper. Res.*, 5(2), 161–166.

Breiman, L. (1957). The individual ergodic theorem of information theory. *Ann. Math. Statist.*, 28(3), 809–811. Correction (1960). 31(3), 809–810.

Geweke, J. F. (1982). Measurement of linear dependence and feedback between multiple time series. *J. Amer. Statist. Assoc.*, 77(378), 304–324. With discussion and with a reply by the author.

Granger, C. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3), 424–438.

Jiao, J., Zhao, L., Permuter, H. H., Kim, Y.-H., and Weissman, T. (2013). Universal estimation of directed information.

Kelly, J. L., Jr. (1956). A new interpretation of information rate. *Bell Syst. Tech. J.*, 35, 917–926.

Kim, Y.-H. (2008). A coding theorem for a class of stationary channels with feedback. *IEEE Trans. Inf. Theory*, 54(4), 1488–1499.

Kramer, G. (1998). *Directed Information for Channels with Feedback*. Hartung-Gorre Verlag, Konstanz. Dr. sc. thchn. Dissertation, Swiss Federal Institute of Technology (ETH) Zurich.

References (cont.)

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics*, 22, 79–86.

MacLean, L. C., Thorp, E. O., and Ziemba, W. T. (2011). *The Kelly Capital Growth Investment Criterion: Thenry and Practice*. World Scientific, Singapore.

Marko, H. (1973). The bidirectional communication theory: A generalization of information theory. *IEEE Trans. Comm.*, 21(12), 1345–1351.

Massey, J. L. (1990). Causality, feedback, and directed information. In *Proc. IEEE Int. Symp. Inf. Theory Appl.*, Honolulu, HI, pp. 303–305.

McMillan, B. (1953). The basic theorems of information theory. *Ann. Math. Statist.*, 24(2), 196–219.

Morvai, G., Yakowitz, S. J., and Algoet, P. (1997). Weakly convergent nonparametric forecasting of stationary time series. *IEEE Trans. Inf. Theory*, 43(2), 483–498.

Ornstein, D. (1978). Guessing the next output of a stationary process. *Israel J. Math.*, 30, 292–296.

Permuter, H. H., Kim, Y.-H., and Weissman, T. (2011). Interpretations of directed information in portfolio theory, data compression, and hypothesis testing. *IEEE Trans. Inf. Theory*, 57(6), 3248–3259.

Permuter, H. H., Weissman, T., and Goldsmith, A. J. (2009). Finite state channels with time-invariant deterministic feedback. *IEEE Trans. Inf. Theory*, 55(2), 644–662.

Schreiber, T. (2000). Measuring information transfer. *Phys. Rev. Lett.*, 85, 461–464.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3), 379–423, 27(4), 623–656.

References (cont.)

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. In *IRE Int. Conv. Rec.*, vol. 7, part 4, pp. 142–163. Reprint with changes (1960). In R. E. Machol (ed.) *Information and Decision Processes*, pp. 93–126. McGraw-Hill, New York.

Tatikonda, S. and Mitter, S. (2009). The capacity of channels with feedback. *IEEE Trans. Inf. Theory*, 55(1), 323–349.

Weissman, T., Kim, Y.-H., and Permuter, H. H. (2013). Directed information, causal estimation, and communication in continuous time. *IEEE Trans. Inf. Theory*, 59(3), 1271–1287.

Willems, F. M. J., Shtarkov, Y. M., and Tjalkens, T. J. (1995). The context-tree weighting method: Basic properties. *IEEE Trans. Inf. Theory*, 41(3), 653–664.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory*, IT-24(5), 530–536.