Universal Estimation of Directed Information

Jiantao Jiao
Tsinghua/Stanford

Lei Zhao
Jump Operation

Haim Permuter
Ben-Gurion

Young-Han Kim
UCSD

Tsachy Weissman
Stanford

ISIT
July 2012, Boston

Definition of Directed Information (Discrete Time)

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

$$H(Y^n|X^n) \quad \triangleq \quad E[-\log P(Y^n|X^n)]$$

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

Definition of Directed Information (Discrete Time)

Directed Information [Marko73,Massey90]

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

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

Causal Conditioning [Kramer98]

$$H(Y^n\|X^n) \triangleq E[-\log P(Y^n\|X^n)]$$

$$H(Y^n|X^n) \triangleq E[-\log P(Y^n|X^n)]$$

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

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

Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]

Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]

Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]
3. channel capacity with feedback [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]

Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]
3. channel capacity with feedback [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. networks capacity with feedback [Kramer98]

Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]
3. channel capacity with feedback [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. networks capacity with feedback [Kramer98]
5. rate distortion with feedforward [Venkataramanan/Pradhan07]

Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]
3. channel capacity with feedback [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. networks capacity with feedback [Kramer98]
5. rate distortion with feedforward [Venkataramanan/Pradhan07]
6. causal MMSE for additive Gaussian noise [Weissman/P/Kim11]

Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]
3. channel capacity with feedback [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. networks capacity with feedback [Kramer98]
5. rate distortion with feedforward [Venkataramanan/Pradhan07]
6. causal MMSE for additive Gaussian noise [Weissman/P/Kim11]
7. stock investment with causal side information [P/Kim/Weissman11]

Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]
3. channel capacity with feedback [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. networks capacity with feedback [Kramer98]
5. rate distortion with feedforward [Venkataramanan/Pradhan07]
6. causal MMSE for additive Gaussian noise [Weissman/P/Kim11]
7. stock investment with causal side information [P/Kim/Weissman11]
8. actions with causal constraint such as “to feed or not to feed back” [Asnani/P/Weissman11]

Directed information and causal conditioning characterize

1. rate reduction in lossless compression due to causal side information at the decoder [Simeone/P12]
2. gain in growth rate in horse-race gambling due to causal side information [P/Kim/Weissman11]
3. channel capacity with feedback [Kim08, Tatikonda/Mitter09, P/Weissman/Goldsmith09]
4. networks capacity with feedback [Kramer98]
5. rate distortion with feedforward [Venkataramanan/Pradhan07]
6. causal MMSE for additive Gaussian noise [Weissman/P/Kim11]
7. stock investment with causal side information [P/Kim/Weissman11]
8. actions with causal constraint such as "to feed or not to feed back" [Asnani/P/Weissman11]

Can be optimized using convex optimization tools [Naiss/P11]

In this talk

- Contribution: We present several estimator for estimation the directed information using universal compression algorithms.
- Idea: Use a universal compression algorithm to induce a universal probability assignment, then plug in the probability assignment into the estimator.
- Result: The proposed estimator are all consistent and provide different properties (such as range, smoothness, convergence guarantees).

Casual condition, the Chain rule

Causal Conditioning

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

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

Chain Rule

$$P(x^n, y^n) = Q(x^n||y^{n-1})P(y^n||x^{n-1})$$

Conservation Law

$I(X^n;Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n)$ [Massey06]

Recall $P(x^n, y^n) = P(x^n||y^{n-1})P(y^n||x^n)$

$$\begin{aligned}
I(X^n;Y^n) &= \mathbf{E}\left[\ln \frac{P(Y^n, X^n)}{P(Y^n)P(X^n)}\right] \\
&= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)P(X^n||Y^{n-1})}{P(Y^n)P(X^n)}\right] \\
&= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)}{P(Y^n)}\right] + \mathbf{E}\left[\ln \frac{P(X^n||Y^{n-1})}{P(X^n)}\right] \\
&= I(X^n \to Y^n) + I(Y^{n-1} \to X^n).
\end{aligned}$$

Conservation Law

$I(X^n;Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n)$ [Massey06]

Recall $P(x^n, y^n) = P(x^n||y^{n-1})P(y^n||x^n)$

$$
\begin{aligned}
I(X^n;Y^n) &= \mathbf{E}\left[\ln \frac{P(Y^n, X^n)}{P(Y^n)P(X^n)}\right] \\
&= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)P(X^n||Y^{n-1})}{P(Y^n)P(X^n)}\right] \\
&= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)}{P(Y^n)}\right] + \mathbf{E}\left[\ln \frac{P(X^n||Y^{n-1})}{P(X^n)}\right] \\
&= I(X^n \to Y^n) + I(Y^{n-1} \to X^n).
\end{aligned}
$$

In case that we have $X_i - (X^{i-1}, Z^{i-1}) - Y^{i-1}$:

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

Conservation Law $I(X^n;Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n)$ [Massey06]

Recall $P(x^n, y^n) = P(x^n||y^{n-1})P(y^n||x^n)$

$$
\begin{aligned}
I(X^n;Y^n) &= \mathbf{E}\left[\ln \frac{P(Y^n, X^n)}{P(Y^n)P(X^n)}\right] \\
&= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)P(X^n||Y^{n-1})}{P(Y^n)P(X^n)}\right] \\
&= \mathbf{E}\left[\ln \frac{P(Y^n||X^n)}{P(Y^n)}\right] + \mathbf{E}\left[\ln \frac{P(X^n||Y^{n-1})}{P(X^n)}\right] \\
&= I(X^n \to Y^n) + I(Y^{n-1} \to X^n).
\end{aligned}
$$

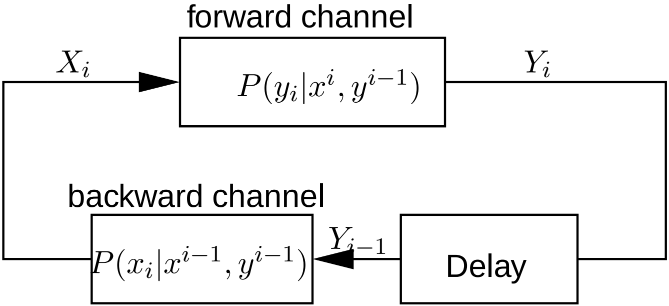
In case that we have $X_i - (X^{i-1}, Z^{i-1}) - Y^{i-1}$:

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

If there is no feedback, $z_i = null$, then

$$
I(X^n;Y^n) = I(X^n \to Y^n) + 0.
$$

Causal influence/relevance between two sequences



The model implies an order $X_1, Y_1, X_2, Y_2, X_3, Y_3, ...$

- The forward link exists if and only if $I(X^n \to Y^n) > 0$ (X_i is "causing" Y_i)
- The backward link exists if and only if $I(Y^{n-1} \to X^n) > 0$. (Y_i is "causing" X_i)

Previous work on Causality

- Granger Causality [Granger69]
- Bidirectional communication [Marko73]
- Gourieroux/Monfort/Renault87

$$I(X^{n-1} \to Y^n) + I(Y^{n-1} \to X^n) + \sum_i I(X_i; Y_i | X^{i-1}, Y^{i-1})$$

- "Measures of mutual and causal dependence between two time series" [Rissanen/Max 87]
- Relation between Granger Causality and directed information [Quinn/Coleman/Kiyavash/Hatsopoulos10] [Quinn/Coleman/Kiyavash11] [Amblard/Michel10]
- Directed information estimation has been applied to
  - Neurobiology [Quinn/Coleman/Kiyavash/Hatsopoulos10]
  - Gene Network [Rao/Hero/States/Engel08]
  - Video Indexing [Chen/Savarese/Hero12]

Universal sequential probability assignment

- A sequential probability assignment Q consists of a set of conditional probabilities $\{Q_{X_i|x^{i-1}}(\cdot), \forall x^{i-1} \in \mathcal{X}^{i-1}\}_{i=1}^{\infty}$.
- A sequential probability assignment Q is universal if

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

  for any stationary probability measure P.
- A source code is universal if each code is uniquely decodable and

$$\limsup_{n\to\infty} \frac{1}{n} \mathrm{E}\left[l_n(X^n)\right] = \mathrm{H}(\mathbf{X}).$$

  for every stationary ergodic source $\mathbf{X}$
- Univ. compressors $\Longleftrightarrow$ Univ. sequential probability assign.

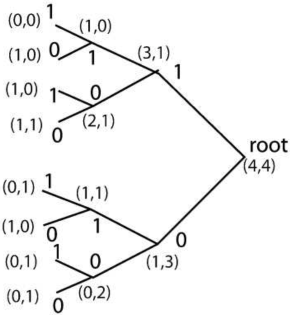
Estimating information measures

The idea of estimating information measures using universal compressor has been used

- **LZ** Lempel-Ziv , [Wyner/Ziv89], [Ziv/Merhav93]
- **BWT** Burrows-Wheeler Transform [Cai/Kulkarni/Verdú04,06]
- **CTW** Context Tree Weighting [Yu/Verdú06]

Context Tree Weighting (CTW)

[Willems, Shtarkov, Tjalkens, 1995], [Willems, 1998]



x=(000)11010010 with D=3

- Universal compressor
- Optimal convergence rates
- Linear complexity
- Explicit sequential probability assignment

Estimator 1

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

where

$$\widehat{H}_1(Y^n||X^n) \quad \triangleq \quad -\frac{1}{n}\log Q(Y^n\|X^n) = -\frac{1}{n}\sum_{i=1}^{n}\log Q(Y_i|Y^{i-1},X^i)$$

Theorem

Let Q be a universal sequential probability assignment, then

$$\lim_{n\to\infty}\widehat{I}_1(X^n \to Y^n) = \mathrm{I}(\mathbf{X}\to\mathbf{Y}) \text{ in } L_1.$$

Further, if Q is a pointwise universal probability assignment, then the convergence of $\widehat{I}_1(X^n \to Y^n)$ to $\mathrm{I}(\mathbf{X}\to\mathbf{Y})$ holds almost surely.

Estimator 1 - convergence rate

Theorem

Let Q be the universal probability assignment induced by basic CTW method, if $(\mathbf{X}, \mathbf{Y})$ then there exists a constant C_1 such that

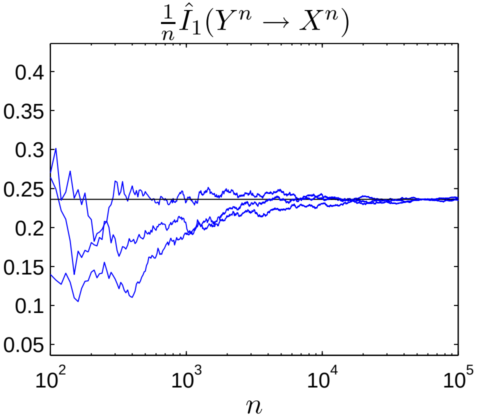
$$\mathbb{E}\left|\widehat{I}_1(X^n \to Y^n) - \mathrm{I}(\mathbf{X} \to \mathbf{Y})\right| \leq C_1 n^{-1/2} \log n$$

and $\forall \epsilon > 0$*,*

$$\widehat{I}_1(X^n \to Y^n) - \mathrm{I}(\mathbf{X} \to \mathbf{Y}) = o(n^{-1/2}(\log n)^{5/2+\epsilon}) \textit{ a.s.}$$

Under a minimax criteria (Rissanen lower bound) this is the best one can do (One can not guarantee that the error can not decrease faster than $O(n^{-1/2})$).

Estimator 1



- merits: algorithmic and theoretical
- erratic for small n
- unbounded range including negative values.

Second approach

Consider an estimation of entropy rate

$$\lim \frac{1}{n} H(X^n)$$

First,

$$-\frac{1}{n} \log Q(X^n) = -\frac{1}{n} \sum_{i=1}^{n} \log Q(X_i | X^{i-1})$$

Second approach

Consider an estimation of entropy rate

$$\lim \frac{1}{n} H(X^n)$$

First,

$$-\frac{1}{n} \log Q(X^n) = -\frac{1}{n} \sum_{i=1}^{n} \log Q(X_i | X^{i-1})$$

Second,

$$\frac{1}{n} \sum_{i=1}^{n} h(Q(\cdot | X^{i-1})),$$

where

$$h(P(\cdot)) = \sum_{x} -P(x) \log P(x).$$

Estimator 2

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

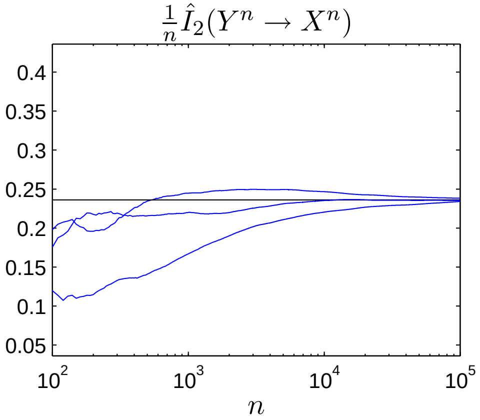
where

$$\widehat{H}_2(Y^n \| X^n) \triangleq \frac{1}{n} \sum_{i=1}^{n} f(Q_{X_{i+1}, Y_{i+1} | X^i, Y^i}(\cdot, \cdot))$$

$$f(P_{X,Y}) \triangleq - \sum_{x,y} P_{X,Y}(x, y) \log P_{Y|X}(y|x)$$

performance guarantees for $\widehat{I}_2$ similar to those for $\widehat{I}_1$

Estimator 2



- merits: algorithmic, theoretical, smooth, bounded range
- can be negative

Third approach

Estimate directed information using divergence,

$$\begin{aligned} I(X^i;Y_i) &= D(P_{X^i,Y_i}||P_{X^i,Y_i} \times P_{X^i,Y_i}) \\ &= D(P_{Y_i|X^i}||P_{Y_i}|P_{X^i}) \end{aligned}$$

And the directed information

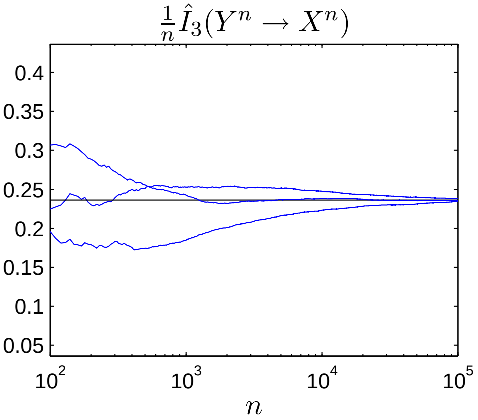
$$I(X^i;Y_i|Y^{i-1}) = D(P_{Y_i|X^i,Y^{i-1}}||P_{Y_i|Y^{i-1}}|P_{X^i,Y^{i-1}})$$

Estimator 3:

$$\widehat{I}_3(X^n \to Y^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} D(Q_{Y_i|X^i,Y^{i-1}}(\cdot)\|Q_{Y_i|Y^{i-1}}(\cdot))$$

Similar performance guarantees, though weaker (Stationary ergodic Markov needed) than for the previous two.

Estimator 3



- merits: algorithmic, theoretical, bounded range, nonnegative
- smoothness can be improved for small n

Estimator 4

$$\begin{aligned}
&\widehat{I}_4(X^n \to Y^n) \\
&\quad \triangleq \quad \frac{1}{n}\sum_{i=1}^{n} D(Q_{X_{i+1},Y_{i+1}|X^i,Y^i}(\cdot,\cdot) \| Q_{Y_{i+1}|Y^i}(\cdot) Q_{X_{i+1}|X^i,Y^i}(\cdot))
\end{aligned}$$

Estimator 4

$$\begin{aligned}
&\widehat{I}_4(X^n \to Y^n) \\
&\triangleq \quad \frac{1}{n}\sum_{i=1}^{n} D(Q_{X_{i+1},Y_{i+1}|X^i,Y^i}(\cdot,\cdot)\|Q_{Y_{i+1}|Y^i}(\cdot)Q_{X_{i+1}|X^i,Y^i}(\cdot))
\end{aligned}$$

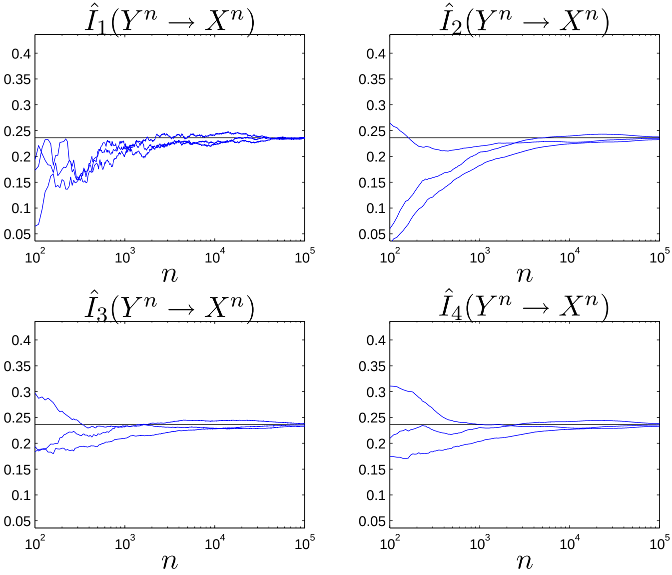
Recall the other Estimators:

$$\widehat{H}_1(X^n||Y^n) \triangleq -\frac{1}{n}\sum_{i=1}^{n}\log Q(Y_i|Y^{i-1},X^i)$$

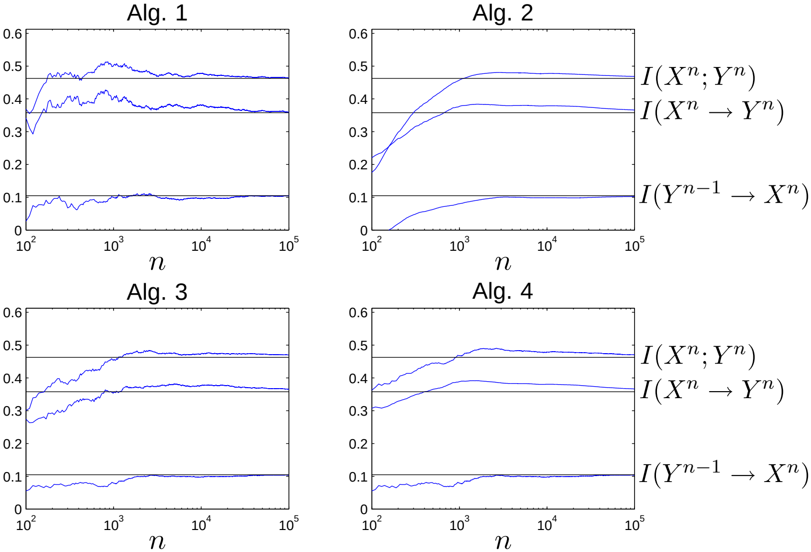
$$\widehat{H}_2(Y^n||X^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} f(Q_{X_{i+1},Y_{i+1}|X^i,Y^i}(\cdot,\cdot))$$

$$\widehat{I}_3(X^n \to Y^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} D(Q_{Y_i|X^i,Y^{i-1}}(\cdot)\|Q_{Y_i|Y^{i-1}}(\cdot))$$

All estimators

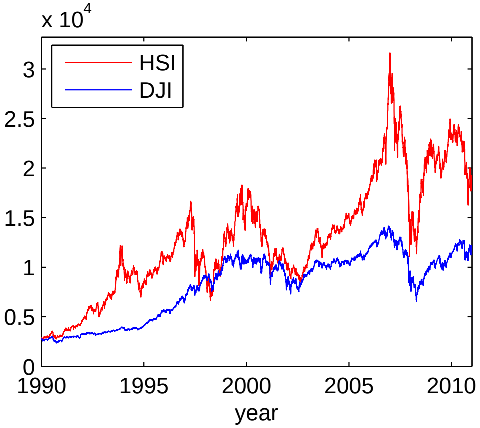


Computation of directed and reverse-directed information



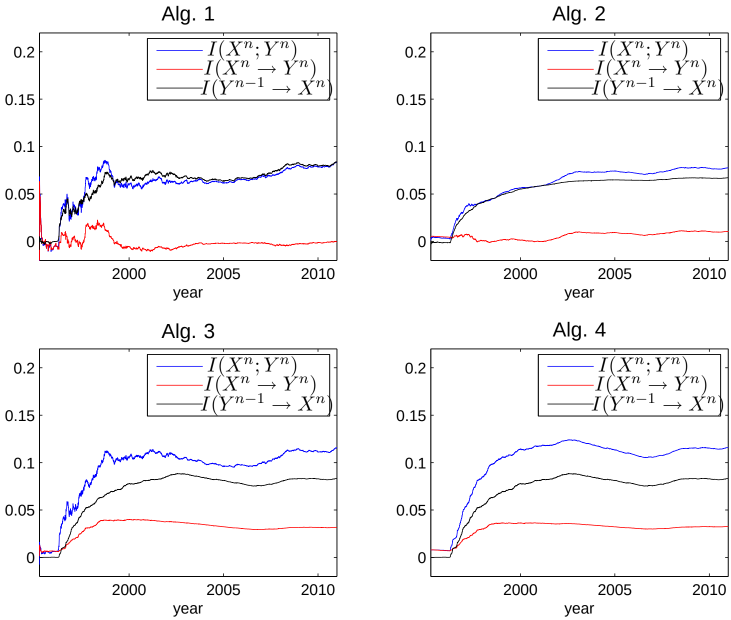
Stock market example

The Hang Seng Index (HSI) and Dow Jones Index (DJI) indexes between 1990-2011



We would like to determine who is causally influencing whom.

Mutual influence of HSI, $\{X_i\}$, and DJI , $\{Y_i\}$.



Summary and Future Work

- Universal compressor used to estimate the directed information via the assignment probability that it induces
- Four different algorithms are suggested
- All Use the probability assignment $Q(X_i, Y_i|X^{i-1}, Y^{i-1})$ and $Q(Y_i|Y^{i-1})$.
- Different properties (smoothness, range, nonnegative)
- CTW is a good universal compressor
- L_1 and $a.s.$ convergence is guaranteed
- Future work:
  - Continuous alphabet and Larger alphabet
  - Low number of samples
  - applications [page ranking, biology]
- Code available at http://www.stanford.edu/~tsachy/DIcode/

Summary and Future Work

- Universal compressor used to estimate the directed information via the assignment probability that it induces
- Four different algorithms are suggested
- All Use the probability assignment $Q(X_i, Y_i|X^{i-1}, Y^{i-1})$ and $Q(Y_i|Y^{i-1})$.
- Different properties (smoothness, range, nonnegative)
- CTW is a good universal compressor
- L_1 and $a.s.$ convergence is guaranteed
- Future work:
  - Continuous alphabet and Larger alphabet
  - Low number of samples
  - applications [page ranking, biology]
- Code available at `http://www.stanford.edu/~tsachy/DIcode/`

Thank you very much!