Maximum-entropy distributions for a neuronal network from subnetwork data

P.G.L. Porta Mana
<pgl@portamana.org>

Y. Roudi
<yasser.roudi@ntnu.no>

V. Rostami
<vrostami@uni-koeln.de>

E. Torre
<torre@ibk.baug.ethz.ch>

Draft of 25 July 2019 (first drafted 4 November 2015)

This work shows how to build a maximum-entropy probabilistic model for the total activity of a network of neurons, given only some activity data or statistics – for example, empirical moments – of a *subnetwork* thereof. This kind of model is useful because neuronal recordings are always limited to a very small sample of a network of neurons. The model is applied to two sets of neuronal data available in the literature. In some cases it makes interesting forecasts about the larger network – for example, two low-regime modes in the frequency distribution for the total activity – that are not visible in the sample data or in maximum-entropy models applied only to the sample. For the two datasets, the maximum-entropy probability model applied only to the subnetwork is compared with the marginal probability distribution obtained from the maximum-entropy model applied to the full network. On a linear probability scale no large differences are visible, but on a logarithmic scale the two distributions show very different behaviours, especially in the tails.

Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

� comment about the possibility of drawing conclusions about a brain area using different sets of neurons (eg because of recording across many sessions)

1 Introduction

What correlations are important for the description of the multi-neuronal activity in a specific brain area? How does such activity change when external stimuli or experimental conditions change? Does such activity range over all its mathematically possible values, or only over a subset thereof?

Answering this kind of questions always engages an element of uncertainty. We can't say 'the answer is such and such', but can at best assign a degree of reasonable belief – that is, probability – to every possible answer. The assessment of this distribution of belief involves experimental data, such as recordings of neuronal activity from specific

brain areas, and pre-data knowledge about biological conditions and mechanisms. Our pre-data degrees of belief are often simplified to be mathematically more tractable, and are therefore called 'models'.

Despite remarkable advances in recording technologies, the best experimental measurements of neuronal activity can still only record a very small sample of neurons compared to the numbers that constitute a functionally distinguished brain region. Many probabilistic models focus on such samples only, somehow neglecting, in their assumptions, that the recorded neurons are a sample from a larger network. This kind of isolation assumptions sometimes escape attention, being subtly hidden in the mathematics. Some probabilistic models try to take also unrecorded neurons into account individually, and therefore become very complex 🧩 refs. It would be useful to have models that operate in between: addressing the larger brain area whence the sample comes, but without adding too much complexity and detail about it. Such intermediate models would be useful, for example, for preliminary investigations, to help us to decide which hypotheses to discard and which to consider for more complex and costly studies.

In the present work we propose such an intermediate model. It answers this question: How much can the *total* activity of a large neuronal network have been, given the observation of the activity of a very small sample thereof? This model addresses a larger brain area, avoiding the assumption of isolation of the sample; and by focusing on the total activity, rather than the activity of individual neurons, it remains simple and numerically tractable.

The proposed model is based on the straightforward combination of the maximum-entropy method and basic sampling relations from the probability calculus, discussed in § 2. The maximum-entropy or minimum-relative-entropy method (Jaynes 1957a; much clearer in Jaynes 1963; Sivia 2006; Hobson et al. 1973; Jaynes 1985; Grandy 1980) has been used for different kinds of estimations of the neuronal activity of various brain areas and about other phenomena of importance to the neurosciences, for example gene and protein interaction (for example Martignon et al. 1995; Bohte et al. 2000; Shlens et al. 2006; Schneidman et al. 2006; Tkačik et al. 2006; Macke et al. 2009; Tkačik et al. 2009; Roudi et al. 2009; Barreiro et al. 2010; Gerwinn et al. 2010; Macke et al. 2011; Ganmor et al. 2011; Cohen et al. 2011; Granot-Atedgi et al. 2013; Macke et al. 2013; Tkačik et al. 2014; Shimazaki et al. 2015; Mora et al. 2015;

Lezon et al. 2006; Weigt et al. 2009).

We illustrate possible uses of our proposed model in § 3, by applying it to two concrete data sets: (a) the activity of 65 neurons recorded for 20 min from a rat's Medial Entorhinal Cortex (Stensola et al. 2012), (b) the activity of 159 neurons recorded for 15 min from a macaque's Motor Cortex (Rostami et al. 2017). For each data set the model gives the most plausible frequency distribution with which all levels of total activity of a larger network of 1 000–10 000 neurons appeared during the recording. These guessed frequency distributions are distinctly different from the recorded ones for the sampled neurons. For the first data set, for example, the frequency distribution for the full network has two very distinct modes, both at low activities, see fig. 1, whereas the frequency distribution for the sample is monotonically decreasing with its maximum at zero activity. The frequency distribution for the second data set doesn't have two modes but still presents one prominent shoulder in its low-activity mode. Note that these guessed features of the full network cannot be inferred by the application of maximum-entropy *to the sample alone*. Although we don't formulate any hypotheses on why the larger network could have two distinct low-activity regimes – the purpose of the present work is only methodological – these results shows that the proposed method can lead to the formulation or preliminary assessment of interesting hypotheses.

Maximum-entropy models have often been used to test whether a smaller part of the sample data, for example time averages and second-order correlations, is sufficient for quantifying our degree of belief about some aspects of the neuronal activity of the sample. This use of maximum-entropy at the sample level, however, has an important issue. We discuss this issue and explain how our approach bypasses it in § 4. Moreover, our proposed model gives clearer answers to questions of correlation sufficiency: we show this by applying it to the two data set above addressing an example question of sufficiency, two moments versus four moments.

We obviously don't know whether the frequency distribution obtained with our approach is the *actual* one which the activity levels of the full network had during the recording; it's only the most plausible. In high dimensions, however, the features of the most plausible distribution may *not* be typical of the majority of most plausible distributions; and the set of all possible frequency distributions, if the full network for example

comprises 1 000 neurons, is a 1 000-dimensional space. In § 6 we therefore try to assess which features of the frequency distribution delivered by our method may be typical and therefore expected of the actual one. We find that general features such as the bimodality of the first data set are indeed typical. Maximum-entropy models can be considered as approximations of Bayesian models based on various assumptions of inferential sufficiency (Jaynes 1996; Porta Mana 2017a). How do our guesses change if we modify our pre-data assumptions? We show, in the same section, that the typical features indicated by our method are robust against such changes.

     ♟ Say something more about advantages of such question/answer: for example we can make statements about total activity of brain area even across recordings when sampled neurons aren't the same.

     ♟ Add this?: from sampling theory we know that important features of the full network may not be visible in a sample because smoothed out. But using sampling theory in the inverse direction we can infer such full-network features from the sample.

Our notation and terminology follow ɪꜱᴏ (1993; 2006a,b) standards and Jaynes (2003) for degrees of belief. We often simply say 'belief' for 'degree of belief'.

2   Model: maximum-entropy and sampling

Let's introduce some context and notation for our problem.

The context we consider is as follows. During an experimental session we have recorded the spiking activities of n neurons for a certain amount of time. These neurons are our 'sample' or 'subnetwork'. Their spikes are binned into T time bins and binarized to $\{0, 1\}$ values in each bin. Call a_t the number of neurons that fire during time bin t: this is the *total activity* of the sample, or just 'activity' for short. Obviously $a_t \in \{0, 1, \ldots, n\}$; if $a_t = 0$, no neuron spikes during bin t; if $a_t = n$, all spike at some point during bin t. For brevity, let's say 'at t' for 'during time bin t'. If we divide the total activity by the population size, $a/n \in \{0, 1/n, \ldots, 1\}$ we have the normalized total activity or population-averaged activity. From the activities $\{a_t\}$ we can count how often the activity levels $a = 0$, $a = 1$, and so on appeared during the recording, obtaining the distribution of measured relative frequencies $(f_a) =: f$. We can also consider the

sample activity at time bins *outside* of the recorded range. Such activity is unknown to us, of course.

♟ maybe move this paragraph to the intro?Present-day technologies enable the recording of neuronal activity from small brain areas ♟ be more specific. For many animal species, the neurons that are recorded within the area are not – and at present cannot be – specifically chosen from among the rest, owing to several limiting factors; for example, limitations in how precisely electrodes are inserted or neurons are targeted by viruses. In fact, the set of recorded neurons may even change during very long recordings or across experimental sessions. We assume that there's an area, comprising a network of N neurons, for which we believe that any other sample of size n could have equally plausibly been recorded instead of the sample of n neurons that was actually recorded. This is our 'full network'. ♟ how about calling it 'the pool'? The total activity of these N neurons at t is A_t. The relative frequencies of the various activity levels during the recording were $(F_A) =: F$. We don't know the values A_t at each t, or the frequency distribution F. We only know for certain that $A_t \in \{0, 1, \ldots, N\}$ and that $A_t \geqslant a_t$ for obvious reasons. For the time being we assume that we know N; in §*** we discuss the consequences of our lack of precise knowledge about this number.

Our questions concern general features of the total activity A of the full network during and after the recording, and across sessions under the same study conditions. For example: what was its frequency distribution during the recording? How much does this frequency distribution change across sessions? How much total activity should we expect at any time bin during a recording? We cannot answer these questions with certainty; we can only give distributions of probability or degrees of belief over their possible answers. The approach presented here gives such probability distributions.

♟ Move this to intro?We want to stress the usefulness of making a quantified guess about the full-network activity. First of all, this guess about a brain area seems to be the primary idea behind recording a sample from that area. Second, it allows us to make comparisons across experimental sessions; such comparisons would me difficult or meaningless if made with the recorded samples, which generally comprise non-overlapping sets of neurons and differ in size. Third, only at the full-network level can we meaningfully assess whether correlations

of some order are informationally sufficient: as explained in § 4, any such assessment at the sample level leads to conclusions opposite to the ones we think we're reaching.

The idea behind our approach is easily summarized:

(a) we build a distribution $p_{\mathrm{me}}(A)$ for the total activity of the full network using the maximum-entropy method;

(b) the constrained averages used in the maximum-entropy method for the full network are, in turn, determined via sampling theory from the constrained averages for the sample.

Let's discuss these points in detail.

Regarding (a), we assume that you're familiar with the maximum-entropy method. We actually use the *minimum-relative-entropy* method (Hobson et al. 1973), but call it 'maximum-entropy' for brevity. It amounts to a pair of prescriptions: choose the distribution, among those satisfying specific convex constraints, such as fixed expectations, that has minimum relative entropy with respect to a reference distribution, often taken to be the uniform one; and judge those expectations to be equal to measured averages. We add two remarks about this method that are seldom made in the literature. First remark: the distribution $p_{\mathrm{me}}(A)$ given by this method is the zeroth-order approximation (in the sense of Laplace's method: De Bruijn 1961 ch. 4; Tierney et al. 1986; Strawderman 2000) of four different distributions for the full network:

- the most probable *frequency* distribution for the total activity across the *recorded* bins,

- the *belief* distribution for the value of the total activity at any time bin among those *recorded*,

- the most probable *frequency* distribution for the total activity in a very long run of *new* time bins,

- the *belief* distribution for the value of the total activity at a *new* time bin.

The maximum-entropy distribution is thus an approximation of our belief distribution about four completely different quantities. Note that the four distributions above numerically differ in higher-order approximations. Second remark: the maximum-entropy method based on the Shannon entropy implicitly makes some assumptions about the probabilities for

the long-run frequency distributions (Jaynes 1996; Porta Mana 2009; 2017a). We say more about these two points in §***.

In our case, to apply the maximum-entropy prescriptions to the total activity of the full network we need to fix some averages of its belief distribution, for example the distribution's moments. But we don't have any measured moments for the full network to equate the distribution moments to. Here enters point (b): the probability calculus gives an exact, linear relation between the first m moments for the full network and the first m for the sample (Porta Mana et al. 2015 eqs (16)); the ones determine the others and vice versa at every time bin. This relation is a classical result of sampling theory (Whitworth 1965 chs I–IV; Feller 1968 ch. II; Jaynes 2003 ch. 3; see also Whitworth 1897).

Combining this result with the maximum-entropy prescription 'moments = measured moments' for the sample, we have that the measured moments for the sample determine the moments for the full network:

$$\overbrace{\text{measured moments} \rightarrow \text{sample moments}}^{\text{maximum-entropy prescription}} \underbrace{\rightarrow \text{full-network moments}}_{\text{sampling theory}}$$

These two steps are more straightforward if instead of power moments we use *normalized factorial moments* (Potts 1953). The mth normalized factorial moment of a distribution $p(a)$ for the activity of the sample neurons is defined as the average

$$\sum_a \binom{a}{m}\binom{n}{m}^{-1} p(a), \qquad 1 \leqslant m \leqslant n \tag{1}$$

This moment can be interpreted as the expectation of the number of distinct m-tuples of simultaneously spiking neurons (within a bin's time width), normalized by the number of distinct m-tuples. For example, with $m = 2$, if $na = 4$ neurons spike in a network of $n = 5$, we have $\binom{4}{2} = 6$ distinct pairs of simultaneously spiking neurons, and the total number of distinct pairs is $\binom{5}{2} = 10$. The normalized number of spiking pairs is therefore 6/10. Note that the first m factorial moments provide the same information as the first m power moments and vice versa: they linearly determine each other because $\binom{a}{m}$ is a polynomial in a of degree m. Fixing the ones is therefore equivalent to fixing the others. But the normalized factorial moments have this extremely convenient property:

the first n normalized factorial moments for a sample and for the full network are numerically identical:

$$\sum_a \binom{a}{m}\binom{n}{m}^{-1} p(a) = \sum_A \binom{A}{m}\binom{N}{m}^{-1} p(A), \qquad 1 \leqslant m \leqslant n, \qquad (2)$$

where $p(A)$ is our distribution of belief about the full-network activity.

We can therefore apply the maximum-entropy method to obtain a distribution for the full network by constraining its mth factorial moment to be equal to the sample's recorded average of $\binom{a}{m}\binom{n}{m}^{-1}$, for as many m as we please with $1 \leqslant m \leqslant n$. In formulae these constraints on $p_{\mathrm{me}}(A)$ are

$$\underbrace{\frac{1}{T}\sum_t \binom{a_t}{m}\binom{n}{m}^{-1} \equiv \sum_a \binom{a}{m}\binom{n}{m}^{-1} f_a}_{\text{measured moments}} = \underbrace{\sum_A \binom{A}{m}\binom{N}{m}^{-1} p_{\mathrm{me}}(A)}_{\text{distribution moments}} \qquad (3)$$

for all m we wish.

The calculation amounts to a convex optimization (Mead et al. 1984; Press et al. 2007 ch. 10; Fang et al. 1997; Boyd et al. 2009; Porta Mana 2017b) and for the numbers N, n, m considered in the present work it can be done on a modern computer without approximations of normalization constants or potential functions.

The number of moments used with this method depends on the questions and hypotheses that a researcher is exploring; for example, hypotheses of sufficient statistics, such as the sufficiency of pairwise correlations to quantify our degree of belief about network activity. We discuss this kind of use in § 4.

The particular case in which n moments are constrained is especially important: it corresponds to fully constraining the marginal frequency distribution for the activity of the sample neurons, f. In this case, our belief about the full-network activity is based on all available measured frequency data. Note that the application of the maximum-entropy method *at the sample level* is trivial and meaningless in this case – it just gives back the measured frequency distribution. But application of the method *at the level of the full network* is not trivial.

Using the full frequency distribution of the sample may be a bad idea, however, because the maximum-entropy distribution may become

a bad approximation of some of the four distributions described above. It is preferable to use a moderately high number of moments smaller than n. We explain this point in §***.

In the next section we apply the method just described to the data sets from two actual recordings, using six moments, and discuss the properties of the resulting distributions.

3   Application: two data sets

We apply the approach just described to two data sets publicly available in the literature:

• The **first data set**, from Stensola et al. (2012 rat 14147), consists of $n = 65$ neurons (27 of which classified as grid cells) from rat Medial Entorhinal Cortex, recorded for about 20 minutes. Their spikes are binned into $T = 417\,641$ bins of 3 ms width.

• The **second data set**, from Rostami et al. (2017), consists of $n = 159$ neurons from macaque Motor Cortex, recorded for about 15 minutes. Their spikes are binned into $T = 300\,394$ bins of 3 ms width.

For concreteness's sake we'll consider the maximum-entropy distribution $p_{\mathrm{me}}(A)$ as *the most probable frequency distribution* for the full-network activity during the recording; but remember that it is also the approximation of three other distributions, as discussed in the previous section.

We first calculate the distribution by using six moments. This number already provides almost as much information as the full frequency distribution of the sample, and at the same time illustrates the use of the approach in questions of statistic sufficiency (typically limited to two or three moments). Figure 1 shows the resulting densities (that is, distribution $\times N$) for full-network sizes $N = n$, $N = 1\,000$, $N = 5\,000$, $N = 10\,000$ 🧩 motivate?. The case $N = n$ corresponds to applying the maximum-entropy method at the sample level; it can be observed that with six moments it reproduces almost exactly the measured frequency distribution.

The distribution for the full-network is sharper than the measured frequency distribution for the sample; the sharper the larger N is. Most remarkably, it has two distinct low-activity modes for the first data set. But also the second data set presents, upon closer inspection, a
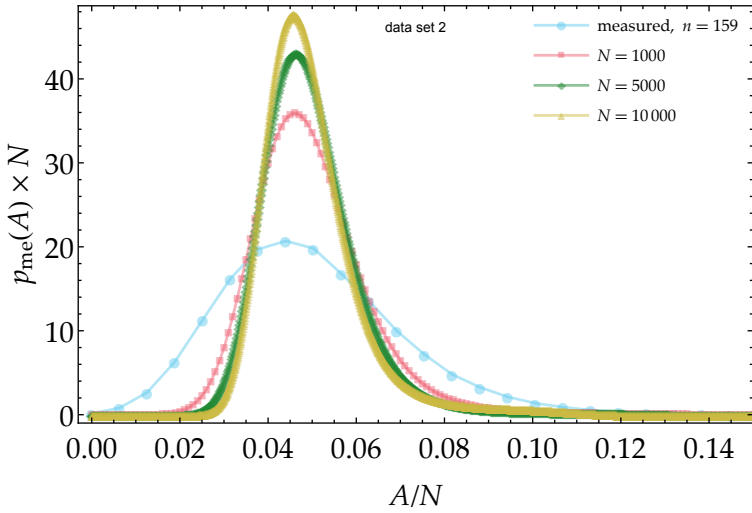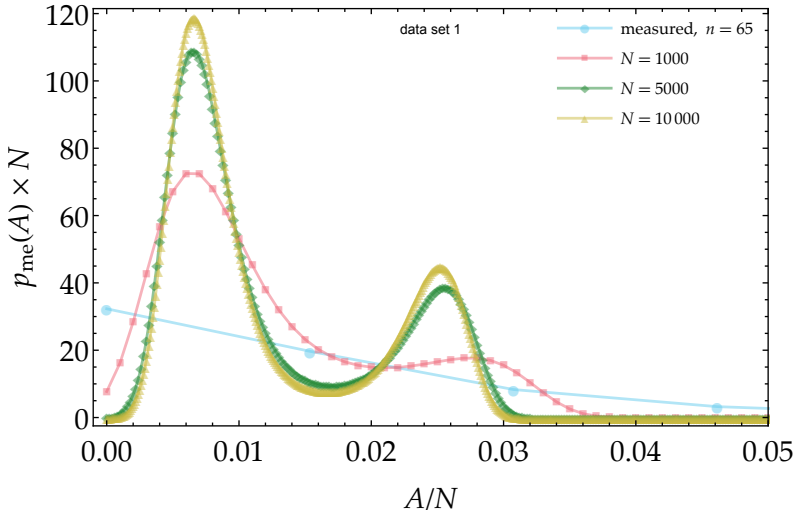
Figure 1   ***

small shoulder on the right of the mode, suggestive of two activity regimes. We frame no hypotheses about the biological cause of these two modes (they could stem from the presence of different kinds of cells or modules). These features are clearly not present in the sample or in the maximum-entropy distribution at the sample level. The application of the probability calculus thus reveals interesting possible features of the full network.

To illustrate how our approach can be applied to studies of sufficient statistics, fig. 2 shows the results for two constrained moments (equivalent to constraining means and correlations only) and for four constrained moments, with $N = 10\,000$. In either data set we obtain two very distinct distributions. For the first data set in particular, four moments lead to a bimodal distribution, whereas two to a unimodal one, showing that two moments would not be sufficient statistics. For comparison, the two distributions obtained applying maximum-entropy *at the sample level* are shown as an inset in the plot for the first data set: their differences aren't so glaring as in the full-network application.

4   Marginalization to the sample level:
the issue with pairwise correlations and other statistics

In the neurosciences the maximum-entropy method is often applied to questions of sufficient statistics; for example: do pairwise correlations in the neuronal activity provide the same information as the full frequency distribution of activities? In this section we discuss how our proposed application bears on this kind of questions.

Let's make the question more precise first. We want to assess our degree of belief about the frequencies of the activities of the sample, or about the activity of the sample in a new time bin. For this assessment we should use all measured data we have. But it sometimes happens that dropping part of the data – for example, the measured third- and higher-order moments – leaves our assessment almost unchanged: the dropped data are *informationally irrelevant*, or almost so. The remaining data – for example, first and second moments – are then *informationally sufficient*. This informational quasi-sufficiency is interesting because it may hint at peculiar biological mechanisms or dependencies.

We can approximately check how much a part of the data is irrelevant by simply dropping it from the conditional of our degree of belief,
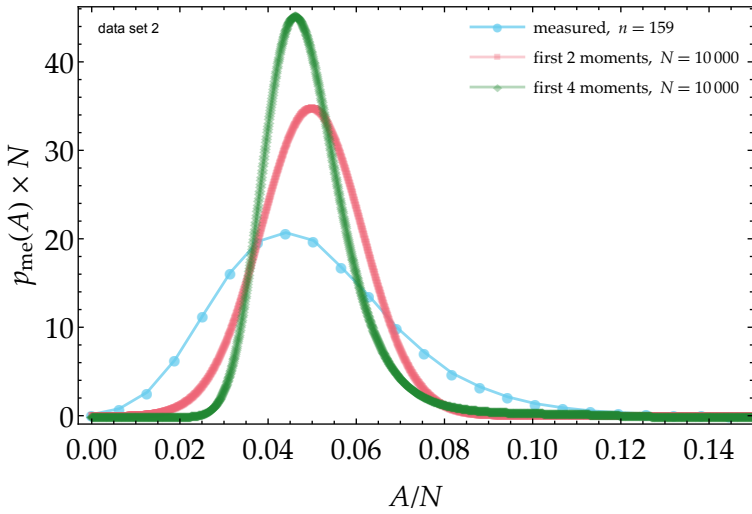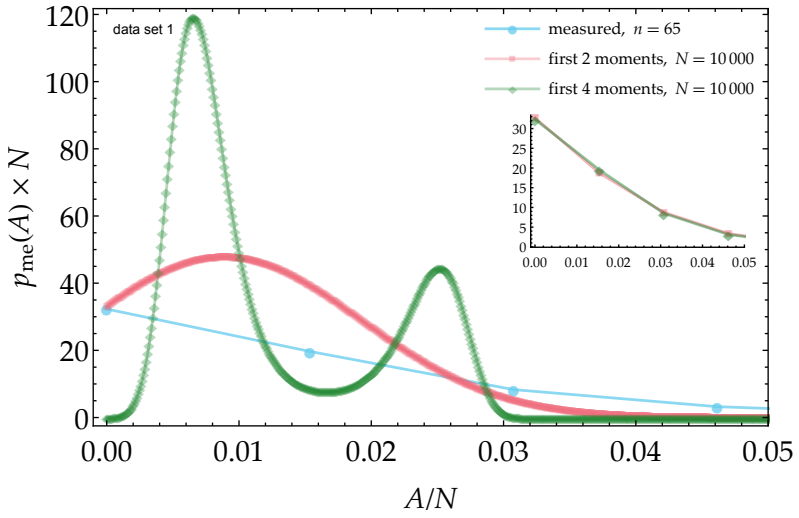
Figure 2   ***

and see if the latter changes appreciably. When our degree of belief is built with the maximum-entropy method, we simply drop part of the data used as constraints and see how the resulting distribution changes. For example, we may only retain the first two moments (means and correlations) or the first four moments.

In the neurosciences this check is often done with maximum-entropy distributions constructed *for the sample*, ignoring the full network it's sampled from. The method introduced in the present work, however, reveals an issue with this application at the sample level. The issue doesn't come from the method itself but from a general fact of probability theory. If a probability distribution has some sufficient statistics (Bernardo et al. 1994 § 4.5; Fortini et al. 2000), then its marginals *cannot* have the same sufficient statistics, and vice versa, except for trivial cases such as uniform distributions. This generally also holds among different marginals: if a marginal has some sufficient statistics, another won't have it. This impossibility is known in statistical mechanics: marginals of Gibbs states aren't Gibbs states (Maes et al. 1999). Mathematically this impossibility translates into a system of n independent equations in m unknowns with $n > m$ (Porta Mana et al. 2015 § 3.1; 2018 § 3).

In our example this means that if means and pairwise correlations or the first four moments are informationally sufficient for a particular sample from a brain area, then they *cannot* be sufficient for the full network of neurons in that area, or for a different sample. Vice versa, if we assume that they are informationally sufficient for the full network, then we must expect them *not* to be sufficient for any recorded sample thereof. Now, questions about sufficient statistics are meant to be addressed to a whole brain area, not just to some specific, casually recorded sample. Thus, paradoxically, maximum-entropy methods applied *at the sample level* give an answer opposite to the one we might think they're giving.

Questions about informational sufficiency are therefore correctly addressed by applying the maximum-entropy method at the full-network level. If a comparison with the measured frequencies of the sample activities is desired, then the marginal distribution to the sample size should be used.

But does the application at the full-network level lead to appreciably different results from that at the sample level? after all we're interested in an approximate informational sufficiency, not in an analytically exact one. A correct answer can only be given case by case. Figure 3 gives a graphical
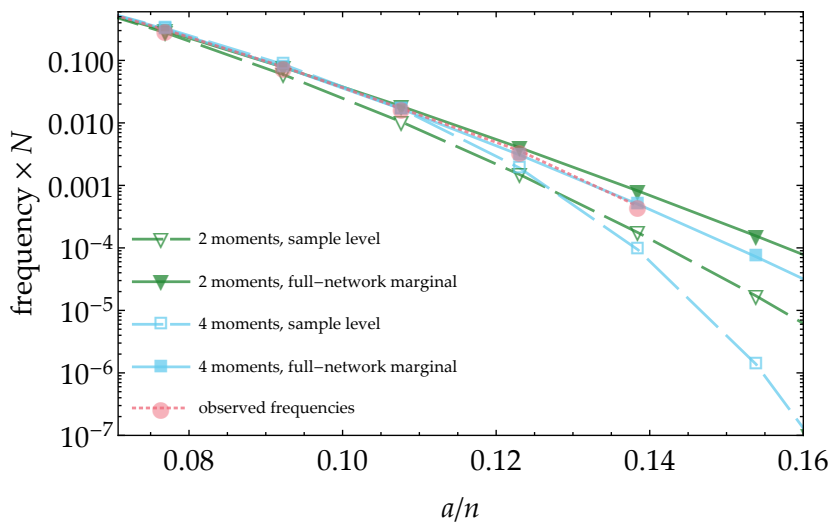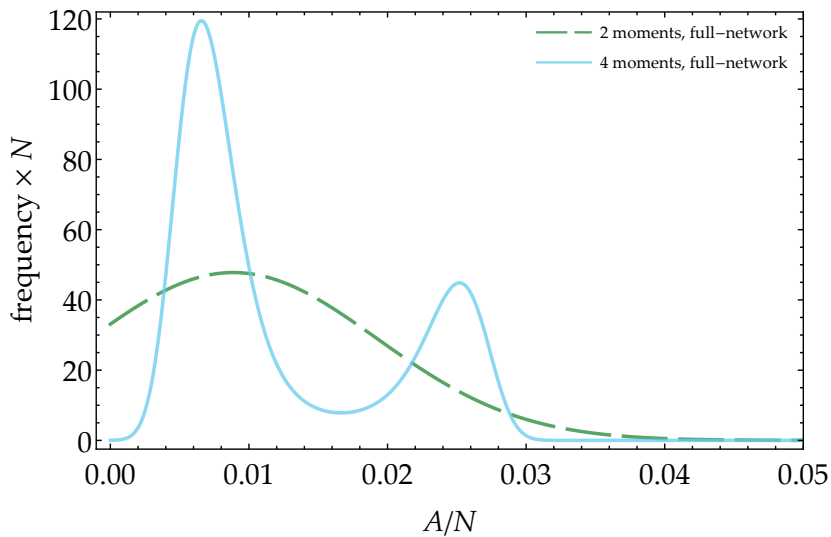
Figure 3   ***

answer to this question in the case of our first data set. The upper plot shows the maximum-entropy distributions for a full network of 10 000 neurons, constructed from two moments (– – –) and from four moments (——). The lower plot shows the maximum-entropy distributions for the sample, from two moments (▽) and from four moments (▢), and the distributions obtained by marginalizing the full-network distribution to the sample size, from two moments (▼) and from four moments (■). The measured frequency distribution (●) is also shown. We note the following:

- The application to the full-network (upper plot) leads to two completely different distributions (even different number of modes); clearly the two sets of statistics are not even approximately equivalent for inferential purposes.

- The application at the sample level leads to deceivingly similar distributions (▽ and ▢), which can only be distinguished on a logarithmic scale. This could lead to the erroneous conclusion that the two statistics are approximately equivalent.

- The difference between marginals from the full-network distribution and the distributions obtained at the sample level (filled vs empty markers) is larger than the difference between different statistics (triangles vs squares).

- The marginals of the application to the full network are closer to the measured frequencies than the distributions obtained at the sample level (although this closeness is not a valid criterion for their goodness).

Therefore, the maximum-entropy application at the sample level and at the full-network level lead to different results; the latter is not only more meaningful but also superior because it shows clearly the different informational sufficiency of the two sets of statistics.

5   The full-network size N

The calculations and conclusions presented in the preceding sections depend on the size of the full network, N. The full network can't be the whole brain, of course; that would be silly. How large can N be?

The crucial point is formula (2), on which our method is based, and which asserts the equality of the factorial moments of the full network

and its sample, (or, equivalently, asserting that the power moments for the one are linear functions of the power moments for the other). This formula is only valid if our beliefs about the activity of the sample $p(a)$ and that of the full network $p(A)$ are related by a hypergeometric distribution:

$$p(a) = \binom{n}{a} \binom{N-n}{A-a} \binom{N}{A}^{-1} p(A). \qquad (4)$$

This is a relation of 'drawing without replacement' (Jaynes 2003 ch. 3; Ross 2010 § 4.8.3; Feller 1968 § II.6). In other words, N must be such that we equally believe other samples of size n could have been recorded by the electrodes or the probe used. This is indeed what we stated in § 2.

This requirement delimits an area around the probe. Its number of neurons depends on the animal species and on the brain region from which the recording was made, as the neuron density can be very different. It's practically impossible to specify the exact number of neurons, but its order of magnitude is enough for making qualitative inferences. As we'll discuss in the next section, even if the exact number were known the maximum-entropy distribution ought still to be interpreted qualitatively.

It would be possible to derive a distribution $p(A)$ for the full network based on an unknown N, by expressing our belief distribution $p(N)$ about N and using this to marginalize N out:

$$p(A \mid N \text{ unknown}) = \sum_{N} p_{\text{me}}(A \mid N) \, p(N). \qquad (5)$$

But this would be overkill, owing to the qualitative character of the maximum-entropy distribution constructed with exact N.

6   Limitations and assumptions

In the study or use of the frequency distribution obtained with the procedure here presented we must take into account two important points.

The first point is that there are many possible frequency distributions which we believe, to different degrees, could be the true one that happened during the recording. The one given by our procedure is simply the one with the largest degree of belief, the mode of the belief distribution. The space of possible frequency distributions has many dimensions, however – thousands or tens of thousands. We must remember that belief distributions in high dimensions have counter-intuitive

properties. For example, the mode or mean can have *atypical* features when compared with the features of most other points of the space. The mode and mean can also be very different from each other.

The question, then, is *which features of the maximum-entropy frequency distribution are typical of the majority of plausible frequency distributions?* We can only answer for sure by using the full-fledged probability calculus. A more complete study (in preparation) with the first data set reveals that most of the plausible frequency distributions have three important features in common with the maximum-entropy one:

- all activities $A/N \gtrsim 5\%$ have practically zero frequencies;

- there are two regions of activity levels with high frequencies, roughly separated by a trough of lower frequencies.

- The frequencies of the region on the left ($A/N \lesssim 1.8\%$) are higher than those of the region on the right ($A/N \gtrsim 1.8\%$).

But there are also differences. For example, many plausible frequency distributions have three or four modes instead of just two; these modes are higher than those of the maximum-entropy distribution; and the bump of high frequencies on the right is slightly shifted towards lower activities than the corresponding maximum in the maximum-entropy distribution.

The second point is that our degrees of belief about the frequency distribution for the full network depend not only on the measured data in the sample, but also on our pre-data beliefs I about the distribution. Which assumptions lead to the maximum-entropy result? This distribution appears when our initial belief about the possible frequency distributions F is quantified by an entropic prior (Neumann 2007; Rodríguez 1991; Skilling 1998; Caticha et al. 2004; Porta Mana 2017a):

$$\mathrm{p}(F \mid I) \propto \exp[-L\,\mathrm{H}(F;R)] \approx \binom{L}{LF_0,\dots,LF_N} \prod_A R_A{}^{LF_A} \qquad (6)$$

where $\mathrm{H}(F;R) \coloneqq \sum_A F_A \ln(F_A/R_A)$ is the relative entropy or discrimination information (Kullback 1987; Jaynes 1963; Hobson 1969; Hobson et al. 1973), R is the reference distribution, and L a positive parameter. The approximate equality (obtained through Stirling's approximation), where the large parentheses denote a multinomial coefficient, shows that this prior belief is proportional to the number of ways in which the

distribution F can be realized in L time bins. The parameter L roughly quantifies how many time bins our data set must have to affect our initial belief. The maximum-entropy approximation is valid when L is large, but small compared to the sharpness of the constraints on F; in our case this means $L \approx 10$, give or take an order of magnitude.

We could obviously consider pre-data beliefs different from (6), for example one quantified by a Dirichlet distribution (which is equivalent to the above but with F and R switched), or a uniform distribution in F-space. Would these lead to markedly different post-data beliefs? A full-fledged probabilistic analysis shows that the three typical features listed above still appear with these different initial beliefs. They are therefore robust.

7 Summary and discussion

We have presented a procedure to construct the most plausible frequency distribution of population-averaged activities of a network of neurons, given the recording about a small sample thereof. This procedure combines the maximum-entropy method and basic identities from sampling theory. From the application to two real data sets we saw that the frequency distributions obtained with our procedure can have features very different from the one measured in the sample, such as multiple modes. This procedure can also be used with moment constraints of different order – means, population-averaged pairwise correlations, or higher-order correlations – thus giving an approximate assessment of the informational sufficiency of specific subsets of moments. In fact, we saw that the application of maximum-entropy only at the sample level leads to misleading results about this kind of sufficiency questions.

Thanks

Open Science Framework, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible.

Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

Barreiro, A. K., Gjorgjieva, J., Rieke, F. M., Shea-Brown, E. T. (2010): *When are microcircuits well-modeled by maximum entropy methods?* arXiv:1011.2797.

Bernardo, J.-M., Smith, A. F. (1994): *Bayesian Theory*. (Wiley, Chichester).

Bohte, S. M., Spekreijse, H., Roelfsema, P. R. (2000): *The effects of pair-wise and higher-order correlations on the firing rate of a postsynaptic neuron*. Neural Comp. **12**[1], 153–179.

Boyd, S., Vandenberghe, L. (2009): *Convex Optimization*, 7th printing with corrections. (Cambridge University Press, Cambridge). http://www.stanford.edu/~boyd/cvxbook/. First publ. 2004.

Caticha, A., Preuss, R. (2004): *Maximum entropy and Bayesian data analysis: entropic prior distributions*. Phys. Rev. E **70**[4], 046127.

Cohen, M. R., Kohn, A. (2011): *Measuring and interpreting neuronal correlations*. Nat. Neurosci. **14**[7], 811–819. http://marlenecohen.com/pubs/CohenKohn2011.pdf.

De Bruijn, N. G. (1961): *Asymptotic Methods in Analysis*, 2nd ed. (North-Holland, Amsterdam). First publ. 1958.

Erickson, G. J., Rychert, J. T., Smith, C. R., eds. (1998): *Maximum Entropy and Bayesian Methods*. (Springer, Dordrecht).

Fang, S.-C., Rajasekera, J. R., Tsao, H.-S. J. (1997): *Entropy Optimization and Mathematical Programming*, reprint. (Springer, New York).

Feller, W. (1968): *An Introduction to Probability Theory and Its Applications. Vol. I*, 3rd ed. (Wiley, New York). First publ. 1950.

Ford, K. W., ed. (1963): *Statistical Physics*. (Benjamin, New York).

Fortini, S., Ladelli, L., Regazzini, E. (2000): *Exchangeability, predictive distributions and parametric models*. Sankhyā A **62**[1], 86–109.

Ganmor, E., Segev, R., Schneidman, E. (2011): *Sparse low-order interaction network underlies a highly correlated and learnable neural population code*. Proc. Natl. Acad. Sci. (USA) **108**[23], 9679–9684. http://www.weizmann.ac.il/neurobiology/labs/schneidman/The_Schneidman_Lab/Publications.html.

Gerwinn, S., Macke, J. H., Bethge, M. (2010): *Bayesian inference for generalized linear models for spiking neurons*. Front. Comput. Neurosci. **4**, 12.

Grandy Jr., W. T. (1980): *Principle of maximum entropy and irreversible processes*. Phys. Rep. **62**[3], 175–266.

Granot-Atedgi, E., Tkačik, G., Segev, R., Schneidman, E. (2013): *Stimulus-dependent maximum entropy models of neural population codes*. PLoS Comput. Biol. **9**[3], e1002922.

Haken, H., ed. (1985): *Complex Systems – Operational Approaches: in Neurobiology, Physics, and Computers*. (Springer, Berlin).

Hobson, A. (1969): *A new theorem of information theory*. J. Stat. Phys. **1**[3], 383–391.

Hobson, A., Cheng, B.-K. (1973): *A comparison of the Shannon and Kullback information measures*. J. Stat. Phys. **7**[4], 301–310.

iso (1993): *Quantities and units*, 3rd ed. International Organization for Standardization.

— (2006a): *ISO 3534-1:2006: Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability*. International Organization for Standardization.

iso (2006b): *ISO 3534-2:2006: Statistics – Vocabulary and symbols – Part 2: Applied statistics*. International Organization for Standardization.

Jaynes, E. T. (1957a): *Information theory and statistical mechanics*. Phys. Rev. **106**[4], 620–630. http://bayes.wustl.edu/etj/node1.html, see also Jaynes (1957b).

— (1957b): *Information theory and statistical mechanics. II*. Phys. Rev. **108**[2], 171–190. http://bayes.wustl.edu/etj/node1.html, see also Jaynes (1957a).

— (1963): *Information theory and statistical mechanics*. In: Ford (1963), 181–218. Repr. in Jaynes (1989), ch. 4, 39–76. http://bayes.wustl.edu/etj/node1.html.

— (1985): *Macroscopic prediction*. In: Haken (1985), 254–269. Updated version 1996 at http://bayes.wustl.edu/etj/node1.html.

— (1989): *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, reprint. (Kluwer, Dordrecht). Edited by R. D. Rosenkrantz. First publ. 1983.

— (1996): *Monkeys, kangaroos, and N*. http://bayes.wustl.edu/etj/node1.html. First publ. 1986. (Errata: in equations (29)–(31), (33), (40), (44), (49) the commas should be replaced by gamma functions, and on p. 19 the value 0.915 should be replaced by 0.0915).

— (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Kullback, S. (1987): *The Kullback-Leibler distance*. American Statistician **41**[4], 340–341.

Lezon, T. R., Banavar, J. R., Cieplak, M., Maritan, A., Fedoroff, N. V. (2006): *Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns*. Proc. Natl. Acad. Sci. (USA) **103**[50], 19033–19038.

Macke, J. H., Buesing, L., Cunningham, J. P., Yu, B. M., Shenoy, K. V., Sahani, M. (2011): *Empirical models of spiking in neural populations*. Adv. Neural Information Processing Systems (NIPS) **24**, 1350–1358.

Macke, J. H., Murray, I., Latham, P. E. (2013): *Estimation bias in maximum entropy models*. Entropy **15**[8], 3109–3129.

Macke, J. H., Opper, M., Bethge, M. (2009): *The effect of pairwise neural correlations on global population statistics*. Tech. rep. 183. (Max-Planck-Institut für biologische Kybernetik, Tübingen). http://www.kyb.tuebingen.mpg.de/publications/attachments/MPIK-TR-183_%5B0%5D.pdf.

Maes, C., Redig, F., Van Moffaert, A. (1999): *The restriction of the Ising model to a layer*. J. Stat. Phys. **96**[1], 69–107.

Martignon, L., Von Hassein, H., Grün, S., Aertsen, A., Palm, G. (1995): *Detecting higher-order interactions among the spiking events in a group of neurons*. Biol. Cybern. **73**[1], 69–81.

Mead, L. R., Papanicolaou, N. (1984): *Maximum entropy in the problem of moments*. J. Math. Phys. **25**[8], 2404–2417. http://bayes.wustl.edu/Manual/MeadPapanicolaou.pdf.

Mora, T., Deny, S., Marre, O. (2015): *Dynamical criticality in the collective activity of a population of retinal neurons*. Phys. Rev. Lett. **114**[7], 078105.

Neumann, T. (2007): *Bayesian inference featuring entropic priors*. Am. Inst. Phys. Conf. Proc. **954**, 283–292. http://www.tilman-neumann.de/docs/BIEP.pdf.

Porta Mana, P. G. L. (2009): *On the relation between plausibility logic and the maximum-entropy principle: a numerical study*. arXiv:0911.2197. Presented as invited talk at the 31st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering 'MaxEnt 2011', Waterloo, Canada.

— (2017a): *Maximum-entropy from the probability calculus: exchangeability, sufficiency*. Open Science Framework doi:10.17605/osf.io/xdy72, arXiv:1706.02561.

Porta Mana, P. G. L. (2017b): *Geometry of maximum-entropy proofs: stationary points, convexity, Legendre transforms, exponential families*. Open Science Framework doi:10.17605/osf.io/vsq5n, arXiv:1707.00624.

Porta Mana, P. G. L., Rostami, V., Torre, E., Roudi, Y. (2018): *Maximum-entropy and representative samples of neuronal activity: a dilemma*. Open Science Framework doi:10.17605/osf.io/uz29n, bioRxiv doi:10.1101/329193, arXiv:1805.09084.

Porta Mana, P. G. L., Torre, E., Rostami, V. (2015): *Inferences from a network to a subnetwork and vice versa under an assumption of symmetry*. bioRxiv doi:10.1101/034199.

Potts, R. B. (1953): *Note on the factorial moments of standard distributions*. Aust. J. Phys. **6**^{4}, 498–499.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (2007): *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge). First publ. 1988.

Rodríguez, C. C. (1991): *Entropic priors*. http://omega.albany.edu:8008/.

Ross, S. (2010): *A First Course in Probability*, 8th ed. (Pearson, Upper Saddle River, USA). First publ. 1976.

Rostami, V., Porta Mana, P. G. L., Grün, S., Helias, M. (2017): *Bistability, non-ergodicity, and inhibition in pairwise maximum-entropy models*. PLoS Comput. Biol. **13**^{10}, e1005762. See also the slightly different version arXiv:1605.04740. Data available at https://doi.org/10.5061/dryad.n9f77.

Roudi, Y., Tyrcha, J., Hertz, J. (2009): *Ising model for neural data: model quality and approximate methods for extracting functional connectivity*. Phys. Rev. E **79**^{5}, 051915.

Schneidman, E., Berry II, M. J., Segev, R., Bialek, W. (2006): *Weak pairwise correlations imply strongly correlated network states in a neural population*. Nature **440**^{7087}, 1007–1012. http://www.weizmann.ac.il/neurobiology/labs/schneidman/The_Schneidman_Lab/Publications.html.

Shimazaki, H., Sadeghi, K., Ishikawa, T., Ikegaya, Y., Toyoizumi, T. (2015): *Simultaneous silence organizes structured higher-order interactions in neural populations*. Sci. Rep. **5**, 9821.

Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., Litke, A. M., Chichilnisky, E. J. (2006): *The structure of multi-neuron firing patterns in primate retina*. J. Neurosci. **26**^{32}, 8254–8266. See also correction in Shlens, Field, Gauthier, Grivich, Petrusca, Sher, Litke, Chichilnisky (2008).

— (2008): *Correction, the structure of multi-neuron firing patterns in primate retina*. J. Neurosci. **28**^{5}, 1246. See Shlens, Field, Gauthier, Grivich, Petrusca, Sher, Litke, Chichilnisky (2006).

Sivia, D. S. (2006): *Data Analysis: A Bayesian Tutorial*, 2nd ed. (Oxford University Press, Oxford). Written with J. Skilling. First publ. 1996.

Skilling, J. (1998): *Massive inference and maximum entropy*. In: Erickson, Rychert, Smith (1998), 1–14. http://www.maxent.co.uk/documents/massinf.pdf.

Stensola, H., Stensola, T., Solstad, T., Frøland, K., Moser, M.-B., Moser, E. I. (2012): *The entorhinal grid map is discretized*. Nature **492**^{7427}, 72–78. Data available at https://doi.org/10.11582/2018.00027.

Strawderman, R. L. (2000): *Higher-order asymptotic approximation: Laplace, saddlepoint, and related methods*. J. Am. Stat. Assoc. **95**^{452}, 1358–1364. http://stat.smmu.edu.cn/STONE/jasa/56_Higher-Order%20Asymptotic%20Approximation%20Laplace,%20Saddlepoint%20and%20Related%20Methods.pdf.

Tierney, L., Kadane, J. B. (1986): *Accurate approximations for posterior moments and marginal densities*. J. Am. Stat. Assoc. **81**^{393}, 82–86.

Tkačik, G., Mora, T., Marre, O., Amodei, D., Palmer, S. E., Berry II, M. J., Bialek, W. (2014): *Thermodynamics and signatures of criticality in a network of neurons*. Proc. Natl. Acad. Sci. (USA) **112**[37], 11508–11513.

Tkačik, G., Schneidman, E., Berry II, M. J., Bialek, W. (2006): *Ising models for networks of real neurons*. arXiv:q-bio/0611072.

— (2009): *Spin glass models for a network of real neurons*. arXiv:0912.5409.

Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., Hwa, T. (2009): *Identification of direct residue contacts in protein-protein interaction by message passing*. Proc. Natl. Acad. Sci. (USA) **106**[1], 67–72.

Whitworth, W. A. (1897): *DCC Exercises: Including Hints for the Solution of All the Questions in Choice and Chance*. (Deighton Bell & Co., Cambridge).

— (1965): *Choice and Chance: With One Thousand Exercises*, repr. of 5th ed. (Hafner, New York and London). First publ. 1867.