Problems with first-order infotheoretic measures on short sequences

Nithin Nagaraj[1,*]

[1]Consciousness Studies Programme, National Institute of Advanced Studies (NIAS), Bengaluru, India
* Corresponding author: nithin@nias.res.in

NIAS

Abstract: Shannon entropy (H) and Mutual Information (MI) are practically estimated using first-order statistics as it is easier and convenient. However, the estimated first-order H and MI values could be grossly incorrect as we shall demonstrate with 3 carefully designed short linearly independent sequences X1, X2 and X3. These are carefully constructed to take values from the set {0, +1, -1} such that the time stamps of zeroes exactly coincide. Being linearly independent, the pairwise correlation coefficients are zeros. X2 and X3 are cyclic permutations of each other and X1 is independent of both X2 and X3. The estimated pair-wise first-order MI values turn out to be the same for all the pairs, thus showing its inability to capture the additional mutual dependence between X2 and X3. After puncturing the zero-values from all the sequences (along with the time stamps) the first-order MIs turn out to be zeros for all pairs - wrongly implying that the sequences are independent, whereas X2 and X3 continue to be cyclic permutations of each other (and hence completely dependent). Compression-complexity measures such as the Effort-To-Compress complexity measure can correctly capture the non-linear dependencies in this case. First-order estimation of H and MI is thus fraught with danger in practical applications, especially on short data lengths and in such situations, it is preferable to employ compression-complexity measures.

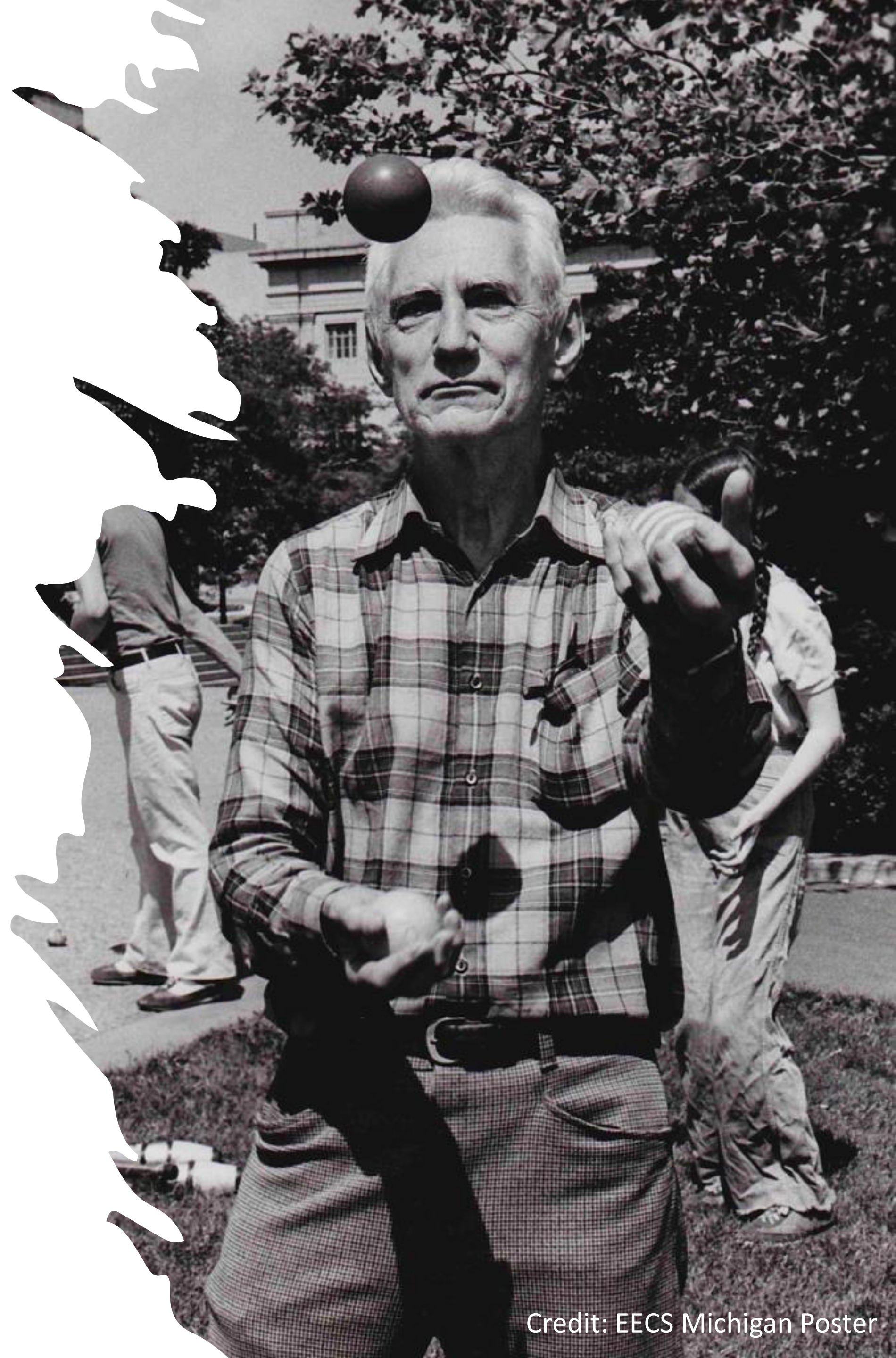
Keywords: Shannon entropy; Mutual Information; first order; compression-complexity; Effort-To-Compress; short sequences

IECI 2021

Claude E Shannon

(1916-2001)

The Father of
Information Theory



Credit: EECS Michigan Poster

IECI 2021

Shannon's 1948 Masterpiece*

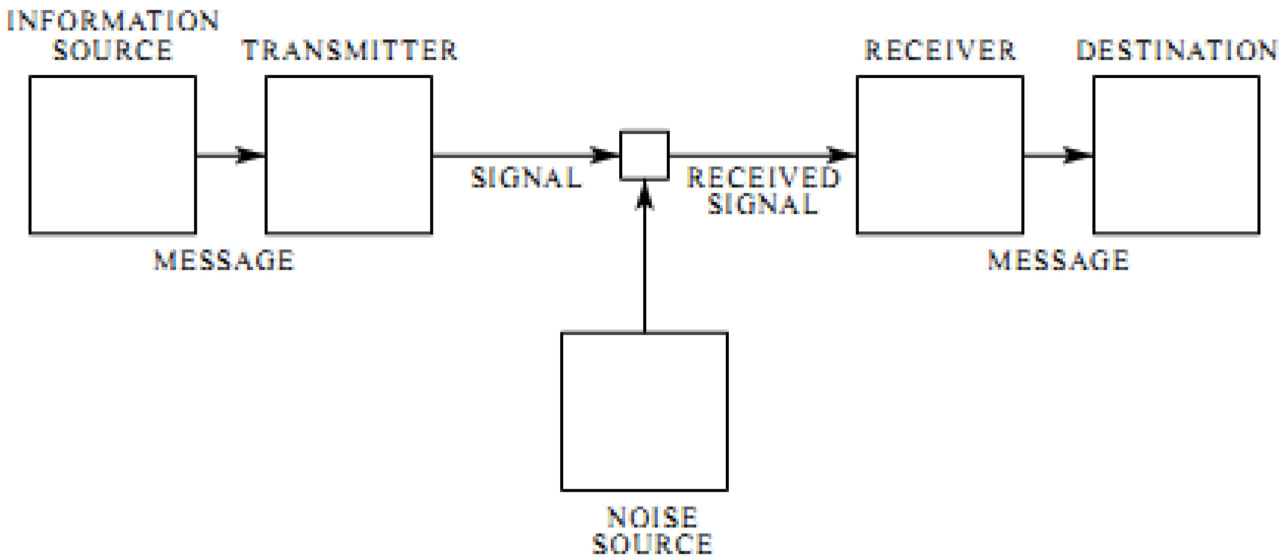


Fig. 1—Schematic diagram of a general communication system.

Entropy

$$\boxed{H(X) = -\sum_{i=1}^{M} p_i \log_2(p_i) \quad \text{bits/symbol}}$$

Channel capacity

$$\boxed{C = \sup_{p_X(x)} I(X;Y)}$$

*Claude E Shannon, ***A Mathematical Theory of Communication***, Bell Sys. Tech. Journal, Vol. 27, pp. 379-423, 1948.

Entropy of a stochastic process*

- For a sequence of n discrete random variables, entropy (if the limit exists):

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

- If these are independent and identically distributed (i.i.d), then:

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{H(X_1, X_2, \ldots, X_n)}{n} = \lim_{n \to \infty} \frac{nH(X_1)}{n} = H(X_1)$$

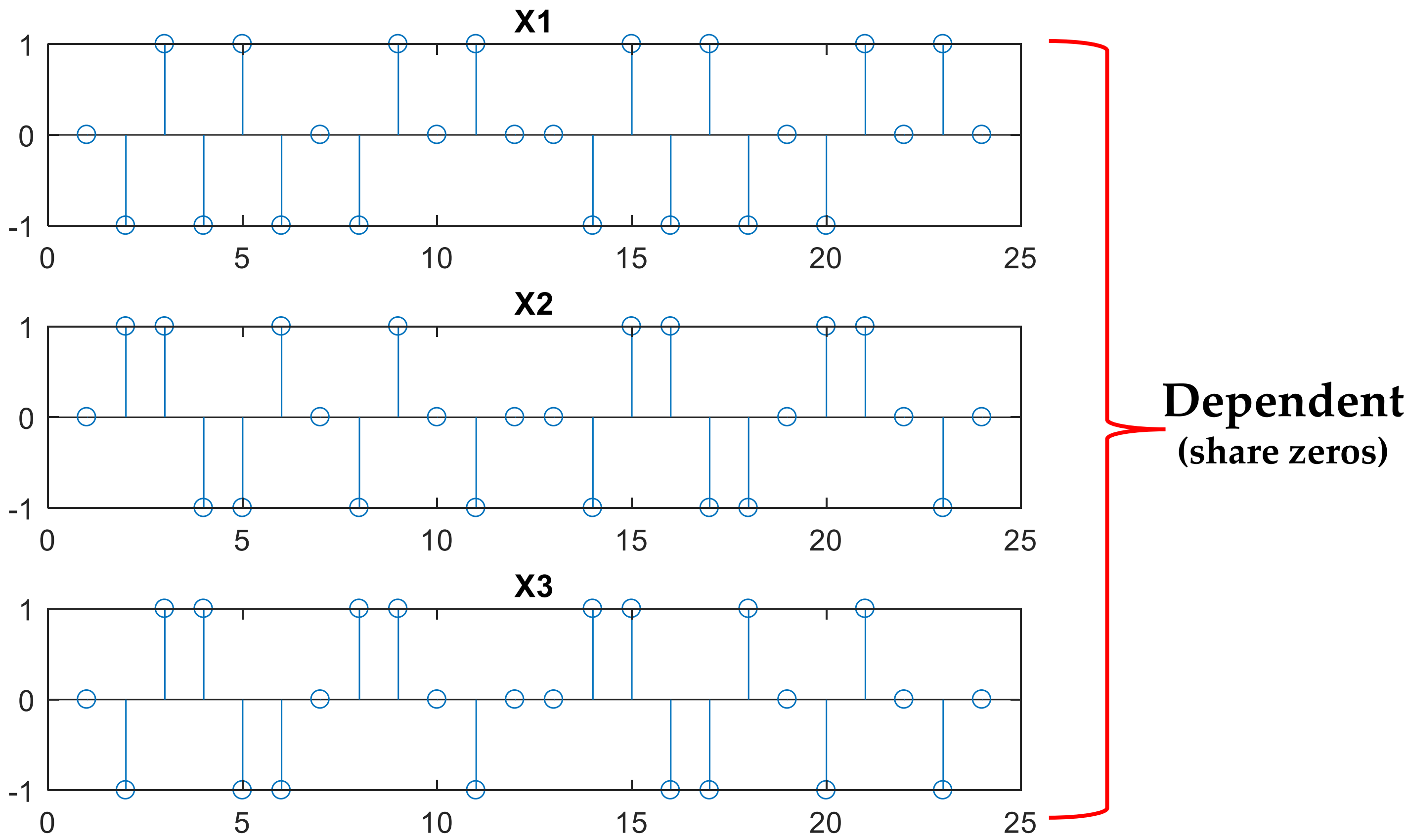
- In most practical scenarios, i.i.d is assumed and hence first-order statistics is used to estimate the entropy of a time series.

*Cover, T. M. (1999). ***Elements of information theory***. John Wiley & Sons.

IECI 2021

Three Short Time Series X1, X2 and X3:

take values from the set {-1,0,1}

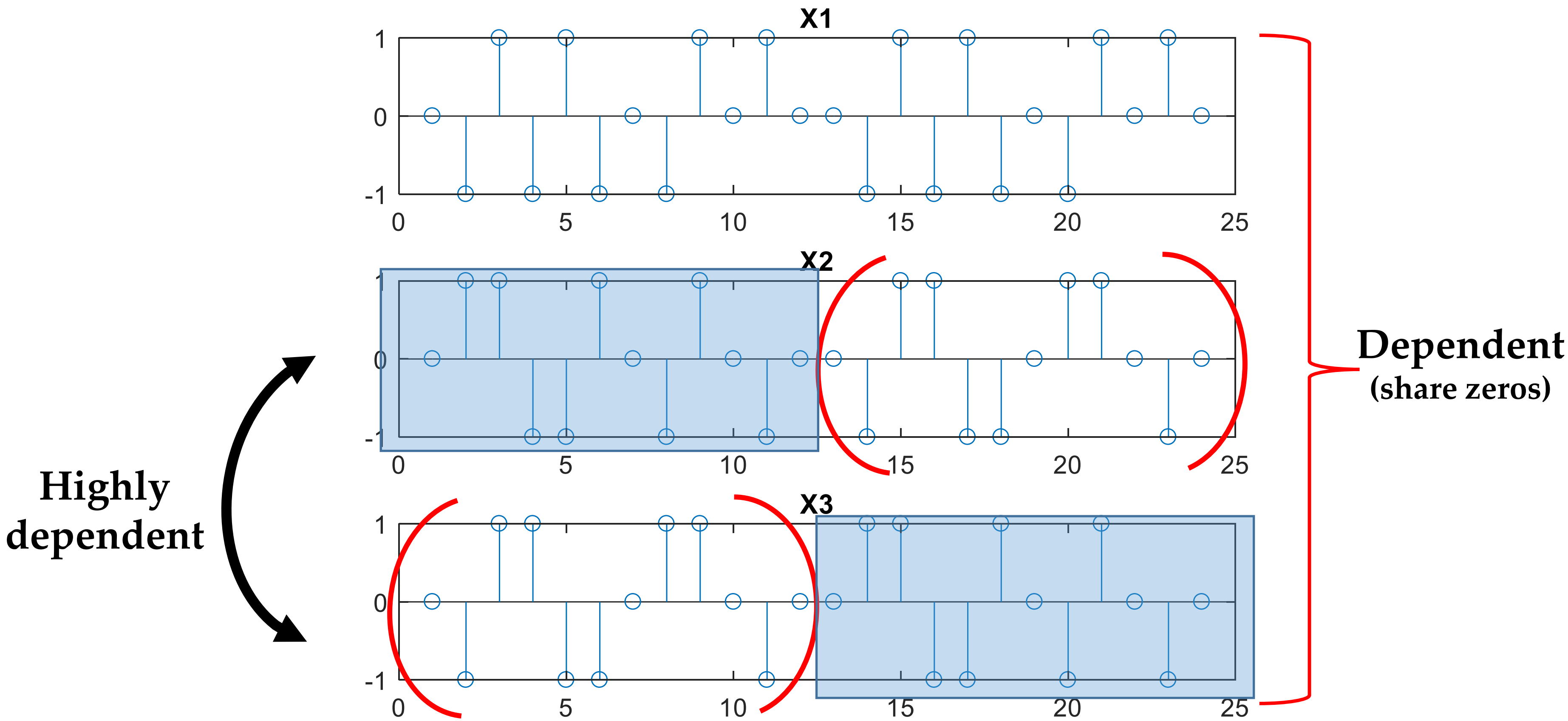


Length = 24 samples

IECI 2021

Three Short Time Series X1, X2 and X3:

take values from the set {-1,0,1}



Length = 24 samples

Linear correlation, Mutual Information (MI) and Mutual Effort-To-Compress (METC*)

$$MI(X,Y) = H(X) + H(Y) - H(X,Y)$$

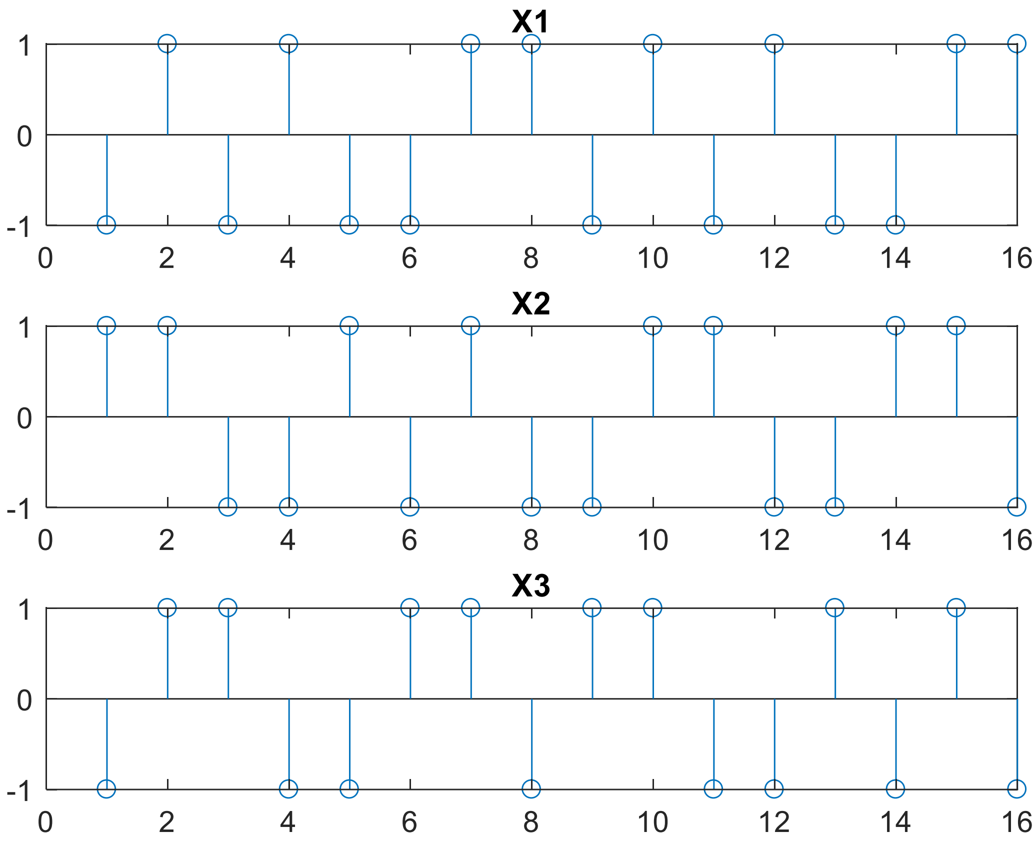
$$METC(X,Y) = ETC(X) + ETC(Y) - ETC_{2D}(X,Y)$$

Pair	ρ (Correlation Coefficient)	MI First-order (bits)	METC
(X1, X2)	0	0.9183	0.2609
(X1,X3)	0	0.9183	0.2609
(X2,X3)	0	0.9183	0.5652

*Nagaraj, N., Balasubramanian, K., & Dey, S. (2013). **A new complexity measure for time series analysis and classification**. *The European Physical Journal Special Topics*, *222*(3), 847-860.

IECI 2021

Zeros removed



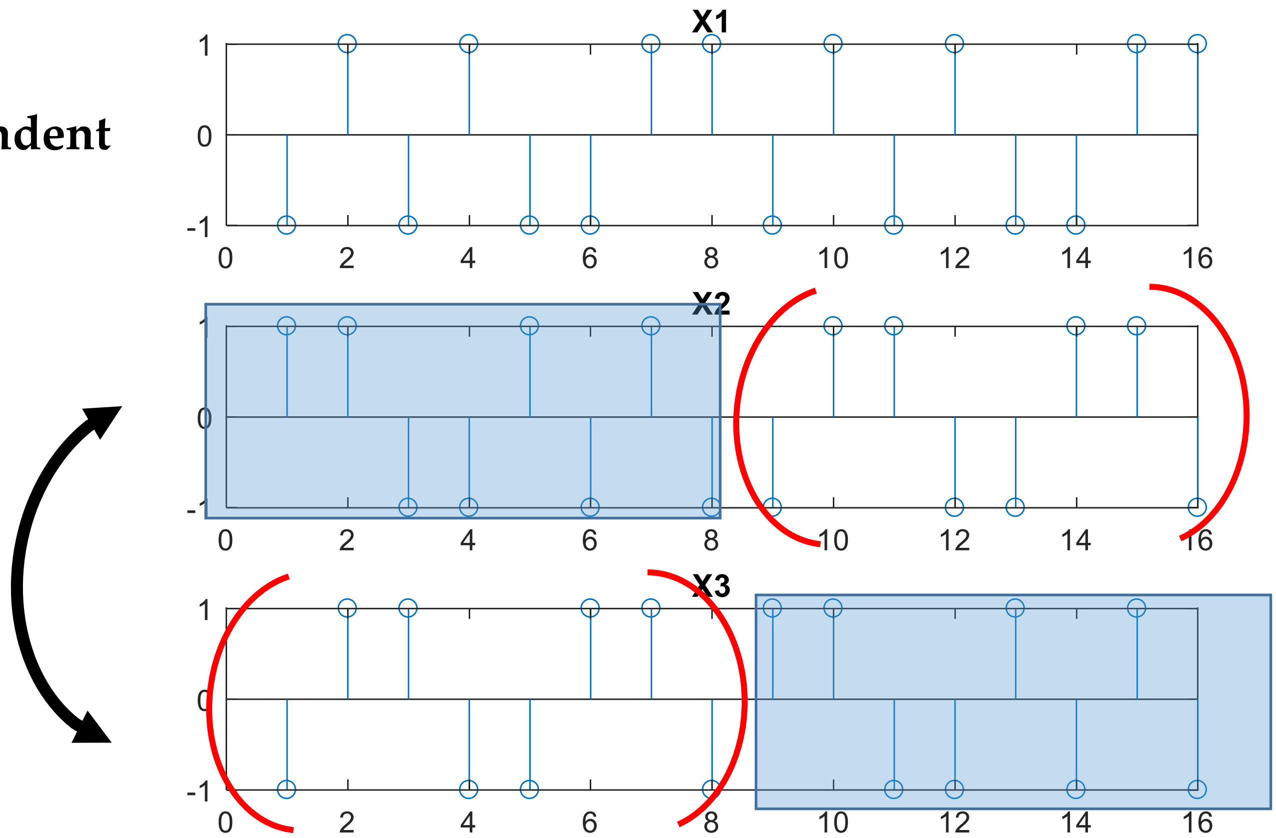
Length = 16 samples

IECI 2021

Zeros removed

Independent

Highly dependent



Length = 16 samples

Linear correlation, Mutual Information (MI) and Mutual Effort-To-Compress (METC*) on the new time series with zeros removed

Pair	ρ (Correlation Coefficient)	MI First-order (bits)	METC
(X1, X2)	0	0	0
(X1,X3)	0	0	0
(X2,X3)	0	0	0.2

Discussion & Conclusion

- Assumption of i.i.d is problematic

- For **short** sequences, first-order statistics is typically used, but as it was demonstrated, this may lead to erroneous conclusions

- Effort-To-Compress, a ***compression-complexity measure***, is able to capture the nonlinear dependence between the time series better than MI without making any further assumptions on the data

- This has implications in causal analysis as well*

*****Nithin Nagaraj. "**Problems with information theoretic approaches to causal learning**." *arXiv preprint arXiv:2110.12497v1 [cs.IT]* (2021).

IECI 2021

The Bandwagon

CLAUDE E. SHANNON

INFORMATION theory has, in the last few years, become something of a scientific bandwagon. Starting as a technical tool for the communication engineer, it has received an extraordinary amount of publicity in the popular as well as the scientific press. In part, this has been due to connections with such fashionable fields as computing machines, cybernetics, and automation; and in part, to the novelty of its subject matter. As a consequence, subject are aimed in a very specific direction, a direction that is not necessarily relevant to such fields as psychology, economics, and other social sciences. Indeed, the hard core of information theory is, essentially, a branch of mathematics, a strictly deductive system. A thorough understanding of the mathematical foundation and its communication application is surely a prerequisite to other applications. I personally believe that many of the concepts

Ref: Shannon, C. E. (1956). The bandwagon. *IRE transactions on Information Theory*, *2*(1), 3.

A warning from Shannon!

IECI 2021

Shannon’s words of caution:

“...it has perhaps been ballooned to an importance beyond its actual accomplishments.‘’

“...it carries at the same time an element of danger‘’

“...it is certainly no panacea for the communication engineer or, *a fortiori*, for anyone else.‘’

“It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized that the use of a few exciting words like *information, entropy*, *redundancy*, do not solve all our problems.‘’

So have we been forewarned!

IECI 2021

For further reading:

Nithin Nagaraj. "**Problems with information theoretic approaches to causal learning**." *arXiv preprint arXiv:2110.12497v1 [cs.IT]* (2021).

IECI 2021

Acknowledgments

The financial support of Tata Trusts is gratefully acknowledged.

Many thanks to Prof. Sangeetha Menon, Consciousness Studies Programme, NIAS for all the support.

TATA TRUSTS

NIAS CSP

NIAS

IECI 2021

Please email nithin@nias.res.in with any questions, suggestions, and feedback.

Thank you for your patient listening

IECI 2021

Appendix

Working of ETC

Effort-to-Compress (ETC)*

It measures the effort required to compress an input sequence using a lossless data compression algorithm, *Non-sequential recursive pair substitution*.

Example: Consider the Input '11011010' with length L = 8.
'11' occurs most frequently. Replace it with '2'
'11011010' => '202010'
Now, '20' occurs most frequently. Replace it with '3'
'202010' => '3310'
Similarly, '3310' => '410' => '50' => '6'. **STOP!**

ETC is defined as number of steps needed to convert input sequence to a constant sequence. For this example, $n = 5$.

$ETC_{\text{normalized}} = n/(L-1) = 5/(8-1) = 5/7 = 0.7142.$
$0 \leq ETC_{\text{normalized}} = \leq 1$

*Nagaraj, Nithin, Karthi Balasubramanian, and Sutirth Dey. "A new complexity measure for time series analysis and classification." *The European Physical Journal Special Topics* 222.3 (2013): 847-860.

IECI 2021