On sampling and modeling complex systems

J. Stat. Mech. (2013) P09003

(http://iopscience.iop.org/1742-5468/2013/09/P09003)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

On sampling and modeling complex systems

Matteo Marsili[1], **Iacopo Mastromatteo**[2] **and Yasser Roudi**[3,4]

[1] The Abdus Salam International Centre for Theoretical Physics, Strada Costiera 11, I-34014 Trieste, Italy
[2] Capital Fund Management, 21-23 Rue de l'Université, F-75007 Paris, France
[3] Kavli Institute for Systems Neuroscience, NTNU, Trondheim, Norway
[4] Nordita, KTH Royal Institute of Technology and Stockholm University, Stockholm, Sweden
E-mail: marsili@ictp.trieste.it, Iacopo.Mastromatteo@cfm.fr and yasserroudi@gmail.com

Received 18 April 2013
Accepted 3 August 2013
Published 6 September 2013

Online at stacks.iop.org/JSTAT/2013/P09003
doi:10.1088/1742-5468/2013/09/P09003

Abstract. The study of complex systems is limited by the fact that only a few variables are accessible for modeling and sampling, which are not necessarily the most relevant ones to explain the system behavior. In addition, empirical data typically undersample the space of possible states. We study a generic framework where a complex system is seen as a system of many interacting degrees of freedom, which are known only in part, that optimize a given function. We show that the underlying distribution with respect to the known variables has the Boltzmann form, with a temperature that depends on the number of unknown variables. In particular, when the influence of the unknown degrees of freedom on the known variables is not too irregular, the temperature decreases as the number of variables increases. This suggests that models can be predictable only when the number of relevant variables is *less* than a critical threshold. Concerning sampling, we argue that the information that a sample contains on the behavior of the system is quantified by the entropy of the frequency with which different states occur. This allows us to characterize the properties of *maximally informative samples*: within a simple approximation, the most informative frequency size distributions have power law behavior and Zipf's law emerges at the crossover between the under sampled regime and the regime where the sample contains enough statistics to make inferences on the behavior of the system. These ideas are illustrated in some applications, showing that they can be used to identify

1742-5468/13/P09003+21$33.00

relevant variables or to select the most informative representations of data, e.g. in data clustering.

Keywords: critical phenomena of socio-economic systems, protein function and design (theory), clustering techniques, statistical inference

ArXiv ePrint: 1301.3622

Contents

1. Introduction **2**

2. The setup **4**

2.1. Gibbs distribution on $\underline{s}$ 6

3. Learning from sampling a complex system **7**

3.1. Most informative samples 9

3.2. Criticality and Zipf's law 11

4. Applications **12**

4.1. Protein sequences 13

4.2. Clustering and correlations of financial returns 14

4.3. Keywords in a text 16

5. Discussion **17**

Acknowledgments **18**

Appendix. When are models predictive? The Gaussian case **18**

References **20**

1. Introduction

Complex systems such as cells, the brain, the choice behavior of an individual or the economy can generally be regarded as systems of many interacting variables. Their distinguishing feature is that, contrary to generic random systems, they perform a specific function and exhibit non-trivial behaviors. Quantitative science deals with collecting experimental or empirical data that reveal the inherent mechanisms and organizing principles that suffice to reproduce the observed behavior within theoretical models. The construction of machines or the design of intervention which achieve a desired outcome, such as for example in drug design [25] or for the regulation of financial markets [26], crucially depend on the accuracy of the models.

This endeavor has intrinsic limits: our representations of complex systems are not only approximate, they are incomplete. They take into account only a few variables—that are at best the most relevant ones—and the interactions among these. By necessity they neglect a host of other variables that also affect the behavior of the system, even though

J. Stat. Mech. (2013) P09003

on a weaker scale. These are not only variables we neglect, but *unknown unknowns* we do not even know exist and have an effect.

This is not necessarily a problem as long as (i) the phenomenon depends on a few relevant variables and (ii) one is able to identify and to probe them[5]. Yet, even if advances in IT and experimental techniques have boosted our ability to probe complex systems to an unprecedented level of detail, we are typically in the situation where the state space of the system at hand is severely under sampled and relevant variables (e.g. the expression of a gene) are in many cases inferred from indirect measurements.

In addition, there are intriguing statistical regularities that arise frequently when probing complex systems. Frequency counts in large samples often exhibit the so-called Zipf's law, according to which the kth most frequent observation occurs with a frequency that is roughly proportional to $1/k$, an observation that has attracted considerable interest over several decades now[6]. Model systems in physics, e.g. for ferromagnetism, exhibit similar scale-free behavior only at special 'critical' points, where the system undergoes a phase transition. This leads one to wonder about mechanisms by which nature would self-organize to a critical point [3] or on the generic features of systems that share this property [4]. Yet, the fact that Zipf's law occurs in a wide variety of different systems suggests that it does not convey specific information about the mechanism of self-organization of any of them.

Here we address the general problem of modeling and sampling a complex system from a theoretical point of view. We focus on a class of complex systems which are assumed to maximize an objective function depending on a large number of variables. Only some of the variables are known, whereas the others are unknown. Accordingly, only the part of the function that depends solely on the known variables is known, for the rest one can at best know its statistics. The assumption that complex systems optimize some function, even if it is widely used in modeling (e.g. utility/fitness maximization in economics/biology), may be debatable. Still, it allows us to address two related issues: first, under what conditions do models based on a subset of known variables reproduce systems behavior? How many variables should our models account for and how relevant should they be? Second, can we quantify how much information a given sample contains on the behavior of a complex system? What is the maximal amount of information that a finite data set can contain and what are the properties of optimally informative samples in the strongly under sampled regime?

In section 2, after constructing a mathematically well defined set up, we first discuss the issue of model's predictability: given some knowledge about how the objective function depends on the observed variables, what is the probability that we correctly predict the behavior of these variables? We show that, under very broad conditions, the dependence of the probability to observe a given outcome on the (observable part of the) objective function takes a Gibbs–Boltzmann form. In particular, if the dependence on unknown variables is not too irregular—i.e. if the distribution of the unknown part of the objective function has thin tails—then the 'temperature' parameter *decreases* with the number of unknown variables. This suggests that models are predictable only when the number of

[5] Indeed, as Wigner argues: 'It is the skill and ingenuity of the experimenter which show him phenomena which depend on a relatively narrow set of relatively easily realizable and reproducible conditions' [1].

[6] The literature on this finding is so vast that a proper account would require a treatise of its own. We refer to recent reviews [2] and papers [4, 11, 15] and references therein.

unknown variables is *large* enough. This is illustrated for a particular case, drawing from results on the Random Energy Model [5], which is worked out in the Appendix. There we find that models are predictable only when the number of known variables is *less* than a critical threshold. This suggests a general argument for the non-trivial fact that 'in spite of the baffling complexity of the world, $[\cdots]$ phenomena which are independent of all but a manageably small set of conditions' exist, which makes science possible [1].

In section 3 we will then be concerned with what can be called an inverse problem: if we choose some variables to observe, and collect a number of samples, how much do we learn about the objective function? We argue that (i) the information that the sample contains on the behavior of the system is quantified by the entropy of the frequency with which different states occur. On the basis of this, (ii) we characterize most informative samples and we find that their frequency size distributions, in the under sampled regime, have power law behavior. Within our approximated treatment, we find that the under sampling regime can be distinguished from the regime where the sample contains enough statistics to make inferences on the underlying distribution. Finally, (iii) the distribution with the highest information content coincides with Zipf's law, which is attained at the crossover between these two regimes.

Finally section 4 gives evidence, based on concrete applications in proteins, finance and language, that these insights can be turned into practical criteria for studying complex systems, in particular for selecting relevant variables and/or the most informative representation of them.

2. The setup

We consider a system which optimizes a given function $U(\vec{s})$ over a certain number of variables $\vec{s} = (\underline{s}, \bar{s})$. Only a fraction of the variables—the 'knowns' $\underline{s}$—are known to the modeler, as well as that part of the objective function $u_{\underline{s}}$ that depends solely on them. The objective function also depends on other variables $\bar{s}$—the 'unknowns'—in ways that are unknown to the modeler. Formally, we can define $u_{\underline{s}} = E_{\bar{s}}[U(\vec{s})]$, where $E_{\bar{s}}[\cdots]$ stands for the expected value over a prior distribution on the dependence of $U(\vec{s})$ on the unknown variables, which encodes our ignorance on them. In other words,

$$U(\vec{s}) = u_{\underline{s}} + v_{\bar{s}|\underline{s}} \tag{1}$$

where $v_{\bar{s}|\underline{s}} = U(\vec{s}) - E_{\bar{s}}[U(\vec{s})]$ is an unknown function of $\bar{s}$ and $\underline{s}$, which we assume to be drawn randomly and independently for each $\vec{s} = (\underline{s}, \bar{s})$ from a given distribution $p(v)$. Hence $E_{\bar{s}}[\cdots]$ denotes the expectation with respect to this distribution. The fact that $v_{\bar{s}|\underline{s}}$ are independent draws from $p(v)$ here translates into the fact that knowledge of $\bar{s}$ does not provide any information on $\underline{s}$ as long as $v_{\bar{s}|\underline{s}}$ is unknown. This is what would be dictated by the maximum entropy principle [6], in the absence of other information on the specific dependence of U on $\vec{s}$.[7] This also corresponds to the most complex model we could think of for the unknown part of the system, as its full specification requires a number of parameters that grows exponentially with the number of unknown variables.

[7] Indeed, if the variables were not independent, we should have some information on their mutual dependence and if they were not identical we should have some clue of how they differ.

J. Stat. Mech. (2013) P09003

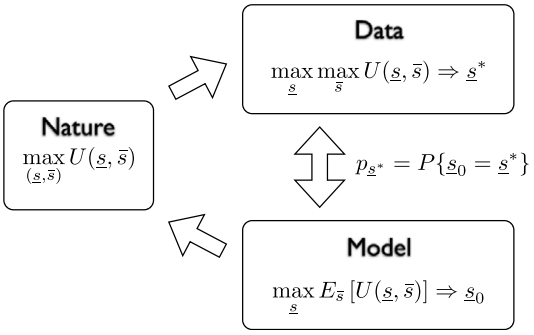

Figure 1. Sketch of the setup: $\underline{s}$ are the known variables. The behavior of the system is encoded in the optimal choice $\underline{s}^*$. This results from the maximization of a function $U(\underline{s}, \bar{s})$, which also depends on unknown variables $\bar{s}$. Assuming it is possible to model the dependence of the objective function on the known variables $\underline{s}$, i.e. that $u_{\underline{s}} = E_{\bar{s}}[U(\underline{s}, \bar{s})]$ is known, what is the probability that the model's prediction $\underline{s}_0$ matches the observed behavior of the system? How relevant and how many should the known variables be?

Therefore, the behavior of the system is given by the solution

$$\vec{s}^* = (\underline{s}^*, \bar{s}^*) \equiv \arg\max_{\vec{s}} U(\vec{s}) \tag{2}$$

whereas the behavior predicted by the model, on the known variables, is given by

$$\underline{s}_0 \equiv \arg\max_{\underline{s}} u_{\underline{s}}. \tag{3}$$

Within this simplified description, the predictability of the model is quantified by the probability

$$p_{\underline{s}_0} = P\{\underline{s}_0 = \underline{s}^*\} \equiv E_{\bar{s}}[\delta_{\underline{s}_0,\underline{s}^*}] \tag{4}$$

that the model reproduces the behavior of the system. This setup is sketched in figure 1.

Let us give a few examples:

- The choice of the city (i.e. $\underline{s}$) in which individuals decide to live, does not only depend on the characteristics of the city—which may be encoded in some index $u_{\underline{s}}$ of city's living standards—but also on unobserved factors ($\bar{s}$) in unknown individual specific ways. Here $v_{\bar{s}|\underline{s}}$ is a different function for each individual—encoding the value of other things $\bar{s}$ he/she cares about (e.g. job and leisure opportunities, personal relations, etc), in the particular city $\underline{s}$.
- A plant selects its reproductive strategy depending on the environment where it lives. This ends up in measurable phenotypic characteristics, e.g. of its flowers, that can be classified according to a discrete variables $\underline{s}$. The variables the species is optimizing over $\vec{s} = (\underline{s}, \bar{s})$, also include unobserved variables $\bar{s}$ that influence other traits of the phenotype in unknown ways.
- A text is made of words $\underline{s}$ in a given language. Each word $\underline{s}$ in the text has been chosen by the writer, depending on the words $\bar{s}$ that precede and follow it, in order to efficiently represent concepts in the most appropriate manner. We assume that this can be modeled by the writer maximizing some function $U(\vec{s})$.

J. Stat. Mech. (2013) P09003

- Proteins are not random hetero-polymers. They are optimized to perform a specific function, e.g. transmit a signal across the cellular membrane. This information is encoded in the sequence $\vec{s}$ of amino acids; however, only a part of the chain ($\underline{s}$) is directly involved in the function (e.g. binding of some molecules at a specific site). The rest ($\bar{s}$) may have evolved to cope with issues that have nothing to do with the function, and that depend on the specific cellular environment the protein acts in.

Within this set up, in section 2.1 we address the following question: if we only have access to $\underline{s}$, how well can we predict the behavior of the system? More precisely: what is the functional dependence of the probability for a configuration $\underline{s}$ to be the true maximum $\underline{s}^*$?

2.1. Gibbs distribution on $\underline{s}$

The functional dependence of the probability for a generic configuration $\underline{s}$ to be the true maximum $\underline{s}^*$, which we have denoted as $p_{\underline{s}} = P\{\underline{s} = \underline{s}^*\}$, can be derived under very general conditions. We focus here on the case where all the moments are finite: $E_{\bar{s}}[v_{\bar{s}|\underline{s}}^m] < +\infty$ for all $m > 0$. Without loss of generality, we can take $\underline{s} = (s_1, \ldots, s_n)$ and $\bar{s} = (s_{n+1}, \ldots, s_N)$, with the variables $s_i = \pm 1$ taking two values for $i = 1, \ldots, N$. The system would not be that complex if n and N were small, so we focus on the limit where both n and N are very large (ideally $n, N \to \infty$).

For all $\underline{s}$, extreme value theory [7] shows that

$$\max_{\bar{s}} v_{\bar{s}|\underline{s}} \cong a + \frac{\eta_{\underline{s}}}{\beta}, \tag{5}$$

where a is a constant, $\eta_{\underline{s}}$ are i.i.d. Gumbel distributed, i.e. $P\{\eta_{\underline{s}} < x\} = \mathrm{e}^{-\mathrm{e}^{-x}}$ and β depends on the tail behavior of the distribution of $v_{\bar{s}|\underline{s}}$ (see later). Therefore

$$p_{\underline{s}} \equiv P\{\underline{s} = \underline{s}^*\} = P\{\beta u_{\underline{s}} + \eta_{\underline{s}} \geq \beta u_{\underline{s}'} + \eta_{\underline{s}'}, \forall \underline{s}' \neq \underline{s}\} \tag{6}$$

$$= \int_{-\infty}^{\infty} \mathrm{d}\eta_{\underline{s}}\, \mathrm{e}^{-\eta_{\underline{s}} - \mathrm{e}^{-\eta_{\underline{s}}}} \prod_{\underline{s}' \neq \underline{s}} \int_{-\infty}^{\eta_{\underline{s}} + \beta(u_{\underline{s}} - u_{\underline{s}'})} \mathrm{d}\eta_{\underline{s}'}\, \mathrm{e}^{-\eta_{\underline{s}'} - \mathrm{e}^{-\eta_{\underline{s}'}}} \tag{7}$$

$$= \frac{1}{Z(\beta)} \mathrm{e}^{\beta u_{\underline{s}}}, \qquad Z(\beta) = \sum_{\underline{s}'} \mathrm{e}^{\beta u_{\underline{s}'}} \tag{8}$$

which is the Boltzmann distribution, also called Logit model in choice theory. The derivation of the Logit model from a random utility model under the assumption of Gumbel distributed utilities is well known [8, 9]. Limit theorems on extremes dictate the form of this distribution for the whole class of models for which $v_{\bar{s}|\underline{s}}$ have all finite moments. This result extends to the case where $v_{\bar{s}|\underline{s}}$ are weakly dependent, as discussed in [7].

The result of equation (8) could have been reached on the basis of maximum entropy arguments alone: on the true maximum, $\underline{s}^*$, the model's utility attains a value $u_{\underline{s}^*}$ that will generally be smaller than $u_{\underline{s}_0}$. Without further knowledge, the best prediction for $p_{\underline{s}}$ is given by the distribution of maximal entropy consistent with $E[u_{\underline{s}}] = u_{\underline{s}^*}$. It is well known that the solution of this problem yields a distribution of the form (8). While this is reassuring, maximum entropy alone does not predict how the value of β depends on the number of unknown unknowns. By contrast, extreme value theory implies that if the

J. Stat. Mech. (2013) P09003

asymptotic behavior of $p(v)$ for large v is given by $\log p(v) \sim -|v|^{\gamma}$, then one can take

$$\beta = [(N-n)\log 2]^{1-1/\gamma}\,. \tag{9}$$

One may naïvely expect that the predictability of the model $p_{\underline{s}_0}$ gets smaller when the number $N-n$ of unknown variables increases. This is only true for $\gamma < 1$, as indeed β decreases as the number of unknown unknowns increases in this case. When $p(v)$ decays faster than exponential ($\gamma > 1$), which includes the case of Gaussian variables, β diverges with the number of unknowns. If the number n of observed variables stays finite, we expect that $p_{\underline{s}_0} \to 1$ in the limit of an infinite number of unknown variables.

A manifestation of this non-trivial behavior is illustrated by the Gaussian case ($\gamma = 2$), where also $u_{\underline{s}}$ are assumed to be i.i.d. draws from a Gaussian distribution with variance[8] σ^2. There, as shown in the appendix, for a given value of σ, the prediction of the model is reliable only as long as the fraction $f = n/N$ of known variables is *smaller* than a critical value $f_c = \sigma^2/(1+\sigma^2)$.

Summarizing, in this section we have shown that, given the form of $u_{\underline{s}}$, the probability to correctly predict a certain state $\underline{s}$ follows a Gibbs–Boltzmann form with a 'temperature' that depends on the number of unknown variables. A natural question one may ask at this point is the inverse problem to this: how much can we tell about $u_{\underline{s}}$ by observing the system? This is the question that we will address in section 3.

3. Learning from sampling a complex system

Given a sample $(\underline{s}^{(1)}, \ldots, \underline{s}^{(M)})$ of M observations of the state of a system, what can we learn on its behavior? As before, our working hypothesis is that $\underline{s}^{(i)}$ is the outcome of an optimization of an unknown function $U(\vec{s})$ on a set of variables $\vec{s}$ that we observe only in part. In order to connect to the *direct problem* discussed in section 2.1, we note that one can also define $u_{\underline{s}}$, as $u_{\underline{s}} = E_{\bar{s}}[U(\vec{s})]$, where the expected value, now, is an average over experiments carried out under the same experimental conditions, as far as the variables $\underline{s}$ are concerned. Therefore, the function $u_{\underline{s}}$, while unknown, is the same across the sample. The part of the objective function that depends on the unknown variables can again be defined as $v_{\bar{s}|\underline{s}} = U(\vec{s}) - u_{\underline{s}}$. However, since by definition there is no way to control the unknown variables, we cannot assume, *a priori*, that the influence of the unknowns on the observed variables is the same across the sample. Rather, this is consistent with the function $v_{\bar{s}|\underline{s}}$ being a different independent draw from some distribution $p(v)$, for each $\vec{s}$ and for each point of the sample[9]. Thus we shall think of the sample $(\underline{s}^{(1)}, \ldots, \underline{s}^{(M)})$ as being M independent configurations drawn from a distribution of the Gibbs–Boltzmann form as in equation (8).

[8] σ quantifies the relevance of the known variables. Note indeed that the typical variation ΔU of the objective function when a known variable is flipped is $\sqrt{1+\sigma^2}$ times larger than the change ΔU due to flipping an unknown variable. Hence known variables are also the most relevant ones.

[9] Therefore we shall think of the sample as being the solution of the maximization problem: $\underline{s}^{(i)} = \arg\max_{\underline{s}}[u_{\underline{s}} + \max_{\bar{s}} v^{(i)}_{\bar{s}|\underline{s}}]$ for $i = 1, \ldots, M$. For example, the choice of the city where Mr i decides to live, also depends on individual circumstances, captured by the function $v^{(i)}_{\bar{s}|\underline{s}}$. Note furthermore that the number of unknown variables is assumed to be the same for all points of the sample. This implies that the unknown parameter β in equation (5) is the same for all $i = 1, \ldots, M$.

J. Stat. Mech. (2013) P09003

Let $K_{\underline{s}}$ be the number of times $\underline{s}$ was observed in the sample, that is

$$K_{\underline{s}} = \sum_{i=1}^{M} \delta_{\underline{s}^{(i)},\underline{s}}. \tag{10}$$

In view of the discussion of section 2.1, the relation between the distribution $p_{\underline{s}}$ that our data is sampling and the function $u_{\underline{s}}$ is given by the Gibbs–Boltzmann form of equation (8). This has two consequences:

(1) Since the observed frequency $K_{\underline{s}}/M$ samples the unknown distribution $p_{\underline{s}} \sim \mathrm{e}^{\beta u_{\underline{s}}}$, it also provides a noisy estimate of the unknown function

$$u_{\underline{s}} \approx c + \frac{1}{\beta} \log K_{\underline{s}} \tag{11}$$

for some c and $\beta > 0$.

(2) Even without knowing what $u_{\underline{s}}$ is, we know that $p_{\underline{s}}$ is the maximal entropy distribution subject to an unknown constraint $E_{\underline{s}}[u] = \bar{u}$, or the distribution of maximal $E_{\underline{s}}[u] = \sum_{\underline{s}} p_{\underline{s}} u_{\underline{s}}$ with a given information content $H[\underline{s}] = \bar{H}$.

The first observation highlights the fact that the information that we can extract from the sample on the function the system performs is given by the information contained in $K_{\underline{s}}$ and *not* in $\underline{s}$ itself. In order to make this observation more precise in information theoretic terms, we remark that, *a priori* all of the M points i in the sample should be assigned the same probability $P\{i\} = 1/M$. With respect to this measure, the random variables $\underline{s}$ and $K_{\underline{s}}$ acquire distributions, respectively, given by $P\{\underline{s}^{(i)} = \underline{s}\} = K_{\underline{s}}/M$ and $P\{K_{\underline{s}^{(i)}} = k\} = km_k/M$ where

$$m_k = \sum_{\underline{s}} \delta_{k,K_{\underline{s}}} \tag{12}$$

is the number of states $\underline{s}$ that are sampled exactly k times. Therefore their associated entropies are:

$$\hat{H}[\underline{s}] = -\sum_{\underline{s}} \frac{K_{\underline{s}}}{M} \log \frac{K_{\underline{s}}}{M} = -\sum_{k} \frac{km_k}{M} \log \frac{k}{M} \tag{13}$$

$$\hat{H}[K] = -\sum_{k} \frac{km_k}{M} \log \frac{km_k}{M} = \hat{H}[\underline{s}] - \sum_{k} \frac{km_k}{M} \log m_k \tag{14}$$

where the notation $\hat{H}$ denotes empirical entropies. Since $K_{\underline{s}}$ is a noisy observation of the function $u_{\underline{s}}$, we conclude that the information that the data contains on the function $u_{\underline{s}}$ that the system optimizes is quantified by $\hat{H}[K]$. This conclusion is consistent with the fact that $\hat{H}[K]/\log 2$ is the (minimal) number of bits per state that is necessary to optimally encode the output of the experiment (see [12] chapter 5).

In order to gain intuition, it is instructive to consider the case of extreme under sampling, where each state is sampled at most once, i.e. $K_{\underline{s}} = 1$ for all states $\underline{s}$ in the sample and $K_{\underline{s}} = 0$ otherwise. This corresponds to considering the regime $\beta \approx 0$ in equation (8), where the data does not allow us to distinguish different observations in the sample and yields a uniform distribution on $\underline{s}$. At the other extreme, when the same state $\underline{s}_0$ is observed

J. Stat. Mech. (2013) P09003

M times, i.e. $K_{\underline{s}} = M\delta_{\underline{s},\underline{s}_0}$, the data samples the function $u_{\underline{s}}$ in just one point $\underline{s}_0$. In both cases the statistical range of the observed $K_{\underline{s}}$ does not allow us to learn much about the function $u_{\underline{s}}$ that is optimized. Notice that $\hat{H}[K] = 0$ in both these extreme cases, whereas $\hat{H}[\underline{s}] = \log M$ in the first case and $\hat{H}[\underline{s}] = 0$ in the latter. Our intuition that in both these extreme cases we do not learn anything about the behavior of the system is precisely quantified by the value of $\hat{H}[K]$.[10] For intermediate cases, $\hat{H}[\underline{s}]$ will take an intermediate value in $[0, \log M]$ and we expect that different distributions are possible, which might provide a positive amount of information $\hat{H}[K] > 0$ on the system's behavior. Notice that, $K_{\underline{s}} > K_{\underline{s}'}$ suggests that state $\underline{s}$ is optimal under broader conditions than $\underline{s}'$. But if $K_{\underline{s}} = K_{\underline{s}'}$ the sample does not allow one to distinguish the two states. In this sense, $\hat{H}[K]$ quantifies the number of states that the sample allows us to distinguish.

3.1. Most informative samples

Observation (2) above states that the distribution $p_{\underline{s}}$ can be seen as a distribution of maximal $E_{\underline{s}}[u] = \sum_{\underline{s}} p_{\underline{s}} u_{\underline{s}}$ with a given $H[\underline{s}] = \bar{H}$. The choice of which and how many variables to model, effectively fixes the number of unknown variables, which controls the inverse temperature parameter β in equation (8), and ultimately tunes the entropy $\bar{H}$ to different values between zero and $n \log 2$. Since[11] $\hat{H}[\underline{s}] \leq H[\underline{s}]$, we should look at empirical distributions with bounded $\hat{H}[\underline{s}] \leq \bar{H}$. Among these, those with maximal information content are those whose distribution $\mathbf{m} = \{m_k, k > 0\}$ is such that $\hat{H}[K]$ is maximal[12]:

$$\mathbf{m}^* = \arg \max_{\mathbf{m}:\hat{H}[\underline{s}]\leq\bar{H}} \hat{H}[K] \tag{15}$$

subject to the additional constraint $\sum_k k m_k = M$. The solution to this problem is made non-trivial by the fact that m_k should be a positive integer. Here we explore the solution within a very rough approximation where we consider m_k a positive real number. This

[10] To get an intuitive understanding of the information content of the two variables, imagine you want to find Mr X in a population of M individuals (this argument parallels the one in Ki Baek *et al* [11]). Without any knowledge, this requires $\log M$ bits of information. But if you know that Mr X lives in a city of size k, then your task is that of finding one out of $k \cdot m_k$ individuals, which requires $\log(km_k)$ bits. Averaging over the distribution of K, we find that the information gain is given by $\hat{H}[K]$. How informative is the size of the city? Clearly if all individuals live in the same city, e.g. $m_k = \delta_{k,M}$, then this information is not very useful. At the other extreme, if all cities are formed by a single individual, i.e. $m_k = M\delta_{k,1}$, then knowing the size of the city where Mr X lives is of no use either. In both cases $\log[km_k] = \log M$. Therefore there are distributions m_k of city sizes that are more informative than others. Notice that, in any case, the size k of the city cannot provide more information than knowing the city $\underline{s}$ itself, i.e. $\hat{H}[K] \leq \hat{H}[\underline{s}]$.

[11] This follows from the asymptotic equipartition property (AEP) [12] that derives from the law of large numbers and states that, when $M \gg 1$ is large

$$-\frac{1}{M} \log P\{\underline{s}^{(1)}, \ldots, \underline{s}^{(M)}\} = -\frac{1}{M} \sum_{i=1}^{M} \log p_{\underline{s}^{(i)}} \simeq H[\underline{s}].$$

Using $P\{\underline{s}^{(1)}, \ldots, \underline{s}^{(M)}\} = p_{\underline{s}^{(1)}} \cdots p_{\underline{s}^{(M)}}$, this leads to

$$\hat{H}[\underline{s}] + D_{\mathrm{KL}}(\hat{p} \parallel p) \simeq H[\underline{s}],$$

where $\hat{p}_{\underline{s}} = K_{\underline{s}}/M$ and $D_{\mathrm{KL}}(\hat{p} \parallel p) = \sum_{\underline{s}} \hat{p}_{\underline{s}} \log(p_{\underline{s}}/\hat{p}_{\underline{s}})$ is the Kullback–Leibler divergence. Note that $\hat{H}[\underline{s}] \leq \log M$, so if M is not large enough $\hat{H}[\underline{s}]$ is not a good estimate of $H[\underline{s}]$. Since $D_{\mathrm{KL}}(\hat{p} \parallel p) \geq 0$, then $\hat{H}[\underline{s}] \leq H[\underline{s}]$.

[12] A similar argument can be found in Baek *et al* [11], though the analysis and conclusions presented here differ substantially from those [11].

J. Stat. Mech. (2013) P09003

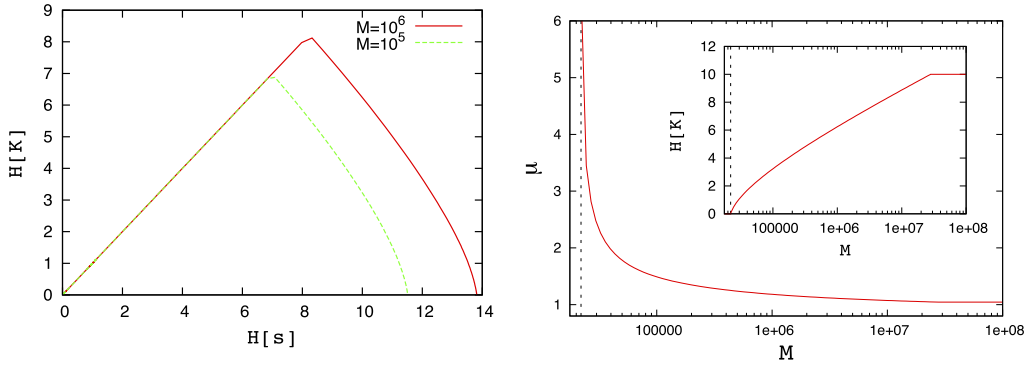

Figure 2. (Left) Maximal entropy $\hat{H}[K]$ plotted as a function of the system entropy $\hat{H}[\underline{s}]$ for $M = 10^5$ and 10^6. The under sampled regime corresponds to the right region, while the left region for which $\hat{H}[K] \approx \hat{H}[s]$ represents the regime in which the distribution $p_{\underline{s}}$ is well sampled. The peak separating the two regimes is associated with a Zipf distribution for m_k. (Right) Exponent μ as a function of M within the approximated solution presented in the text, with $H[\underline{s}] = 10$. In the inset we represent $\hat{H}[K]$ as a function of M. The vertical dashed line corresponds to $\log M = H[\underline{s}]$.

provides an upper bound to the entropy $\hat{H}[K]$ that we combine with the upper bound $\hat{H}[K] \leq \hat{H}[\underline{s}]$ implied by the data processing inequality [12], which arises from the fact that the random variable $K_{\underline{s}}$ is a function of $\underline{s}$.

In the region where $\hat{H}[K] < \hat{H}[\underline{s}]$, the solution to the approximated problem is readily found by maximizing

$$\hat{H}[K] + \mu \hat{H}[\underline{s}] + \lambda \sum_{k>1} k m_k \tag{16}$$

over $m_k \in \mathbb{R}^+$, where μ and λ are Lagrange multipliers that are used to enforce the constraints $\hat{H}[\underline{s}] = \bar{H}$ and $\sum_{k=1}^{M} k m_k = M$. The solution reads:

$$m_k^* = c k^{-1-\mu}, \qquad 1 \leq k \leq M \tag{17}$$

where $c > 0$ is a constant that is adjusted in order to enforce normalization. As μ varies, the upper bound draws a curve in the $\hat{H}[K]$ versus $\hat{H}[\underline{s}]$ plane, as shown in figure 2 (left) for two values of M. In particular, the slope of the curve is exactly given by $-\mu$. Therefore we see that at the extreme right, $\hat{H}[K] \to 0$ as $\hat{H}[\underline{s}] \to \log M$ with infinite slope $\mu \to \infty$, corresponding to a distribution $m_k = M\delta_{k,1}$. As μ decreases, the distribution m_k spreads out and $\hat{H}[K]$ increases accordingly.

There is a special point where the upper bound $\hat{H}[K]$ derived from the solution with $m_k \in \mathbb{R}$ matches the data processing inequality line $\hat{H}[\underline{s}] = \hat{H}[K]$. We find that the slope of the line at this point (see figure 2) approaches $\mu = 1$ from above, which corresponds to a distribution $m_k \sim k^{-2}$.

In the regime where $\hat{H}[K] < \hat{H}[\underline{s}]$, the true distribution $p_{\underline{s}}$ is undersampled and a number of states $\underline{s}$ are all sampled an equal number of times. When $\hat{H}[K] = \hat{H}[\underline{s}]$, instead, almost all states are sampled a different number of times. Therefore knowing the frequency $K_{\underline{s}}/M$ of a state is equivalent to knowing the state $\underline{s}$ itself. Notice that, in this regime, m_k

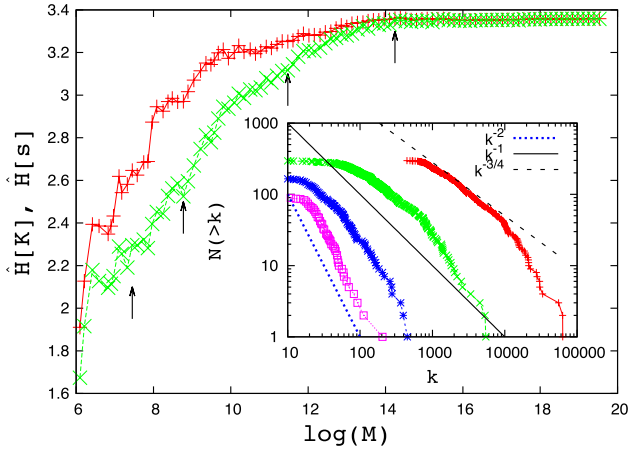

Figure 3. Distribution in cities for subsamples of M households of the IPUM database (http://usa.ipums.org). Main figure: $\hat{H}[\underline{s}]$ and $\hat{H}[K]$ as a function of M. Inset: cumulative distribution $N\ (>k) = \sum_{q>k} m_q$ of city distribution for subsamples of $M = 1721$, 6452, 96 118 and 1535 956 (from left to right, corresponding to the arrows in the main figure.

is *not* given by the solution of the above optimization problem, since $\hat{H}[K]$ is bound by the data processing inequality. Indeed, in this regime, the empirical distribution converges to whatever the underlying distribution is[13], with $m_k = 0$ or 1 for almost all the values of k.

These results provide a picture of how most informative samples behave as the sample size M increases, and the curve in the left part of figure 2 moves upward (see figure 2 right). As long as $\log M$ is smaller than the entropy $\bar{H}$ of the unknown distribution, we expect that all states in the sample will occur at most once, i.e. $\hat{H}[K] = 0$. When $M \approx \mathrm{e}^{\bar{H}}$, we start sampling states more than once. Beyond this point, $\hat{H}[K]$ will increase and $m_k \sim k^{-1-\mu}$ will take a power law form, with an exponent that decreases with M (see figure 2 right). When M is large enough the entropy $\hat{H}[K]$ will saturate to the value $\bar{H}$ of the underlying distribution and μ will draw closer to one. Further sampling will provide closer and closer approximations of the true distribution $p_{\underline{s}}$ (see figure 3).

The above argument suggests that power law distributions are the frequency distributions with the largest information content *in the undersampled regime* (i.e. to the right of the cusp in figure 2 left). The value of the exponent μ can be read from the slope of the curve. The maximum, which corresponds to a cusp, has $\mu \simeq 1$, hence a distribution that is close to the celebrated Zipf's law $m_k \sim k^{-2}$. Actually, the plot of μ versus M in figure 2 suggests that there is a broad range of M over which μ takes values very close to one.

3.2. Criticality and Zipf's law

The results above suggest that Zipf's law ($\mu = 1$) emerges as the most informative distribution which is compatible with a fixed value of the entropy $H[\underline{s}]$. Here we want to

[13] There is an interesting duality between the distribution of $\underline{s}$ and that of K: when the former is under sampled (e.g. all states are seen only a few times) the distribution m_k is well sampled (i.e. $m_k \propto M$), whereas when $\underline{s}$ is well sampled, m_k is under sampled, i.e. $m_k = 0$ or 1.

show how this is consistent with the approach in [4]. Mora and Bialek [4] draw a precise relation between the occurrence of Zipf's law and criticality in statistical mechanics. In brief, given a sample and an empirical distribution $\hat{p}_{\underline{s}} = K_{\underline{s}}/M$, it is always possible to define an energy function $E_{\underline{s}} = -\log \hat{p}_{\underline{s}}$ and a corresponding entropy, $\Sigma(E)$ through the usual relation $\mathrm{e}^{\Sigma(E)} = \mathrm{d}\mathcal{N}(E)/\mathrm{d}E$ with the number $\mathrm{d}\mathcal{N}(E)$ of energy states between energy E and $E + \mathrm{d}E$. For $E = -\log(k/M)$, $\mathrm{d}\mathcal{N}(E) = m_k\,|\mathrm{d}k/\mathrm{d}E| = km_k$. Therefore, $\Sigma(E) = \log(km_k)$, which means that Zipf's law $m_k \sim k^{-2}$ corresponds to linear relation $\Sigma(E) \simeq \Sigma_0 + \beta E$ with slope $\beta = 1$. The relation with criticality in statistical mechanics arises because the vanishing curvature in $\Sigma(E)$ corresponds to an infinite specific heat [4].

The linearity of the $\Sigma(E)$ relation is not surprising. Indeed, the range of variation of entropy and energy in a sample of M points is limited by $\delta\Sigma, \delta E \leq \log M$. For intensive quantities $\sigma = \Sigma/n$ and $\epsilon = E/n$, this corresponds to a linear approximation of the $\sigma(\epsilon) \simeq \sigma_0 + \beta\epsilon$ relation over an interval $\delta\sigma, \delta\epsilon \sim (\log M)/n$ that can be relatively small. The fact that the coefficient takes the particular value $\beta \approx 1$ is, instead, non-trivial and it corresponds to the situation where the entropy versus energy relation enjoys a wider range of variation.

The results of section 3.1[14] provide an alternative perspective on the origin of Zipf's law: imagine a situation where we can choose the variables $\underline{s}$ with which to probe the system. Each choice corresponds to a different function $u_{\underline{s}}$ or to a different $\sigma(\epsilon)$ relation, of which the sample probes a small neighborhood of size $(\log M)/n$. For each choice of $\underline{s}$, this relation will likely look linear $\sigma(\epsilon) \simeq \sigma_0 + \beta\epsilon$ with a different coefficient β. How should one choose the variables $\underline{s}$? It is clear that probing the system along variables for which $\beta \ll 1$ results in a very noisy dataset, whereas if $\beta \gg 1$ one would be measuring constants. In contrast, probing the system on 'critical' variables, i.e. those for which $\beta \approx 1$, provides more information on the system's behavior. Zipf's law, in this perspective, is a consequence of choosing the known variables as those that reveal a wider range of variability in the $\sigma(\epsilon)$ relation.

4. Applications

Are the findings above of any use?

As we have seen, the distribution m_k conveys information on the internal self-organization of the system. In the case of city size distribution, the occurrence of a broad distribution suggests that the city $\underline{s}$ is a relevant variable that enters in the optimization problem that individuals solve. Indeed, individuals could be clustered according to different criteria (electoral districts, population living in areas of equal size, etc) and we do not expect broad distributions in general. Furthermore, we expect that if we progressively sample a population of individuals, the resulting city size distribution would 'evolve' approximately as described above. Figure 3 shows the result of such an exercise for a data set of US citizens (see caption). Interestingly, we find that for small samples the distribution takes a power law form $m_k \sim k^{-\mu-1}$ with exponent $\mu > 1$, and as M increases

[14] We remark an interesting formal analogy between the picture above and the statistical mechanics analogy of [4], within the simplified picture provided by our approximation. Upon defining $Z_\mu = \sum_k k^{-\mu}$, it is easy to check that $\hat{H}[\underline{s}] = \log M + \partial_\mu \log Z_\mu$ and $\hat{H}[K] = \log Z_\mu - \mu\partial_\mu \log Z_\mu$. Thus, identifying Z_μ with a partition function, $\hat{H}[\underline{s}]$ and $\hat{H}[K]$ stand precisely in the same relation as the energy and the entropy of a statistical mechanical system.

J. Stat. Mech. (2013) P09003

the distribution gets broader (i.e. μ decreases) and converges to the city size distribution, when only 0.5% of the individuals are sampled[15].

In most applications the relevant variables are not known. In this case, the maximization of $\hat{H}[K]$ can be used as a guiding principle to select the most appropriate variables or to extract them from the data. We illustrate the problem with three examples.

4.1. Protein sequences

A protein is defined in terms of its amino-acid sequence[16] $\vec{s}$ but its functional role in the cell, as well as its 3d structure, is not easily related to it. The sequences $\vec{s}$ of homologous proteins—i.e. those that perform the same function—can be retrieved from public databases [13]. Mutations across sequences of homologous proteins are such that they preserve that function but otherwise might be optimized in order to cope with their particular cellular environment. This suggests that there may be relevant amino acids $\underline{s}$ that are optimized for preserving the function and less relevant ones.

How to find relevant variables? One natural idea is to look at the subsequence of the n evolutionarily most conserved amino acids[17]. Figure 4 shows the information content $\hat{H}[K]$ as a function of $\hat{H}[\underline{s}]$ as the number n of 'relevant' amino acids varies for the family PF000072 of response regulator receiver proteins[18] [13]. For n large, most of the sequences are seen only once (small $\hat{H}[K]$), and $\hat{H}[\underline{s}] \propto \log M$, whereas for $n < 25$ the entropy $\hat{H}[\underline{s}]$ decreases steeply as n decreases. Correspondingly, $\hat{H}[K]$ exhibits a maximum at $n = n_c = 22$ and then approaches $\hat{H}[\underline{s}]$.

Even if the empirical curve does not saturate the theoretical bound, the frequency distribution exhibits Zipf's law around the point n_c where $\hat{H}[K]$ is maximal. Figure 5 shows that for $n \approx n_c$ the number m_k of sequences that are sampled k times falls off as $m_k \sim k^{-2}$, characteristic of a Zipf's law, whereas for $n \approx N$ it falls off faster and for $n \sim O(1)$ it is dominated by one large value of $k \approx M$.

Alternatively, one may use the maximization of $\hat{H}[K]$ as a guide for identifying the relevant variables. We do this by an agglomerative algorithm, where we start from a sequence $\underline{s}$ of length zero and iteratively build subsequences of an increasing number n of sites. At each step, we add the site i that makes the information content $\hat{H}[K]$ of the resulting subsequence as large as possible[19]. The result, displayed in figure 4, shows that this procedure yields subsequences with a higher $\hat{H}[K]$, which are also shorter. In particular, the maximal $\hat{H}[K]$ is achieved for subsequences of just three amino acids.

[15] Cristelli *et al* [15] have shown that Zipf's law does not hold if one restricts the statistics to a subset of cities which is different from the set over which self-organization takes place. This points to a notion of *coherence* of the sample, which is consistent with our framework, where the sample is thought of being the outcome of an optimization problem. Note that our subsampling differs from the one in [15], as we are sampling individuals rather than cities.

[16] Each s_i takes 21 values rather than 2, but that is clearly an non-consequential difference with respect to the case where $s_i = \pm 1$.

[17] For any given subset $\underline{s}$ of the $\vec{s}$ variables, the frequency $\hat{p}_{\underline{s}}$ can computed and, from this the entropies $\hat{H}[\underline{s}]$ and $\hat{H}[K]$. As a measure of conservation, we take the entropy of the empirical distribution of amino acids in position i.

[18] Our analysis is based on $M = 62\,074$ sequences, that after alignment, are $N = 112$ amino acids long. The same data was used in [14].

[19] Notice that the algorithm is not guaranteed to return the subset of sites that maximizes $\hat{H}[K]$ for a given $n > 1$.

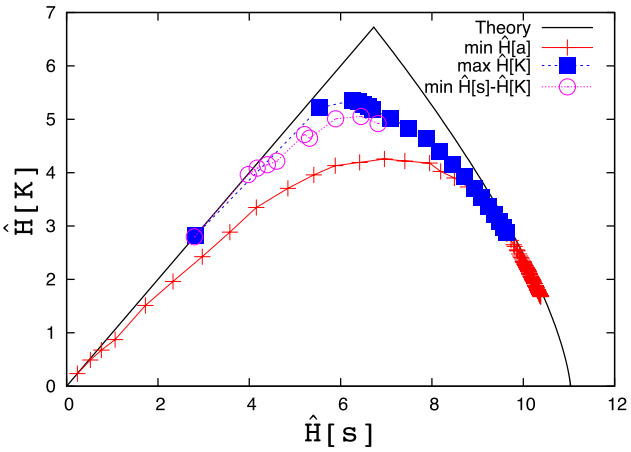

Figure 4. Entropy $\hat{H}[K]$ as a function of $\hat{H}[\underline{s}]$ for the protein family PF000072. Subsequence of the n most conserved positions (red +); subsequences of n positions with maximal $\hat{H}[K]$ (blue ■) and with minimal $\hat{H}[\underline{s}] - \hat{H}[K]$ (pink ○). n increases from left to right in all cases.

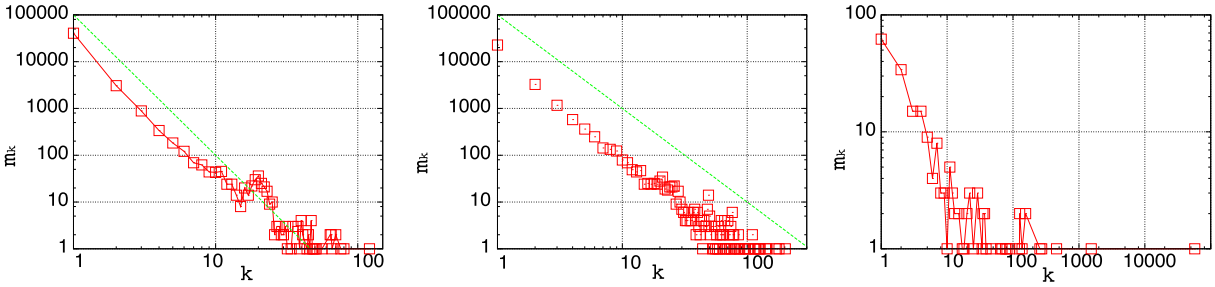

Figure 5. Frequency distribution m_k for $n = N = 112$ (left), $n = 22 \approx n_c$ (center) and $n = 2$ (right). Lines are proportional to k^{-3} (left) and k^{-2} (center).

Interestingly, if one looks at the subsequence of sites that are identified by this algorithm one finds that the first two sites of the subsequence are among the least conserved ones: they are those that allow one to explain the variability in the dataset in the most compact manner—loosely speaking, they are 'high temperature' variables ($\beta \ll 1$). The following ten sites identified by the algorithm are instead 'low temperature' variables, as they are the most conserved ones. This hints at the fact that relevant variables should not only encode a notion of optimality, but also account for the variability within the data set, under which the system is (presumably) optimizing its behavior.

4.2. Clustering and correlations of financial returns

In many problems data is noisy and high dimensional. It may consist of M observations $\hat{x} = (\vec{x}^{(1)}, \ldots, \vec{x}^{(M)})$ of a vector of features $\vec{x} \in \mathbb{R}^{\mathrm{T}}$ of the system under study. Components of $\vec{x}$ may be continuous variables, so the analysis of previous sections is not applicable. In these cases a compressed representation $\underline{s}^{(i)}$ of each point $\vec{x}^{(i)}$ would be desirable, where $\underline{s}$ takes a finite number of values and can be thought of as encoding a relevant description of the system. There are several ways to derive a mapping $\underline{s} = F(\vec{x})$, such as

J. Stat. Mech. (2013) P09003

quantization [12] or data clustering. The general idea is that of discretizing the space of $\vec{x}$ in cells, each labeled by a different value of $\underline{s}$, so 'similar' points $\vec{x}^{(i)} \approx \vec{x}^{(j)}$ fall in the same cell, i.e. $\underline{s}^{(i)} = \underline{s}^{(j)}$. The whole art of data clustering resides in what 'similar' exactly means, i.e. on the choice of a metrics in the space of $\vec{x}$. Different data clustering algorithms differ on the choice of the metrics, as well as on the choice of the algorithm which is used to group similar objects in the same cluster, and on the resolution, i.e. on the number of clusters. Correspondingly, different clustering algorithms extract a different amount of information on the internal structure of the system. In practice, how well the resulting cluster structure reflects the internal organization of the data depends on the specific problem, but there is no unambiguous way, to the best of our knowledge, to compare different methods.

The point we want to make here is that the discussion of section 4.1 allows us to suggest a universal method to compare different data clustering algorithms and to identify the one that extracts the most informative classification. The idea is simple: for any algorithm A, compute the variables $K^A_{\underline{s}}$ and the corresponding entropies $\hat{H}[\underline{s}^A]$ and $\hat{H}[K^A]$ and plot the latter with respect to the former, as the number n of clusters varies from 1 to M. If such curve for algorithm A lies above the corresponding curve for algorithm B, we conclude that A extracts more information on the systems behavior and hence it is to be preferred to B.

This idea is illustrated by the study of financial correlations of a set of $M = 4000$ stocks in the NYSE in what follows[20]. Financial markets perform many functions, such as channeling private investment to the economy, allowing inter-temporal wealth transfer and risk management. Time series of the price dynamics carry a signature about such complex interactions, and have been studied intensively [16]–[18]: the principal component in the singular value decomposition largely reflects portfolio optimization strategies whereas the rest of the correlations exhibit a structure which is highly correlated with the structure of economic sectors, down to a scale of 5 min [18]. Since we are borrowing this example to make a generic point, we shall not enter into further details, and refer the interested reader to [16]–[18]. Several authors have applied single linkage data clustering method to this problem [16], which consists in building minimal spanning trees (MST) where the links between the most correlated stocks, which do not close loops, are iteratively added to a forest. Clusters are identified by the disconnected trees that, as links are sequentially added, merge one with the other until a single cluster remains. The resulting curve $\hat{H}[K]$ versus $\hat{H}[\underline{s}]$ is shown in figure 6.

A different data clustering scheme has been proposed in [19, 18] based on a parametric model of correlated random walks for stock prices. The method is based on maximizing the likelihood with a hierarchical agglomerative scheme [19]. The curve $\hat{H}[K]$ versus $\hat{H}[\underline{s}]$ lies clearly above the one for the MST (see figure 6). Reference [18] has shown that the structure of correlation is revealed more clearly if the principal component dynamics is subtracted from the data[21]. This is reflected by the fact that the resulting curve $\hat{H}[K]$ versus $\hat{H}[\underline{s}]$ shifts further upward. In the present case, it is possible to compare these results with the classification given by the US Security and Exchange Commission (SEC),

[20] Here $\vec{x}^{(i)} = (x_1^{(i)}, \ldots, x_T^{(i)})$ consists of daily log returns $x_t^{(i)} = \log(p_t^{(i)}/p_{t-1}^{(i)})$, where $p_t^{(i)}$ is the price of stock i on day t, and t runs from 1 January 1990 to 30 April 1999.

[21] If x_t^0 is the principal component in the singular value decomposition of the data set, this amount to repeating the analysis for the modified dataset $\tilde{x}_t^{(i)} = x_t^{(i)} - x_t^0$.

J. Stat. Mech. (2013) P09003

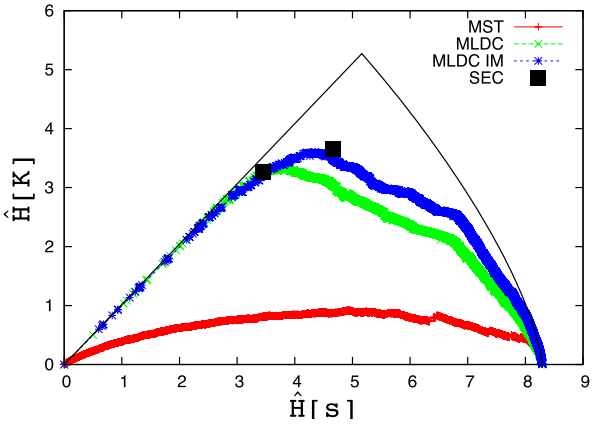

Figure 6. Entropy $\hat{H}[K]$ as a function of $\hat{H}[\underline{s}]$ as the number n of clusters increases (from left to right), for different data clustering schemes. From bottom to top, single linkage (MST), maximum likelihood with (MLDC) and without (MLDC IM) the principal component. The SEC classification at two and three digits of the stocks is also shown as black squares.

which is given by the black squares in figure 6 for two and three digits SEC codes. This classification codifies the information on the basis of which agents trade, so it enters into the dynamics of the market. The curve obtained removing the principal component draws remarkably close to these points, suggesting that the clustering method extracts a large fraction of the information on the internal organization of the market. Again, the rank plot of cluster sizes reveals that Zipf's law occurs where $\hat{H}[K]$ is close to its maximum, whereas marked deviations are observed as one moves away from it.

4.3. Keywords in a text

A written text can be thought of as the result of a design, by the writer: there are tens of thousands of words in the vocabulary of a given language, but in practice the choice is highly constrained by syntax and semantics, as revealed by the fact that the frequency distribution in a typical text is highly peaked on relatively few words, and it roughly follows Zipf's law.

The frequency with which a given word w occurs in a given section $\underline{s}$ of a manuscript should contain traces of the underlying optimization problem. This insight has been exploited by Montemurro and Zanette [20] in order to extract keywords from a text. The idea in [20] is: (i) split the text into parts $\underline{s}$ of L consecutive words; (ii) compute the fraction $\hat{p}_{\underline{s}}^{(w)}$ of times word w appears in part $\underline{s}$; (iii) compute the difference $\Delta H[\underline{s}]$ between the entropy $\hat{H}[\underline{s}]$ of a random reshuffling of the words in the parts and the actual word frequency. Keywords are identified with the least random words, those with the largest $\Delta H[\underline{s}]$.

From our perspective, for each choice of L and each word w, one can compute $\hat{H}^w[K]$ and $\hat{H}^w[\underline{s}]$. Figure 7 shows the resulting curve as L varies for Darwin's '*On the Origin of Species*'. Among all words that occur at least 100 times, we select those that achieve a maximal value of $\hat{H}[K]$ as well as some of those whose maximal value of $\hat{H}[K]$ (on L) is

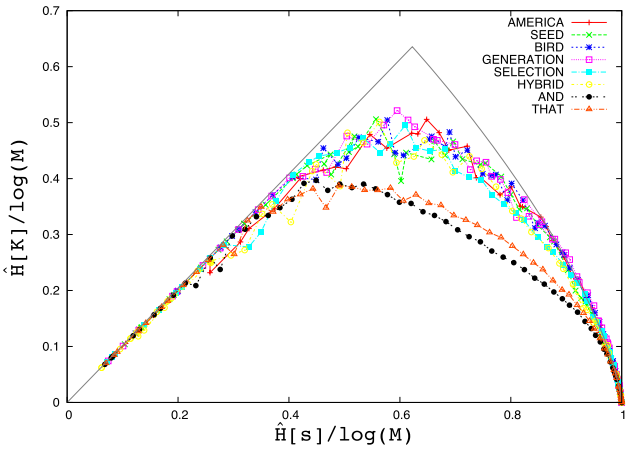

Figure 7. Entropy $\hat{H}[K]$ as a function of $\hat{H}[\underline{s}]$ for the occurrence of different words (see legend) of Darwin's '*On the Origin of Species*' in segments of L consecutive words (L increasing from right to left).

the smallest. The latter turn out to be generic words ('and', 'that') whereas among the former we find words (e.g. 'generation', 'seed', 'bird') that are very specific of the subject discussed in the book. Whether this observation can be used to derive a more efficient extractor of keywords than the one suggested in [20] or not, is a question that we leave for future investigations. For our present purposes, we merely observe that $\hat{H}[K]$ allows us to distinguish words that are 'mechanically' chosen from those that occur as a result of a more complex optimization problem (the keywords).

5. Discussion

Advances in IT and experimental techniques have boosted our ability to probe complex systems to unprecedented levels of detail. Increased performance in computing, at the same time, has paved the way for reproducing *in silico* the behavior of complex systems, such as cells [21], the brain [22] or the economy [23].

However, it is not clear whether this approach will ultimately deliver predictive models of complex systems. Interestingly [24] observes that efforts in Artificial Intelligence to reproduce *ab initio* human capabilities in intelligent tasks have completely failed: search engines, recommendation systems and automatic translation [24] have been achieved by unsupervised statistical learning approaches that harvest massive data sets, abandoning altogether the ambition to understand the system or to model it in detail. At the same time, problems such as drug design [25] and the regulation of financial markets [26] still remain elusive, in spite of the increased sophistication of techniques deployed.

This calls for understanding the limits of modeling complex systems and devising ways to select relevant variables and compact representations. The present contribution is an attempt to address these concerns. In doing that, we uncover a non-trivial relation between 'criticality', which in this context is used to refer to the occurrence of broad distributions in the frequency of observations (Zipf's law), and the relevance of the measured variables. We make this relation precise by quantifying the information content of a sample: most informative data, which sample relevant variables, exhibit power law

J. Stat. Mech. (2013) P09003

frequency distributions, in the undersampling regime. Conversely, a description in terms of variables which are not the ones the system cares about will not convey much information. Mostly, informative data sets are those for which the frequency of observations covers the largest possible dynamic range, providing information on the system's optimal behavior in the wider range of possible circumstances. This corresponds to a linear entropy–energy relation, in the statistical mechanics analogy discussed in [4].

Our results point in the same direction as the recent finding that inference of high dimensional models is likely to return models that are poised close to 'critical' points [28]. This builds on the observation [27] that the mapping between the parameter space of a model and the space of distributions can be highly nonlinear. In particular, it has been shown in simple models [28] that regions of parameter space of models that have a vanishing measure (critical points) concentrate a finite fraction of the possible (distinguishable) empirical distributions. This suggests that 'optimally informative experiments' that sample uniformly the space of empirical distributions are likely to return samples that look 'close to a critical point' when we see them through the eyes of a given parametric model.

Our findings are also consistent with the observation [15] that Zipf's law entails some notion of 'coherence of the sample' in the sense that typical subsamples deviate from it. In our setting, the characteristic that makes the sample homogeneous is that it refers to systems 'doing the same thing' under 'different conditions'.

As shown in section 4, the ideas in this paper can be turned into a criterion for selecting mostly informative representations of complex systems. This, we believe, is the most exciting direction for future research. One particular direction in which our approach could be useful is that of the identification of hidden variables, or *unknown unknowns*. In particular, the identification of relevant classification of the data can be turned into the specification of hidden variables, whose interaction with the observed ones can be inferred. This approach would not only predict how many hidden variables one should consider, but also how they specifically affect the system under study. Progress along these lines will be reported in future publications.

Acknowledgments

We gratefully acknowledge William Bialek, Andrea De Martino, Silvio Franz, Thierry Mora, Igor Prunster, Miguel Virasoro and Damien Zanette for various inspiring discussions, which we have taken advantage of.

Appendix. When are models predictive? The Gaussian case

In this appendix, we consider the setup of section 2 in the case of a Gaussian distribution of $v_{\underline{\bar{s}}|\underline{s}}$, for which $\beta = \sqrt{2N(1-f)\log 2}$. Here, and in the rest of the appendix, $f = n/N$ is the fraction of known variables, and we shall focus on the asymptotic behavior in the limit $n, N \to \infty$ with $f = n/N$ finite.

We assume that the dependence of the objective function $u_{\underline{s}}$ on known variables $\underline{s} = (s_1, \ldots, s_n)$ is known and we concentrate on the specific example where $u_{\underline{s}}$ are also i.i.d. draws from a Gaussian distribution with zero mean and variance σ^2. This is the most complex system one could think of, as its specification requires an exponential number

of parameters. As argued in section 2.1, this is also a particular case where the subset of known variables coincides with the subset of the most relevant ones. The question we address is: does the knowledge of the function $u_{\underline{s}}$ allow us to predict the optimal behavior $\underline{s}^*$?

As a prototype example, consider the problem of reverse engineering the choice behavior of an individual that is optimizing an utility function $U(\vec{s})$. For a consumer, $\vec{s}$ can be thought of as a consumption profile, specifying whether the individual has bought good i ($s_i = +1$) or not ($s_i = -1$) for $i = 1, \ldots, N$. However, consumer behavior can be observed only over a subset $\underline{s} = (s_1, \ldots, s_n)$ of the variables, and only the part $u_{\underline{s}}$ of the utility function that depends solely on the observed variables can be modeled[22]. Under what conditions the predicted choice $\underline{s}_0$ is informative on the actual behavior $\underline{s}^*$ of the agent? Put differently, how relevant and how many (or few) should the relevant variables be in order for $\underline{s}_0$ to be informative on the optimal choice $\underline{s}^*$?

In light of the result of section 2.1, the answer depends on how peaked is the distribution $p_{\underline{s}}$. For $\beta \to \infty$ the probability distribution concentrates on the choice $\underline{s}_0$ that maximizes $u_{\underline{s}}$, whereas for $\beta \to 0$ it spreads uniformly over all 2^n possible choices $\underline{s}$. Our problem, in the present setup, reverts to the well known REM, which is discussed in detail, e.g. in [5, 10]. We recall here the main steps.

The entropy of the distribution $p_{\underline{s}}$ is given by:

$$H[\underline{s}] = -\sum_{\underline{s}} p_{\underline{s}} \log p_{\underline{s}} = \log Z(\beta) - \beta \frac{\mathrm{d}}{\mathrm{d}\beta} \log Z(\beta), \qquad Z(\beta) = \sum_{\underline{s}} \mathrm{e}^{-\beta u_{\underline{s}}} \tag{A.1}$$

where the above equality is easily derived by a direct calculation.

In order to estimate $Z(\beta)$ let us observe that $2^{-n} Z(\beta)$ is an average and the law of large numbers suggests that it should be close to the expected value of $\mathrm{e}^{\beta u_{\underline{s}}}$

$$\frac{1}{2^n} Z(\beta) \simeq E\left[\mathrm{e}^{\beta u_{\underline{s}}}\right] = \mathrm{e}^{\beta^2 \sigma^2 / 2} \equiv \frac{1}{2^n} Z_{\mathrm{ann}}(\beta) \tag{A.2}$$

which depends on the fact that $u_{\underline{s}}$ is a Gaussian variable with zero mean and variance σ^2. Therefore, if we use Z_{ann} instead of Z in equation (A.1), we find

$$H[\underline{s}] \simeq n \log 2 - \frac{\beta^2 \sigma^2}{2} = N\left[f - (1-f)\sigma^2\right] \log 2. \tag{A.3}$$

One worrying aspect of this result is that if

$$\sigma \geq \sigma_{\mathrm{c}} = \sqrt{\frac{f}{1-f}} \tag{A.4}$$

the entropy is negative. The problem lies in the fact that the law of large number does not hold for $\sigma \geq \sigma_{\mathrm{c}}$ due to the explicit dependence of β on N, in the limit $N \to \infty$. In order to see this, notice that the expected value of $u_{\underline{s}}$ over $p_{\underline{s}}$ is given by

$$u_{\underline{s}^*}^{(\mathrm{ann})} = \sum_{\underline{s}} p_{\underline{s}} u_{\underline{s}} = \frac{\mathrm{d}}{\mathrm{d}\beta} \log Z \simeq \beta \sigma^2 = \sigma^2 \sqrt{2N(1-f)\log 2} \tag{A.5}$$

[22] This setup is the one typically considered in random utility models of choice theory in economics [8].

where the second relation holds when the law of large numbers holds. However, this cannot be larger than the maximum of $u_{\underline{s}}$, which, by extreme value theory of Gaussian variables, is given by

$$u_{\underline{s}_0} = \max_{\underline{s}} u_{\underline{s}} \simeq \sigma\sqrt{2Nf\log 2}. \tag{A.6}$$

Indeed the estimate in equation (A.5) gets larger than the maximum given in equation (A.6) precisely when $\sigma \geq \sigma_c$, i.e. when $H[\underline{s}]$ becomes negative. It can be shown that the law of large numbers, and hence the approximation used above, holds only for $\sigma < \sigma_c$ [5, 10]. The basic intuition is that for $\sigma < \sigma_c$ the sum in Z is dominated by exponentially many terms (indeed $e^{H[\underline{s}]}$ terms) whereas for $\sigma \geq \sigma_c$ the sum is dominated by the few terms with $u_{\underline{s}} \simeq \max u_{\underline{s}}$.

For $\sigma < \sigma_c$ we can use equations (A.2) and (A.6) to compute

$$p_{\underline{s}_0} = P\{\underline{s}_0 = \underline{s}^*\} \simeq e^{-N(1-f)(\sigma-\sigma_c)^2}, \qquad \sigma^2 < \sigma_c^2, \tag{A.7}$$

which is exponentially small in N. Therefore the model prediction $\underline{s}_0$ carries no information on the systems' behavior $\underline{s}^*$ for $\sigma < \sigma_c$.

On the other hand, for $\sigma > \sigma_c$, $Z(\beta)$ is dominated by $u_{\underline{s}_0}$ and it can be estimated by expanding the number $\mathcal{N}(u) = 2^n e^{-u^2/(2\sigma^2)}/\sqrt{2\pi\sigma^2}$ of choices $\underline{s}$ with $u_{\underline{s}} = u$ around $u_{\underline{s}_0}$. Simple algebra and asymptotic analysis reveals that

$$p_{\underline{s}_0} \simeq 1 - \frac{\sigma_c}{2\sqrt{\pi f\log 2}(\sigma-\sigma_c)+\sigma_c} + O(N^{-1}). \tag{A.8}$$

In words, the transition from the region $p_{\underline{s}_0} \simeq 0$ to the region where $p_{\underline{s}_0} \simeq 1$ is rather sharp, and it takes place in a region of order $|\sigma - \sigma_c| \sim 1/\sqrt{N}$.

The most remarkable aspect of this solution is that σ_c increases with f: for a given value of σ the correct solution $\underline{s}^*$ is recovered only if the fraction of known variables is *less* than a critical value

$$f_c = \sigma^2/(1+\sigma^2). \tag{A.9}$$

This feature is ultimately related to the fact that the effect of unknown unknowns is a decreasing function of the number $N(1-f)$ of them (see equation (5)). This, in turn, is a consequence of the Gaussian nature of the variables $v_{\underline{s}|\underline{s}}$ or in general of the fact that the distribution of u and v falls off faster than exponential.

References

[1] Wigner E P, *The unreasonable effectiveness of mathematics in the natural sciences*, 1960 *Comm. Pure Appl. Math.* **13** 1

[2] Newman M E J, *Power laws, Pareto distributions and Zipfs law*, 2005 *Contemp. Phys.* **46** 323351
Clauset A, Shalizi C R and Newman M E J, *Power-law distributions in empirical data*, 2009 *SIAM Rev.* **51** 661703

[3] Bak P, 1996 *How Nature Works* (New York: Springer)

[4] Mora T and Bialek W, 2011 *J. Stat. Phys.* **144** 268

[5] Cook J and Derrida B, 1991 *J. Stat. Phys.* **63** 5050

[6] Jaynes E T, 1957 *Phys. Rev.* II **106** 620630

[7] Galambos J, 1978 *The Asymptotic Theory of Extreme Order Statistics* (New York: Wiley)

[8] McFadden D and Zarembka in P (ed), 1974 *Frontiers in Econometrics* (New York: Academic) pp 105–42

[9] For a recent review of different derivation of probabilistic choice models, see Bouchaud J P, 2012 arXiv:1209.0453

J. Stat. Mech. (2013) P09003

[10] Mezard M and Montanari A, 2009 *Information, Physics and Computation* (Oxford: Oxford University Press)
[11] Ki Baek S *et al,* 2011 *New J. Phys.* **13** 043004
[12] Cover T M and Thomas J A, 1991 *Elements of Information Theory* (New York: Wiley)
[13] www.sanger.ac.uk/resources/databases/pfam.html
[14] Lunt B H *et al,* 2010 *Meth. Enzymol* **471** 17
[15] Cristelli M, Batty M and Pietronero L, *There is more than a power law in Zipf*, 2012 *Nature Sci. Rep.* **2** 812
[16] Onnela J P, Chakraborti A, Kaski K, Kertesz J and Kanto A, *Dynamics of market correlations: taxonomy and portfolio analysis*, 2003 *Phys. Rev.* E **68** 056110
[17] Potters M, Bouchaud J P and Laloux L, *Financial applications of random matrix theory: old laces and new pieces*, 2005 *Acta Phys. Pol.* B **36** 2767
[18] Borghesi C, Marsili M and Miccichè S, *Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode*, 2007 *Phys. Rev.* E **76** 026104
[19] Giada L and Marsili M, *Algorithms of maximum likelihood data clustering with applications*, 2002 *Physica* A **315** 650664
[20] Montemurro M A and Zanette D H, *Entropic analysis of the role of words in literary texts*, 2002 *Adv. Complex Syst.* **5** 7
[21] Karr J R *et al,* 2012 *Cell* **150** 389
Tomita M, 2001 *Trends Biotechnol.* **19** 205
[22] Lichtman J W and Sanes J R, 2008 *Curr. Opin. Neurobiol.* **18** 34653
[23] Among the projects that aim at reproducing macro-economic behavior from agent behavior, see www.eurace.org/index.php?TopMenuId=2, www.crisis-economics.eu/home and http://ineteconomics.org/grants/agent-based-model-current-economic-crisis, or the more ambitious Living Earth Simulator of the FuturICT project (www.futurict.eu/)
[24] Halevy A, Norvig P and Pereira F, 2009 *IEEE Intell. Syst. Archive* **24** 8
Cristianini N, 2010 *Neural Netw.* **23** 466
[25] Munos B, 2009 *Nature Rev. Drug Disc.* **8** 963
[26] Haldane A G and Madouros V, *The dog and the Frisbee*, 2012 *BIS Central Bankers' Speech at Federal Reserve Bank of Kansas City's 36th Economic Policy Symp. (The Changing Policy Landscape); (Jackson Hole, WY, Aug. 2012)*
[27] Myung I J, Balasubramanian V and Pitt M A, *Counting probability distributions: differential geometry and model selection*, 2000 *Proc. Nat. Acad. Sci.* **97** 11170
[28] Mastromatteo I and Marsili M, *On the criticality of inferred models*, 2011 *J. Stat. Mech.* P10012

J. Stat. Mech. (2013) P09003