Universal Estimation of Directed Information via Sequential Probability Assignments

Jiantao Jiao
Tsinghua University
xajjt1990@gmail.com

Haim H. Permuter
Ben Gurion University
haimp@bgu.ac.il

Lei Zhao
Jump Operations
zhaolei122@gmail.com

Young-Han Kim
UCSD
yhk@ucsd.edu

Tsachy Weissman
Stanford University
tsachy@stanford.edu

***Abstract*—We propose four approaches to estimating the directed information rate between a pair of jointly stationary ergodic processes with the help of universal probability assignments. The four approaches yield estimators with different merits such as nonnegativity and boundedness. We establish consistency of these estimators in various senses and derive near-optimal rates of convergence in the minimax sense under mild conditions. The estimators carry over directly to estimating other information measures of stationary ergodic processes, such as entropy rate and mutual information rate, and provide alternatives to classical approaches in the existing literature. Guided by the theoretical results, we use context tree weighting as the vehicle for the implementations of the proposed estimators. Experiments on synthetic and real data are presented, demonstrating the potential of the proposed schemes in practice and the efficacy of directed information estimation as a tool for detecting and measuring causality and delay.**

***Index Terms*—Causal influence, context tree weighting, directed information, rate of convergence, universal probability assignment**

I. Introduction

First introduced by Marko [1] and Massey [2], directed information arises as a natural counterpart of mutual information for channel capacity when causal feedback from the receiver to the sender is present. In [3] and [4], Kramer extended the use of directed information to discrete memoryless networks with feedback, including the two-way channel and the multiple access channel. Tatikonda and Mitter [5] used directed information spectrum to establish a general feedback channel coding theorem for channels with memory. Kim [6] established the feedback capacity for a class of stationary channels using directed information. In [7], Permuter, Weissman, and Goldsmith considered the capacity of discrete-time channels with feedback where the feedback is a time-invariant deterministic function of the output, and used directed information to describe the capacity under mild conditions. Recently, Permuter, Kim, and Weissman [8] showed that directed information plays an important role in portfolio theory, data compression, and hypothesis testing, in the presence of causality constraints.

Beyond information theory, directed information is a valuable tool in biology, for it provides an alternative to identify causal inferences between two processes. In Mathai, Martins, and Shapiro [9], directed information was used to identify pairwise influence. Rao, Hero, States, and Engel [10] used directed information to test the direction of influence in gene networks.

Since directed information has significance in various fields, it is of both theoretical and practical importance to develop efficient ways for estimating it. The problem of estimating information measures, such as entropy, relative entropy and mutual information, has been extensively studied in the literature. Verdú [11] gave an overview of universal estimation of information measures. Wyner and Ziv [12] applied the idea of Lempel–Ziv parsing to estimate the entropy rate, which converges in probability for all stationary ergodic processes. Ziv and Merhav [13] used Lempel–Ziv parsing to estimate relative entropy (Kullback–Leibler divergence) and established consistency under the assumption that the observations are generated by independent Markov sources. Cai, Kulkarni, and Verdú [14] proposed two universal divergence estimators for finite-alphabet sources, one based on the Burrows–Wheeler transform (BWT) [15] and the other based on the context tree weighting method (CTW) [16]. The BWT-based estimator was applied in universal entropy estimation in Cai, Kulkarni, and Verdú [17], while the CTW-based one was applied in universal erasure entropy estimation in Yu and Verdú [18].

For the problem of estimating directed information, Quinn, Coleman, Kiyavashi, and Hatspoulous [19] developed an estimator to infer causality in ensemble neural spike train recordings. Based on parametric generalized linear model (GLM) assumption and stationary ergodic Markov assumption [19], they showed strong consistency results. Compared to [19], Zhao, Kim, Permuter, and Weissman [20] focused on universal methods and showed L_1 consistency for all jointly stationary ergodic process pairs with finite alphabet.

As an improvement and further development of [20], and a reflection of Jiao, Permuter, Zhao, Kim, and Weissman [21], the main contribution of this paper is a general framework for estimating information measures of stationary ergodic processes, using “single-letter” information-theoretic functionals. Although our methods can be applied in estimating a number of information measures, we focus—for concreteness and relevance to emerging applications—on estimating the directed information rate between a pair of jointly stationary ergodic processes. The first proposed estimator is adapted from the universal divergence estimator in [14] using the CTW method, and we give a refined analysis yielding strong consistency

978-1-4673-2579-0/12/$31.00 ©2012 IEEE

results. We further propose three additional estimators in a unified framework to estimate the directed information rate, present both weak and strong consistency results, and establish near-optimal rates of convergence under mild conditions. We then employ our estimators on both simulated and real data, showing their effectiveness in measuring channel delays and causal influences between processes. In particular, we use these estimators to establish significant causal influence from the Dow Jones Industrial Average to the Hang Seng Index, but relatively low causal influence in the reverse direction, based on the daily market data in the period from 1990 to 2011.

The rest of the paper is organized as follows. Section II reviews some preliminaries and Section III presents our proposed estimators and some of their basic properties. Section IV is dedicated to performance guarantees for the proposed estimators, rates of convergence results under mild conditions, and minimax optimality. Section V shows experimental results applying the proposed estimators, both on simulated and real data, and demonstrates the effectiveness of these estimators in inferring delay of channels and causal influences between processes. For proofs of stated results please see [21].

II. Preliminaries

We begin with definitions of directed information, universal and pointwise universal probability assignments. We then introduce the context tree weighting (CTW) method used in our implementations.

We use uppercase letters $X, Y, \ldots$ to denote random variables, and lowercase letters $x, y, \ldots$ to denote values they assume. We denote the n-tuple $(X_1, X_2, \ldots, X_n)$ as X^n and $(x_1, x_2, \ldots, x_n)$ as x^n. Calligraphic letters $\mathcal{X}, \mathcal{Y}, \ldots$ denote alphabets of $X, Y, \ldots$, and $|\mathcal{X}|$ denotes the cardinality of $\mathcal{X}$. Given a probability law P, $P(x_i|X^{i-1})$ denotes the conditional pmf $P(x_i|x^{i-1})$ evaluated for the random sequence X^{i-1}, while $P(X_i|X^{i-1})$ is the random variable denoting the X_ith component of $P(x_i|X^{i-1})$. Throughout this paper, $\log(\cdot)$ means $\log_2(\cdot)$.

A. Directed Information

Introduce the notation of causal conditionsl pmf as follows:

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

the directed information from X^n to Y^n is defined as

$$I(X^n \to Y^n) \triangleq \sum_{i=1}^{n} I(X^i; Y_i|Y^{i-1}) = H(Y^n) - H(Y^n \| X^n), \tag{2}$$

where $H(Y^n \| X^n)$ is the *causally conditional entropy* [3], concretely,

$$H(Y^n \| X^n) \triangleq \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X^i). \tag{3}$$

Compared with the definition of mutual information,

$$I(X^n; Y^n) = H(Y^n) - H(Y^n|X^n), \tag{4}$$

directed information has the causally conditional entropy in place of the conditional entropy. Unlike mutual information, directed information is not symmetric, i.e., $I(Y^n \to X^n) \neq I(X^n \to Y^n)$ in general.

We have the conservation law for directed information

$$I(X^n; Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n), \tag{5}$$

where

$$\begin{aligned} I(Y^{n-1} \to X^n) &= I((\emptyset, Y^{n-1}) \to X^n) && (6)\\ &= H(X^n) - \sum_{i=1}^{n} H(X_i|X^{i-1}, Y^{i-1}) && (7) \end{aligned}$$

denotes the *reverse* directed information. Other interesting properties of directed information can be found in [3], [22].

The directed information rate [3] between a pair of jointly stationary random processes $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$\bar{I}(\mathbf{X} \to \mathbf{Y}) \triangleq \lim_{n\to\infty} \frac{1}{n} I(X^n \to Y^n), \tag{8}$$

and we can easily check that

$$\bar{I}(\mathbf{X} \to \mathbf{Y}) = \overline{H}(\mathbf{Y}) - \overline{H}(\mathbf{Y} \| \mathbf{X}), \tag{9}$$

where $\overline{H}(\mathbf{Y}) = H(Y_0|Y_{-\infty}^{-1})$ is the entropy rate of process $\mathbf{Y}$, $\overline{H}(\mathbf{Y}\|\mathbf{X})$ is the causally conditional entropy rate defined as $\overline{H}(\mathbf{Y}\|\mathbf{X}) \triangleq \lim_{n\to\infty}(1/n)H(Y^n\|X^n) = H(Y_0|X_{-\infty}^{0}, Y_{-\infty}^{-1})$.

Equation (9) shows that we can estimate $\overline{H}(\mathbf{Y})$ and $\overline{H}(\mathbf{Y}\|\mathbf{X})$ separately to estimate the directed information rate.

B. Universal Probability Assignment

A probability assignment Q consists of a set of conditional pmfs $Q(x_i|x^{i-1})$ for every $x^{i-1} \in \mathcal{X}^{i-1}$. Note that Q induces a probability measure on a random process $\mathbf{X}$.

Definition 1 (Universal probability assignment): A probability assignment Q is said to be *universal for a class* $\mathscr{P}$ if the normalized Kullback–Leibler divergence satisfies

$$\lim_{n\to\infty} \frac{1}{n} D(P(x^n)||Q(x^n)) = 0 \tag{10}$$

for every probability measure P in $\mathscr{P}$. A probability assignment Q is said to be *universal* (without a qualifier) if it is universal for the class of stationary probability measures.

Definition 2 (Pointwise universal probability assignment): A probability assignment Q is said to be *pointwise universal for a class* $\mathscr{P}$ if

$$\limsup_{n\to\infty} \left(\frac{1}{n} \log \frac{1}{Q(X^n)} - \frac{1}{n} \log \frac{1}{P(X^n)} \right) \leq 0 \quad P\text{-a.s.} \tag{11}$$

for every probability measure P in $\mathscr{P}$. A probability assignment Q is said to be *pointwise universal* (without a qualifier) if it is pointwise universal for the class of stationary ergodic probability measures.

It is well known that there exist universal and pointwise universal probability assignments, see, for example, [23].

C. Context Tree Weighting Method

One particularly celebrated universal probability assignment is the context tree weighting (CTW) algorithm by Willems, Shtarkov, and Tjalken [16]. The computational complexity of the CTW is linear in the block length n, and the algorithm provides the probability assignments Q directly, which is the weighted probability at the root node. For details, see [16], [24] and [21]. Note that here we use the extended version of CTW for non-binary alphabets, which is discussed in [25].

The probability assignment Q in CTW is both universal and pointwise universal for the class of stationary ergodic Markov processes. For the proof of universality, see, [16], and for the pointwise universality, please see [21].

III. Four Estimation Algorithms

In this section, we introduce four algorithms to estimate the directed information rate $\bar{I}(\mathbf{X} \to \mathbf{Y})$ of a pair of jointly stationary ergodic processes $\mathbf{X}$ and $\mathbf{Y}$. Let $\mathcal{M}(\mathcal{X}, \mathcal{Y})$ be the set of all probability distributions on $\mathcal{X} \times \mathcal{Y}$. Define f as the function that maps a joint pmf $P(x, y)$ of a random pair (X, Y) to the corresponding conditional entropy $H(Y|X)$, i.e.,

$$f(P) \triangleq -\sum_{x,y} P(x, y) \log P(y|x), \tag{12}$$

where $P(y|x)$ is the conditional pmf induced by $P(x, y)$. Take Q as a universal probability assignment.

Define four estimators as follows:

$$\hat{I}_1(X^n \to Y^n) \triangleq \hat{H}_1(Y^n) - \hat{H}_1(Y^n \| X^n), \tag{13}$$

$$\hat{I}_2(X^n \to Y^n) \triangleq \hat{H}_2(Y^n) - \hat{H}_2(Y^n \| X^n), \tag{14}$$

$$\hat{I}_3(X^n \to Y^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} D(Q(y_i|X^i, Y^{i-1}) \| Q(y_i|Y^{i-1})), \tag{15}$$

$$\hat{I}_4(X^n \to Y^n) \triangleq$$

$$\frac{1}{n}\sum_{i=1}^{n} D(Q(x_{i+1}, y_{i+1}|X^i, Y^i) \| Q(y_{i+1}|Y^i) Q(x_{i+1}|X^i, Y^i)), \tag{16}$$

where

$$\hat{H}_1(Y^n \| X^n) \triangleq -\frac{1}{n} \log Q(Y^n \| X^n), \tag{17}$$

$$\hat{H}_2(Y^n \| X^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} f(Q(x_{i+1}, y_{i+1}|X^i, Y^i)), \tag{18}$$

and $\hat{H}_1(Y^n) = \hat{H}_1(Y^n \| \emptyset)$, $\hat{H}_2(Y^n) = \hat{H}_2(Y^n \| \emptyset)$. Note that an entropy estimate such as $\hat{H}_1(Y^n \| X^n)$ is a *random variable* (since it is a function of (X^n, Y^n)), as opposed to entropy terms such as $H(Y^n \| X^n)$, which are deterministic and depend on the *distribution* of (X^n, Y^n).

Note that the universal probability assignments conditioned on different data are calculated separately. For example, $Q(y_i|Y^{i-1})$ is not computed from $Q(x_i, y_i|X^{i-1}, Y^{i-1})$, but by running the universal probability assignment algorithm again on dataset Y^{i-1}. Of course, $Q(y_i|x_i, X^{i-1}, Y^{i-1})$ is computed from $Q(x_i, y_i|X^{i-1}, Y^{i-1})$.

The estimator $\hat{I}_1$ is adapted from one universal divergence estimator in [14]. One disadvantage of $\hat{I}_1(X^n \to Y^n)$ is that it has a nonzero probability of being very large, which is overcome by $\hat{I}_2$, the estimator introduced in [20], by using information-theoretic functionals to "smooth" the estimate. Evidently we can show $|\hat{I}_2| \le \log |\mathcal{Y}|$.

The common disadvantage of $\hat{I}_1$ and $\hat{I}_2$ is that they are computed by subtraction of two quantities, and have a nonzero probability of being negative. $\hat{I}_3$ and $\hat{I}_4$ are introduced to overcome this, since they take the form of a Kullback–Leibler divergence and are always nonnegative.

IV. Performance Guarantees

In this section, we present consistency results for the proposed estimators. Under some mild conditions, we derive near-optimal rates of convergence in the minimax sense. For proofs, please see [21].

Theorem 1: Let Q be a universal probability assignment and $(\mathbf{X}, \mathbf{Y})$ be jointly stationary ergodic. Then

$$\lim_{n\to\infty} \hat{I}_1(X^n \to Y^n) = \bar{I}(\mathbf{X} \to \mathbf{Y}) \quad \text{in } L_1. \tag{19}$$

Furthermore, if Q is also a pointwise universal probability assignment, then the limit in (19) holds almost surely as well.

If $(\mathbf{X}, \mathbf{Y})$ is a stationary ergodic aperiodic Markov process, we can say more about the performance of $\hat{I}_1$ using the probability assignment in CTW method.

Proposition 1: Let Q be the probability assignment in CTW. If $(\mathbf{X}, \mathbf{Y})$ is a jointly stationary ergodic aperiodic Markov process whose order does not exceed the prescribed maximum depth in CTW, then there exists a constant C_1 such that

$$\mathbb{E}\left|\hat{I}_1(X^n \to Y^n) - \bar{I}(\mathbf{X} \to \mathbf{Y})\right| \le C_1 n^{-1/2} \log n, \tag{20}$$

and $\forall \epsilon > 0$,

$$\left|\hat{I}_1(X^n \to Y^n) - \bar{I}(\mathbf{X} \to \mathbf{Y})\right| = o(n^{-1/2}(\log n)^{5/2+\epsilon}) \ P\text{-a.s.} \tag{21}$$

We can establish similar consistency results for the second estimator $\hat{I}_2$ in (14).

Theorem 2: Let Q be a universal probability assignment, and $(\mathbf{X}, \mathbf{Y})$ be jointly stationary ergodic. Then

$$\lim_{n\to\infty} \hat{I}_2(X^n \to Y^n) = \bar{I}(\mathbf{X} \to \mathbf{Y}) \text{ in } L_1. \tag{22}$$

As was the case for $\hat{I}_1$, if the process $(\mathbf{X}, \mathbf{Y})$ is a jointly stationary ergodic aperiodic Markov process, we can say more about the performance of $\hat{I}_2$ as follows:

Proposition 2: Let Q be the probability assignment in CTW. If $(\mathbf{X}, \mathbf{Y})$ is a jointly stationary ergodic Markov process whose order does not exceed the prescribed maximum depth in CTW, then

$$\lim_{n\to\infty} \hat{I}_2(X^n \to Y^n) = \bar{I}(\mathbf{X} \to \mathbf{Y}) \quad P\text{-a.s. and in } L_1. \tag{23}$$

Furthermore, if $(\mathbf{X}, \mathbf{Y})$ is also aperiodic, there exists a constant C_2 such that

$$\mathbb{E}\left|\hat{I}_2(X^n \to Y^n) - \bar{I}(\mathbf{X} \to \mathbf{Y})\right| \le C_2 n^{-1/2} (\log n)^{3/2}. \tag{24}$$

The rates of convergence for the first two estimators are optimal within a logarithmic factor in the minimax sense.

Proposition 3: Let $\mathcal{P}(\mathbf{X}, \mathbf{Y})$ be any class of processes that includes the class of i.i.d. processes. Then, there exists a positive constant C_3 such that

$$\inf_{\hat{I}} \sup_{\mathcal{P}(\mathcal{X},\mathcal{Y})} \mathbb{E}|\hat{I} - \bar{I}(\mathbf{X} \to \mathbf{Y})| \ge C_3 n^{-1/2}, \tag{25}$$

where the infimum is over all estimators $\hat{I}$ of the directed information rate based on (X^n, Y^n).
Evidently, convergence rate better than $O(n^{-1/2})$ is not attainable even with respect to the class of i.i.d. sources and thus, a fortiori, in our setting of a much larger uncertainty set.

For the third and fourth estimators, we establish the following results.

Theorem 3: Let Q be the probability assignment in CTW. If $(\mathbf{X}, \mathbf{Y})$ is a stationary ergodic Markov process whose order does not exceed the prescribed maximum depth in CTW, then

$$\lim_{n\to\infty} \hat{I}_3(X^n \to Y^n) = \bar{I}(\mathbf{X} \to \mathbf{Y}) \quad P\text{-a.s. and in } L_1. \tag{26}$$

Theorem 4: Let Q be the probability assignment in CTW. If $(\mathbf{X}, \mathbf{Y})$ is a stationary ergodic Markov process whose order does not exceed the prescribed maximum depth in CTW, then

$$\lim_{n\to\infty} \hat{I}_4(X^n \to Y^n) = \bar{I}(\mathbf{X} \to \mathbf{Y}) \quad P\text{-a.s. and in } L_1. \tag{27}$$

V. Experimental Results

In this section, we show how one can use the directed information estimator to detect delay of a channel, and to measure the "causal influence" of one sequence on another. We generate simulated data to detect the channel delay and use real stock market data to detect and measure the causal influence that exists between the Chinese and the US stock markets.

A. Channel Delay Estimation via Shifted Directed Information

Assume a setting depicted as follows: in Fig. 1.

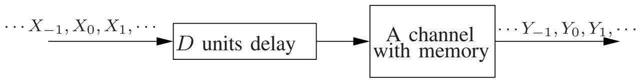
$\cdots X_{-1}, X_0, X_1, \cdots$
D units delay
A channel
with memory
$\cdots Y_{-1}, Y_0, Y_1, \cdots$

Fig. 1. Using the shifted directed information estimation to find the delay D.

Our goal is to find the delay D. We use the shifted directed information $I(Y^{n+d} \to X^n)$ to estimate D, where $I(Y^{n+d} \to X^n)$ is defined as

$$I(Y^{n+d} \to X^n) \triangleq \sum_{i=1}^{n} H(X_i|X^{i-1}) - H(X_i|X^{i-1}, Y^{i+d}). \tag{28}$$

To illustrate the idea, suppose that the binary processes $\mathbf{X}$ and $\mathbf{Y}$ are related as

$$Y_i = X_{i-D} + X_{i-D-1} + W_i, \tag{29}$$

where $W_i \sim$ Bernouli(ϵ) and addition in (29) is modulo 2. Note that the mutual information rate $\lim \frac{1}{n} I(Y^n; X^n)$ is not influenced by D. However, the shifted directed information rate $\lim \frac{1}{n} I(Y^{n+d} \to X^n)$ is highly influenced by D. Assuming that there is no feedback, for $d < D$ we have the Markov chain $Y^{i+d} \to X^{i-1} \to X_i$ due to (29), and therefore $I(Y^{n+d} \to X^n) = 0$. However, for $d \ge D$, $I(Y^{n+d} \to X^n) > 0$. For instance, in the channel example (29), if $W_i = 0$ almost surely, then for $d \ge D$, $I(Y^{n+d} \to X^n) = H(X^n)$. Therefore, we can use the shifted directed information $I(Y^{n+d} \to X^n)$ to estimate D.

For the sake of simplicity, we only show the estimation results using $\widehat{I}_2$, other estimators have similar outputs. Fig. 2 depicts $\hat{I}_2(Y^{n+d} \to X^n)$ where $n = 10^6$ for the setting in Fig. 1, where the input is a binary stationary Markov process of order one and the channel is given by (29). The delay of the channel is $D = 2$. One can note clearly that for $d < D$, $\hat{I}_2(Y^{n+d} \to X^n)$ is very close to zero and for $d \ge D$, $\hat{I}_2(Y^{n+d} \to X^n)$ is significantly larger than zero.

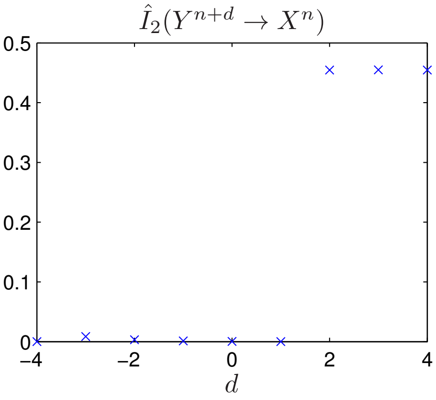

Fig. 2. The value of $\hat{I}_2(Y^{n+d} \to X^n)$ where $n = 10^6$ for the setting depicted in Fig. 1 with $D = 2$. One can observe that when $d < D$, $\hat{I}_2(Y^{n+d} \to X^n) = 0$ and for $d \ge D$, $\hat{I}_2(Y^{n+d} \to X^n) > 0$.

B. Causal Influence Measurement

There is extensive literature on detecting and measuring causal influence. See, for example, [26] for a recent survey of some of the common tools and approaches in biomedical informatics. One particularly celebrated tool - in both the life and economics sciences - for assessing whether and to what extent one time series influences another is the Granger causality test [27]. It is a simple exercise to verify that under jointly Gauss-Markov assumptions, the Granger causality coincides with the directed information (up to a multiplicative constant).

Assuming for every pair (X_i, Y_i), X_i happens earlier than Y_i. It can be easily verified that $I(X^n \to Y^n) = 0$ if and only if $P(y_i|x^i, y^{i-1}) = P(y_i|y^{i-1})$ for $i \ge 1$, and $I(Y^{n-1} \to$

$X^n) = 0$ if and only if $P(x_i|x^{i-1}, y^{i-1}) = P(x_i|x^{i-1})$ for $i \geq 1$. More generally, the directed information $I(X^n \to Y^n)$ quantifies how much $\mathbf{X}$ causally influences $\mathbf{Y}$, while the directed information in the reverse direction $I(Y^{n-1} \to X^n)$ quantifies how much $\mathbf{Y}$ influences $\mathbf{X}$.

To illustrate this idea, we compute the directed information rate between the Hang Seng Index (HSI) and the Dow Jones Index (DJIA) using data from 1990 and 2011 on a daily scale. Since everyday the HSI changes before DJIA, HSI should play the role of process $\mathbf{X}$ in the estimation. We discretize the value of stock market into three values: -1, 1, and 0, by going down by more than 0.8%, going up by more than 0.8%, and changes between them, respectively.

We denote by X_i and Y_i the (quantized ternary valued) change in the HSI and the DJIA in day i, respectively, and estimate $\frac{1}{n}I(X^n;Y^n)$, $\frac{1}{n}I(X^n \to Y^n)$, and $\frac{1}{n}I(Y^{n-1} \to X^n)$, using all four algorithms. Fig. 3 plots our estimates of these information-theoretic measures.

Evidently, the reverse directed information is much higher than the directed information; hence we can say that between 1990 and 2011, it was the Chinese market that was influenced more by the US market rather than the other way around.

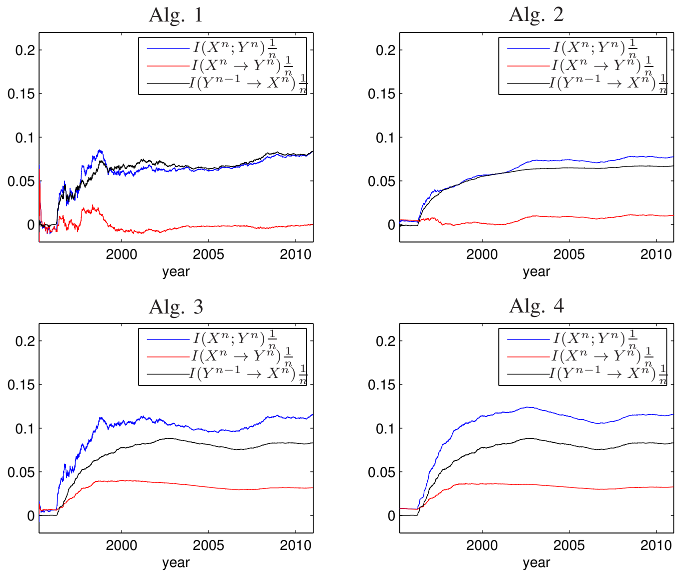

Fig. 3. Estimates of information-theoretic measures between HSI denoted by $\mathbf{X}$, and DJI denoted by $\mathbf{Y}$. It is clear that the reverse directed information is much higher than the directed information, hence it is DJI that causally influences HSI rather than the other way around.

References

[1] H. Marko, "The bidirectional communication theory–a generalization of information theory," *IEEE Trans. Commum.*, vol. COM-21, pp. 1345–1351, 1973.

[2] J. L. Massey, "Causality, feedback, and directed information," in *Proc. IEEE Int. Symp. Inf. Theory Appl.*, Honolulu, HI, Nov. 1990, pp. 303–305.

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

[4] ——, "Capacity results for the discrete memoryless network," *IEEE Trans. Inf. Theory*, vol. 49, no. 1, pp. 4–21, 2003.

[5] S. Tatikonda and S. Mitter, "The capacity of channels with feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 1, pp. 323–349, 2009. [Online]. Available: http://dx.doi.org/10.1109/TIT.2008.2008147

[6] Y.-H. Kim, "A coding theorem for a class of stationary channels with feedback," *IEEE Trans. Inf. Theory*, vol. 54, no. 4, pp. 1488–1499, 2008.

[7] H. H. Permuter, T. Weissman, and A. J. Goldsmith, "Finite state channels with time-invariant deterministic feedback," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 644–662, 2009.

[8] H. H. Permuter, Y.-H. Kim, and T. Weissman, "Interpretations of directed information in portfolio theory, data compression, and hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 3248–3259, Jun. 2011.

[9] P. Mathai, N. C. Martins, and B. Shapiro, "On the detection of gene network interconnections using directed mutual information," in *Proc. UCSD Inf. Theory Appl. Workshop*, 2007.

[10] A. Rao, A. O. Hero, D. J. States, and J. D. Engel, "Using directed information to build biologically relevant influence networks," *Journal on Bioinformatics and Computational Biology*, vol. 6, no. 3, pp. 493–519, 2008.

[11] S. Verdú, "Universal estimation of information measures," in *Proc. of IEEE ISOC ITW2005 on Coding and Complexity*, 2005.

[12] A. D. Wyner and J. Ziv, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. Inf. Theory*, vol. 35, no. 6, pp. 1250–1258, 1989.

[13] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1270–1279, 1993.

[14] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal divergence estimation for finite-alphabet sources," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3456–3475, 2006.

[15] M. Burrows and D. J. Wheeler, *A block-sorting lossless data compression algorithm*. Digital Systems Research Center, Tech. Rep. 124, 1994.

[16] 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, pp. 653–664, 1995.

[17] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal entropy estimation via block sorting," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1551–1561, 2004.

[18] J. Yu and S. Verdú, "Universal erasure entropy estimation," in *Proc. IEEE Int. Symp. Inf. Theory*, 2006.

[19] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of Computational Neuroscience: Special Issue on Methods of Information Theory in Computational Neuroscience*, 2011. [Online]. Available: http://dx.doi.org/10.1007/s10827-010-0247-2

[20] L. Zhao, Y.-H. Kim, H. H. Permuter, and T. Weissman, "Universal estimation of directed information," in *Proc. IEEE Int. Symp. Inf. Theory*, 2010, pp. 230–234.

[21] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *IEEE Trans. Inf. Theory*, submitted. [Online]. Available: http://arxiv.org/abs/1201.2334

[22] J. L. Massey and P. C. Massey, "Conservation of mutual and directed information," in *Proc. IEEE Int. Symp. Inf. Theory*, 2005, pp. 157–158.

[23] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[24] F. Willems and T. Tjalkens, *Complexity Reduction of the Context-Tree Weighting Algorithm: A Study for KPN Research*. Tech. Rep. Univ. Eindhoven, Eindhoven, The Netherlands, EIDMA Rep. RS.97.01, 1997.

[25] T. J. Tjalkens, Y. M. Shtarkov, and F. M. J. Willems, "Sequential weighting algorithms for multi-alphabet sources," in *6th Joint Swedish-Russian International Workshop on Information Theory*, 1993, pp. 230–234.

[26] S. Kleinberg and G. Hripcsak, "A review of causal inference for biomedical informatics," *Journal of Biomedical Informatics*, vol. 44, no. 6, pp. 1102–1112, 2011.

[27] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.