Physica A 456 (2016) 112–119

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

The likely determines the unlikely

Xiaoyong Yan [a,b], Petter Minnhagen [c,*], Henrik Jeldtoft Jensen [d]

[a] Systems Science Institute, Beijing Jiaotong University, Beijing 100044, China

[b] Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China

[c] IceLab, Department of Physics, Umeå University, 901 87 Umeå, Sweden

[d] Centre for Complexity Science and Department of Mathematics, Imperial College London, South Kensington Campus, SW7 2AZ, United Kingdom

HIGHLIGHTS

- A MaxEnt-method is shown to predict the event-distribution for broad class of real systems.
- Human heights, wind-speed, length of travels and frequency of words are included.
- The complete event-distributions are obtained from a small amount of accessible data.
- Rare events are predicted with high accuracy.
- The likely determines the unlikely for this class of systems.

ARTICLE INFO

Article history:
Received 6 November 2015
Available online 29 March 2016

Keywords:
Complex systems
Frequency distributions
Maximum entropy
Predictions
Real data

ABSTRACT

We point out that the functional form describing the frequency of sizes of events in complex systems (e.g. earthquakes, forest fires, bursts of neuronal activity) can be obtained from maximal likelihood inference, which, remarkably, only involve a few available observed measures such as number of events, total event size and extremes. Most importantly, the method is able to predict with high accuracy the frequency of the rare extreme events. To be able to predict the few, often big impact events, from the frequent small events is of course of great general importance. For a data set of wind speed we are able to predict the frequency of gales with good precision. We analyse several examples ranging from the shortest length of a recruit to the number of Chinese characters which occur only once in a text.

1. Introduction

A detailed understanding of the mechanisms controlling a certain phenomena can often lead to reliable predictions of what to expect. When one considers complex phenomena, say the weather or language, such a very detailed level of description is typically not possible. Despite of this lack of detail it can still be possible to establish a statistical accurate account of possible behaviours [1,2]. The maximum entropy, or likelihood, principle can be applied to a very broad range of phenomena [3–9]. The method consists in estimating the probabilistic description, which is statistically most likely to be consistent with the observations available. It is important to keep in mind that no causal mechanistic description is invoked. Rather one assume that the underlying combinatorial multitude will make happen what is most plausible under given observed constraints. That is, the macro-events generated by the largest number of micro-events are most likely to

* Corresponding author.
E-mail address: Petter.Minnhagen@physics.umu.se (P. Minnhagen).

http://dx.doi.org/10.1016/j.physa.2016.03.027
0378-4371/

Table 1
Likelihood predictions.

Known data	M	N	$k*$	Prediction	Obtained	Measured
Recruits[a]	8 770 975 cm	48 907	$k_{\max} = 207$ cm	$\Longrightarrow$	74.88%	69.08%
Wind speed[b]	232 695 m/s	23 332	5.5% is 5 m/s	$\Longrightarrow$	0.0058%	0.0084%
Bus-trips[c]	35 770 210 km	7 083 210	$10\% \in [0, 1]$ km	$\Longrightarrow$	$\in [2, 3]$ km	$\in [2, 3]$ km
				$\Longrightarrow$	≥ 9 km $= 10\%$	≥ 9 km $= 9.8\%$
Car-drives[d]	448 778 km	48 569	$k_c = 90$ km	$\Longrightarrow$	14.7%	14.5%
Signs[e]	17 915	1 552	$k_{\max} = 747$	$\Longrightarrow$	40.51%	29.12%
English[f]	60 181	6 570	$k_{\max} = 3300$	$\Longrightarrow$	50.44%	52.69%

[a] Swedish recruits born 1975 (see Section 3): M = total length, N = total number, $k_{\max}$ = tallest, $\Longrightarrow$ obtained = shortest in % of tallest.
[b] Wind speed in Öland island, Sweden (see Section 3): M = total wind speed observed, N = total observation days, $k* = 5.5\%$ of days the observed wind speed is 5 m/s, $\Longrightarrow$ obtained = % of wind speed equal to or larger than 32 m/s.
[c] Bus-trips in Shijiazhuang (see Section 3): M = total distance, N = total number of trips, $k* = 10\%$ of trips in interval [0, 1] km, $\Longrightarrow$ obtained = position of maximum and % of trips larger than 9 km.
[d] Car-drives in Detroit (see Section 3): M = total distance, N = total number, $k*$ = the longest 10 trips longer than $k_c = 90$ km, $\Longrightarrow$ obtained = % of trips in interval [0, 1] km.
[e] The Chinese novel *A Q Zheng Zhuan* by Xun Lu written in Chinese characters (see Section 3): M = total number of characters, N = number of different characters, $k_{\max}$ = occurrence of most the frequent character, $\Longrightarrow$ obtained = % of characters occurring only once.
[f] The English novel *Under the Greenwood Tree* by Thomas Hardy (see Section 3): M = total number of English words, N = number of different words, $k_{\max}$ = occurrence of most the frequent word, $\Longrightarrow$ obtained = % of words occurring only once.

occur. Say, throw two dice, it is more likely that the sum of the eyes is equal to 7 than equal to 2, since 6 micro-events lead to 7 eyes and only one to 2 eyes.

It is therefore to be expected that the method will work for stochastic phenomena like lotteries or dice games. However, in the present work we demonstrate that even for causal highly interdependent and deterministic situations the maximum entropy principle leads often, but not always, to predictions of high precision. Below we will comment on the conditions under which reliable predictions may be expected.

The maximum entropy method combined with Bayesian inference is very well established and used routinely, see Ref. [4]. Here we describe how the methodology can be developed to obtain accurate estimates of the entire distribution and predictions about extreme behaviour based on just three numbers: a measure of the total "size", the number of elements and a single measure of most frequent events or the extreme observed. To emphasise the broad applicability we study six different phenomena: heights of humans, wind speed, bus trips, car drives and English and Chinese language.

In many applications rare extreme events are of particular importance, e.g. gales; while their rarity makes it difficult to estimate their frequency of occurrence from observation of the past. We therefore focus on how the method can extract the statistics of the unlikely from the easily observed most frequent events.

Table 1 shows how we in all the considered cases from only three observables, all which are typically easy to access, are able to extract good predictions about various types of extreme or marginal behaviour. One may wonder how a stochastic procedure like the MaxEnt analysis underlying the predictions in Table 1 is able to handle presumably rational and fairly deterministic phenomena like the choice made when travelling a certain distance or the words chosen to express thoughts in a written text. The conclusion is obviously not that some unrecognised stochasticity is in reality governing our choice of words, our travel needs or the growth of recruits or the speed of the wind.

The reason is that despite each individual choice of journey, expression in terms of words or growth of a person may very well be entirely deterministic, in each case large numbers of possible choices exist which leads to a huge number of combinations. So when considering an large collection of realisations of these choices, we cannot distinguish between underlying proper stochastic processes or deterministic processes with a very large sample space. The situation is not very different from when we use statistics to analyse the throw of dice. Each throw is controlled by deterministic mechanics, different throws are subject to slightly different conditions and therefore a large set of throws manifests the combinatorial possibilities available to each deterministic throw.

Section 2 gives a brief review and motivation of the predictive method by which the results are obtained. Results for six explicit examples are given and discussed in some detail in Section 3. Finally, a sum-up and a broader perspective is given in Section 4.

2. Predictive method

To recall how it is typically applied we consider a set of boxes containing N balls [10]. There are N boxes and M unnumbered (indistinguishable) balls. The balls are scrambled by randomly picking two boxes and then moving one ball from the first to the second. The scrambling will produce $N(k)$ boxes which contain k balls. A stationary probability distribution, $P(k)$, describing an ensemble of boxes and balls, is reached after many swaps of balls. In this ensemble the average number of boxes with k balls is given by $NP(k)$. The Shannon entropy of the probability distribution is given by the functional $S[P(k)] = -\sum_k P(k) \ln(P(k))$. To find the most likely distribution $P(k)$ subject to the relevant constraints, one maximises the functional $G[P(k)] = S - b\langle k\rangle - b\langle 1\rangle$, which imposes the constraint $M/N = \langle k\rangle = \sum_k kP(k)$, i.e. the