Testing Statistical Laws in Complex Systems

Martin Gerlach\textsuperscript{1} and Eduardo G. Altmann\textsuperscript{2}

\textsuperscript{1}Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, USA
\textsuperscript{2}School of Mathematics and Statistics, University of Sydney, 2006 NSW, Australia

(Received 19 September 2018; revised manuscript received 19 December 2018; published 26 April 2019)

The availability of large datasets requires an improved view on statistical laws in complex systems, such as Zipf’s law of word frequencies, the Gutenberg-Richter law of earthquake magnitudes, or scale-free degree distribution in networks. In this Letter, we discuss how the statistical analysis of these laws are affected by correlations present in the observations, the typical scenario for data from complex systems. We first show how standard maximum-likelihood recipes lead to false rejections of statistical laws in the presence of correlations. We then propose a conservative method (based on shuffling and undersampling the data) to test statistical laws and find that accounting for correlations leads to smaller rejection rates and larger confidence intervals on estimated parameters.

DOI: 10.1103/PhysRevLett.122.168301

Introduction.—Statistical regularities collected in the form of “universal laws” play a central role in complex systems [1–3]. Zipf’s law of word frequencies [4], the Gutenberg-Richter law of earthquake magnitudes [5], scale-free degree distributions in networks [6], and inter-event time distributions between bursty events [7–10] are prominent examples that triggered entire research lines devoted to explaining the origin and to exploring the consequences of these laws.

Recently, the empirical support of such laws has been heavily questioned. The best known example is the case of scale-free degree distribution of networks; after the seminal work of Barabasi and Albert in 1999 [6], the early 2000s were marked by findings of power law distributions in various network datasets, while in the last five years the trend has reversed and it is now common to read that networks with power law degree distribution are rare [11,12] (see Ref. [13] for a journalistic account). This recent shift in conclusions, which appears in the analysis of Zipf’s law in language [3,14,15] and also in other areas [16,17], is partially due to new (larger) datasets but mostly due to the improved statistical methods: least-squared fitting and visual inspection of double-logarithmic plots (used since Zipf) have been replaced by maximum likelihood methods made popular in the influential article by Clauset, Shalizi, and Newman [17], see Refs. [18–21] for variations. A point often ignored in the interpretations of the recent findings is that these methods rely on two hypotheses:

H1: The observations $x$ are distributed as $p(x;\alpha)$, where $\alpha$ are parameters, e.g., for a power law

$$p(x;\alpha) = Cx^{-\alpha}.$$  \hspace{1cm} (1)

H2: The empirical observations $x_i$, $i = 1, \ldots, N$ are independent (e.g., of $i$ or $x_{i-1}$).

While the statistical laws correspond to H1, the statistical tests rely also on H2 (implicitly assumed, e.g., when the log-likelihood is computed as $\sum_{i=1}^{N} \log p(x_i)$ [5,12,17,21,22]). Complex systems are characterized by strong (temporal and spatial) interdependencies [23] and it is thus not clear whether the recent claims [11,12,16] of violation of the statistical laws arise from systematic deviations of the law itself (H1) or, instead, whether they are due to the well-known fact that observations are not independent (H2).

In this Letter we show that dependencies in the data (violation of H2) have a strong impact on the empirical analysis of statistical laws, leading to rejections even in processes that satisfy the law (H1), and to overconfident selection of models and parameters. We then propose an alternative method that distinguishes between H1 and H2, yielding an upper bound on the degree of correlations for which the statistical law is rejected.

General setting.—Let $\{x_i\} = x_1, x_2, \ldots, x_N$ be an ordered sequence obtained from a measurement process that asymptotically has a well-defined distribution $p(x) = (\#x_i = x/N)$ as $N \to \infty$. In observations of dynamical systems (or time series), $x_i$ will typically depend on the observations at previous times so that for all times $\tau$ smaller than some (relaxation) time $\tau^*$ we find $p(x_i|x_{i-\tau}) \neq p(x)$. Violations of H2 happen also when data is not measured as a time series. In the case of Zipf’s law of word frequencies, syntax restrict the valid sequences of word tokens, in violation of H2 (both in the rank-frequency and frequency distribution pictures [3,15]). In the case of networks, H2 can be violated because of the generative process or because of the sampling employed to observe the nodes and links (typically a subsample of an underlying network). In fact, it has been shown that the degree distribution of networks is sensitive to the sampling procedure [24–26].
Moreover, the hypotheses H1 and H2 of the standard tests for power law distribution do not build a proper probabilistic network model [27], are thus not suitable to a rigorous statistical analysis [29], and the analysis of the degree distribution of networks requires further assumptions about the sampling or generative process.

More generally, strong correlations are ubiquitous in complex systems [23] and it is hard to imagine a case for which H2 holds. In Fig. 1 we show how previously proposed statistical laws and correlations appear together in paradigmatic complex systems; the Gutenberg-Richter law for earthquakes (exponential [5]), interevent times of words (stretched exponential [7–9]), Zipf’s law for word frequencies (power law [4]), and scale-free distribution for the node degree in networks (power law [6]). While earthquake events and interevent times naturally occur as time series data, we mapped word frequencies in texts and the network data into ordered sequences \( \{ x_i \} \) based on a simple sampling process (see caption of Fig. 1) in order to illustrate and quantify the violation of H2 in an unified framework.

**Constructed example.**—We now show that the traditional methods [17] lead to a rejection of a power law distribution [Eq. (1)] even for data which are power law distributed for \( N \to \infty \). This is done by building a Markov process [36,37] in which H1 is satisfied but H2 is violated i.e., \( x_i \) depends on \( x_{i-1} \) and \( p(x) = Cx^\alpha \) for \( N \to \infty \), see the Supplemental Material [30, Sec. III].

In Fig. 2 we show that the violations of H2 have a strong influence on the analysis of statistical laws formulated in H1. In particular, the application of the traditional recipes [17] lead to the wrong conclusion that the data are not compatible with a power law distribution: the probability of rejecting the null hypothesis at a 5% significance level is much larger than 5% even for small sample sizes \( N \) [inset of Fig. 2(b)]. This corresponds to a type-I error because, by construction, the data satisfy H1. The origin of this failure thus originates from the fact that correlations lead to an effective reduction of the number of independent observations implying larger fluctuations which lead to larger deviations from the fitted model. Specifically, we recall that the test employed in Ref. [17] consists of comparing the Kolmogorov-Smirnov (KS) distance between the correlated data and the fitted curve, KS\(_{\text{correlated}}\) (blue curve), and the KS distance between independent samples of the model (H1 + H2) and the fitted curve, KS\(_{\text{model}}\) (orange curve). More precisely, the statistical law is rejected at 5% significance level if KS\(_{\text{correlated}} > \) KS\(_{\text{model}}\) in 95% realizations (samplings) of the model. While in our artificial data KS\(_{\text{correlated}} \propto 1/\sqrt{N}\) (as expected) and thus KS\(_{\text{correlated}} \to 0\) for \( N \to \infty \), this convergence is shifted from the convergence of KS\(_{\text{model}}\) [Fig. 2(b)] due to the correlations. This shift leads to an increased rejection rate (\( \approx 1, p \) value \( \approx 0 \)).

Violations of H2 are important not only in the hypothesis-testing setting discussed above, they also lead to increased systematic and statistical errors (bias and fluctuations) in the fitting of the parameter \( \hat{\alpha} \) [Fig. 2(c)] and, thus, in the selection between models [38,39].

**Real data.**—In order to confirm that the results discussed above are also relevant in real datasets—which have a fixed size \( N \)—we consider two types of undersampling of data to
sizes $n < N$: taking $n$ points either randomly or preserving the structures or correlations by taking consecutive portions of the time series (the network and word-frequency databases are first mapped to a time series, as in Fig. 1). In order to distinguish between the effect of the shape of the distribution (H1) and correlations (H2) we compare the distribution of the $n$ points with (i) the proposed statistical law and (ii) the empirical distribution (i.e., the one obtained for $n = N$). Our results (see the Supplemental Material [30], Sec. IV) confirm that correlated data show higher rejection rate and fluctuations of parameters.

Alternative approach.—In the vast literature of statistical methods for dependent data, two general approaches can be identified. The first approach is to incorporate the violation of independence in more sophisticated (parametric) models, e.g., in a time series one could consider Gaussian Markov processes [40]. This is of limited use in our case because statistical laws aim to provide a coarse-grained description (stylized facts) valid in many systems, instead of different detailed models of particular cases. The second (nonparametric) approach, which we pursue here, is to decorrelate or decluster the data, leading to a dataset with an “effective” sample size $N^* \leq N$ [41–43]. In practice, the analysis consists of multiple realizations of the following three steps. (i) Randomize (shuffle) the original sequence and select randomly $n$ points, for different $n \in [1, N]$. (ii) Apply the traditional statistical analysis (i.e., the hypothesis test, model comparison, and fitting based on H1 + H2) to the randomized dataset obtained in (i), investigating their dependence on $n$. (iii) Estimate the correlation $\tau^*$, defined as the time after which two observations (in the time series) are independent from each other. Out of the total $N$ samples we thus estimate $N^* = N/\tau^*$ to be the number of independent samples and therefore we select the results from step (ii) for $n \approx N^*$.

The determination of $\tau^*$—or the effective sample size $N^*$—in step (iii) requires knowledge or assumptions about how the data were generated. For the case of temporal sequences we propose to compute the autocorrelation and take as $\tau^*$ the lag for which it reaches an interval around zero (1 percentile of the random realizations, as in Fig. 1). In the constructed example (Fig. 2), we obtain $\tau^* = 407 \Rightarrow N^* = 245$, which leads to a rejection rate (at $p$ value $= 0.05$) equal to 5% for all $n < N^*$. For the case of networks, the determination of the effective sample size $N^*$ depends on the generative process and/or the sampling used to measure the data (here we assumed a specific edge sampling method, as described in Fig. 1.) In Fig. 3, we show evidence of the effectiveness of our approach through a systematic analysis of the $p$ value distribution as a function of $n$ for both the constructed and empirical datasets. This is further corroborated in artificial data (see the Supplemental Material [30], Sec. V) showing that (i) our method for the selection of $\tau^*$ is superior to the one proposed in Ref. [41] (sum of the autocorrelation function) and (ii) can be equally applied to data with other types of correlation: a Markov process with negative correlation and a Gaussian process with long-range correlations. In all cases our approach shows an uniform distribution of $p$ values under the null hypothesis.

An important message of our analysis is that conclusions about the statistical law can be obtained even when the precise value of $\tau^*$ (or the effective sample size $N^*$) is unknown in step (iii). By shuffling and undersampling the sequence at different sizes $n$—steps (i) and (ii)—we can investigate how the results depend on $n$ and obtain the range in $\tau^*$ for which the different conclusions hold. For instance, in the case of earthquakes [Fig. 3(c)] we see that the rejection increases dramatically around $n \approx 10^5$. We thus conclude that, in this dataset of size $N \approx 10^5$, we
Traditional methods based on the hypothesis of the data to identify the strongest assumption about correlations of law (H1). Here we proposed a methodology which allows provide much stronger evidence of the rejection of the process so that rejections of the compound hypothesis should make weaker assumptions about the generative the estimated parameters. Stronger tests of statistical laws data, and (ii) are over-optimistic regarding uncertainties of (i) lead to wrong rejections of the laws because of correlated rejection. In fact, here we have shown how these methods a strong hypothesis that is easily violated, therefore favoring independent data (H2) are weak tests because they include in the choice of best models. The need to account for correlations (and violations of H2) actual data show much larger fluctuations than expected under the hypothesis of independence are important and have been mostly ignored in the analysis of statistical laws in complex systems (see Refs. [5,12,25] for exceptions). As shown above, due to correlations (and violations of H2) actual data show much larger fluctuations than expected under the hypothesis of independent observations. By using a shuffled and undersampled dataset, we obtain larger uncertainties in the estimated parameters; we expect similar lack of certainty in the choice of best models. The need to account for violations of the independence assumption, shown in this Letter, applies much more broadly than the cases treated above. Correlations should be accounted for whenever testing statistical laws in complex systems, such as linguistic laws [3], scaling laws with system size—maximum likelihood methods based on H2 have been applied to biological allometric laws [46] and to city data [47]—and different distributions of interevent time (burstiness) [7–10].

The code and data shown in this Letter can be obtained by following the link in Ref. [48].

We thank F. Font-Clos and J. Moore for the careful reading of the manuscript.