Data based Intervention Approach for Complexity-Causality Measure Supplementary Material

Aditi Kathpalia, Nithin Nagaraj

In this supplementary material, we provide details of our proposed method Compression-Complexity Causality (CCC), which are not covered in the main paper. We explain how compression-complexity is computed for individual and a pair of time series as well as the way dictionary construction is done for estimating conditional CCC for multi-variate measurements. We also describe the criteria and rationale for choosing the parameters of CCC and details of our MATLAB implementation that is made available for free download and use.

1 Individual and Joint Compression Complexities

In this section, we define how individual and joint compression-complexities are computed using the Effort-To-Compress (ETC) measure [1].

1.1 ETC measure for a time series: $ETC(X)$

Since ETC expects a symbolic sequence as its input (of length > 1), the given time series should be binned appropriately to generate such a sequence. Once such a symbolic sequence is available, ETC proceeds by parsing the entire sequence (from left to right) to find that pair of symbols in the sequence which has the highest frequency of occurrence. This pair is replaced with a new symbol to create a new symbolic sequence (of shorter length). This procedure is repeated iteratively and terminates only when we end up with a constant sequence (whose entropy is zero since it consists of only one symbol). Since the length of the output sequence at every iteration decreases, the algorithm will surely halt. The number of iterations needed to convert the input sequence to a constant sequence is defined as the value of ETC complexity. For example, the input sequence '12121112' gets transformed as follows: $12121112 \mapsto 33113 \mapsto 4113 \mapsto 513 \mapsto 63 \mapsto 7$. Thus, $ETC(12121112) = 5$. ETC achieves its minimum value (0) for a constant sequence and maximum value $(m-1)$ for a m length sequence with distinct symbols. Thus, we normalize the ETC complexity value by dividing by $m-1$. Thus normalized $ETC(12121112) = \frac{5}{7}$. Note that normalized ETC values are always between 0 and 1 with low values indicating low complexity and high values indicating high complexity.

1.2 Joint ETC measure for a pair of time series: $ETC(X,Y)$

We perform a straightforward extension of the above mentioned procedure ($ETC(X)$) for computing the joint ETC measure $ETC(X,Y)$ for a pair of input time series X and Y of the same length. At every iteration, the algorithm scans (from left to right) simultaneously X and Y sequences and replaces the most frequent jointly occurring pair with a new symbol for both the pairs. To illustrate it by an example, consider, $X = 121212$ and $Y = abacac$. The pair (X,Y) gets transformed as follows: $(121212, abacac) \mapsto (1233, abdd) \mapsto (433, edd) \mapsto (53, fd) \mapsto (6, g)$. Thus, $ETC(X,Y) = 4$ and normalized value is $\frac{4}{5}$. It can be noted that $ETC(X,Y) \leq ETC(X) + ETC(Y)$.

2 Dictionary building for conditional CCC

To estimate causality from time series Y to X, amidst the presence of other variables (say Z and W), two time varying dictionaries are built — D that encodes information from all variables (X, Y, Z, W) and D' that encodes information from all variables except Y (X, Z, W only). Suppose the time series blocks being considered at a time t are X_{past}, Y_{past}, Z_{past} and W_{past}, then the dictionary at that time D_{past} is built as follows. Suppose (for example)

$$\begin{pmatrix} X_{past} \\ Y_{past} \\ Z_{past} \\ W_{past} \end{pmatrix}$$

blocks of length 4 time points take values

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix},$$

after each time series block (such as X_{past}) is binned using 2 bins. Then encoding in D_{past} is done based on assigning a particular value to each column. As each row in the first column can take 2 values, there exists a total of 16 possible combinations that the 4 rows can take together in a column. We encode information in 4 rows to a single row by assigning combinations of different values in the 4 rows an encoding from '0' to '15'. In the dictionary D_{past}, the above sequences are encoded as a single sequence —

$$\begin{pmatrix} 6 & 3 & 15 & 3 \end{pmatrix}.$$

The second dictionary D'_{past} at the same time constructed using all variables except Y similarly encodes blocks

$$\begin{pmatrix} X_{past} \\ Z_{past} \\ W_{past} \end{pmatrix}$$

taking values

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix}$$

as

$$\begin{pmatrix} 2 & 3 & 7 & 3 \end{pmatrix}$$

assigning each column one particular state out of 8 possible states. Thus, for the above example, $D = (6, 3, 15, 3)$ and $D' = (2, 3, 7, 3)$. ETC can now be applied on the two dictionaries D and D' as these sequences are now just 1-dimensional symbolic sequences.

3 Parameter selection for CCC: Criteria and Rationale

In Table 1, we summarize the criteria and rationale for choosing the four parameters (w, δ, B, L) of the proposed measure CCC. We have described the measure of Compression-Complexity Causality in the main paper with the idea of intervention. Appropriate parameter selection criteria is done with the view to find out the correct intervention point for a time series to check its causal influence on another given time series . Put more specifically, the main task is choosing the correct value for the length of the time series block Y_{past} and accordingly for X_{past}.

The parameter w which is the length of the moving window ΔX is fixed to 15 for all the datasets used in this work. It is chosen such that it contains sufficient number of data points over which CC rate can be reliably estimated. Earlier studies have revealed that ETC is able to reliably capture complexity of even very short time series (as small as length of 10 samples) [2]. δ, the step size by which the ΔX as well as X_{past} window is moved, is chosen based on the criteria of sufficient overlap $(20-50\%)$ between successive X_{past} windows of length L. B, the number of bins used to generate the symbolic sequence of the input time series is chosen such that it is sufficient to capture the underlying dynamics. It was found that for the AR processes, $B \geq 2$ is sufficient whereas the time series from the chaotic tent map requires at least $B = 8$.

Once w, δ, B are chosen, we choose L, the window length of X_{past}. For this, we analyze the curves of ETC measure as they vary with L, for different time series blocks as appropriate for a given dataset. A detailed description of selection criteria for L is discussed below.

Table 1: Criteria and rationale for choosing the parameters (w, δ, B, L) for CCC. Values of each parameter chosen for Autoregressive (AR), Tent Map (TM), Squid Giant Axon System (SA) and Predator Prey Ecosystem (PP) are enlisted in the rightmost column. Please refer to the main paper for details of these four systems.

Parameter	Description	Criteria	Rationale	Values Chosen
w	Window length ΔX	Minimal data length over which CC rate can be reliably estimated.	Earlier studies have revealed that ETC is able to reliably capture complexity of even very short time series [2].	AR: 15 TM: 15 SA: 15 PP: 15
δ	Step-size	An overlap of $20 - 50\%$ between successive time series windows (X_{past} of length L) over which CC is estimated.	To capture the continuity of time series dynamics.	AR: 80 TM: 80 SA: 50 PP: 4*
B	Number of bins	Smallest number of symbols that capture the time series dynamics.	CCC requires symbolic sequences that represent the underlying dynamics.	AR: 2 TM: 8 SA: 2 PP: 8
L	Window length of immediate past to ΔX (X_{past}) and (Y_{past})	After choosing w, δ, B as above, to check causal influence from Y_{past} to ΔX, we plot $ETC(X_{past} + \Delta X)$ and $ETC(Y_{past} + \Delta X)$ vs. L. ***First criteria***: Choose a value of L at which the two curves are well separated. If the above criteria fails (there is an overlap in the ETC curves for all L), we plot $ETC(X_{past}, Y_{past})$ and $ETC(X_{past} + \Delta X, Y_{past} + \Delta X)$ vs. L. ***Second criteria***: Choose a value of L such that the two curves are well separated.	Well separation of the complexity values of time series blocks $(X_{past} + \Delta X)$ and $(Y_{past} + \Delta X)$ is taken to give maximum possible opportunity to Y_{past} to influence ΔX as against X_{past}. This L is hence the best intervention point. If no such value of L can be found, the maximum separation of curves (X_{past}, Y_{past}) and $(X_{past} + \Delta X, Y_{past} + \Delta X)$, gives the maximum opportunity to (X_{past}, Y_{past}) jointly to affect ΔX.	AR: 150 TM: 100 SA: 75 PP: 40

*This was an exception with 90% overlap as very short data length was available.

3.1 Selection Criteria for L

As discussed in Table 1, for given time series X and Y, we first plot $ETC(X_{past} + \Delta X)$ and $ETC(Y_{past} + \Delta X)$ vs. L when causality is to be checked from Y_{past} to ΔX. We choose a value of L at which the two curves are well separated. In this work, we start with an $L = 20 (> w)$ and go up to $L = 300$ (in case of the predator prey ecosystem data, only 62 data points were available and thus we go up to $L = 40$). In Figs. 1, 2, 3 and 4 which show these curves plotted for linearly and non-linearly coupled tent maps, predator prey and squid giant axon systems respectively, there exists some range of values of L for which the two curves are well separated. A value of L can thus be chosen from within this range. The choice of L for these curves is based on averaged ETC values for referred blocks over the entire time series. However, the choice of L may vary with time if we expect to have causality at different temporal scales with varying time. Moreover, for all the cases taken we have chosen the same values of L for checking causality from Y_{past} to ΔX and for checking causality from X_{past} to ΔY. These values can however be different depending on

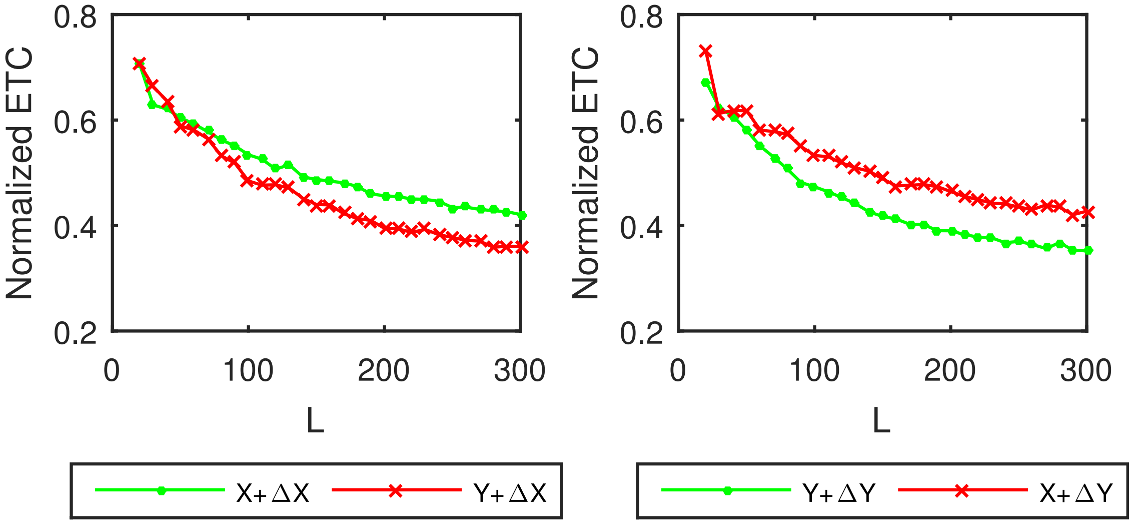

Figure 1: (color online). Averaged $ETC(X_{past} + \Delta X)$, $ETC(Y_{past} + \Delta X)$ curves in the left subfigure and $ETC(Y_{past} + \Delta Y)$, $ETC(X_{past} + \Delta Y)$ curves in the right subfigure for linearly coupled tent maps ($\epsilon = 0.2$) with Y causing X (simulated as per Eq. 17, 18 of the main manuscript). $w = 15, \delta = 100, B = 8$ and L is incremented by a value of 5 data points each time. Using the first criteria for selection of L, $L = 100$ to 300.

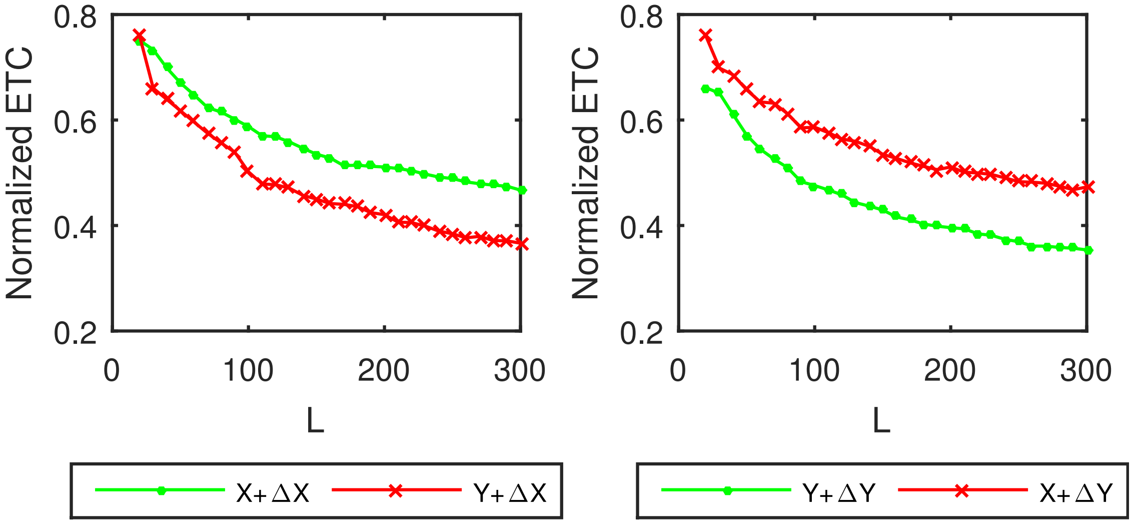

Figure 2: (color online). Averaged $ETC(X_{past}+\Delta X)$, $ETC(Y_{past}+\Delta X)$ curves in the left subfigure and $ETC(Y_{past} + \Delta Y)$, $ETC(X_{past} + \Delta Y)$ curves in the right subfigure for non linearly coupled tent maps ($\epsilon = 0.2$) with Y causing X (simulated as per Eq. 17, 19 of the main manuscript). $w = 15, \delta = 100, B = 8$ and L is incremented by a value of 5 data points each time. Using the first criteria for selection of L, $L = 75$ to 300.

the curves of $ETC(X_{past} + \Delta X)$, $ETC(Y_{past} + \Delta X)$ and $ETC(Y_{past} + \Delta Y)$, $ETC(X_{past} + \Delta Y)$ respectively.

The separation between the curves $ETC(X_{past} + \Delta X)$ and $ETC(Y_{past} + \Delta X)$ is taken to give Y_{past} the maximum opportunity to cause ΔX. The complexities of these time series blocks will be very different at the scale at which there is an influence from past block of Y to the present block of X. Thus the choice of L is about adaptive determination of the temporal scale at which causality exists from Y to X.

If the above criteria fails (there is an overlap in the curves), it means that at no temporal scale can Y intervene to make visible its dynamical influence on ΔX (by change of complexity) as against the dynamical influence due to past of X. We then plot $ETC(X_{past}, Y_{past})$ and $ETC(X_{past} + \Delta X, Y_{past}+\Delta X)$ vs. L. We choose a value of L such that the two curves are well separated. In case

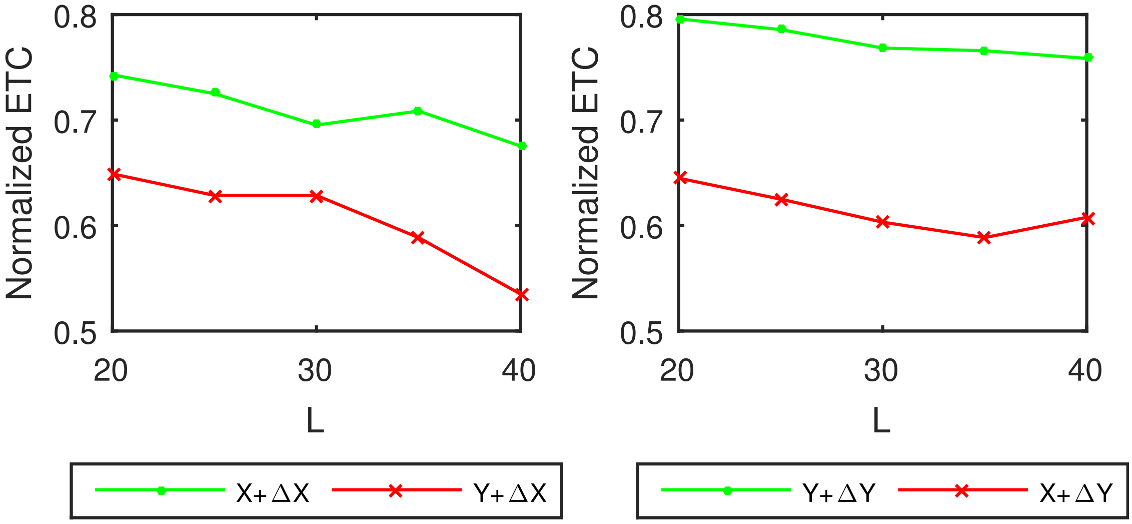

Figure 3: (color online). Averaged $ETC(X_{past} + \Delta X)$, $ETC(Y_{past} + \Delta X)$ curves in the left subfigure and $ETC(Y_{past} + \Delta Y)$, $ETC(X_{past} + \Delta Y)$ curves in the right subfigure for predator prey ecosystem with Y representing Didinium (predator) population and X representing Paramecium (prey) population. $w = 15, \delta = 1, B = 8$ and L is incremented by a value of 5 data points each time. Using the first criteria for selection of L, $L = 20$ to 40.

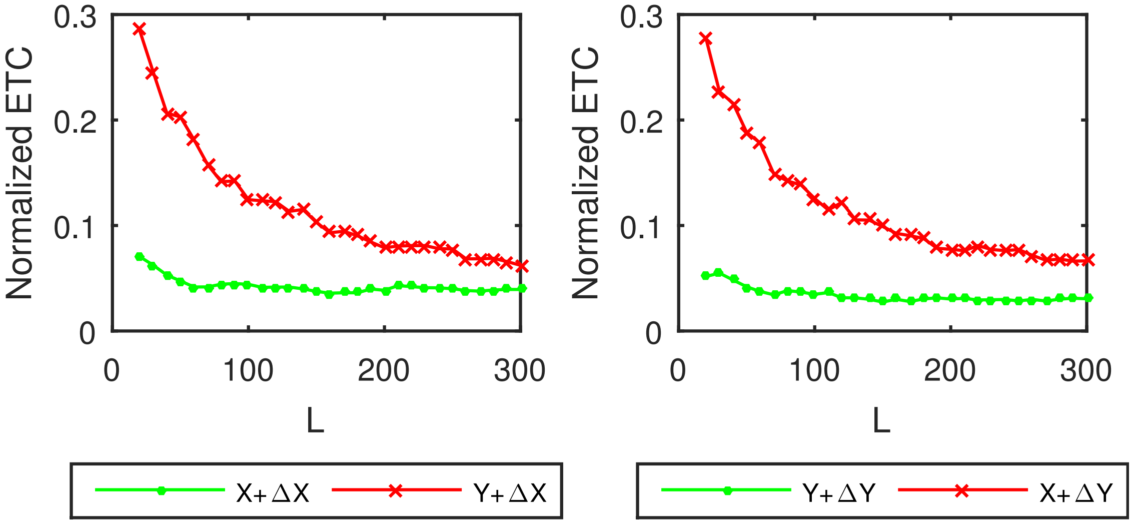

Figure 4: (color online). Averaged $ETC(X_{past}+\Delta X)$, $ETC(Y_{past}+\Delta X)$ curves in the left subfigure and $ETC(Y_{past}+\Delta Y)$, $ETC(X_{past}+\Delta Y)$ curves in the right subfigure for squid giant axon system ('a5t01') with Y representing the applied stimulus current and X representing observed voltage. $w = 15, \delta = 100, B = 2$ and L is incremented by a value of 5 data points each time. Using the first criteria for selection of L, $L = 75$ to 300. Lower values of L are not used despite sufficient separation so as to avoid making computation based on the transient stage values.

of AR processes where the first criteria is not met due to the overlap between $ETC(X_{past} + \Delta X)$ and $ETC(Y_{past} + \Delta X)$, the second pair of curves is plotted as shown in Figs. 5. The rationale behind this criteria is to see at which intervention point L do X_{past}, Y_{past} jointly begin to have an influence on the dynamical evolution of ΔX.

If the two time series are independent or are constant in time and identical, both the above criteria are bound to fail. This implies that there exists no temporal scale at which there is an influence from one of these time series to the other. For the case of two independent and uniformly distributed real time series the curves for both criteria are shown in Figs. 7 and 8. There exists no value of L at which there is a causality from Y to X or vice versa.

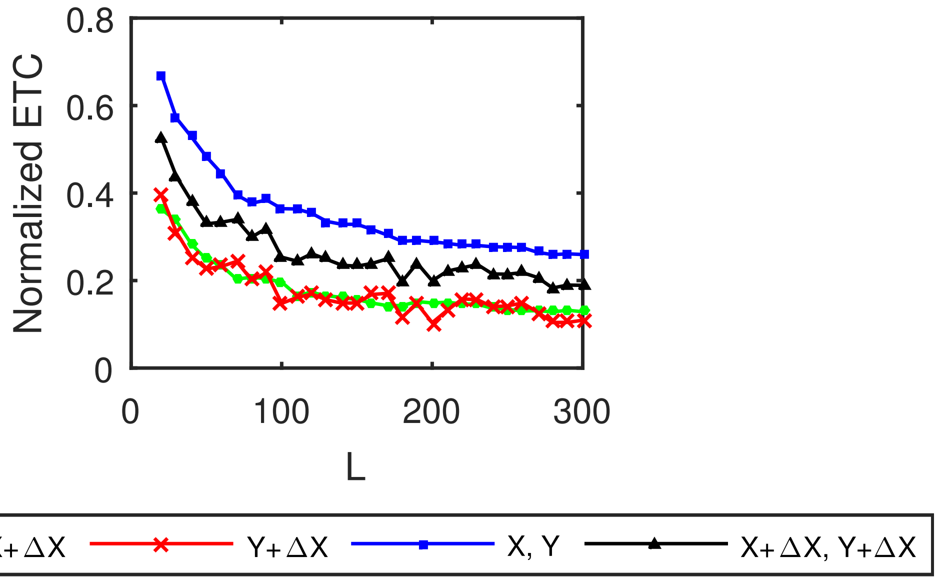

Figure 5: (color online). Averaged $ETC(X_{past} + \Delta X)$, $ETC(Y_{past} + \Delta X)$, $ETC(X_{past}, Y_{past})$, $ETC(X_{past} + \Delta X, Y_{past} + \Delta X)$ curves for coupled AR processes with Y causing X (simulated as per Eq. 15 with all settings as in Section 5.1.1 of the main manuscript with $\epsilon = 0.8$). $w = 15, \delta = 100, B = 2$ and L is incremented by a value of 5 data points each time.Using the second criteria for selection of L, $L = 100$ to 300.

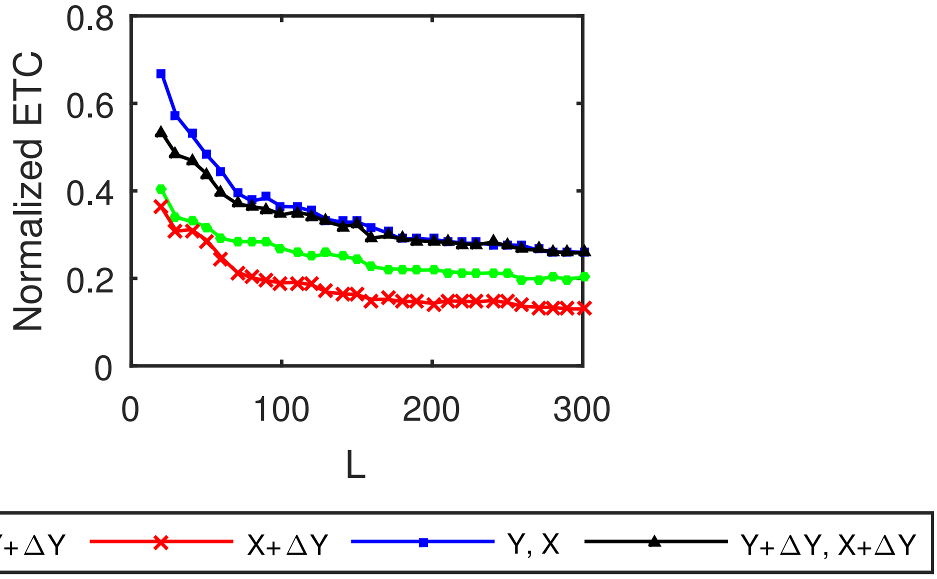

Figure 6: (color online). Averaged $ETC(Y_{past} + \Delta Y)$, $ETC(X_{past} + \Delta Y)$, $ETC(Y_{past}, X_{past})$, $ETC(Y_{past} + \Delta Y, X_{past} + \Delta Y)$ curves for coupled AR processes with Y causing X (simulated as per Eq. 15 with all settings as in Section 5.1.1 of the main manuscript with $\epsilon = 0.8$). $w = 15, \delta = 100, B = 2$ and L is incremented by a value of 5 data points each time. Using the first criteria for selection of L, $L = 100$ to 300.

4 Description of CCC Toolbox

The accompanying CCC toolbox, implemented in MATLAB contains the following files:

1. **demo_2processes.m** calls functions to simulate a system of two coupled AR processes or tent maps to estimate the value of Compression-Complexity Causality between them.

2. **demo_3processes.m** calls functions to simulate a system with three AR processes with cou-

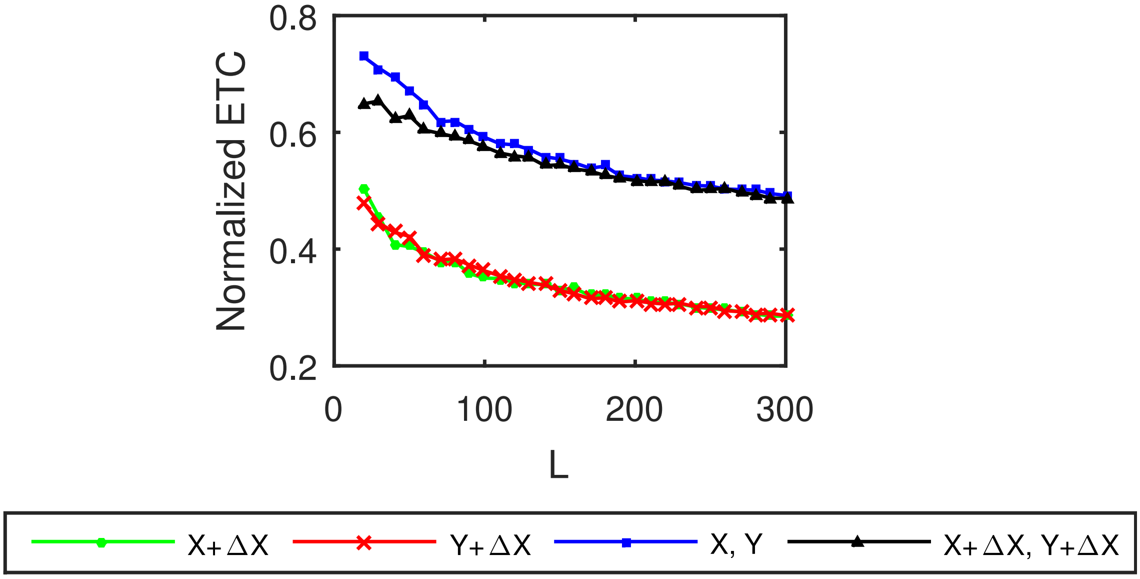

Figure 7: (color online). Averaged $ETC(X_{past} + \Delta X)$, $ETC(Y_{past} + \Delta X)$, $ETC(X_{past}, Y_{past})$, $ETC(X_{past} + \Delta X, Y_{past} + \Delta X)$ curves for independent processes Y and X. $w = 15, \delta = 100, B = 2$ and L is incremented by a value of 5 data points each time.

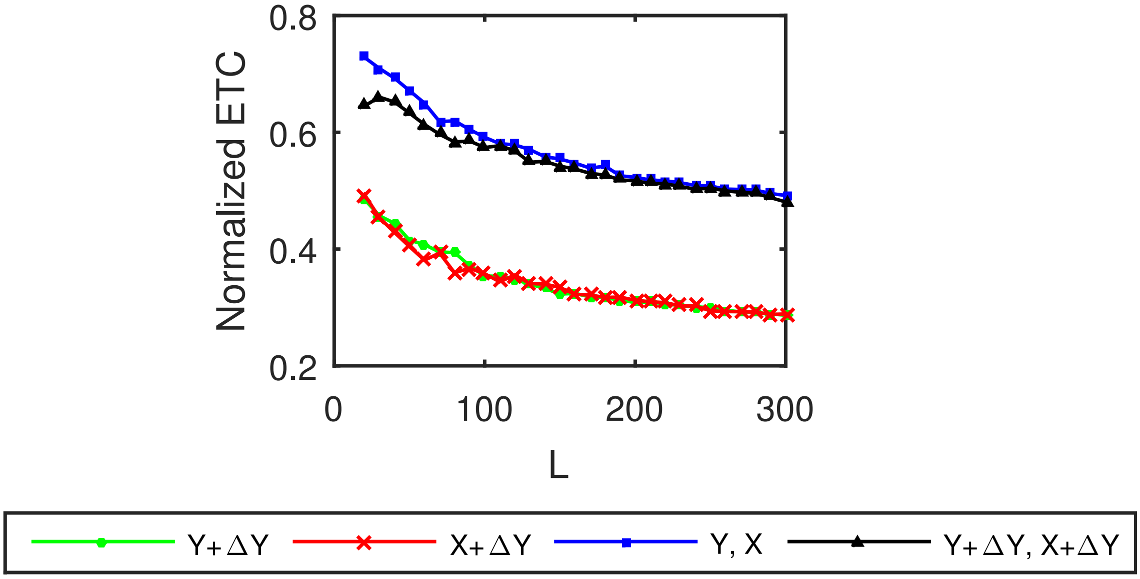

Figure 8: (color online). Averaged $ETC(Y_{past} + \Delta Y)$, $ETC(X_{past} + \Delta Y)$, $ETC(Y_{past}, X_{past})$, $ETC(Y_{past} + \Delta Y, X_{past} + \Delta Y)$ curves for independent processes Y and X. $w = 15, \delta = 100, B = 2$ and L is incremented by a value of 5 data points each time. Using the second criteria for selection of L, based on this figure and Fig. 7, $L = 100$ to 300, avoiding the range of L giving transient values of CCC.

pling between them and estimates the value of conditional Compression-Complexity Causality between any two variables chosen.

3. **coupled_AR.m** simulates a system of two unidirectionally coupled AR processes with a desired level of noise or percentage of non-uniform sampling.

4. **puncture.m** introduces non-uniform sampling/non-synchronous measurements in the data.

5. **coupled_tent.m** simulates a system of two unidirectionally non-linearly coupled tent maps.

6. **UpdateTent.m** updates the values of the tent map at every iteration.

7. **coupled_AR_3processes.m** simulates a system of three coupled AR processes.

8. **conditional_CCC.m** estimates conditional Compression-Complexity Causality between any two input variables (time series) from a given multivariate system.

9. **ETC.m** estimates individual/joint ETC values. Dn_to_D1.m subroutine called by the ETC function performs the task of dictionary building.

10. **Partition.m** bins the given time series before estimating ETC values.

References

[1] N. Nagaraj, K. Balasubramanian, and S. Dey, "A new complexity measure for time series analysis and classification," *The European Physical Journal Special Topics*, vol. 222, no. 3-4, pp. 847–860, 2013.

[2] K. Balasubramanian and N. Nagaraj, "Aging and cardiovascular complexity: effect of the length of RR tachograms," *PeerJ*, vol. 4, p. e2755, 2016.