Understanding Different Notions of Causality and Estimating Causality Testing Methods such as Transfer Entropy

Submitted To

Dr. Harikrishnan

Course Name: Study Oriented Project

BITS Pilani

Submitted By

Anurag Tendulkar - 2020A7PS1010G

Date - 02/05/2023

Birla Institute of Technology and Science, KK Birla Goa Campus, India

Table of Contents

Introduction

We humans realize that the world is not made up of just dry data, rather it is linked together by an intricate web of cause-effect relations. It is causal explanations, not dry data that make up most of our knowledge. Determining and measuring the cause-effect relationships is fundamental to most scientific studies of natural phenomenon. The notion of causation is different from correlation which studies the association of trends or patterns in data (Kathpalia and Nagaraj 2019, 15). It is necessary to imagine the consequences of events to understand their causes. To be a causal learner, an entity must possess three distinct levels of cognitive ability: seeing, doing, and imagining (Pearl & Mackenzie, 2018, 15). These levels are the rungs of the Ladder of Causation which are required for a machine to make causal inferences.

The first level - seeing is about making predictions based on past data. Questions can be answered by collecting and analyzing past data. An eg of questions on this rung can be - How likely is a customer who bought toothpaste likely to buy floss? We can answer this question by plotting a best fit line but we cannot conclude the cause variable and the effect variable. Does toothpaste influence floss or vice versa? Despite the advances of machines, the author places them on this rung.
On the second rung of the ladder, the aim is to compare the outcomes arising under different interventions, given two or more (possible) interventions in a system. There is no past data relating to the intervention to draw conclusions from. Eg of a question from this rung can be - What will happen to floss sales when the price of toothpaste doubles? The author says that such interventions can be answered by incorporating a causal model such as recording the market conditions with the sales.
The third rung is counterfactuals is used for explaining the dependencies or the underlying generating process for the causal relationships observed. To answer these questions we must take a step back and think about the different outcomes of the problem. Eg. What is the probability that a customer who bought toothpaste would still have bought it if we had doubled the price. Eg. Was aspirin the reason the headache stopped?

Having a causal model helps us understand the reason behind the problem thus preventing it the next time it occurs. There is some difficulty while dealing with counterfactuals as we cannot exactly tell what will happen in an imaginary world where some facts are excluded.

Storing causal information can be challenging as this will require incorporating information from previous studies in the field or by doing laboratory experiments. Attempts in this regard started with graphical representations of causal path diagrams by Sewall Wright. Currently, non-parametric structural equation models (NPSEMs) which provide a very general data generating mechanism suitable for encoding causation, dominate the field. (Pearl and Mackenzie 2018, 15)

There are different methods to test causality, one of them being Transfer Entropy. It is recognized as a powerful tool to detect the transfer of information between joint processes. It has a solid foundation in information theory and naturally detects directional and dynamical information. Moreover, the formulation of transfer entropy does not assume any underlying process, making it sensitive to all types of dynamical information. The estimation of Transfer Entropy is complicated by a number of practical issues leading to several ways to estimate it. In this project we discuss the approach where we use binning to evaluate the probability distribution function.

We will deal with causality as estimated from collected time-series measurements where it is not possible to intervene on the experimental setup.

Problem Definition

Expressing Transfer Entropy in Terms of Shannon Entropy and Estimating it Using Uniform Binning

Methodology

The average number of bits needed to optimally encode independent draws of the discrete variable - I following a probability distribution p(i) is given by the Shannon entropy, where it extends to all states the process can assume. (Schreiber and Thomas 2000, 15)

$$H_I = -\sum_i p(i) \log_2 p(i)$$

The excess number of bits that will be coded if a different distribution q(i) is used instead of p(i) is given by the Kullback entropy.

$$K_I \ = \ \sum_i p(i) \ \log \ p(i)/q(i).$$

The mutual information of two processes I and J with joint probability pIJ (i, j) can be seen as the excess amount of code produced by erroneously assuming that the two systems are independent, i.e. assuming qIJ (i, j) = pI (i) pJ (j) instead of pIJ (i, j).

$$M_{IJ} = \sum p(i,j) \log \frac{p(i,j)}{p(i)\, p(j)}$$

Note that mutual information is symmetrical, has no directional sense and can be made asymmetrical if we add a delay embedding.

Henceforth we will use the shorthand notation i (k) n = (in, . . . , in−k+1) for words of length k, or k dimensional delay embedding vectors.
The average number of bits needed to encode one additional state of the system if all previous states are known is given by the entropy rate

$$h_I = -\sum p(i_{n+1}, i_n^{(k)}) \log p(i_{n+1}|i_n^{(k)}) \,.$$

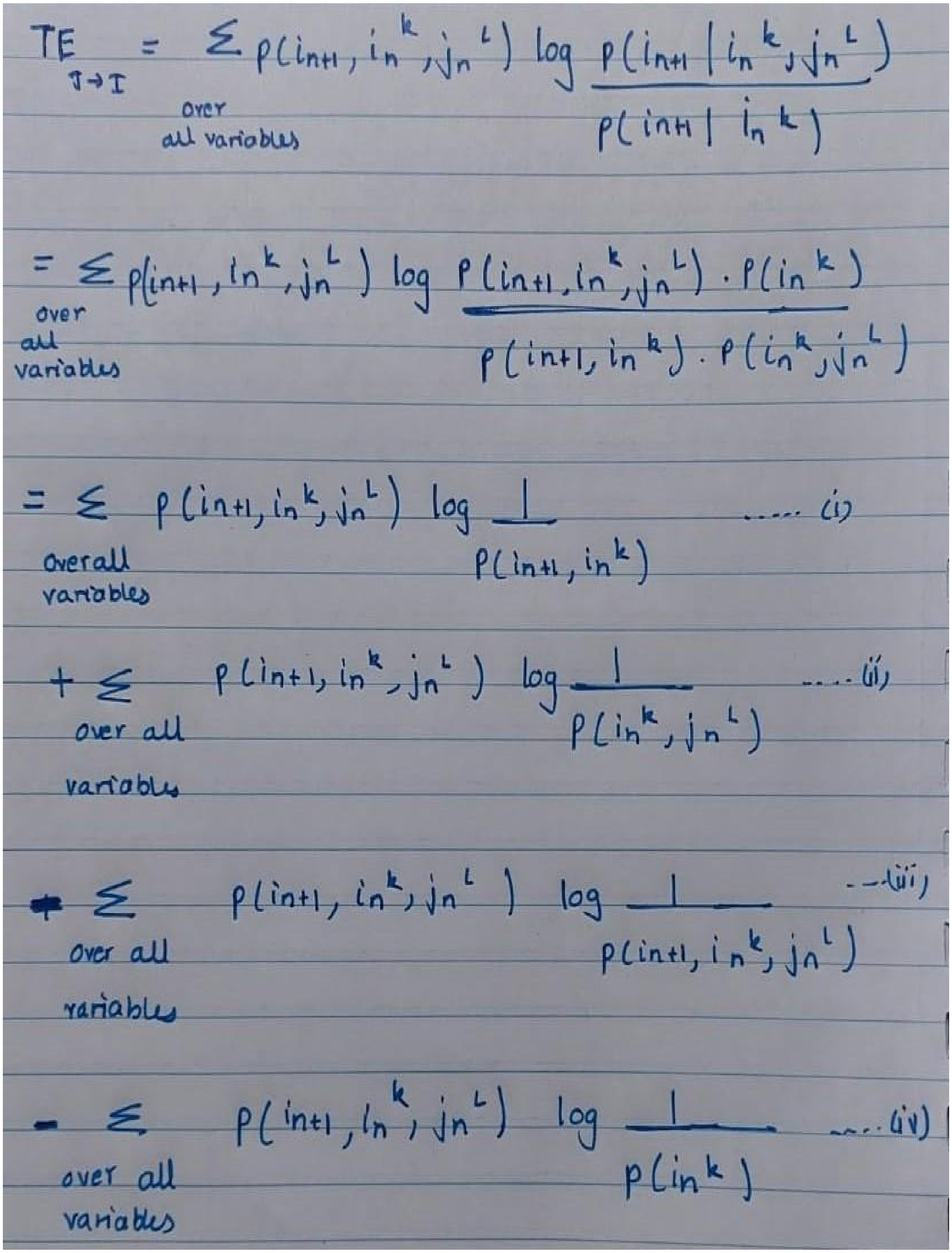
In the absence of information flow from J to I, the state of J has no influence on the transition probabilities on system I. The incorrectness of this assumption can again be quantified by a Kullback entropy by which we define the transfer entropy:

$$T_{J\to I} = \sum p(i_{n+1}, i_n^{(k)}, j_n^{(l)}) \log \frac{p(i_{n+1}|i_n^{(k)}, j_n^{(l)})}{p(i_{n+1}|i_n^{(k)})} \,.$$

TJ->I becomes non-symmetric thus it measures the degree of dependence of I on J.

TE J->I can be expressed in terms of shannon entropy-
TE J->I = H(in+1, in(k)) + H(in(k), jn(l)) - H(in+1, in(k), jn(l)) - H(in(k))

The proof of expressing transfer entropy in terms of Shannon entropy is shown below.



$$TE_{J\to I} = \sum_{\substack{\text{over} \\ \text{all variables}}} p(i_{n+1}, i_n^k, j_n^L) \log \frac{p(i_{n+1} | i_n^k, j_n^L)}{p(i_{n+1} | i_n^k)}$$

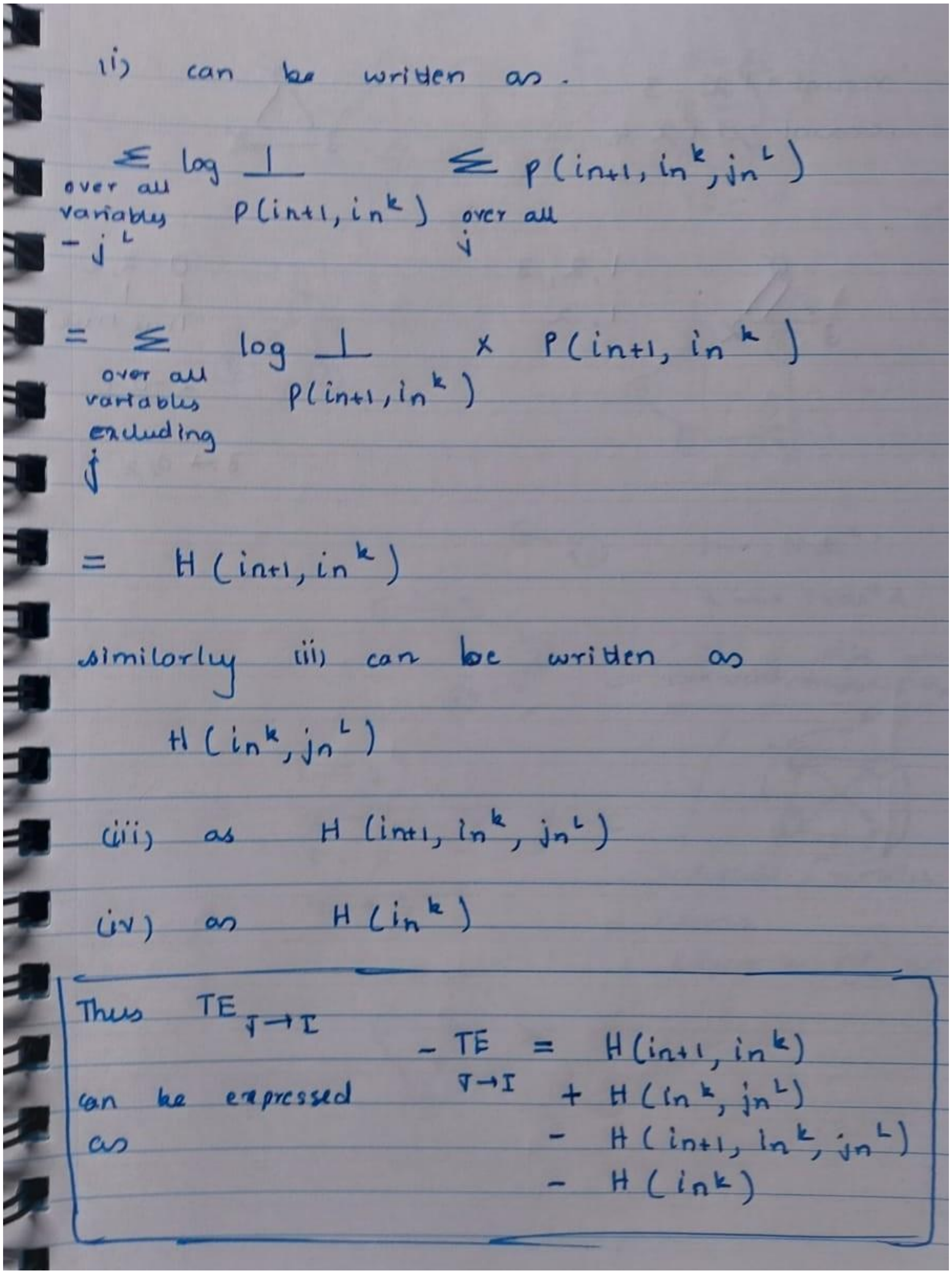
$$= \sum_{\substack{\text{over} \\ \text{all} \\ \text{variables}}} p(i_{n+1}, i_n^k, j_n^L) \log \frac{p(i_{n+1}, i_n^k, j_n^L) \cdot p(i_n^k)}{p(i_{n+1}, i_n^k) \cdot p(i_n^k, j_n^L)}$$

$$= \sum_{\substack{\text{overall} \\ \text{variables}}} p(i_{n+1}, i_n^k, j_n^L) \log \frac{1}{p(i_{n+1}, i_n^k)} \quad \text{.....}(i)$$

$$+ \sum_{\substack{\text{over all} \\ \text{variables}}} p(i_{n+1}, i_n^k, j_n^L) \log \frac{1}{p(i_n^k, j_n^L)} \quad \text{.....}(ii)$$

$$- \sum_{\substack{\text{over all} \\ \text{variables}}} p(i_{n+1}, i_n^k, j_n^L) \log \frac{1}{p(i_{n+1}, i_n^k, j_n^L)} \quad \text{---}(iii)$$

$$- \sum_{\substack{\text{over all} \\ \text{variables}}} p(i_{n+1}, i_n^k, j_n^L) \log \frac{1}{p(i_n^k)} \quad \text{.....}(iv)$$



(i) can be written as.

$$\sum_{\substack{\text{over all}\\ \text{variables}\\ -j^L}} \log \frac{1}{p(i_{n+1}, i_n^k)} \quad \sum_{\substack{\text{over all}\\ j}} p(i_{n+1}, i_n^k, j_n^L)$$

$$= \sum_{\substack{\text{over all}\\ \text{variables}\\ \text{excluding}\\ j}} \log \frac{1}{p(i_{n+1}, i_n^k)} \times p(i_{n+1}, i_n^k)$$

$$= H(i_{n+1}, i_n^k)$$

similarly (ii) can be written as

$$H(i_n^k, j_n^L)$$

(iii) as $H(i_{n+1}, i_n^k, j_n^L)$

(iv) as $H(i_n^k)$

Thus $TE_{J \to I}$ can be expressed as

$$TE_{J \to I} = H(i_{n+1}, i_n^k) + H(i_n^k, j_n^L) - H(i_{n+1}, i_n^k, j_n^L) - H(i_n^k)$$

The problem reduces to estimating the four Shannon entropies and the corresponding joint probability distribution.

Binning

Binning is a technique used to estimate the probability density function of given random variables. This approach consists of an uniform quantization of the time series followed by estimation of the entropy approximating probabilities with the frequency of visitation of the quantized states. A time series y, realization of the generic process Y, is first normalized to have zero mean and unit variance, and then coarse grained spreading its dynamics over E quantization levels of amplitude r=(ymax - ymin)/E, where ymax and ymin represent minimum and maximum values of the normalized series. (E - Epsilon) (Montalto, Faes, and Marinazzo 2014, 15)

Quantization assigns to each sample the number of the level to which it belongs, so that the quantized time series yE takes values within the alphabet A = (0,1, ... , E-1). Uniform quantization of embedding vectors of dimension d results in an uniform partition of the d-dimensional state space into E^d disjoint hypercubes of size r, such that all vectors V falling within the same hypercube are associated with the same quantized vector Vj, and are thus indistinguishable within the tolerance r. The entropy is then estimated as:

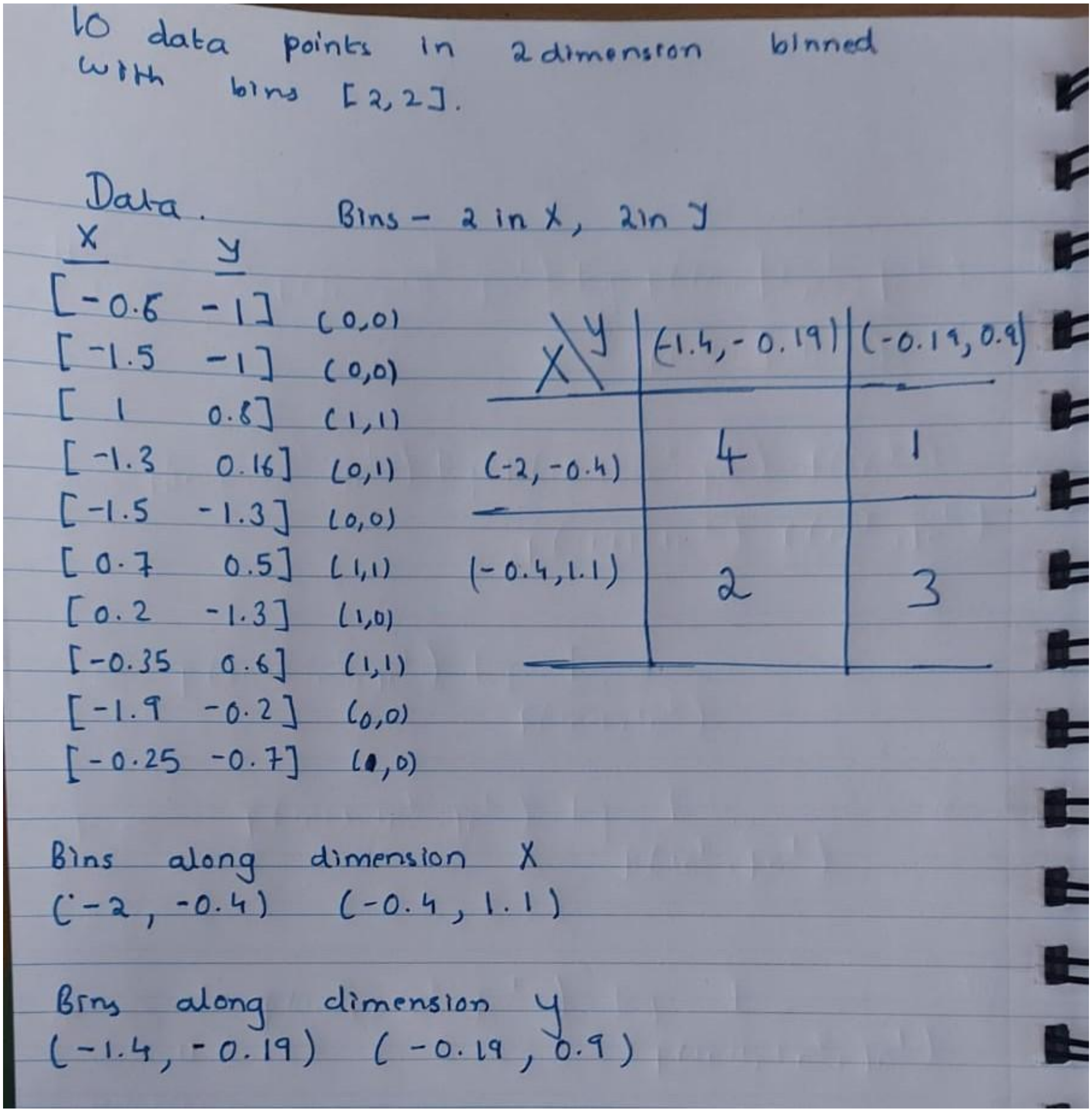
$$H(V_{\xi}) = -\sum_{V_{\xi} \in A^{d}} p(V_{\xi}) log\, p(V_{\xi})$$

where the sum is extended over all vectors found in the available realization of the quantized series, and the probabilities p(VE) are estimated for each hypercube simply as the fraction of quantized vectors VE falling into the hypercube (E - Epsilon).

The number of Bins(E) is a Hyperparameter and was calculated using Strudges formulae - n = log2(series length).

Function used - np.histogramdd()

The example shown below bins 10 data points in 2d using the function np.histogramdd() with 2 bins in X and 2 bins in y. Dividing by the total number of bins gives the probability of each vector falling in the bin.



10 data points in 2 dimension binned with bins [2,2].

Data.

Bins - 2 in X, 2 in Y

X	Y	
[-0.6	-1]	(0,0)
[-1.5	-1]	(0,0)
[1	0.8]	(1,1)
[-1.3	0.16]	(0,1)
[-1.5	-1.3]	(0,0)
[0.7	0.5]	(1,1)
[0.2	-1.3]	(1,0)
[-0.35	0.6]	(1,1)
[-1.9	-0.2]	(0,0)
[-0.25	-0.7]	(0,0)

X \ Y	(-1.4, - 0.19)	(-0.19, 0.9)
(-2, -0.4)	4	1
(-0.4, 1.1)	2	3

Bins along dimension X
(-2, -0.4) (-0.4, 1.1)

Bins along dimension y
(-1.4, -0.19) (-0.19, 0.9)

Data Generation and Description

Data was generated from a coupled autoregressive time series

Xt = 0.5 * Xt-1 + et
Yt = 0.5 * Yt-1 + c * Xt-1 + et

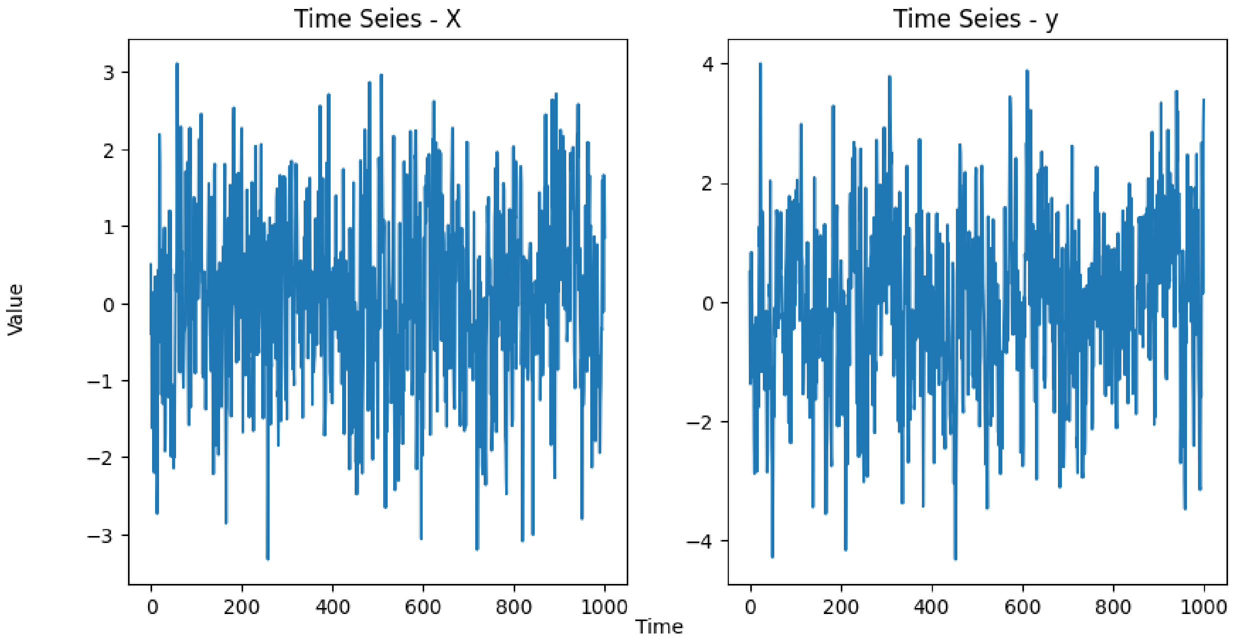
Where et - is normal noise and c is the coupling coefficient.

X is the driving equation and y depends on X. The order of the autoregressive time series is 1. This is done to simplify the calculations while estimating the entropy. The time step used for generating the data is 1 unit. The initial conditions used to generate the data were :

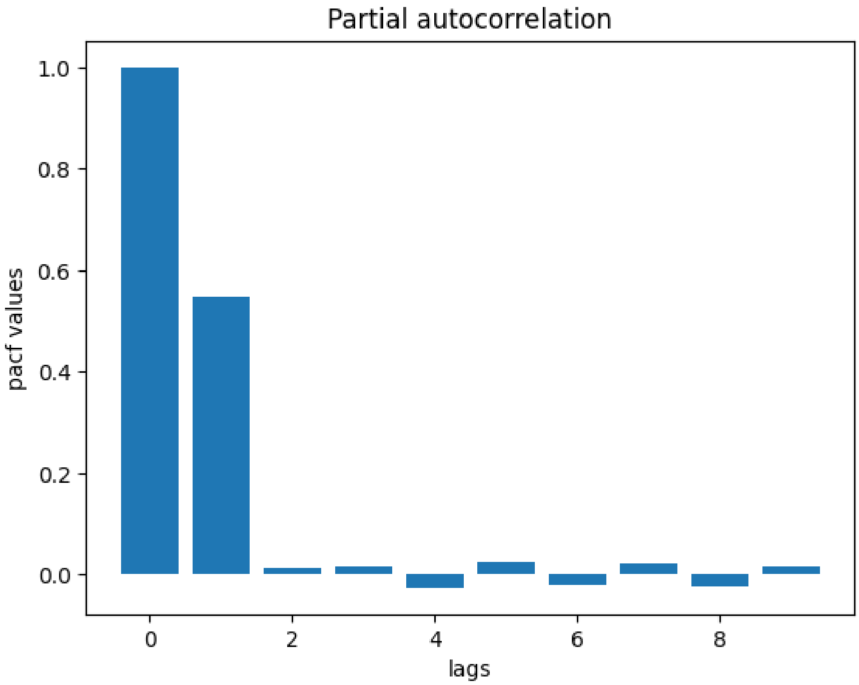
X[0] = 0.5
Y[0] = 0.5

10,000 data instances were generated. Large number of data instances were generated as it reduces the chance of empty bins during probability estimations.

Data when c = 0.5



Partial Autocorrelation to get lags of y - (pacf)



The pacf of lag1 is significant as compared to others. Thus we use l = 1 in the TE equation

Experiments

As per the above equations, the coupling coefficient c was varied from 0 to 0.9 with a step size of 0.15. The transfer entropy was estimated for each of the coupling values. Parameters used -

1. Number of Bins - log2(length of series) by Strudges Rule
2. Length of series - 10,000

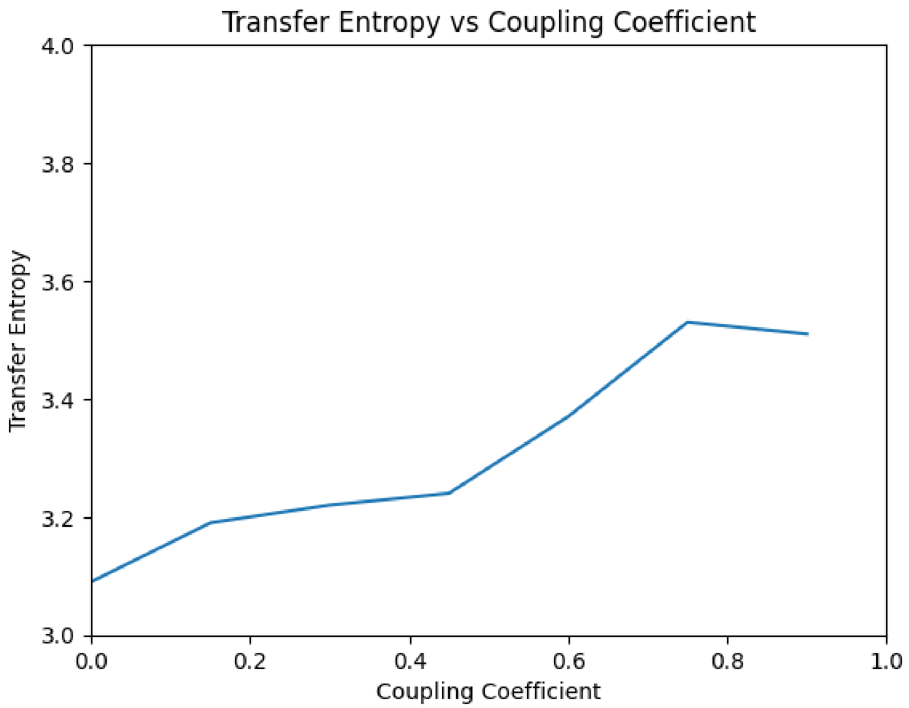
The estimated values of Transfer Entropy were plotted against the values of coupling coefficient.

Next to see the sensitivity of the estimated Transfer Entropy with respect to the bin count, the number of bins were varied as per follows

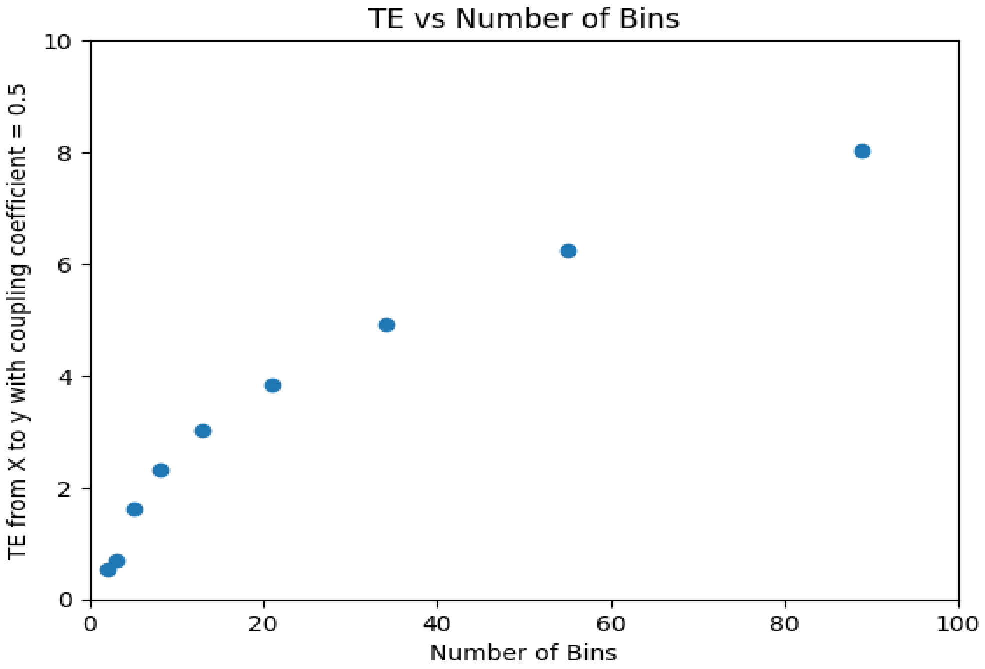
bin_count = [2, 3, 5, 8, 13, 21, 34, 55, 89]
The coupling coefficient was fixed to 0.5 f and the number of datapoints generated was 10000 for all the bin values

To see the effect of the number of datapoints used on the estimation of the joint probability and then the Transfer Entropy, the number of datapoints generated was varied from 10 to 10^7.

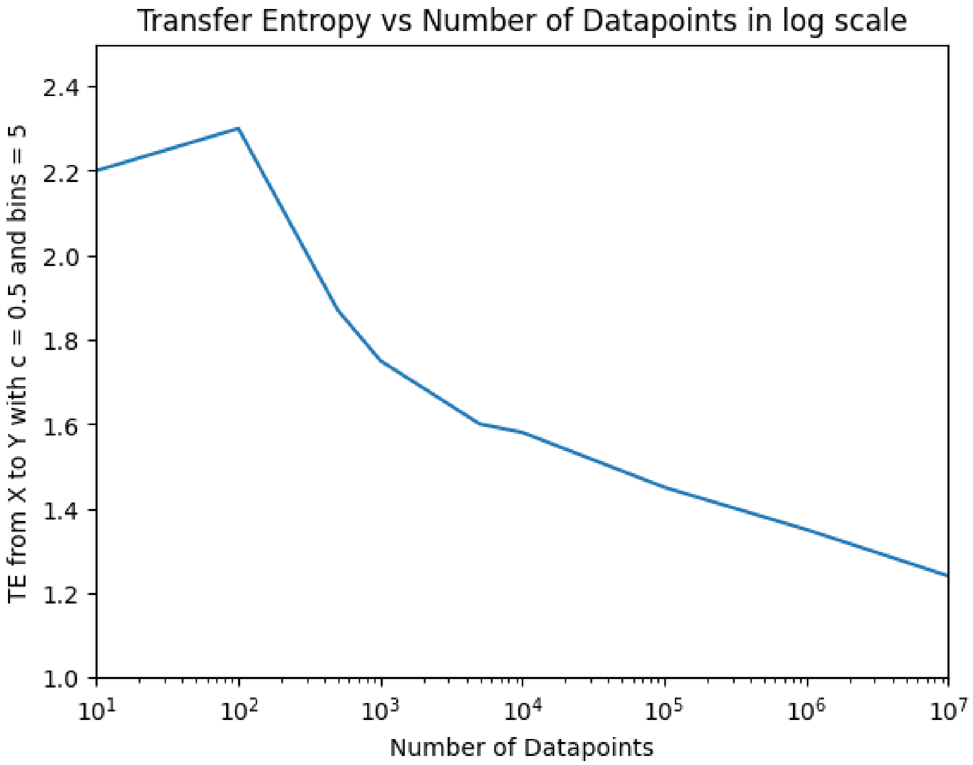
Results



The values of Transfer Entropy ranged from 3.09 to 3.51 as the coupling coefficient values were increased



The plot shows that TE is extremely sensitive to the number of bins. Thus the number of bins is an important parameter that has to be tuned while estimating TE by uniform binning.



Corresponding TE values obtained from the above plot were
[2.2, 2.3, 1.87, 1.75, 1.6, 1.58, 1.45, 1.35, 1.24]

Discussion

Transfer Entropy is a difficult metric to estimate as there can be errors while estimating the joint probability distribution of the variables involved. Uniform Binning is one of the simplest methods to estimate the probability distribution but has several drawbacks. There may be bins which are empty which affect the calculation of entropy. To correctly estimate the probability distribution the bin sizes should be different and should be calculated optimally. The TE calculated above for the coupling coefficient 0 should ideally be approximately 0 and should increase as the coupling coefficient increases. As per the calculations, the TE increased when the coupling coefficients were increased. This is a good sign as we know that X causes y and as the coupling coefficient increases, the effect of X on y should increase.

The number of bins and the amount of data generated are important parameters that need to be tuned while estimating the transfer entropy. As per the 2nd experiment when the number of bins was increased it resulted in a significant increase in the TE estimation. This is because there is more information as bins increase and there is more randomness in the pdf. Thus the entropy of the data becomes larger and TE values increase. The number of datapoints generated increases the probability distribution values as more data is available. The number of vectors in each bin increases. There is less information as there are less bins. Thus the entropy decreases and TE values drop.

There are more estimators for the probability distribution such as Linear estimator(LIN) and Nearest Neighbour Estimator(NN). (Montalto, Faes, and Marinazzo 2014, #) These estimators are out of the scope of this project but further work can be done in these areas.

Conclusion

Transfer Entropy is a rung one estimator for causality as it uses past data to make inferences and there is no intervention done during the estimation. The estimation of Transfer Entropy is a careful process and the parameters must be tuned to obtain meaningful results. Uniform Binning is a basic method to estimate the probability distribution and further work can be done in using better estimators for the joint distribution.

Acknowledgements

I would like to express my gratitude to Dr. Harikrishnan N B for his support and assistance throughout the research project. Dr. Harikrishnan N B took time out of his busy schedule to hold weekly meetings to track the progress of the project and made sure we got the required resources and assistance. Through this project I learned the methodology and approach behind conducting research. I would like to thank my peers Yash Yelmame, Siddhant Satyanaik, Baibhav Padhy, Neel Khare, Snehashish for their presentations and their work. This project motivates me to learn more about probability estimation techniques, Transfer Entropy and has provided me with a wonderful experience in research.

References

1. Kathpalia, Aditi, and Nithin Nagaraj. 2019. “Measuring Causality.” arXiv:1910.08750v1 [stat.ME].
2. Montalto, Allesandro, Luca Faes, and Daniele Marinazzo. 2014. “MuTE: A MATLAB Toolbox to Compare Established and Novel Estimators of the Multivariate Transfer Entropy.” *PLOS One*, (October). 10.1371/journal.pone.0109462.
3. Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why*. N.p.: Basic Books. http://bayes.cs.ucla.edu/WHY/.
4. Schreiber, and Thomas. 2000. “Measuring Information Transfer.” *Phys. Rev. Lett.* 85, no. 2 (Jul): 461--464. 10.1103/PhysRevLett.85.461.

Links to code -

1. For Estimation of Transfer Entropy:

https://colab.research.google.com/drive/16UwriMcZRyQ0p0ixt0GgKFcnA61cZEc5?usp=sharing

2. For Autoregressive Time Series:

https://github.com/Anurag-Tendulkar/AutoRegressive-Models/blob/master/AR.ipynb