Issues in data expansion in understanding criticality in biological systems

Vaibhav Wasnik
Department of Biochemistry, University of Geneva, Geneva, Switzerland

Received 5 August 2017 and Received in final form 25 December 2017
Published online: 31 January 2018 – © EDP Sciences / Società Italiana di Fisica / Springer-Verlag 2018

Abstract. At the point of a second-order phase transition also termed as a critical point, systems display long-range order and their macroscopic behaviors are independent of the microscopic details making up the system. Due to these properties, it has long been speculated that biological systems that show similar behavior despite having very different microscopics, may be operating near a critical point. Recent methods in neuroscience are making it possible to explore whether criticality exists in neural networks. Despite being large in size, many datasets are only a minute sample of the neural system and methods have to be developed to expand these datasets to study criticality. In this work we develop an analytical method of expanding a dataset to the large $N$ limit to make statements about the critical nature of the dataset. We show that different ways of expanding the dataset while keeping its variance and mean fixed yield different results regarding criticality. This hence casts doubts on the established procedures for deducing criticality of biological systems through expansion of finite-sized datasets.

Introduction

Many biological systems display self-organization, arising from specific local interactions between the various constituents. One of the ways that global behaviors can emerge because of local interactions is if biological systems were poised at criticality. Long-range order also implies robustness of the biological system to respond to stimuli. This intuition has led to research into the question whether biological systems are poised at criticality.

Works on the subject of criticality in biological systems could be broken down into two directions. One direction points to how criticality could emerge based on certain characteristics of the underlying system, without looking at a specific biological system. For example, Zipfs law distributions and hence criticality are shown to arise naturally when one of the fluctuating variables in the system is hidden [1]. Criticality of datasets has been shown to also be linked to the exact inference of the probability distributions describing the dataset [2]. The understanding of criticality by taking into account the temporal dynamics of biological systems is discussed in [3].

The other direction involves looking at specific biological datasets to ascertain their critical nature. Gene expression dynamics in the macrophage has been shown to exhibit criticality [4]. Criticality in evolutionary ecology has been suggested to come from network-like structure of multispecies communities that are close to instability [5]. Human brain functional systems have also been suggested to exist in the endogenous state of dynamical criticality [6].

Advances in neuroscience have led to the availability of neuronal datasets which could be used to study criticality. One of the first works in this direction which attempted to understand criticality in neuronal datasets were [7,8]. Since then many works have appeared in the literature on studying criticality in neuronal systems. For example, [9] constructed models that are consistent with the distribution of global network activity. New ideas in modeling efforts to understand criticality in vertebrate retina have been proposed in [10,11].

Since criticality is a concept borrowed from physics, a way to map neuronal systems to concepts in statistical physics is required. In [8] the spiking of salamander retinal neurons was recorded when it was subject to external stimulus. The spikes were binned in appropriate time intervals, leading to patterns made up of binary bits, akin to the up and down state of spin, with neuron firing corresponding to $s_i = 1$ and non-firing to $s_i = -1$. In this work we will use the words spin and spikes interchangeably. $k_BT$ which is temperature in statistical mechanics is simply considered as a parameter in the neuronal problem, with $k_BT = 1$ representing the parameter when the neuronal measurements are made. Since criticality is only observed in statistical mechanical systems in the large $N$ limit, there was a need to expand the datasets. In [7] the construction of larger datasets from smaller ones involved sampling from a distribution of the average spiking ($s_i$)
with \( i \in [1, N] \) and the correlation between the spikes \( c_{ij} = \langle s_i s_j \rangle - \langle s_i \rangle \langle s_j \rangle \), with \( i \neq j \) and \( i, j \in [1, N] \), where \( N \) is the number of neurons in the dataset. Since this expansion of data was done computationally, the expansion could not be carried to the limit \( N \to \infty \). However, it was noticed that as the size of the constructed dataset increased, the specific heat peaked closer and closer to \( k_B T = 1 \). According to the authors, this suggested a possible divergence in the specific heat at \( k_B T = 1 \) as the system size increased to the large \( N \) limit, implying that the retinal neuronal network was operating at or near the critical point.

This method of expansion followed the wisdom of finite-size scaling. In this work we first highlight the fact that finite-size scaling arguments depend on the fact that except for one coupling, the rest of the couplings in the problem are poised at their critical value, which is not guaranteed a priori in fitting the expanded datasets. We then develop a method to expand the datasets analytically keeping the distribution of \( \langle s_i \rangle \) and \( c_{ij} \) of the original dataset, so that the critical temperature can be evaluated. We then analyze the salamander retinal neuronal dataset used by [7], to show that the values of the critical temperature can be evaluated. The fact that our results differ from [7], even though our method of expansion of the dataset obeys the same constraints, implies that these constraints are not sufficient to establish the critical nature of the system under study.

**The computational method**

Since expanding a system to the large \( N \) limit is not feasible computationally, [7] tried to expand the system computationally to a larger but still finite size. Label the neurons in the dataset by \( i \in [1, M] \), where \( M \) is the number of neurons. Let \( \langle s_i \rangle \) \( i \in [1, M] \) be the average spiking of the neurons. Let \( P_{s_i} \) be the probability distribution. Let \( c_{ij} \) with \( i, j \in [1, M] \) and \( i \neq j \) be the correlation between the spikes and \( P_{c_{ij}} \) be the probability distribution that is fit to the correlations. The larger dataset labeled by neurons \( \alpha \in [1, N] \), \( M < N \) is produced by picking the neuronal averages \( \langle s_\alpha \rangle \) from the probability distribution \( P_{s_\alpha} \) and the correlations \( c_{\alpha \beta} \) with \( \alpha, \beta \in [1, N] \), \( \alpha \neq \beta \) from the probability distribution \( P_{c_{ij}} \). These are then fit with a Boltzmannian distribution with a Hamiltonian

\[
H = \sum_{\alpha=1}^{N} h_\alpha s_\alpha + \sum_{\alpha, \beta=1}^{N} J_{\alpha \beta} s_\alpha s_\beta. \tag{1}
\]

The \( h, J \)'s are evaluated using

\[
\langle s_\alpha \rangle = \frac{\partial Z}{\partial h_\alpha}, \tag{2}
\]

\[
\langle c_{\alpha \beta} \rangle = \frac{\partial Z}{\partial h_\alpha \partial h_\beta},
\]

where \( Z \) is the partition function evaluated for \( k_B T = 1 \). This was done computationally. The specific heats were then evaluated as a function of different values of \( k_B T \).

It was found that as the system size \( N \) increased the specific heat peaks seemed to get closer to \( k_B T = 1 \). This led [7] to conjecture that in the large \( N \) limit the peak would become a divergence, implying that \( k_B T = 1 \) was the critical point. This method is the template for how criticality is assessed in neuronal datasets in the literature.

**Why the computational method may not work?**

The computational method outlined in the previous section, implicitly uses the concept of finite-size scaling in guessing the critical temperature. Let us outline what finite-size scaling is. Let us assume a system has size \( L \) with lattice spacing \( b \). A renormalization group analysis involves summing over a fraction of lattice sites, so that we are left with a lattice of size \( L/b \), \( b > 1 \). The free energies would transform as

\[
f\left(\{K_i\}, L\right) = b^{-d} f\left(\{K'_i\}, L/b\right) \tag{3}
\]

where couplings \( K_i \)'s get transformed into \( K'_i \). If by repeated application of the renormalization group iteration we reach a stage where \( K_i = K'_i = K_c \), it would imply that we have reached a fixed point, also called a critical point. Consider small deviations away from the critical point. Let these deviations transform under one iteration of the renormalization group as \( k'_i = b^{\nu_i} k_i \). This gives

\[
f(t, h, \{k_i\}, L) = b^{-d} f\left((t/b^{\nu_0}, h b^{\nu_2}, \{b^{\nu_i} k_i\}, L/b\right), \tag{4}
\]

where we have explicitly written the deviation from the critical values of the magnetic field \( h \) and temperature \( t = \frac{T_b}{L} - 1 \). If we were to iterate the renormalization group \( \ln(L/L_0)/\ln b \) times we then have

\[
f(t, h, \{k_i\}, L) = (L/L_0)^{-d} f\left((t(L/L_0)^{\nu_0}, h(L/L_0)^{\nu_2}, \{L(L/L_0)^{\nu_i} k_i\}, L_0\right). \tag{5}
\]

Now when one makes a central assumption (and only under this assumption) that the terms \( (L/L_0)^{\nu_i} k_i \) can be ignored, that quantities such as magnetic susceptibility scale as

\[
\left. \frac{\partial^2 f}{\partial h^2} \right|_{h=0} = L^{\gamma/\nu} G(L^{1/\nu} t), \tag{6}
\]

where \( \nu = \frac{1}{\nu_0} \). This is the origin of the finite-size scaling hypothesis. The maximum of the susceptibility occurs at \( L^{1/\nu} t = v_0 \). This would imply a relationship \( T = T_c + v_0 T_c L^{-1/\nu} \). A plot of temperature \( T \) where the susceptibility peaks versus \( L \) would peak closer to the critical temperature as the system size is increased and using this plot one can evaluate the critical temperature. This is the logic behind the evaluation of the critical temperature using the computational method. The computational method in the literature works with the peaks of specific heat instead of susceptibility, but the same logic as above carries through.
However, finite-size scaling is known to work in only some situations. For example it is shown to not work in the case of \( d > 4 \) Ising-like models [12]. The assumption of finite-size scaling requires that the deviation of the couplings from their critical value \((L/L_0)^b k_i \) can be ignored. If \((L/L_0)^b k_i \) cannot be ignored, there is a breakdown in the finite-size scaling hypothesis. Requiring \((L/L_0)^b k_i \) to be small would imply that for some reason anytime a \( J_{ij} \) is fit to reproduce the sample’s \( (s_i) \) and \( c_{ij} \), these \( J_{ij} \)’s are close to the critical value, which \textit{a priori} is not justified. Since we cannot ignore the deviations of couplings from the critical value, it implies a breakdown of the finite-size scaling hypothesis. Hence the computational method of guessing the critical temperature may not work in the problem of evaluating the critical temperature. In order to show this explicitly, we consider an analytical way of expanding the dataset keeping the distribution of the \( (s_i) \) and \( c_{ij} \) fixed below. We then use this method to evaluate the critical temperature of a dataset analyzed in the literature using the computational method and show that the resulting evaluation of the critical temperature does not match with the critical temperature evaluated using the computational method.

**Our construction**

First consider a subset of the original dataset. This subset being representative of the dataset would have the same distribution of \( (s_i) \) and \( c_{ij} \) as the original dataset. This subset is made up of neurons labeled with \( i \in [1, M] \). For each spin \( s_i \), now consider \( N \) spins \( s_i^{(\alpha)} \) with \( \alpha \in [1, N] \), which are statistically similar to spin \( s_i \), \textit{i.e.}

\[
\langle s_i^{(\alpha)} \rangle = (s_i), \quad \alpha \in [1, N],
\]

\[
\langle s_i^{(\alpha)} s_j^{(\beta)} \rangle - \langle s_i^{(\alpha)} \rangle \langle s_j^{(\beta)} \rangle = \langle s_is_j \rangle - \langle s_i \rangle \langle s_j \rangle, \quad \alpha, \beta \in [1, N], \quad i, j \in [1, M], i \neq j.
\]

Such spins \( s_i^{(\alpha)} \) would now have the same distribution of the average spikings and average correlation between the spikes as the spins \( s_i \). The above constraints in eq. (7) can be obtained using a Boltzmannian distribution with a Hamiltonian

\[
H = \sum_{i=1}^{M} h_i \sum_{\alpha=1}^{N} s_i^{(\alpha)} + \sum_{ij=1}^{M} \frac{J_{ij}}{2N} \sum_{\alpha=1}^{N} s_i^{(\alpha)} \sum_{\beta=1}^{N} s_j^{(\beta)}. \tag{8}
\]

We can estimate \( h_i \) and \( J_{ij} \) in the large \( N \) limit. To do this note that

\[
H = \sum_{i=1}^{M} h_i \sum_{\alpha=1}^{N} s_i^{(\alpha)} + \sum_{ij=1}^{M} \frac{J_{ij}}{2N} \sum_{\alpha=1}^{N} s_i^{(\alpha)} \sum_{\beta=1}^{N} s_j^{(\beta)} \tag{9}
\]

can be written as

\[
H = N \left[ \sum_{i=1}^{M} h_i m_i + \sum_{ij=1}^{M} J_{ij} m_i m_j \right], \tag{10}
\]

where

\[
m_i = \frac{s_i^{(1)} + s_i^{(2)} + \ldots + s_i^{(N)}}{N}. \tag{11}
\]

Hence the partition function can be written down as

\[
Z = \sum_{s^{(\alpha)} \in [1, M], \alpha \in [1, N]} e^{-\beta H} s^{(\alpha)} = \sum_{m_i} f(m_i) e^{-N \beta \sum_{i=1}^{M} h_i m_i + \sum_{ij=1}^{M} J_{ij} m_i m_j}, \tag{12}
\]

where we are summing over all possible values taken by \( m_i \) and \( \beta = \frac{1}{k_B T} \). \( f(m_i) \) is the number of ways of getting the value \( m_i \) by all possible combinations of \( s_i^{(\alpha)} \) for \( \alpha \in [1, N] \) and is given by

\[
f(m_i) = \frac{N!}{(\frac{N}{2}(1 + m_i))(\frac{N}{2}(1 - m_i))} \tag{13}
\]

which for \( N \to \infty \) becomes

\[
f(m_i) = e^{N(1 + m_i) \ln(1 + m_i) + N(1 - m_i) \ln(1 - m_i)} \tag{14}
\]

and hence the partition function becomes

\[
Z = \sum_{m_i} e^{N \beta \sum_{i=1}^{M} h_i m_i + \sum_{ij=1}^{M} J_{ij} m_i m_j} \times e^{\frac{1}{2} \sum_{i=1}^{M} (1 + m_i) \ln(1 + m_i) + \sum_{i=1}^{M} (1 - m_i) \ln(1 - m_i)} \tag{15}
\]

In the large \( N \) approximation the partition function is dominated by the saddle point and hence the solution is

\[
\frac{\partial}{\partial m_i} \left[ -\beta \sum_{i=1}^{M} h_i m_i + \sum_{ij=1}^{M} J_{ij} m_i m_j \right] + \frac{1}{2} \sum_{i=1}^{M} (1 + m_i) \ln(1 + m_i) \tag{16}
\]

\[
+ \sum_{i=1}^{M} (1 - m_i) \ln(1 - m_i) = 0
\]

which then gives us

\[
m_i = \tanh \beta \left( \sum_{j=1}^{M} J_{ij} m_j + h_i \right) \tag{17}
\]

or

\[
tanh^{-1} m_i = \beta \left( \sum_{j} J_{ij} m_j + h_i \right). \tag{18}
\]

Take the derivative with respect to \( m_j \). This gives us

\[
\frac{\delta}{\delta m_j} \left[ \frac{\delta h_i}{\delta m_j} \right] = \beta \left( J_{ij} + \frac{\partial h_i}{\partial m_j} \right). \tag{19}
\]
Fig. 1. The $\sum_{ij} C_{ij}$ plotted against $k_B T$ for different subsets of neurons from the original dataset, which have been expanded to the large $N$ limit as described in the text. As we can see $k_B T = 1$ is not a temperature where $\sum_{ij} C_{ij}$ diverges for any of the subsets. This is unlike what has been stated in [7] which suggests that the specific heat should diverge at $k_B T = 1$.

Now,\[ \frac{\partial m_i}{\partial h_i} = C_{ij}, \] which is the correlation between the $m_i$'s. Hence,\[ \frac{\partial h_i}{\partial m_j} = [C^{-1}]_{ij}, \] which gives us\[ \frac{\delta}{1 - m_i^2} = \beta(J_{ij} + [C^{-1}]_{ij}) \] or\[ J_{ij} = -[C^{-1}]_{ij} \]
if $i \neq j$ and\[ J_{ii} = \frac{\beta}{1 - m_i^2} - [C^{-1}]_{ii} \]
using these values of $J_{ij}$ in the equation\[ m_i = \tanh \beta \left( \sum_{j=1,M} J_{ij} m_j + h_i \right), \] we can evaluate $h_i$.

**Analysis of salamander retinal data**

We now use the method outlined above to study the dataset analyzed using the computational method in [7,8].

This data consisted of the neuronal firing from a salamander retina. In [7,8] the data from 40 neurons was binned in 20 ms bins and analyzed using the computational method to suggest that $k_B T = 1$ was the critical point for the system. We similarly bin this data in 20 ms bins. We then consider subsets of 15, 20, 25, 30, 35 and 39 neurons of the 40 neuron sample. Next as suggested above, we make $N$ copies of each neuron in the subset. Finally, we take $N$ to infinity. We then evaluate the $h_i$ and $J_{ij}$'s using the method outlined above and evaluate the $C_{ij}$'s for different temperatures. A salient feature of the presence of long-range order is the divergence of the susceptibility. This would translate into the divergence of $\sum_{ij} C_{ij}$. The values of $k_B T$ where $\sum_{ij} C_{ij}$ diverges correspond to the temperatures where there is long-range order in the system. Because our aim is to test whether the dataset we are working with is critical or not, we would like to test whether $\sum_{ij} C_{ij}$ diverges at $k_B T = 1$. In fig. 1 we plot $k_B T$ versus $\sum_{ij} C_{ij}$ for different subsets. The first thing to observe is that different choices of the subsets lead to different values of $k_B T$ for which $\sum_{ij} C_{ij}$ diverges. Since by construction all expanded subsets have the same distribution of $(s_i)$ and $c_{ij}$, we are led to the implication that the temperature that corresponds to criticality is dependent on how the dataset is expanded and just keeping the same distribution of moments does not guarantee a unique critical temperature. The other thing to observe is that different choices of the subsets lead to different values of $k_B T$ for which $\sum_{ij} C_{ij}$ diverges. Since by construction all expanded subsets have the same distribution of $(s_i)$ and $c_{ij}$, we are led to the implication that the temperature that corresponds to criticality is dependent on how the dataset is expanded and just keeping the same distribution of moments does not guarantee a unique critical temperature. The other thing to observe is that $k_B T = 1$ does not correspond to $\sum_{ij} C_{ij}$ diverging in any choice of the data subsets. This is contrary to the conclusion of [7,8], that suggested that $k_B T = 1$ was the critical point.
In fig. 1 subsets of 15, 20, 25, 30, 35 and 39 neurons of the 40 neuron sample were randomly chosen, but it was also noticed that different ways of choosing a subsample of a particular size, gave different values of critical temperature with no correlation between the critical temperatures and the subsample size. This observation further enforces the conclusion that how the dataset is expanded decides the critical temperature. We hence see evidence of finite-size scaling arguments not working in this specific problem of evaluating the critical temperature as was conjectured above.

Conclusion

The fact that biological datasets are finite sized while criticality could only be ascertained in the large $N$ limit suggests that methods to expand these datasets are required. The computational method has been utilized in the literature to infer criticality of neuronal datasets. This method is based on finite-size scaling. This method involves expanding the neuronal data keeping the distribution of the $\langle s_i \rangle$ and $c_{ij}$ fixed. Next, it looks at how the peak of specific heat changes with system size to evaluate the critical temperature as is customary in a finite-size scaling analysis. However finite-size scaling assumes that except for the case of one coupling, the other couplings in the problem are at the critical value, which a priori is not justifiable in fitting the couplings to an expanded dataset. This implies that the computational method may not be justified in evaluating the critical temperature of a dataset. In order to test this conjecture, we developed an analytical method of expanding the dataset to the large $N$ limit keeping the distributions of the $\langle s_i \rangle$ and $c_{ij}$ fixed. We next analyzed a dataset from the literature which was conjectured to be critical using the computational method. We instead observed that the critical temperature was dependent on how the dataset was expanded. This implies that plainly suggesting a data expansion mechanism that conserves the distribution of $\langle s_i \rangle$ and $c_{ij}$, does not lead one to conclusively arrive at the critical nature of the dataset under consideration. Recently [13] have also shown that specific heat diverges whenever the average correlation strength does not depend on population size, when data with correlations is randomly subsampled during the analysis process. This is irrespective of the detailed structure or origin of correlations. Their findings suggest that specific heat divergence in the computational method of dataset expansion may not even imply criticality of the system from which the original dataset was obtained. These findings hence imply that plainly expanding datasets keeping moments conserved is not sufficient to understand the criticality of datasets and hence better methods have to be researched in order that biological criticality could be ascertained. Maybe, one has to go back to the drawing board and work with datasets which are known to be critical along with datasets that are non-critical and then work with different methods of expanding finite subsets of the same to see where the results of the analysis start to differ. Such an analysis could then possibly conclude the right method of expanding datasets to ascertain criticality. However, until such an analysis is explicitly carried out, this conclusion remains a conjecture.

We would like to thank Dr. Lukas Janssen for discussions on criticality in statistical systems. We would also like to greatly thank Prof. Karsten Kruse for suggesting changes in the manuscript to aid in the logical flow of ideas.

References

1. D. Schwab, I. Nemenman, P. Mehta, Phys. Rev. Lett. 113, 068102 (2014).
2. I. Mastromatteo, M. Marsili, J. Stat. Mech. 10, P10012 (2011).
3. T. Mora, S. Deny, O. Marre, Phys. Rev. Lett. 114, 07815 (2015).
4. M. Nykter, N.D. Price, M. Aldana, S.A. Ramsey, S.A. Kauffman, L.E. Hood, O. Yli-Harja, I. Shmulevich, Proc. Natl. Acad. Sci. U.S.A. 105, 1897 (2008).
5. R.V. Sol, S.C. Manrubia, M. Benton, S. Kauffman, P. Bak, Trends Ecol. Evol. 14, 156 (1999).
6. M.G. Kitzbichler, M.L. Smith, S.R. Christensen, E. Bullmore, PLoS Comput. Biol. 5, 1000314 (2009).
7. G. Tkačik, E. Schneidman, M.J. Berry II, W. Bialek, arXiv preprint q-bio/0611072 (2006).
8. E. Schneidman, M. Berry II, R. Segev, W. Bialek, Nature 440, 1007 (2006).
9. G. Tkačik, O. Marre, T. Mora, D. Amodel, M. Berry II, W. Bialek, J. Stat. Mech. 2013, P03011 (2013).
10. G. Tkačik, O. Marre, T. Mora, D. Amodel, M. Berry II, W. Bialek, Proc. Natl. Acad. Sci. U.S.A. 112, 11508 (2015).
11. G. Tkačik, O. Marre, E. Schneidman, D. Amodel, M. Berry II, W. Bialek, PLoS Comput. Biol. 10, e1003408 (2014).
12. John Cardy (Editor), Finite-Size Scaling, Vol. 2 (Elsevier, 2012).
13. M. Nonnenmacher, C. Behrens, P. Berens, M. Bethge, J. Macke, PLoS Comput. Biol. 13, e1005718 (2017).