Improvement of Symbolic Transfer Entropy

Dimitris Kugiumtzis

Abstract—A number of measures have been proposed for the direction of the coupling between two time series, and transfer entropy (TE) has been found in recent studies to perform consistently well in different settings. Symbolic transfer entropy (STE) has been very recently proposed as a variation of the transfer entropy operating on the ranks of the components of the reconstructed vectors rather than the reconstructed vectors themselves. Here, an improvement of STE is proposed. Specifically, the ranks of the samples of the response system for given time steps ahead are computed with regard to the current reconstructed vector. The grounds of this modification are given and the new measure, called Transfer Entropy on Rank Vectors (TERV), is compared to STE and TE on different settings of state space reconstruction, time series length and observational noise. The results on two simulated systems have shown that the detection of the direction and strength of coupling is improved with TERV over both STE and TE.

Index Terms—bivariate time series, coupling, information measures, transfer entropy, rank vectors

I. INTRODUCTION

The fundamental concept for the dependence of one variable $Y$ measured over time on another variable $X$ measured synchronously is the Granger causality [1]. While Granger defined the direction of interaction in terms of the contribution of $X$ in predicting $Y$, many variations of this concept have been developed, starting with linear approaches in the time and frequency domain (e.g. see [2], [3]) and extending to nonlinear approaches focusing on phase or event synchronization [4], [5], [6], comparing neighborhoods of the reconstructed points from the two time series [7], [8], [9], [10], [11], [13], [14], and measuring the information flow between the time series [15], [16], [17], [18], [19].

Among the different proposed measures we concentrate here on the last class of measures, and particularly on the transfer entropy (TE) [15] and the most recent variant of TE operating on rank vectors, called symbolic transfer entropy (STE) [18] (see also [20] for a similar measure). Other information measures, such as mean conditional mutual information [19] and coarse-grained transinformation rate [16] are rather similar to transfer entropy and are therefore not included in this study. There have been a number of comparative studies of information flow measures and other coupling measures giving varying results. In all the studies where TE was considered, it performed at least as good as the other measures [21], [22], [28]. The STE is proposed as an improvement of TE in real world applications, where noise may mask details of the fine structure, that can be better treated by coarse discretization using ranks instead of samples. We have studied STE and propose here a modification of it in order to improve the correct detection of the direction as well the strength of coupling when it is present. In the following, the TE and STE are presented briefly in Section II, and the proposed measure is described in Section III. Then the results of a simulation study comparing the proposed measure to TE and STE are presented in Section IV, and finally conclusions are given in Section V.

II. INFORMATION MEASURES

Transfer entropy (TE) is a measure of the flow of information from the driving system, denoted $X$, to the response system, denoted $Y$ [15]. Supposing a representative quantity of system $X$ is measured in terms of a scalar time series $\{x_t\}_{t=1}^N$ and respectively $\{y_t\}_{t=1}^N$ for $Y$, TE for the direction from $X$ to $Y$ can be defined in terms of the Shannon entropy $H(x) = \sum p(x) \log p(x)$ as

$$\text{TE}_{X \rightarrow Y} = -H(y_{t+1}, x_t, y_t) + H(x_t, y_t) + H(y_{t+1}, y_t) - H(y_t)$$

or directly in terms of distribution functions as

$$\text{TE}_{X \rightarrow Y} = \sum p(y_{t+1}, x_t, y_t) \log \frac{p(y_{t+1} | x_t, y_t)}{p(y_{t+1} | y_t)}$$

where $p(y_{t+1}, x_t, y_t)$, $p(y_{t+1} | x_t, y_t)$, and $p(y_{t+1} | y_t)$ are the joint and conditional probability mass functions (pmf). The summation is over all the cells of a suitable partition of the joint variable vectors appearing as arguments in the pmfs or $H$. The points $x_t$ and $y_t$ appearing as arguments in eq.(1) and eq.(2) are reconstructed with the method of delays, so that we have $x_t = [x_t, x_{t-\tau_1}, \ldots, x_{t-(m_\tau-1)\tau_1}]$ and $y_t = [y_t, y_{t-\tau_2}, \ldots, y_{t-(m_\tau-1)\tau_2}]$, allowing different delay parameters $\tau_1$, $\tau_2$ and embedding dimensions $m_\tau$, $m_\eta$ for the systems $X$ and $Y$, respectively.

The estimation of TE requires the estimation of the pmfs in eq.(2), or the probability density functions assuming the integral form and no binning. The pmfs are estimated directly by the relative frequency of occurrence of points in each cell, so the only complication is to choose a suitable binning [24], [23]. However, for high-dimensional reconstructions, the binning estimators are data demanding, and therefore estimators of the probability density functions are more appropriate for TE estimation, such as kernels [25], nearest neighbors [26], and correlation sums [27]. Using the latter approach, the TE estimator is given as

$$\text{TE}_{X \rightarrow Y} = \log \frac{C(y_{t+1} | x_t, y_t) C(y_t)}{C(x_t, y_t) C(y_{t+1} | y_t)}$$

where $C(y_{t+1} | x_t, y_t)$, $C(y_t)$, $C(x_t, y_t)$ and $C(y_{t+1} | y_t)$ are the correlation sums, which estimate the probability of
inter-points distances less than some given radius for the points of the form \([yt+1, xt, yt], yt, [xt, yt]\) and \([yt+1, yt]\), respectively. The corresponding vector dimensions are \(1 + m_x + m_y, m_x + m_y\) and \(1 + m_y\). We use the Euclidean norm for the distances and define the radius as the product of 0.1, multiplied with the standard deviation of the data, and the square root of the vector dimension at each case (the latter is used to standardize the Euclidean norm). The use of 0.1 is a trade-off of having enough points within a radius to assure stable estimation of the point distribution and preserving neighborhoods to retain details of the point distribution. Still, for high-dimensional points, even this radius may be insufficient to provide stable estimation.

A. Symbolic transfer entropy

The authors in [18] defined the so-called symbolic transfer entropy (STE) as the transfer entropy defined on rank vectors formed by the reconstructed points. For each point \(y_t\), the ranks of its components in ascending order assign a rank vector \(\hat{y}_t = [r_1, r_2, \ldots, r_m]\), where \(r_j \in [1, 2, \ldots, m]\) for \(j = 1, \ldots, m_y\), is the rank order of the component \(y_{t-(j-1)}\) (for two equal components of \(y_t\) the smallest rank is assigned to the component appearing first in \(y_t\)). Substituting \(y_{t+1}\) in eq.(1) with the rank vector at time \(t+1\), \(\hat{y}_{t+1}\), STE is defined as

\[
\text{STE}_{X \rightarrow Y} = \sum p(\hat{y}_{t+1}, \hat{x}_t, \hat{y}_t) \log \frac{p(\hat{y}_{t+1}, \hat{x}_t, \hat{y}_t)}{p(\hat{y}_{t+1})}.
\]

The estimation of STE from eq.(4) or eq.(5) is straightforward as the pmfs are naturally defined on the rank vectors and no binning or advanced estimator of probability density function is involved. There is a great advantage of using a rank vector \(\hat{y}_t\) over a binning of \(y_t\), say using \(b\) bins for each component: the possible vectors from binning are \(b^m\) while the possible combinations of the rank vectors are \(m!\). For example, for \(b = m_y = 4\), there are 256 cells from binning and only 24 combinations of rank vectors. Still, the estimation of the probability of occurrence of a rank vector becomes unstable as the dimension increases. Especially, for the joint vector of ranks \([\hat{y}_{t+1}, \hat{x}_t, \hat{y}_t]\) the dimension is \(2m + m_y\), for which the equivalent of TE is \([\hat{y}_{t+1}, \hat{x}_t, \hat{y}_t]\) and has dimension \(1 + m_x + m_y\).

III. MODIFICATION OF SYMBOLIC TRANSFER ENTROPY

The conversion of the scalar \(y_{t+1}\) to the rank vector \(\hat{y}_{t+1}\) was chosen rather arbitrarily by the authors in [18] in order to express \(y_{t+1}\) in terms of ranks. Under this conversion, STE is not the direct analogue to TE using ranks instead of samples. The problem is not the use of \(y_{t+1}\) instead of \(\hat{y}_{t+1}\) in the definition of TE in eq.(1) or eq.(2) because \(p(\hat{y}_{t+1}, \hat{x}_t, \hat{y}_t) = p(y_{t+1}, x_t, y_t)\), as all components but \(y_{t+1}\) of the vector \(\hat{y}_{t+1}\) are also components of \(y_t\). The same holds for the conditional pmfs (and the same holds also for the two correlation sums in which \(y_{t+1}\) appears in eq.(3)).

Let us first assume that \(\tau_y = 1\). A first problem lies in the fact that when deriving the rank vector \(\hat{y}_{t+1}\), associated with \(y_{t+1}\), the rank of the last component of \(y_{t+1}\), \(y_{t-m_y+1}\), is not considered. As an example, consider the vector \(y_t = [y_{t-1}, y_{t-2}, y_{t-3}]\) with a corresponding rank vector \(\hat{y}_t = [1, 2, 3, 4]\). If \(y_{t+1}\) is between \(y_t\) and \(y_{t-1}\) then \(\hat{y}_{t+1} = [2, 1, 3, 4]\), if it is between \(y_{t-1}\) and \(y_{t-2}\), then \(\hat{y}_{t+1} = [1, 2, 4]\), and finally if \(y_{t+1}\) is larger than \(y_{t-2}\) (the largest of all components in \(y_{t+1}\) then \(\hat{y}_{t+1} = [4, 1, 2, 3]\). The 4 possible scenarios are shown in Fig. 1.

The definition of rank vector \(\hat{y}_{t+1}\) accounts only for the possible rank positions of \(y_{t+1}\) with respect to the last \(m_y - 1\) samples, ignoring the sample \(y_{t-m_y+1}\), here \(y_{t-3}\). With regard to the same example, \(\hat{y}_{t+1} = [1, 2, 3, 4]\) assigns to both cases \(y_{t-2} < y_{t-1} < y_{t-3}\) and \(y_{t-1} < y_{t-2}\) (see Fig. 1). Indeed there are 5 possible rank positions of \(y_{t+1}\) in the augmented vector \([y_{t+1}, y_t, y_{t-1}, y_{t-2}, y_{t-3}]\), but when forming the joint rank vector \([\hat{y}_{t+1}, \hat{y}_t]\) (as in the computation of STE) there are only \(4!\) possible rank orders. In general, there are \((m_y + 1)!\) possible rank orders for the joint vector \([\hat{y}_{t+1}\), \(\hat{y}_t]\), but STE estimation represents them in \(m_y!\cdot(m_y-1)!\) rank orders of \([\hat{y}_{t+1}, \hat{y}_t]\).

The pmf of the rank vector derived from \([y_{t+1}, y_t]\) and the rank vector \([\hat{y}_{t+1}; \hat{y}_t]\) are shown in Fig. 2 for uniform white noise data and \(m_y = 3\). There are \((m_y + 1)! = 24\) equivalent rank orders for \([y_{t+1}, y_t]\) (see Fig. 2a) but only \(m_y! = 6\) are different \([\hat{y}_{t+1}; \hat{y}_t]\) are found, where \(m_y! = 6\) of them have about double probability, each corresponding to two distinct rank orders that could not be distinguished. As a result, the Shannon entropy is underestimated here. Using \(n = 10^{16}\) samples and the ranks of \([y_{t+1}; \hat{y}_t]\) we found \(H = 4.5846\) bits and using \([\hat{y}_{t+1}; \hat{y}_t]\) we found \(H = 4.0865\) bits, while the true Shannon entropy is \(H = \log_2(1/24) = 4.5850\).

The situation changes if a further time step ahead is used. In general, allowing for \(y_{t-T}\), where \(T \geq 1\), the possible rank orders of \([y_{t+T}; \hat{y}_t]\) are again \((m_y + 1)!\), but for \([\hat{y}_{t+T}; \hat{y}_t]\) are \((m_y + 1)!\cdot(m_y-1)!\). For example, \(m_y = 3\) and \(T = 2\) gives 36 possible rank orders \([y_{t+T}; \hat{y}_t]\), while the possible rank orders for 4 samples are 24. This increase holds in general for \(T > 1\). For uniform white noise data, this result is overestimation of the true \(H = 4.5850\) using the rank orders of \([y_{t+T}; \hat{y}_t]\), \(H =\) IC CSA 2009, June 29 – July 2, Le Havre, Normandy, France 339
4.9736 ($n = 10^{10}$) while using the rank orders of $[y_{t+T}, y_t]$ we estimated $H = 4.5849$.

Thus we propose the following modifications to STE:

1) If $T = 1$, then in the definition of STE replace $y_{t+1}$ by $\hat{y}_{t+1}$, i.e. the rank of $y_{t+1}$ in the augmented vector $[y_{t+1}, \ldots, y_{t+T}]$.

2) If $T > 1$, then replace $y_{t+T}$ by $\hat{y}_{t+T} = [\hat{y}_{t+1}, \ldots, \hat{y}_{t+T}]$, the ranks of $y_{t+1}, \ldots, y_{t+T}$ in the augmented vector $[y_{t+1}, \ldots, y_{t+T}, \ldots, y_{t+1}]$.

For $T > 1$, the proposal is to use all the ranks for times $t, t+1, \ldots, t+T$ in order to keep track of the effect of $X$ on the evolution of the time series of $Y$ up to $T$ time steps ahead. Similar reasoning for $T > 1$ was used for the measure of the coarse-grained transinformation rate [22] and we have used $T > 1$ also for TE [28]. Thus the proposed measure of transfer entropy with rank vectors (TERV) for $T$ steps ahead is

$$\text{TERV}_{X \rightarrow Y} = \sum_{t} \left( H(\hat{y}_{t}, \hat{y}_{t+1}, \hat{y}_{t+2}) - H(\hat{y}_{t}, \hat{y}_{t+1}) - H(\hat{y}_{t+1}, \hat{y}_{t+2}) + H(\hat{y}_{t+1}) \right).$$

The TERV measure is the direct analogue to TE using ranks and extends the measure of information flow from $X$ at time $t$ to $Y$ for a range of $T$ time steps ahead $t$.

Note that when a lag $\tau_y > 1$ is used for the state space reconstruction of $y_t$, there are up to $m_y1 \cdot m_y2$ different rank vectors $[\hat{y}, \hat{y}]$ in the computation of STE. On the other hand, for TERV there are $(T + m_y1)!$ different rank vectors $[\hat{y}, \hat{y}]$. Thus for $\tau_y > 1$, the distortion of the domain of the rank vectors by STE may be large, e.g. for $\tau_y = 2$ and $T = 1$, the pmfs and entropies are computed on $(m_y1 + 1)!$ different rank orders for TERV and $m_y1 \cdot m_y2$ for STE.

IV. ESTIMATION OF INFORMATION MEASURES FROM SIMULATED SYSTEMS

As it was shown for the example of uniform white noise the distortion of the domain of the rank vectors $[\hat{y}, \hat{y}]$ instead of directly using the rank vectors $[\hat{y}, \hat{y}]$ instead has a direct effect on the estimation of entropy. While for uncoupled systems $X$ and $Y$ the entropy terms involving $[\hat{y}, \hat{y}]$ cancel out in the expression of TERV (and respectively for STE), in the presence of coupling some bias is introduced in the estimation of the coupling measure by STE. Using TERV instead this bias is removed.

[Fig. 2. (a) Estimated pmf for the ranks of $[y_{t+1}, y_t]$ with $m_y = 3$ (probabilities are in ascending order), where the samples $y_t$ are from a uniform white noise time series of length $n = 10^{10}$. (b) Same as in (a) but for the rank vector $[\hat{y}_{t+1}, \hat{y}_t]$.

[Fig. 3. (a) Median (solid line), 10th and 90th percentiles (dashed lines) of TE computed on 100 noise-free realizations of length $n = 10^4$ from the system of two unidirectionally coupled Henon maps for varying coupling strengths. The other parameters are $T = 1$, $\tau_T = 1$ and $m_X = m_Y = 2$. The direction $X \rightarrow Y$ is shown with black lines and $Y \rightarrow X$ with gray Online cyan) lines. (b) Same as (a) but for STE. (c) Same as (a) but for TERV. (d) AUROC computed on the 100 realizations for each of the two directions and for the measures TE, STE and TERV, as given in the legend.

We compare the estimation of coupling (strength and direction) in a system of two unidirectionally coupled Henon maps

$$x_{t+1} = 1.4 - x_t^2 + 0.3x_{t-1},$$

$$y_{t+1} = 1.4 - cy_{t} + (1-c)y_t^3 + 0.3y_{t-1}$$

with coupling strengths

$$c = 0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6.$$
The performance of the coupling measures changes in the presence of noise. For the same setup as that in Fig. 3, but adding to the bivariate time series 20% Gaussian white noise, we observe that TERV performs best, followed by STE, while TE has larger variance and fails to detect the correct direction of coupling when it is weak (see Fig. 4). It is notable that the discriminating power of TERV has not been affected much by noise. The AUROC for TERV increases faster with \( c \) than for the other two measures and reaches the highest level at \( c = 0.2 \), while both TE and STE reaches this level when coupling gets strong (\( c = 0.5 \)).

We have estimated TE, STE and TERV on the coupled Henon system for different settings of embedding dimensions (keeping \( m_x = m_y \)), time steps ahead \( T \), time series length \( n \) and noise level. For \( n \) small and \( m_y \) large and mostly for noisy time series, the computation of TE was not possible due to the lack of points within the given radius. Therefore the performance of TE was worse than for STE and TERV. This can be seen in Fig. 5, where the AUROC is shown for the three measures as a function of \( m_y \) for two settings of noise-free short time series and noisy but longer time series. It seems that estimating the information flow for \( T = 3 \) time steps ahead increases the detection of correct direction of weak coupling (e.g. \( c = 0.1 \) in the results of Fig. 5) for all but the STE measures. For the noise-free data, the differences in AUROC among the three measures are small, and TERV for \( T = 3 \) performs best giving AUROC=1 for all \( m_y \) (see Fig. 5a). TERV for \( T = 3 \) performs best also for the noisy data (see Fig. 5b) and generally the AUROC for TERV was almost always higher than for STE, which in turn was higher than for TE.

Similar simulations have been run on a Rössler system driving a Lorenz system given as (subscript 1 for Rössler, 2 for Lorenz)

\[
\begin{align*}
\dot{x}_1 &= -6(y_1 + z_1) \\
\dot{y}_1 &= -6(x_1 + 0.2y_1) \\
\dot{z}_1 &= -6(0.2 + z_1(y_1 - 5.7)) \\
\dot{x}_2 &= 10(x_2 + y_2) \\
\dot{y}_2 &= 28x_2 - y_2 - x_2z_2 + cy_2^2 \\
\dot{z}_2 &= x_2y_2 - z_2 + \frac{5}{2}z_2
\end{align*}
\]

for different coupling strengths \( c \). In all parameter settings, TERV was improving the results of STE, especially when \( T > 1 \) was used (we tested for \( T = 2 \) and \( T = 3 \)). An example is shown in Fig. 6 for weak coupling, \( c = 0.5 \). While STE and TERV score about the same in AUROC when \( T = 1 \), the use of a longer time horizon for the information flow \( T = 3 \) in the estimation of TERV gives perfect detection of coupling direction for all but insufficient embeddings, \( m_x = m_y = 2 \) (a small decrease is observed for \( m_y = 7 \)). This high performance of TERV and \( T = 3 \) holds also when noise is added to the data, while in this case STE and TERV with \( T = 1 \) fail to detect the coupling direction also for \( m_y = 3 \).

V. CONCLUSION

The use of ranks of consecutive samples instead of samples themselves in the estimation of the transfer entropy (TE) seems to gain robustness in conditions often met in real world applications, i.e. the presence of noise and the use of large embedding dimensions. This was confirmed by our results in this simulation study. Given that TE based on ranks can be a useful measure of information flow and direction of coupling, we have studied the recently proposed rank-based transfer entropy, termed symbolic transfer entropy (STE), and suggested a modified version of STE, which we termed TERV. The first modification is to use the rank of \( y_{t+1} \) (one time step ahead for the response time series) in the augmented reconstructed state vector \( y_t \), including also \( y_{t+1} \), instead of

Fig. 4. As Fig. 3, but with 20% Gaussian white noise added to the data.

Fig. 5. (a) AUROC computed for different \( m_y (m_x = m_y) \) on 100 realizations of the weakly coupled Rössler–Lorenz system \((c = 0.5)\) for each of the two directions, for the measures TE, STE and TERV and for time steps ahead \( T \) as given in the legend. The time series are noise-free and \( n = 1024 \). (b) As in (a) but for \( n = 4096 \) and 20% additive Gaussian white noise.

Fig. 6. (a) AUROC computed for different \( m_y (m_x = m_y) \) on 100 realizations of the weakly coupled Rössler–Lorenz system \((c = 0.5)\) for each of the two directions and for the measures STE \((T = 1)\) and TERV \((T = 1 \text{ and } T = 3)\), as given in the legend. The time series are noise-free and \( n = 1024 \). (b) As in (a) but when 20% Gaussian white noise is added to the data.
considering a whole rank vector for \( y_{t+1} \) as done in STE. We showed that indeed this correction gives accurate estimation of the true entropy of the rank vector derived from the joint vector of \( y_t \) and \( y_{t+1} \). Further, we suggested to allow the time step ahead to be \( T > 1 \) and use the ranks of all samples at the \( T \) future times \( (y_{t+1}, \ldots, y_{t+T}) \) derived from the augmented vector containing the current vector \( y_t \) and these future samples. The proposed TERV measure was compared to TE and STE on two synthetic systems, a coupled map and a coupled flow, and the level of detection of the coupling direction was assessed by the area under the receiver operating characteristic curve (AUROC). TERV gave consistently higher AUROC than STE, and when the data were noisy also higher than TE. In particular, the use of \( T > 1 \) improved the performance of TERV.

There are other issues that have not been addressed in this study, such as the use of other settings of state space reconstruction (e.g., delays larger than one and different embedding dimensions for the driver and the response system). Also, there are problems in the estimation of the measures, including TERV, that have not been discussed here, such as the statistical significance of a measures when the systems are uncoupled, and the increase of a measure also for the opposite (wrong) coupling direction when the coupling strength increases. This study is by no means extensive or complete and the measures should also be compared in many different systems (identical and non-identical).

REFERENCES