The Title

by

M.E Bugge

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Undergraduate Research Project
in Data Science.

Victoria University of Wellington
2024

Abstract

An abstract of fewer than 500 words must be included.

Contents

Chapter 1

Introduction

Ordinal Patterns have been around for twenty years. They have been used in applications in several domains. Many of these applications involve comparing time series, signals and images thorough features extracted from their OPs. Since results about the distribution of such features are recent, our initial interest was comparing existing conclusions with those supported by the new statistical tests. While approaching this research avenue, we found indications of reproducibility crisis in this field, mostly related to data availability. We then moved into the application of these new tests to already published results, and found evidence of other type of reproducibility crises: diverging results depending on the type of library employed. MAGNUS; then, we decided to study...

We are interested in verifying the reproducibility of results of Ordinal Patterns in the scientific literature. Initially, we wanted to check how many of the 3000

1.1 Ordinal Patterns

History Bandt and Pompe first paper about ordinal patterns is from 2002. In it they introduced the concept of turning time series data into patterns. They furthermore used Shannon entropy on the pattern distribution [**?**].

Shannon Entropy was developed in 1948 by C.E.Shannon [**?**].

The 1995 paper "A statistical measure of complexity" by López-Ruiz, R. and Mancini, H. L. and Calbet, X introduces the concept of a statistical measure of complexity [**?**]. This idea was introduced into ordinal patterns in 2003 [**?**] and in 2004 the version of it using Jensen-Shannon divergence were published [**?**], which is the version being used in this paper.

Definition Given a time series $X = (X_1, X_2, ..., X_{n+D-1})$. D is chosen, and recommend by Bandt and Pompe to be set between 3 and 7. D is the length of the subsequence each pattern represents. n is the number of patterns. A tuple of data points is transformed into a pattern by ranking them by numerical order. The lowest observation gets assigned the number 0 and the highest observation gets the number D-1. The pattern can then be written as a string of these numbers. A tuple $(1, 3, 2)$ will have pattern 021.

The exact order of each pattern becomes redundant, when calculating the entropy later, so it can be easier to think of the patterns simply as $\pi^1, \pi^2, .., \pi^{D!}$. Tuples that have the same ordering gets the same pattern, is the main point. There is D! different patterns. Equal values are ignored and can be offset by adding small random perturbations. The frequency of each pattern is defined as

$$p(\pi) = \frac{\#\{t|t \leq T - n, (x_{t+1}, ..., x_{t+n}) \text{ has type } \pi\}}{T - n + 1}$$

[**?**]

In the above section the delay time, τ, between each observation in each subsequence, has been 1, however this can be set to any reasonable value that still produces enough patterns. In this paper only $\tau = 1$ will be used.

From the pattern distribution entropies can be calculated. In this paper only Shannon entropy will be used.

$$h(n) = -\sum p(\pi) log(p(\pi))$$

Normalized version

$$H(n) = -\frac{\sum p(\pi) ln(p(\pi))}{ln(D!)}$$

There is several statistical complexity measures that can be used. They are all a product between the used entropy and a distance measure. In this paper Martin-Plastino-Rosso intesive Statistical Complexity Measure is used, where the distance measure is Jensen-Shannon divergence and it is measured between the pattern distribution and the uniform distribution. The Q_0 is normalizing term. It is defined as

$$C[\mathscr{P}] = H[\mathscr{P}] \cdot Q_J[\mathscr{P}, \mathscr{P}_e]$$

$$Q_J[\mathscr{P}, \mathscr{P}_e] = Q_0 \cdot \mathscr{J}[\mathscr{P}, \mathscr{P}_e]$$

$$\mathscr{J}[\mathscr{P}, \mathscr{P}_e] = S[\frac{\mathscr{P} + \mathscr{P}_e e}{2}] - S[\frac{\mathscr{P}}{2}] - S[\frac{\mathscr{P}_e}{2}]$$

$$\mathscr{P}_e = \{p_j = \frac{1}{W}; j = 1..., W\}$$

$$Q_0 = -2(\frac{W+1}{W} ln(W+1) - 2ln(2W) + ln(W))^{-1}$$

[?]

Applications Going to the original Bandt and Pompe paper [?] on Semantic Scholar and getting the ten most recent citations, gives a good picture of how broadly this methodology is being used.

The topic varies from: "Walnut crack detection..." [?], Schizophrenia [?], Analysis of Smart Drilling [?], Epileptic Seizure detection [?], mind wandering during video-based learning [?] and Random Numbers generated based on dual-channel chaotic light [?]. The rest that was found [?, ?, ?, ?]

1.2 Reproducibility

Reproducibility crisis In data science, data from many different scientific fields are analysed, as shown in the application section of ordinal patterns above. It therefore increases the difficulty of reproducing the data collection stage, as most cases requires an interdisciplinary study involving multiple people. In some cases it is impossible e.g. if the equipment needed is unavailable or a time series of weather data cannot be recollected, since it is impossible to go back in time. The point of science is to eliminate trust, when sharing knowledge, however the above observations highlights the need for some degree of trust, in cases, where data cannot be reproduced. Only data that is random numbers generated in a script, will be reproduced in this paper.

The COVID pandemic gave birth to the following quote. "We warn against the potential misuse or misleading interpretation of public data of variable quality" [**?**]. Data from different governments does not have the same quality. This necessities good source criticism. In this paper data from peer reviewed studies will be trusted, which have often either produced the data themselves or gotten it from reputable organisations as NASA or NOAA.

In this paper the focus will be on direct computational reproduction.

"Computational reproducibility is most often direct (reproducing particular analysis outcomes from the same data set using the same code and software), but it can also be conceptual (analysing the same raw data set with alternative approaches, different models or statistical frameworks)" [**?**]