A Users' Guide for the MATALB Package of 'Universal Estimation of Directed Information'

Jiantao Jiao
jiantao@stanford.edu

June 7, 2012

Abstract

It is a users' guide for the MATLAB package of the paper 'Universal Estimation of Directed Information'. It explains briefly the theory of why directed information plays a crucial role in causal analysis, and explains how to use this MATLAB package to facilitate experimentations.

1 Brief Intro to Directed Information

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.

Beyond information theory, directed information is a valuable tool in biology, for it provides an alternative to Granger causality [3], which has been perhaps the most widely-established means of identifying causal inference between two processes.

It has been recently demonstrated that directed information theory unifies causality notions that appeared in the history of economics, statistics, physics, and information theory. For a good overview, please see Amblard and Michel [4].

Now we briefly describe directed information and some related theory. We use uppercase letters $X, Y, \ldots$ to denote random variables, and denote the n-tuple $(X_1, X_2, \ldots, X_n)$ as X^n.

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

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

where $H(Y^n \| X^n)$ is the *causally conditional entropy* [5], defined as

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

Compared with the definition of mutual information,

$$I(X^n; Y^n) = H(Y^n) - H(Y^n | X^n), \tag{3}$$

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.

There are some interesting properties of directed information, some can be found in [4–6]. For brevity, here we restrict to showing two enlightening conservation laws.

Massey and Massey [6] show the conservation law

$$I(X^n; Y^n) = I(X^n \to Y^n) + I(Y^{n-1} \to X^n), \tag{4}$$

where

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

denotes the *reverse* directed information.

Equation (4) is particularly enlightening in settings where X_i and Y_i appear alternately, as shown in Figure 1.

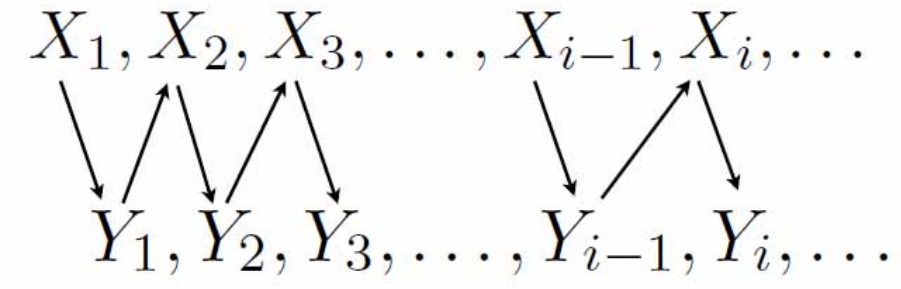

Figure 1: The temporal order of X^n and Y^n

Then, if we note what is captured by $I(X^i; Y_i | Y^{i-1})$ and $I(Y^{i-1}; X_i | X^{i-1})$, and the relationship

$$I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1}) \tag{5}$$

$$I(Y^{n-1} \to X^n) = \sum_{i=1}^{n} I(Y^{i-1}; X_i | X^{i-1}), \tag{6}$$

we can see that $I(X^n \to Y^n)$ reflects how much X^n could help *causally* predict Y^n, and $I(Y^{n-1} \to X^n)$ reflects how much Y^{n-1} could help *causally* predict X^n.

In some situations, X^n and Y^n may happen simultaneously, such as neurological network data and prices of different stock that vary at the same time. The following is another conservation law stated in [4] which, in such situations, may be more insightful than that in Equation (4):

$$I(X^n; Y^n) = I(X^{n-1} \to Y^n) + I(Y^{n-1} \to X^n) + \sum_{i=1}^{n} I(X_i; Y_i | X^{i-1}, Y^{i-1}) \tag{7}$$

The relation in Equation (7) is symmetric and the three terms on the right side have natural interpretations:

1. $I(X^{n-1} \to Y^n)$ reflects how the history of X^n helps causally predict Y^n;
2. $I(Y^{n-1} \to X^n)$ reflects how the history of Y^n helps causally predict X^n;
3. $\sum_{i=1}^{n} I(X_i; Y_i | X^{i-1}, Y^{i-1})$ reflects how data pair (X_i, Y_i) influence each other instantaneously when they appear together.

If we compare Equation (4) with (7), we can see

$$I(X^n \to Y^n) = I(X^{n-1} \to Y^n) + \sum_{i=1}^{n} I(X_i; Y_i | X^{i-1}, Y^{i-1}), \tag{8}$$

that is to say, the *instantaneous* influence can be calculated through directed information $I(X^n \to Y^n)$ and reverse directed information $I(X^{n-1} \to Y^n)$.

2 How to use this MATLAB package

There are two main functions that users may frequently call: compute_DI_MI.m and ctwprob.m.

2.1 compute_DI_MI

[MI, DI, rev_DI]=compute_DI_MI(X,Y,Nx,D,alg,shift_ratio,prob,flag)

Here we briefly explain what these parameters mean.

1. 'X' and 'Y' are input sequences with the same length;
2. 'Nx' is the size of the alphabet of process X, with the assumption that processes X and Y have the same size of alphabets;
3. 'D' is the maximum depth of the context tree used in basic CTW method;
4. 'alg' is a string that indicates which directed information estimator in [7] is used, namely 'E1', 'E2', 'E3', 'E4'.
5. 'prob' is a struct containing probability assignments generated by function 'ctwprob' in case users want to avoid running CTW many times on the same data sequences;
6. 'flag' indicates whether compute_DI_MI calculates the CTW probability assignment. If 'flag' = 0, then the input parameter 'prob' is taken as invalid, function 'compute_DI_MI' calculates CTW probability assignment itself; else the input parameter 'prob' is valid and 'compute_DI_MI' doesn't compute CTW probability assignment but directly take if from input parameter;
7. 'MI' is a vector of estimated mutual information $\hat{I}(X^n;Y^n)$ for a sequence of increasing n;
8. 'DI' is a vector of estimated directed information $\hat{I}(X^n \to Y^n)$ for a sequence of increasing n;
9. 'rev_DI' is a vector of estimated reverse directed information $\hat{I}(Y^{n-1} \to X^n)$ for a sequence of increasing n;
10. shift_ratio determines what proportion of the estimated directed information and mutual information will be discarded when generating the outputs. For example, if 'shift_ratio' = 0.3, and the length of input process X is 1000, then output 'MI' will have length of $(1-0.3)*1000 = 700$, and 'MI' only gives the estimated mutual information $\hat{I}(X^n;Y^n)$ for consecutive n no smaller than 300, so do 'DI' and 'rev_DI'.

2.2 ctwprob

[prob] = ctwprob(X,Y,Nx,D)

Here 'X','Y','Nx','D' are of the same meanings as shown above.

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

[4] P.-O. Amblard and O. J. Michel, "Relating granger causality to directed information theory for networks of stochastic processes," 2011. [Online]. Available: http://arxiv.org/abs/0911.2873

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

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

[7] J. Jiao, H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *submitted to IEEE Trans. Inf. Theory*, 2012.