Learning with unknowns: analyzing biological data in the presence of hidden variables

Claudia Battistin

Kavli Institute for Systems Neuroscience and Centre for Neural Computation, NTNU, Trondheim

Benjamin Dunn

Kavli Institute for Systems Neuroscience and Centre for Neural Computation, NTNU, Trondheim

Yasser Roudi

Kavli Institute for Systems Neuroscience and Centre for Neural Computation, NTNU, Trondheim

Institute for Advanced Studies, Princeton

Abstract

Despite our improved ability to probe biological systems at a higher spatiotemporal resolution, the high dimensionality of the biological systems often prevents sufficient sampling of the state space. Even with large scale datasets, such as gene microarrays or multi-neuronal recording techniques, the variables we are recording from are typically only a small subset, if wisely chosen, representing the most relevant degrees of freedom. The remaining variables, or the so called *hidden variables*, are most likely coupled to the observed ones, and affect their statistics and consequently our inference about the function of the system and the way it performs this function. Two important questions then arise in this context: *which variables should we choose to observe and collected data from? and how much can we learn from data in the presence of hidden variables?* In this paper we suggest that recent algorithmic developments rooting in the statistical physics of complex systems constitute a promising set of tools to extract relevant features from high-throughput data and a fruitful avenue of research for coming years.

Keywords: statistical models, maximum entropy, generalized linear models, Ising model, hidden nodes

1. Introduction

Ongoing technological advancements in high-throughput recordings of biological systems have captured the attention of many scientists who have aimed to come up with methods for analyzing these data. Protein sequence alignments [1], gene expression level measurements [2, 3] and neural recordings [4, 5, 6] constitute extensive samplings of the microstates of the underlying complex systems, and the availability of large datasets calls for statistical tools to analyze them [7]. New methods from machine learning and statistical physics are constantly being proposed to allow us to gain information from these large datasets. In particular, statistical physics, with a set of tools that are built over many years for dealing with high dimensional systems, have come to play an important role for this purpose in recent years. The main focus of this paper is then to discuss and assess some of these tools and the prospect they offer for future developments in the field.

From its origins, statistical physics has dealt with large systems of interacting degrees of freedom [8], describing them in terms of macroscopic variables that ultimately provide a lower dimensional representation of the system and uncover collective behavior. The ability of bridging between different levels of description of a system in a systematic way makes some approaches within statistical physics promising candidates for inspecting data from biological or other complex systems. When engaging in statistical inference, physicists may aim at reconstructing the underlying system from the data, e.g. a network of interactions [9, 10], and exploiting the theoretical understanding they developed on the phenomenon [11]. In order to find insightful structure in noisy data one may adopt some dimensionality reduction scheme, before proceeding to parametric inference. In the deeply undersampled regime, namely when the number of samples is much smaller than the dimensionality of the state space of the system, or even the state space characterized by observed or recorded parts of the system, selecting the relevant degrees of freedom becomes imperative to avoid fitting noise and to gain as much information as possible about the system and its function. This problem can be addressed from an information theoretical point of view [12] and is based on the distribution of mostly informative samples, as we will discuss in section 2. Once the pre-processing of the data is performed, one can incorporate the observed data into rigorous probabilistic models of the data. In some cases, one can then even infer how the hidden variables might influence observed variables: hidden variables, namely those deemed less important to observe and measure from, can then be incorporated in the statistical model using new developments in inference in the presence of hidden variables, whereby one can estimate how much the latent variables affect the inference on the observed subsystem as well as predict their time-varying states, as we will discuss in section 3.

2. Critical Variable Selection

Critical Variable Selection (CVS) is a method recently proposed for selecting variables or degrees of freedom that are important about the function of the system, without having much knowledge a priori about this function. The main idea, proposed in [12], rests on the assumption that without any prior knowledge, and when there is very little data, there is little reason to think that two configurations of the system are functionally different, when they appear in data with similar frequency. The degree of similarity necessary to determine if two configurations of the system are functionally different also depends on the number of samples available: a small difference in the observed frequency of two states become more important when there is more data.

Before going into more details about the issue of variable selection in the under-sampled systems, it is imperative to give a history of what can be seen as a surprising phenomena of similar statistical behavior in widely different complex systems. In the last century, numerous databases have gathered empirical evidence supporting the prevalence of power laws, characterizing samples across diverse domains [13]: from words count in text [14], sand pile avalanches [15], to size of cities [16]. Biology is no exception [13] and given the universality of power laws, general mechanisms have been proposed as generating this distribution [17]. A tantalizing one, is *self-organized criticality*. The relation between this type of criticality and power laws dates back to the 1980s [15]: the idea is that the system is poised at a critical point (at the cross-over between macroscopic states or phases) automatically, without requiring any parameter tuning, and through its own internal dynamics. This is in contrast to other forms of critical phenomena, e.g. water to vapor boiling transition, where an external parameter, namely temperature, should be tuned by an external agent. Criticality implies long range spatial and temporal correlations between the states of the system and complex emergent behavior. The link between power-laws and criticality was originally drawn in statistical physics where the description of equilibrium systems in terms of entropy and energy allows a rigorous derivation. In biology the connection between power law distributions — Zipf's law in particular — and criticality has been extended through the use of maximum entropy models (max-ent) [18]. Max-ent models [19] reflect the current knowledge on the system encoded in the data: the entropy of the inferred distribution is maximized, while being constrained to reproduce some or all of the statistics of the data. Maximum entropy inference naturally introduces an energy function on the states space, that, in case of power law distributed data, relates to the entropy of the data in a way similar to physical systems at the critical point. The interpretation of power laws as a signature of criticality has become very popular in the last decades and has fostered the use of maximum entropy models in statistical inference on biological data [20].

Maximum entropy inference, and its approximate implementations, have proven to be a versatile and useful technique, and to outperform other approaches for biological data analysis in a number of settings e.g. cross-correlation based analysis for inferring interactions in neuronal data or inference of protein residue contacts. For instance, single cell responses can be predicted from the population activity of a network of neurons in the salamander retina [21] using the max-ent approach and this approach appears to predict higher order c correlations [22] when fitted only using lower order correlations (mean firing rate and pairwise). From the activity of simultaneously recorded neurons in the prefrontal cortex of a rat, maximum entropy inference methods reveal task-related changes of the effective couplings which reappear during sleep post-task [23]. In protein contact prediction from multiple sequence alignments [24], fitting maximum entropy models is a step in the powerful *Direct Coupling Analysis* (DCA) method for inferring residue contacts [25]. An important work in this line is the work by Weigt et al [26] who used a relatively computationally expensive message passing method to learn a max-ent model and to use this model to find direct residue contacts in proteins-protein interactions. Later, it was shown that even a rougher (and faster) approximate method can serve the purpose [27]. A further improvement on max-ent DCA performance in contact prediction consists of employing the pseudolikelihood method [28], whose accuracy allows for protein 3D structure reconstruction [29]. This list can be continued and other fruitful applications of the maximum entropy inference exist e.g. for inferring gene regulatory networks [30] or epidemics [31]. The usefulness of maximum entropy approaches can continue to rise, with algorithmic improvements on the maximum entropy approach to match biologically realistic features of networks e.g. by including sparsity [32, 33, 34].

Despite the numerous useful applications and insights gained from the maximum entropy approach, there are, however, a number of problems with this framework and, in particular, the interpretation of criticality when dealing with data collected from large systems. As discussed before, typically when observing large systems, the resulting sample is (a) at best only a small fraction of the system, e.g. hundreds of neurons out of thousands or hundreds of thousands in a cortical circuit, and (b) even for this small fraction, a limited part of the configuration space is sampled, e.g. for hundreds of neurons, one would need to record for years to cover the whole phase space spanned by these neurons. We describe these problems in more details below.

Roudi et al. [35] argued that the success of max-ent models stems from the small number of neurons and low probability of spikes considered in the analysis. It was further shown that applying the max-ent techniques to a random connected network of simulated neuronal network, one obtains similar results as [36], but the usefulness of the max-ent model decreases as the size of the sampled neurons increases. More recently the maximum entropy approach has been criticized in the context of protein contact prediction [37], in particular discussing foundational arguments behind the current the implementation of the max-ent principle and the implications of applying the max-ent approach for uncovering the underlying processes of the systems. Pairwise models are indeed largely employed for inferring interactions but somehow contradict the max-ent spirit: the model is forced to reproduce pairwise correlations of the data, while higher order statistics is neglected. On the other hand, it can be argued that pairwise models are sufficient to infer pairwise interactions, even if they are not sufficient to reproduce protein distribution within domains. The other controversial aspect is the fact that max-ent models assume that the process sampled is at equilibrium, which is certainly debatable for many biological systems, e.g. neural networks.

Even with these problems with the maximum entropy approach, it still had the interesting benefit that it appeared to explain the regularities that many data sets from complex system exhibit, e.g. the Zipfs laws in neural data pattern rank order, stating that this is somewhat related to the criticality of the inferred maximum entropy models. However, recently a number of, in our opinion, more likely explanations have been offered: instead of an actual feature of the system, e.g. being poised at the critical point, the power law distribution has been interpreted as signaling the presence of hidden variables and that experimentalists, searching in this highly under-sampled system, would do best, if they choose the variables that exhibit a Zipf's law [12].

Initially it was shown [12] (see Box 1) that, under mild conditions on the distribution of the states of a system maximizing some unknown utility function, the observed variables exhibit a Boltzmann distribution at an effective temperature that depends on the number of hidden variables. In light of these results, the authors in [12] introduced a new paradigm in data analysis suggesting the entropy of frequencies $H\left[K\right]$ to be used as a *relevance* measure for a sample; here $H[K]$ is the entropy of the counts: defining K_s as the number of times configuration s was seen in M samples, one defines $m_k = \sum_s \delta_{k,K_s}$, namely the number of configurations that were observed k times, and $H[K]$ is the entropy of the variable K which takes value k with probability km_k/M. Identifying a subset of variables that together exhibit the largest $H\left[K\right]$, or broadest distribution of frequencies in the sample, as those carrying information on the function performed by the system, is the backbone of CVS.

Multiple examples of the usefulness of this approach has been given in [12] and later work has proven its power further. For instance [38] used this approach, and found that the algorithm proves to give results that are consistent with the present knowledge on biologically relevant sites (Response Regulator Receiver and Voltage Sensor Domain of Ion Channels) and to outperform state of the art algorithms (Statistical Couplings Analysis) when integrated with DCA. The method has proven to be robust against the number of selected sites and the sample size. In another recent paper [39], the CVS method proposed in [12] has found much firmer theoretical grounds by studying how to optimally cluster the observed configurations of a system depending on the number of samples in the data; see Box. 2 for more details.

Within the same idea of the importance of hidden variables or external factors (which can also obviously be seen as hidden variables), it has been rigorously proven that in the specific case of a high dimensional latent variable model, the frequency of the configurations of observed variables follows the Zipf's power law[40]. This work has subsequently been elaborated and extended in [41], where the authors notice that a broad distribution of frequencies (power law) can be generated by mixing a narrow frequency distribution of the observed variables given the hidden variables across different settings of the hidden ones. Identifying the latent variable allows for the quantification of its explanatory power in terms of the distribution of frequencies in different domains [41]. As an instructive example, we can consider the power law distribution of frequencies in the dataset suggested by a max-ent equilibrium analysis [36] indicating that the system is strongly coupled or even critical. Later, learning a non-stationary model [42], it was proved that the correlations between neurons can be explained by a latent variable (time dependent) external field (presumably the movie to which the retina was exposed during recordings), without requiring significant couplings between neurons. In this case, as in many other cases, the statistical regularity that seems to give insights on the way the complex system coordinate and perform its function, may just be a characteristic of the (incomplete) sampling.

3. Learning the dark side of the network

Whichever method is chosen to select variables to observe, it would also be very useful to infer the effect of the unobserved hidden variables on the system. For instance, can we say how the inferred functional couplings between a set of neurons are going to change if we were to include some of the unrecorded neurons? Parametric models that include hidden variables as part of the network of interactions can be used to do this job. They can, for instance, explain away correlations in the data induced by the hidden variables by disentangling direct from indirect interactions and uncovering external covariates.

Recently the statistical physics community has come up with accurate and efficient algorithms for inference of network interactions in the presence of hidden variables. Analyzing the inference of the symmetric connections in an equilibrium Ising model, [43] shows that the interactions between observed nodes can be retrieved reliably — provided that the fraction of hidden neurons is small enough. But again, more relevant to real biological data are those algorithms that take into account the non-equilibrium nature of the biological systems. Approximate algorithms [44, 45, 46] that reliably reconstruct the connectivity on a partially observed network have been developed for the paradigmatic case of the kinetic Ising model, a Bernoulli generalized linear model [47] with one step time kernel. The algorithms take the form of a *Expectation Maximization* (EM) algorithm [48]. The EM algorithm is a two step recursive algorithm that alternates between: 1) computing the expected value of the states of the hidden units given the data and the current values of the parameters, and 2) updating the parameters to maximize the expected log-likelihood. Its convergence to a local maximum of the likelihood is guaranteed. The main problem for using the EM algorithm for reconstructing networks with hidden nodes is in the first part, and this is where approximate methods rooted in statistical mechanics and field theory has been used in [44, 45, 46]; see also [49, 50] where the problem of inferring the state of hidden variables for continuous variables under Langevin dynamics is studied.

As an example of how one can use these methods to explore the role of hidden nodes in a real life setting, in Figure 1 we show the inference of connectivity between grid cells in the Medial Entorhinal Cortex (MEC) of rats. Grid cells [51] are neurons in mammalian MEC and each grid cell has the property that it fires at a number of locations in the space that the animal navigates, and these locations form a hexagonal pattern tessellating the space. Theoretical network models of how grid cells achieve their hexagonal firing pattern assume that the local network is constructed such that cells that are active in similar spatial locations are more likely to be connected through excitation while pairs with non-overlapping spatial selectivity inhibit each other; see [52, 53] for reviews of grid cells and their theoretical network models.

In [54], the authors demonstrated that this connectivity scheme, known as Mexican hat connectivity, could be recovered with a kinetic Ising model. Furthermore, the trend remained stable with the inclusion of known covariates, including space, head directionality, speed and local theta phase preference. Figure 1 shows that this trend is maintained even when unknown covariates, i.e. hidden units, are assumed, thus strengthening the conclusion that the required Mexican hat connectivity exists in the MEC. While the number of assumed hidden units did not come close to the hundreds of thousands of neurons that were actually unobserved, the stability of the couplings is still interesting. It is

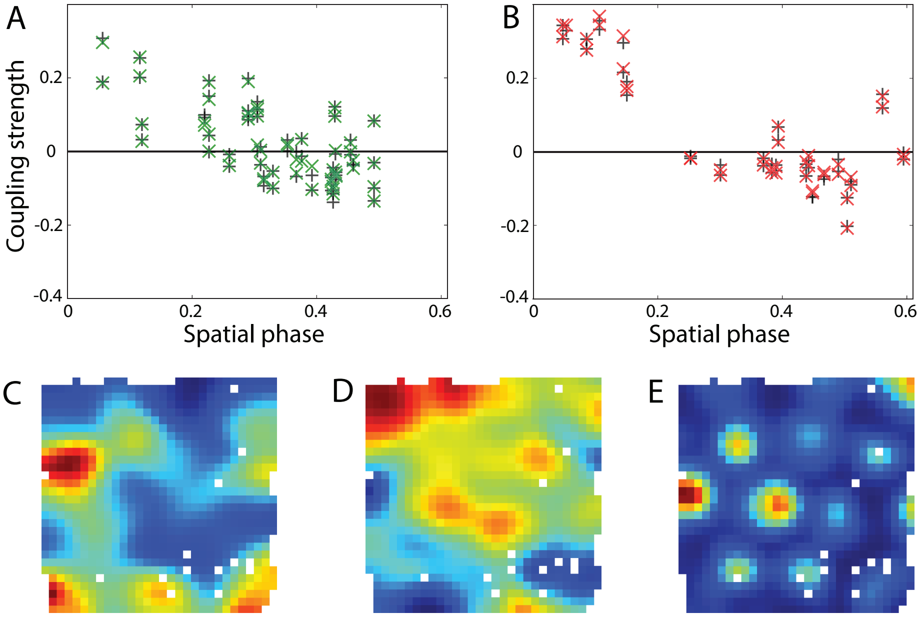

Figure 1: **Hidden units in the entorhinal network:** The phase-dependent functional connectivity for two modules of grid cells captured using the kinetic Ising model with only constant fields (black +s in A and B) is largely maintained when hidden units are assumed (colored ×s in A and B). Panels C, D and E show the smoothed firing rate map of three neurons as a function of the position in a square box that the animal was navigating. Interestingly, the spatial tuning of the hidden units (e.g. C and D) do not show the stereotypical hexagonal firing of the grid cells, exemplified in E. Here four hidden units with no connectivity between them were assumed. Learning has been performed through the approximated EM algorithm from [44]. Time bins of 10ms, a learning rate of 0.1 and 3000 learning steps were used in both the inference with and without hidden units.

also interesting to note that the hidden neurons appear to exhibit a degree of spatial selectivity in their response (Figure 1 C and D), something that would be important to explore further, for instance by studying if with more hidden neurons, one can predict (or post-dict) the existence of other spatially tuned cell types in the MEC. This figure also serves to demonstrate that the methods, although still at their infancy in terms of practical applications, can be applied to real data. There are also opposite cases in the literature: for instance it has been demonstrated that even a single hidden variable can have a dramatic effect in the resulting interpretation of neural data [55], allowing, for example, for the recovery of the correlation structure of the awake animal from data collected under anesthesia.

4. Conclusions

Complex systems such as biological systems are in many ways puzzling us. One of these puzzles is to decide which of their many degrees of freedom are most informative about the system. For instance, if one is aiming to understand the mammalian MEC, one should wonder which cells to record from, given the constraint that one cannot record from all neurons nor can one record from them forever. In our opinion, finding normative and unbiased ways to approach this issue will be an avenue of research which can yield fruitful results in the future for analyzing biological data and systems. We described one new method, the Critical Variable Selection, which we think in recent years has proven very promising. However, we believe more work and different approaches are required to achieve a successful understanding of variable selection from big biological systems and data. We also mentioned the problem of inferring the state of hidden nodes and their effect on the observed ones, and discussed some of the new work that has been done in this direction. Again we think being able to use observed data to talk about the known unknowns or even unknown unknowns is an important topic with great potentials. We are confident that the work we mentioned in this review is just the beginning of a successful story.

5. Acknowledgement

The authors acknowledge financial support from the Kavli Foundation and the Norwegian Research Council's Centre of Excellent scheme (Centre for Neural Computation, grant number 223262). YR is also grateful to the Starr Foundation for financing his position at the IAS.

6. References

[1] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, M. Punta, Pfam: the protein families database, Nucleic Acids Research 42 (D1) (2014) D222–D230.

[2] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, R. Edgar, Ncbi geo: archive for high-throughput functional genomic data, Nucleic Acids Research 37 (suppl 1) (2009) D885–D890.

[3] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, A. Brazma, Arrayexpress update—simplifying data submissions, Nucleic acids research 43 (Database issue) (2015) D1113–6.

[4] M. E. J. Obien, K. Deligkaris, T. Bullmann, D. J. Bakkum, U. Frey, Revealing neuronal function through microelectrode array recordings, Frontiers in neuroscience 8 (2015) 423.

[5] M. A. Nicolelis, Methods for neural ensemble recordings, CRC press, 2007.

[6] C. Grienberger, A. Konnerth, Imaging calcium in neurons, Neuron 73 (5) (2012) 862–885.

[7] R. Braun, Systems analysis of high-throughput data, in: A Systems Biology Approach to Blood, Springer, 2014, pp. 153–187.

[8] K. Huang, Statistical mechanics, New York-London.

[9] J. Hertz, Y. Roudi, J. Tyrcha, Ising model for inferring network structure from spike data, in: R. Quian Quiroga, S. Panzeri (Eds.), Principal of Neural Coding, CRC Press 2013, 2013, pp. 527–546.

[10] M. Timme, J. Casadiego, Revealing networks from dynamics: an introduction, Journal of Physics A: Mathematical and Theoretical 47 (34) (2014) 343001.

[11] P. Grassberger, J.-P. Nadal, From statistical physics to statistical inference and back, Vol. 428, Springer Science & Business Media, 2012.

[12] M. Marsili, I. Mastromatteo, Y. Roudi, On sampling and modeling complex systems, Journal of Statistical Mechanics: Theory and Experiment 2013 (09) (2013) P09003.

[13] M. E. Newman, Power laws, pareto distributions and zipf's law, Contemporary physics 46 (5) (2005) 323–351.

[14] G. K. Zipf, Human behavior and the principle of least effort.

[15] P. Bak, C. Tang, K. Wiesenfeld, Self-organized criticality, Physical review A 38 (1) (1988) 364.

[16] X. Gabaix, Zipf's law for cities: an explanation, Quarterly journal of Economics (1999) 739–767.

[17] M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions, Internet mathematics 1 (2) (2004) 226–251.

[18] T. Mora, W. Bialek, Are biological systems poised at criticality?, Journal of Statistical Physics 144 (2) (2011) 268–302.

[19] E. T. Jaynes, Information theory and statistical mechanics, Physical review 106 (4) (1957) 620.

[20] R. R. Stein, D. S. Marks, C. Sander, Inferring pairwise interactions from biological data using maximum-entropy probability models, PLoS Comput Biol 11 (7) (2015) e1004182.

[21] G. Tkačik, T. Mora, O. Marre, D. Amodei, S. E. Palmer, M. J. Berry, W. Bialek, Thermodynamics and signatures of criticality in a network of

neurons, Proceedings of the National Academy of Sciences 112 (37) (2015) 11508–11513.

[22] S. Cocco, S. Leibler, R. Monasson, Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods, Proceedings of the National Academy of Sciences 106 (33) (2009) 14058–14062.

[23] U. Ferrari, G. Tavoni, F. P. Battaglia, S. Cocco, R. Monasson, Inferred ising model unveils potentiation of pairwise neural interactions and replay of rule-learning related neural activity, BMC Neuroscience 14 (1) (2013) 1.

[24] D. de Juan, F. Pazos, A. Valencia, Emerging methods in protein coevolution, Nature Reviews Genetics 14 (4) (2013) 249–261.

[25] F. Morcos, T. Hwa, J. N. Onuchic, M. Weigt, Direct coupling analysis for protein contact prediction, Protein Structure Prediction (2014) 55–70.

[26] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing, Proceedings of the National Academy of Sciences 106 (1) (2009) 67–72.

[27] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families, Proceedings of the National Academy of Sciences 108 (49) (2011) E1293–E1301.

[28] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, E. Aurell, Improved contact prediction in proteins: using pseudolikelihoods to infer potts models, Physical Review E 87 (1) (2013) 012707.

[29] T. A. Hopf, C. P. Schärfe, J. P. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. Bonvin, D. S. Marks, Sequence co-evolution gives 3d contacts and structures of protein complexes, Elife 3 (2014) e03430.

[30] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, N. V. Fedoroff, Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns, Proceedings of the National Academy of Sciences 103 (50) (2006) 19033–19038.

[31] J. R. Artalejo, M. Lopez-Herrero, The sis and sir stochastic epidemic models: A maximum entropy approach, Theoretical population biology 80 (4) (2011) 256–264.

[32] P. Ravikumar, M. J. Wainwright, J. D. Lafferty, et al., High-dimensional ising model selection using ?1-regularized logistic regression, The Annals of Statistics 38 (3) (2010) 1287–1319.

[33] A. Decelle, F. Ricci-Tersenghi, Pseudolikelihood decimation algorithm improving the inference of the interaction network in a general class of ising models, Physical review letters 112 (7) (2014) 070603.

[34] N. Bulso, M. Marsili, Y. Roudi, Sparse model selection in the highly undersampled regime, arXiv:1603.00952, in press J Stat Mech.

[35] Y. Roudi, S. Nirenberg, P. E. Latham, Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't, PLoS Comput Biol 5 (5) (2009) e1000380.

[36] E. Schneidman, M. J. Berry, R. Segev, W. Bialek, Weak pairwise correlations imply strongly correlated network states in a neural population, Nature 440 (7087) (2006) 1007–1012.

[37] E. Aurell, The maximum entropy fallacy redux?, PLoS Comput Biol 12 (5) (2016) e1004777.

[38] S. Grigolon, S. Franz, M. Marsili, Identifying relevant positions in proteins by critical variable selection, Molecular BioSystems.

[39] A. Haimovici, M. Marsili, Criticality of mostly informative samples: a bayesian model selection approach, Journal of Statistical Mechanics: Theory and Experiment 2015 (10) (2015) P10013.

[40] D. J. Schwab, I. Nemenman, P. Mehta, Zipf's law and criticality in multivariate data without fine-tuning, Physical review letters 113 (6) (2014) 068102.

[41] L. Aitchison, N. Corradi, P. E. Latham, Zipf's law arises naturally in structured, high-dimensional data, arXiv preprint arXiv:1407.7135.

[42] J. Tyrcha, Y. Roudi, M. Marsili, J. Hertz, The effect of nonstationarity on models inferred from neural data, Journal of Statistical Mechanics: Theory and Experiment 2013 (03) (2013) P03005.

[43] H. Huang, Effects of hidden nodes on network structure inference, Journal of Physics A: Mathematical and Theoretical 48 (35) (2015) 355002.

[44] B. Dunn, Y. Roudi, Learning and inference in a nonequilibrium ising model with hidden nodes, Physical Review E 87 (2) (2013) 022127.

[45] C. Battistin, J. Hertz, J. Tyrcha, Y. Roudi, Belief propagation and replicas for inference and learning in a kinetic ising model with hidden spins, Journal of Statistical Mechanics: Theory and Experiment 2015 (5) (2015) P05021.

[46] J. Tyrcha, J. Hertz, Network inference with hidden units, Mathematical biosciences and engineering : MBE 11 (1) (2014) 149–165.

[47] P. McCullagh, J. A. Nelder, Generalized linear models, Vol. 37, CRC press, 1989.

[48] R. Sundberg, Maximum likelihood theory for incomplete data from an exponential family, Scandinavian Journal of Statistics (1974) 49–58.

[49] B. Bravi, M. Opper, P. Sollich, Inferring hidden states in langevin dynamics on large networks: average case performance, arXiv preprint arXiv:1607.01622.

[50] B. Bravi, P. Sollich, Inference for dynamics of continuous variables: the extended plefka expansion with hidden nodes, arXiv preprint arXiv:1603.05538.

[51] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, E. I. Moser, Microstructure of a spatial map in the entorhinal cortex, Nature 436 (7052) (2005) 801–806.

[52] B. L. McNaughton, F. P. Battaglia, O. Jensen, E. I. Moser, M.-B. Moser, Path integration and the neural basis of the'cognitive map', Nature Reviews Neuroscience 7 (8) (2006) 663–678.

[53] E. I. Moser, M.-B. Moser, Y. Roudi, Network mechanisms of grid cells, Phil. Trans. R. Soc. B 369 (1635) (2014) 20120511.

[54] B. Dunn, M. Mørreaunet, Y. Roudi, Correlations and functional connections in a population of grid cells, PLoS Comput Biol 11 (2) (2015) e1004052.

[55] A. S. Ecker, P. Berens, R. J. Cotton, M. Subramaniyan, G. H. Denfield, C. R. Cadwell, S. M. Smirnakis, M. Bethge, A. S. Tolias, State dependence of noise correlations in macaque primary visual cortex, Neuron 82 (1) (2014) 235–248.

[56] T. M. Cover, J. A. Thomas, Elements of information theory, John Wiley & Sons, 2012.

(BOX 1): The entropy of frequencies

Emergent properties of biological complex systems, as large numbers of interacting components carrying out some function as a result of their interactions, can be conceived of as solutions to an optimization problem. This means that the system state $\overrightarrow{s} = (\underline{s}, \overline{s})$, of hidden ($\overline{s}$), and of observed ($\underline{s}$) variables , maximizes an objective function $U(\overrightarrow{s})$. Although the objective function typically depends on all the variables of the system, what the observer measures is $u_{\underline{s}} = E\left[U(\overrightarrow{s})\right]$, an average over the distribution of the hidden variables $\overline{s}$. It was proven [12] that under broad conditions the observed variables $\underline{s}$ then follow the Boltzmann-Gibbs distribution:

$$p(\underline{s}) \propto e^{\beta u_{\underline{s}}}, \tag{1}$$

namely the maximum entropy distribution that constrains $u_{\underline{s}}$. The inverse temperature β is controlled by the number of hidden variables in a way that, for typical distributions, the observed subsystem is predictable only when the number of hidden variables is large enough. In a sample $\hat{s}$ the empirical frequency of states $K_{\underline{s}}$, approximating $p(\underline{s})$, provides a noisy estimate of the unknown function $u_{\underline{s}}$, through equation (1). Consequently the most informative samples are those that maximize the entropy of frequencies $H\left[K\right]$; see the main text and [12] for the definition of $H[K]$. In [12] the authors observed that the maximum entropy of frequencies is a non monotonic function of the entropy of the data and that the latter provides an upper bound to $H\left[K\right]$ through the data processing inequality [56]. In the well sampled regime $H\left[K\right] = H\left[s\right]$, while in the undersampled regime the maximization of $H\left[K\right]$, at fixed $H\left[s\right]$ corresponds to power law distributed sample (when looking at their Zipf's plot), and in particular, at the peak value that the maximum of $H[K]$ can reach, the power law is a Zipf's law. This offers an alternative to physical criticality as an explanation for the observed Zipf's laws: it follows that the Zipf's law distributed samples are the most informative samples that an experimentalist can come up with when the state space is going to be poorly sampled due to experimental limitations, as often is the case in biology. In other words, Zipf's law occur only when one is looking at samples from meaningful variables or degrees of freedom of a complex systems, with usually the majority of variables being unobserved.

(BOX 2): Bayesian model selection for data clustering

The definition of the states of a system is made by the observer that probes the system. It is the available empirical sample $\hat{s}$ of the configurations of the system that dictates the *resolution* of the state space of the system. Loosely speaking, if two states appear in the sample with frequencies that do not differ much, it means that the sample does not permit the distinction of one state from the other. Consequently our inference procedure needs to collapse them into the same state. This problem has been tackled in [39] with a Bayesian model selection approach: find the partition Q of the states space that maximizes the posterior probability $P[Q|\hat{s}]$. The authors observed that the optimal partition induces a power-law distribution on the data, which was proven to characterize maximally informative samples in terms of $H[K]$, the entropy of frequencies [12] (see Box 1). In general, looking for the optimal clustering of the data can be computationally expensive. Fortunately, it turns out that a good approximation for the optimal partition $Q^\star$, is clustering together only states with exactly the same frequency. In practical applications — relevant sites selection in proteins, clustering of financial stocks — the optimal partition $Q^\star$ significantly overlaps with the raw one, further demonstrating that $H[K]$ discussed in the text and Box 1, can be used as a proxy for $H[Q^\star]$.

Highlights

- Ref. [12] (**). This is one o the first paper that relates Zipf's law not to the some processes, critical or otherwise, that happen in a system generating the data, but as a consequence of sampling properties, and, in particular, sampling in the presence of hidden variables. The authors introduce a new approach to optimal variable selection using count distribution that has been later implemented for critical variable selection in selecting important amino-acid sites in a protein [38] and has further been analyzed and used for states clustering [39].

- Ref [37](**). This paper conveys conceptual and methodological arguments against the viability of the max-ent approach. Partly the skepticism is concerned with the preferential employment of pairwise models with respect to max-ent models with higher order interactions. This approach is clearly very far from using all available information on the system, provided by the data, as the maximum entropy spirit would dictate [19].The paper presents even more fundamental criticisms to maximum entropy inference, regarding the implicit assumption that the samples of the system are drawn from an equilibrium distribution, meaning that the process under investigation is supposed to obey detailed balance. If this could possibly be true for protein alignments, with puzzling implications on the reversibility of evolution on long time scales, it cannot definitely hold for systems subjected to strong perturbations as in-vivo neuronal recordings.

- Ref. [44](*). Using generating functional methods, the authors derive an approximate expression for the likelihood of a partially observed kinetic Ising model. The algorithm is able to retrieve the generative network of interactions, even those between the hidden units. Importantly for its potential applications, the number of hidden neurons does not have to be known a priori, but can be inferred from the data.

- Ref. [49](*). In this paper, as well as in [50], the inference error on the time series of hidden nodes in a network of interacting continuous degrees of freedom is studied. Its time course as well as its dependence on the degree of symmetry of the connectivity and on the fraction of observed to hidden degrees of freedom are investigated. Such theoretical analyses are of great importance for assessing the potentialities and limitations of statistical models with hidden variables.

- Ref. [38](*). This paper reports the first systematic application of the ideas proposed in [12]. Critical Variable Selection, as called in this paper, aims at making a ranking of the relevant sites in a protein family in terms of their relevance to the function of the protein. The method has proven to be robust against the number of selected sites and the sample size, and as opposed to other methods of choosing relevant sites, goes beyond single-site conservation or pairwise correlations.

- Ref. [39] (**). Using Bayesian model selection, this paper has provided a solid analysis for deciding how different configurations of a system should or should not be distinguished from each other given the number of available samples form the system. By doing this, it has also provided a more theoretically justified account for the proposal in [12]. See Box 2 for more details.