IECI 2021
The 1st International Electronic Conference on Information
01–15 DECEMBER 2021 | ONLINE
Chaired by DR. MARK BURGIN
information
MDPI

Information-theoretic Underpinnings of the Effort-to-Compress Complexity Measure

Aditi Kathpalia [1,*], Nithin Nagaraj [2]

[1] Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech Republic;
[2] Consciousness Studies Programme, National Institute of Advanced Studies, Bengaluru, India.

* Corresponding author: kathpalia@cs.cas.cz

Abstract:
Effort-to-Compress (ETC) is a measure of complexity based on a lossless data-compression algorithm that has been used extensively in characterization and analysis of time-series. ETC has been shown to give good performance for short and noisy time series data and has found applications in the study of cardiovascular dynamics, cognitive research and regulating the feedback of musical instruments. It has also been used to develop causal inference methods for time series data. In this work, a theoretical analysis helps us to demonstrate the links of ETC measure to the total self-information contained in the joint occurrence of most dominant (shortest) patterns occurring at different scales (of time) in a time-series. This formulation helps us to visualize ETC as a dimension like quantity that computes the effective dimension at which patterns in a time-series (translated to a symbolic sequence) appear. We also show that the algorithm that computes ETC can be used as a means for an analysis akin to 'multifractal analysis' using which the power contained in patterns appearing at different scales of the sequence/ series can be estimated. Multifractal analysis has been used widely in analysis of biomedical signals, financial and geophysical data. Our work provides a theoretical understanding of the ETC complexity measure that links it to information theory and opens up more avenues for its meaningful usage and application.

Keywords: effort-to-compress; complexity; self-information; multifractal-analysis

IECI 2021

Introduction: 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 (NSRPS).*

Input: ‘11011010’
‘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.

Min value: 0, max value: $N-1$, N: length of time series.
If ETC steps required = n, $ETC_{\text{normalized}} = n/N-1$

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

Introduction: ETC Applications and Strengths

- ETC has been employed for complexity estimation, formulating measures for causal inference and temporal irreversibility.
- Applications span cardiovascular research, cognitive studies, characterizing genomic sequences and music.
- Performs better than Shannon entropy and Lempel Ziv Complexity for short and noisy sequences

Nagaraj, Nithin, and Karthi Balasubramanian. "Dynamical complexity of short and noisy time series." *The European Physical Journal Special Topics* 226.10 (2017): 2191-2204.

IECI 2021

Results and discussion: info-theoretic formulation

ETC iterations: a sequence $Y_1 \rightarrow Y_2 \rightarrow \cdots \rightarrow Y_{n+1}$
In Y_1, the pair transformed is X_1, in Y_2: X_2, ... ,in Y_n : X_n
W: joint occurrence of paired patterns $X_1, X_2, \ldots, X_n$

$$
\begin{aligned}
p(W) &= p(X_1, X_2, X_3, \ldots X_n), \\
&= p(X_n | X_1, X_2, \ldots X_{n-1}) \cdot p(X_1, X_2, \ldots X_{n-1}), \\
&= p(X_n | X_1, X_2, \ldots X_{n-1}) \cdot p(X_{n-1} | X_1, X_2, \ldots X_{n-2}) \cdot p(X_1, X_2, \ldots X_{n-2}), \\
&\vdots \\
&= p(X_n | X_1, X_2, \ldots X_{n-1}) \cdot p(X_{n-1} | X_1, X_2, \ldots X_{n-2}) \ldots p(X_2 | X_1) p(X_1).
\end{aligned}
$$

Total self-information or Shannon information, $G(W)$, contained in the joint occurrence of $(X_1, X_2, \ldots, X_n)$:

$$
\begin{aligned}
G(W) &= -\log(p(W)), \\
&= -\log(p(X_1)) - \log(p(X_2 | X_1)) - \log(p(X_3 | X_1, X_2)) \ldots - \log(p(X_n | X_1, X_2, \ldots X_{n-1})).
\end{aligned}
$$

Kathpalia, Aditi, and Nithin Nagaraj. "Time-Reversibility, Causality and Compression-Complexity." *Entropy* 23.3 (2021): 327.

Results and discussion: info-theoretic formulation

$p(X_i|X_1, X_2, \ldots, X_{i-1}) \approx$ the frequency of occurrence of X_i in Y_i (as replacement of $X_1, X_2, \ldots, X_{i-1}$ has been done in Y_i)

$$G(W) = -\log\left(\frac{q_1}{N}\right) - \log\left(\frac{q_2}{N-q_1}\right) - \log\left(\frac{q_3}{N-q_1-q_2}\right) \ldots - \log\left(\frac{q_n}{N-q_1-q_2\ldots-q_{n-1}}\right),$$
$$= -\log\left(\frac{q_1}{N}\right)\left(\frac{q_2}{N-q_1}\right)\left(\frac{q_3}{N-q_1-q_2}\right)\ldots\left(\frac{q_n}{N-q_1-q_2\ldots-q_{n-1}}\right).$$

where $q_1, q_2, \ldots, q_n$ are the frequencies of occurrence of $X_1, X_2, \ldots, X_n$ in $Y_1, Y_2, \ldots, Y_n$ respectively.

Compression achieved by ETC algorithm at any step= fractional reduction in length of sequence at that step. If equivalent compression at each step is x and total number of steps/ iterations is n,

$$x^n = \left(\frac{q_1}{N}\right)\left(\frac{q_2}{N-q_1}\right)\left(\frac{q_3}{N-q_1-q_2}\right)\ldots\left(\frac{q_n}{N-q_1-q_2\ldots-q_{n-1}}\right),$$

IECI 2021

Results and discussion: info-theoretic formulation

Taking logarithm,

$$n \cdot \log(x) = \log\left(\left(\frac{q_1}{N}\right)\left(\frac{q_2}{N-q_1}\right)\left(\frac{q_3}{N-q_1-q_2}\right)\dots\left(\frac{q_n}{N-q_1-q_2\dots-q_{n-1}}\right)\right),$$

$$n = \frac{\log\left(\left(\frac{q_1}{N}\right)\left(\frac{q_2}{N-q_1}\right)\left(\frac{q_3}{N-q_1-q_2}\right)\dots\left(\frac{q_n}{N-q_1-q_2\dots-q_{n-1}}\right)\right)}{\log(x)}.$$

$$\boxed{\begin{aligned} n &= -\frac{1}{\log(x)} \cdot G(W), \\ n &= k \cdot G(W), \end{aligned}}$$

ETC steps = k.(Total self-information)

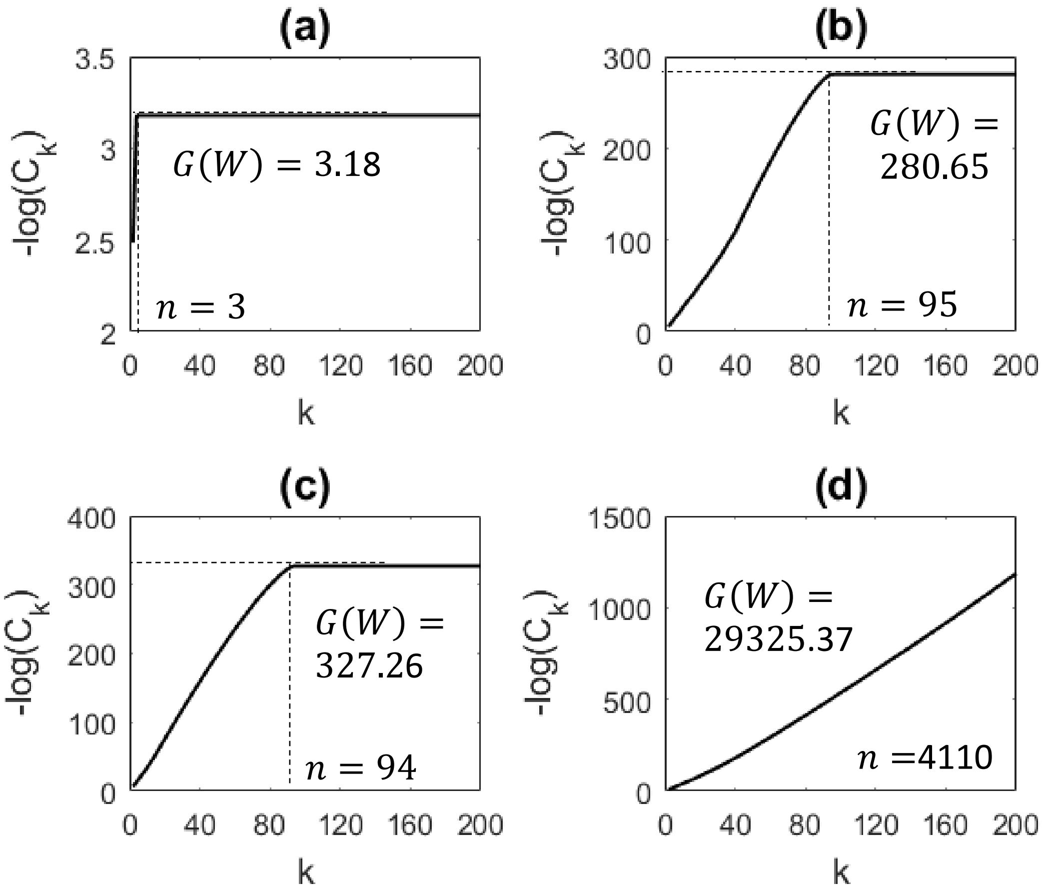
Let $\left(\frac{q_1}{N}\right)\left(\frac{q_2}{N-q_1}\right)\dots\left(\frac{q_n}{N-q_1-q_2\dots-q_{n-1}}\right) = C_n$. Let n_∞ denote the number of ETC steps required for $N \to \infty$ and let x_∞ denote the limit of x.

$$\boxed{n_\infty = \lim_{N\to\infty} \frac{\log(C_{n_\infty})}{\log(x_\infty)}.}$$

Dimension like quantity

IECI 2021

Results and discussion: Simulation examples



Sequences of length 10000,

(a) Repeating periodic: [1 2 3 4]
(b) Repeating periodic: [1 2….1000]
(c) Repeating periodic + random: [1 2 ….25]+ 100 random nos. from [1,2,…,25]
(d) Random: each entry chosen from U(0,1)

k= No. of ETC iterations (no. of bins used=8),
C_k= Total fractional compression until the k^{th} iteration

Results and discussion: Simulation examples

- Highly periodic sequences have low n and $G(W)$ and highly random have high n and $G(W)$.
- $-\log(C_k)$ saturates at $G(W)$ when k reaches n.
- Even though the nature of time-series (b) and (c) is very different, their n is approximately equal. n is the effective dimension at which patterns appear.

IECI 2021

Conclusions and Future Work

- Widely used ETC complexity measure is shown to have a theoretical basis.
- It is related to the *self-information* contained in joint occurrence of most-dominant paired patterns contained in the sequence.
- It can be visualized as a *dimension-like* quantity.
- Measure for **temporal irreversibility** based on ETC has been formulated on the above basis.
- There is a potential to develop a technique akin to **multifractal spectral analysis**.

IECI 2021

For further reading:

Kathpalia, Aditi, and Nithin Nagaraj. "Time-Reversibility, Causality and Compression-Complexity." *Entropy* 23.3 (2021): 327.

Acknowledgements:

The authors are thankful to Dr. M. Paluš for giving useful comments. A. Kathpalia is thankful for the financial support provided by the Czech Science Foundation, Project No. GA19-16066S and by the Czech Academy of Sciences, Praemium Academiae awarded to M. Paluš. N. Nagaraj gratefully acknowledges the financial support of Tata Trusts and Dept. of Science & Tech., Govt. of India (grant nos. DST/CSRI/2017/54 and DST/SATYAM/2017/45).

IECI 2021

Thank You!

Feedback, comments and queries are welcome.
Please write to kathpalia@cs.cas.cz

IECI 2021