

Maximum-entropy distributions for a neuronal network from subnetwork data [draft]

P.G.L. Porta Mana

<pgl@portamana.org>

V. Rostami

<vrostami@uni-koeln.de>

Y. Roudi

<yasser.roudi@ntnu.no>

E. Torre

<torre@ibk.baug.ethz.ch>

Draft of 17 November 2019 (first drafted 4 November 2015)

✚ Abstract must be rewritten once paper is ready 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: a simple model for questions about large neuronal networks

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 instantaneous neuronal activity can still only record a very small sample of neurons – hundreds at most – 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  Yasser & Vahid: refs for here?. 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.

 [Yasser:] In the present work we aim to address the issue of building probability distributions over a large network using recording from subnetworks by considering the distribution of total population activity as an example. [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 model we propose 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 (jaynes1957jaynes1963sivia1996_r2006; hobsonetal1973; jaynes1985b_r1996; grandy1980) 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 (**martignonetal1995; bohteetal2000; shlensetal2006; schneidmanetal2006; tkaciketal2006; mackeetal2009b; tkaciketal2009; roudietal2009c; barreiroetal2010; gerwinnetal2010; mackeetal2011b; ganmoreetal2011; cohenetal2011; granotatedgietal2013; mackeetal2013; tkaciketal2014b; shimazakietal2015; moraetal2015; lezonetal2006; weigtetal2009**).

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 (**stensolaetal2012**), (b) the activity of 159 neurons recorded for 15 min from a macaque’s Motor Cortex (**rostamietal2016_r2017**). 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.

We want to stress the usefulness of making a quantified guess about the activity of a larger brain area. Such a guess seems indeed to be the primary idea behind recording a sample from that area. There’s also an important methodological reason. Maximum-entropy methods have been used to assess the informational sufficiency of pairwise correlations and correlations of higher order (**martignonetal1995; bohteetal2000; schneidmanetal2006; shlensetal2006; barreiroetal2010; ganmoreetal2011; granotatedgietal2013**). But when the difference in size between sample and full network is too large, correlation sufficiency for the sample implies the *lack* of correlation sufficiency for the larger

network, and vice versa. Therefore, maximum-entropy applications at the sample level can be deceptive for questions of statistical sufficiency, and a full-network application is more reliable. We discuss this point in detail in § 4.

How large is the full network addressed by the method proposed here? Its size must have some limit and can't obviously include the full brain. The size is determined by the validity of the formulae from sampling theory and discussed in § 5.

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 (jaynes1986d_r1996; portamana2017). 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.

A summary of all points above is given in the final § 7.

Our notation and terminology follow ISO (iso1993; iso2006; iso2006b) standards and Jaynes (jaynes1994_r2003) for degrees of belief. We often simply say 'belief' for 'degree of belief'.

2 The approach: 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 $a_t \in \{0, 1, \dots, n\}$; 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 we have the normalized total activity or population-averaged activity a/n , ranging from 0 to 1 in $1/n$ steps. 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 period. Such activity is unknown to us, of course.

For many animal species, the neurons that are recorded within a brain area are not 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. The set of recorded neurons may even change slightly across experimental sessions that are very far apart in time. 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. We call this larger network the 'full network'. 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 $A_t \in \{0, 1, \dots, N\}$ and that $A_t \geq a_t$ for obvious reasons. For the time being we assume that we know N ; in § 5 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? The approach presented here gives probability distributions over the possible answers to these questions.

The idea behind our approach is easily summarized:

- (a) we build a distribution $P_{\text{me}}(A)$ 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 familiarity with the maximum-entropy method. We actually use the *minimum-relative-entropy* method (hobsonetal1973), 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 $P_{\text{me}}(A)$ given by this method is the zeroth-order approximation (debruijn1958_r1961tierneyetal1986; strawderman2000) of four different distributions for the full network:

- (i) the most probable *frequency* distribution for the total activity across the *recorded* bins,
- (ii) the *belief* distribution for the value of the total activity at any time bin among those *recorded*,
- (iii) the most probable *frequency* distribution for the total activity in a very long run of *new* time bins,
- (iv) 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 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 (jaynes1986d_r1996; portamana2009; portamana2017). We discuss these two remarks further in § 6.

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 (portamanaetal2015); the ones determine the others and vice versa at every time bin. This relation is a classical result of sampling theory (whitworth1867_r1965feller1950_r1968jaynes1994_r2003whitworth1897).

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:

$$\overbrace{\text{measured moments} \rightarrow \text{sample moments} \rightarrow \text{full-network moments}}^{\text{maximum-entropy prescription}} \quad \underbrace{\hspace{10em}}_{\text{sampling theory}}$$

These two steps are more straightforward if instead of power moments we use *normalized factorial moments* (**potts1953**). The m th normalized factorial moment of a distribution $p(a)$ for the activity of the sample neurons is defined as the average

$$\sum_{a=0}^n \binom{a}{m} / \binom{n}{m} p(a), \quad 1 \leq m \leq n \quad (1)$$

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 $a = 4$ 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 . Specifying one set is therefore equivalent to specifying the other set. But the normalized factorial moments have a convenient property for our analysis (**portamanaetal2015potts1953**): *the first n normalized factorial moments for the sample and for the full network are numerically identical*:

$$\sum_{a=0}^n \binom{a}{m} / \binom{n}{m} p(a) = \sum_{A=0}^N \binom{A}{m} / \binom{N}{m} P(A), \quad 1 \leq m \leq n, \quad (2)$$

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 $P_{\text{me}}(A)$ for the full network by constraining its m th factorial moment to be equal to the sample’s recorded average of $\binom{a}{m} / \binom{n}{m}$, for as

many m as we please with $1 \leq m \leq n$. In formulae, the constraints on $P_{\text{me}}(A)$ are

$$\underbrace{\frac{1}{T} \sum_t \binom{a_t}{m} / \binom{n}{m}}_{\text{measured moments}} \equiv \sum_a \binom{a}{m} / \binom{n}{m} f_a = \underbrace{\sum_A \binom{A}{m} / \binom{N}{m}}_{\text{distribution moments}} P_{\text{me}}(A). \quad (3)$$

Finding the constrained maximum-entropy distribution $P_{\text{me}}(A)$ amounts to a convex optimization (**meadetal1984pressetal1988_r2007fangetal1997; boydetal2004_r2009; portamana2017b**) and for the numbers N, n, m considered in the present work it can be done on a modern computer without approximations of the partition functions that appear in the formalism.

The number of moments used with this method depends on the questions and hypotheses that a researcher is exploring; for example, hypotheses about the ‘cooperativity’ or ‘interaction’ – for example, pairwise or higher-order – of the 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 Example application: two data sets

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

- The **first data set**, from **stensolaetal2012**, 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 **rostamietal2016_r2017**, 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.

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 statistical sufficiency. Figure 1 shows the resulting densities (that is, $\text{distribution} \times N$) for four example values of full-network sizes: $N = n$, $N = 1\,000$, $N = 5\,000$, $N = 10\,000$. 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. We discuss the case of unknown N in § 5.

The figure shows that the distribution for the full-network is less flat than the measured frequency distribution for the sample; their difference increases with N . Most remarkably, for the first data set the distribution has two distinct low-activity modes. For the second data the distribution presents a small shoulder on the right of the mode. 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.

Use with moment constraints

To illustrate how our approach can be applied to studies of sufficient statistics, fig. 2 shows the full-network distributions, for $N = 10\,000$, obtained constraining the first two moments (equivalent to constraining means and correlations) and the first four moments. The frequency distribution of the sample is also shown for comparison. In both data

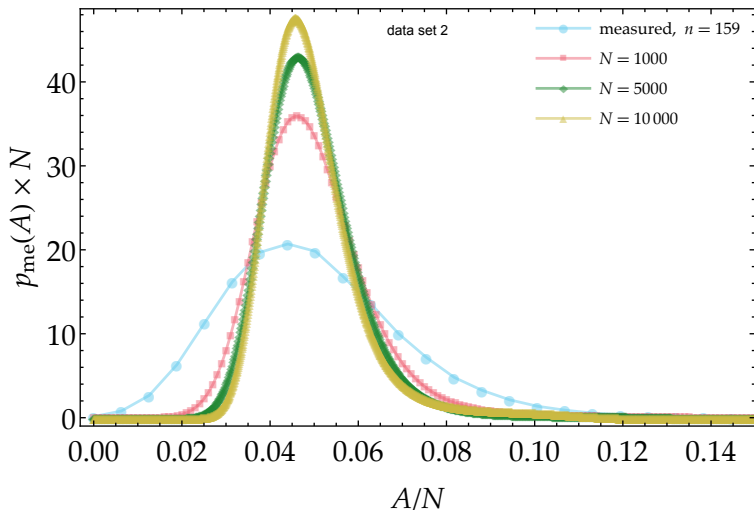
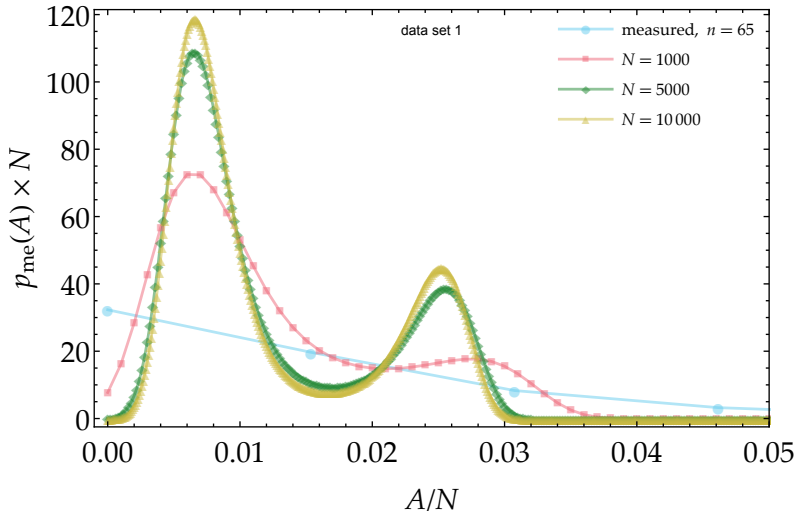


Figure 1 ***

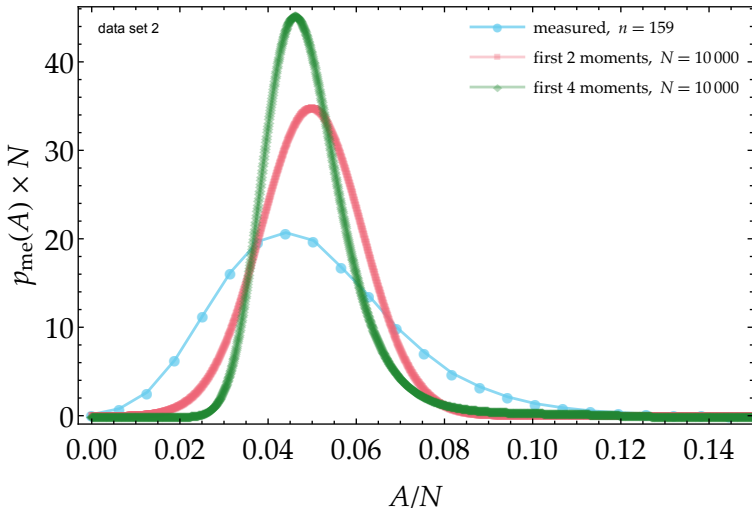
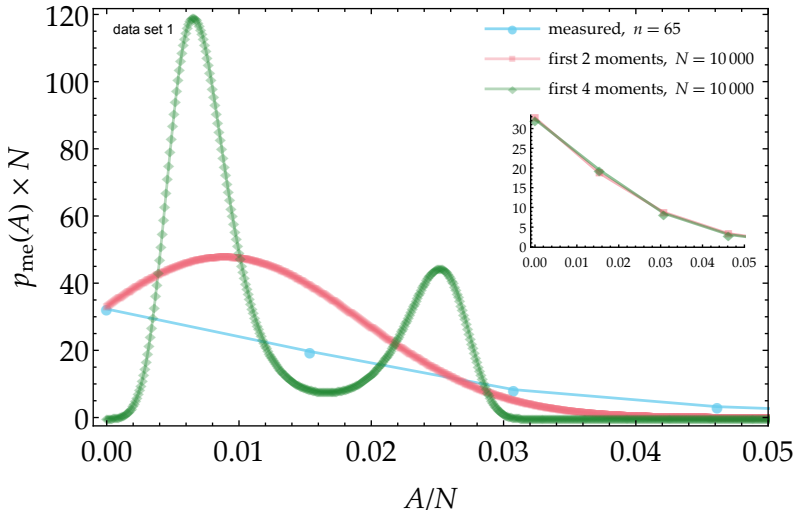


Figure 2 ***

sets the two-moment constraint leads to a distribution quite different from the four-moment constraint. For the first data set in particular, one distribution is bimodal, the other unimodal. This visual difference shows directly – without the need to calculate relative entropies, for example – that the first two moments alone are *not* informationally sufficient for this specific data set.

It is instructive to contrast the conclusions about correlation sufficiency reached through this full-network analysis, with those we would have reached by applying the maximum-entropy method *at the sample level*, as traditionally done. Such a comparison involves an important methodological caveat, so we discuss it in full in the next section.

4 Quantifying the importance of higher-order correlations: the limitations of methods at the sample level

As mentioned in the Introduction (see references there), in the neurosciences the maximum-entropy method has also been used as a way of quantifying the ‘cooperativity’ ([gersteinetal1985](#)) or ‘interaction’ ([martignonetal1995](#); [schneidmanetal2006](#); [shlensetal2006](#)) or ‘synchrony’ ([bohteetal2000](#); [amarietal2003](#)) of neuronal activity. In this section we discuss how our proposed application bears on this kind of quantification.

‘Cooperativity’, ‘interaction’, and similar terms are vague, so we need to translate them into a more precise notion first. Here we use the notion of *informational sufficiency* ([bernardoetal1994](#); [jaynes1994](#); [r2003](#); [cifarellietal1982](#); [kullbacketal1951](#); [fisher1922](#)) because it relates to those terms, is intuitive, and is connected with maximum-entropy distributions. Its idea is as follows. Our probabilities about the frequencies of the activities of the sample, or about the activity of the sample in a new time bin, are in principle conditional on all experimental data and statistics we have. But it can be the case that discarding part of the data or statistics – for example, the measured third- and higher-order moments – leaves our probabilities almost unchanged. This means that the discarded statistics are *informationally irrelevant* or almost so. The remaining statistics – for example, first and second moments – are *informationally sufficient*. (For more technical results and connections with the notion of symmetry see e.g. [darmois1935](#); [neyman1935](#);

koopman1936; pitman1936; halmosetal1949; bahadur1954; berk1972; lauritzen1974; cifarellietal1980; cifarellietal1981; diaconisetal1981; lauritzen1982_r1988; diaconis1992; furmanczyketal1998; fortinietal2000; nogalesetal2000; kallenberg2005; lauritzen2007; ayetal2015.)

There's a tight connection between informational sufficiency and maximum-entropy distributions (jaynes1982bbernardoetal1994): if a probability distribution for repetitive phenomena has a sufficient statistics, then by the Pitman-Koopman theorem (koopman1936; pitman1936; darmois1935hipp1974; andersen1970; denny1967; denny1972; fraser1963; barankinetal1963; barndorffnielsen1978_r2014) it is a mixture of exponential distributions of maximum-entropy form.

We can thus quantify the informational relevance of a subset of statistics, for example first and second moments (means and correlations), with respect to a larger set, for example the first four moments, by comparing the probabilities conditional on the subset and on the full set. For probabilities built with the maximum-entropy method, this means comparing those constrained on the subset and on the full set. This procedure is actually equivalent to comparing the probabilities of the two hypotheses about sufficiency, conditional on the *full* data, assuming their pre-data probabilities to be equal. We show this equivalence below and perform the calculations for our first data set.

Before applying this notion to neuronal activity, however, we must keep in mind that *informational sufficiency is not preserved under sampling*: if a probability distribution has some sufficient statistics, then its marginals, such as the distribution for a sample, *cannot* have the same sufficient statistics, and vice versa; except for trivial cases such as uniform probability distributions. This impossibility is known in statistical mechanics: if a system is described by a Gibbs state, its subsystems cannot be perfectly described by Gibbs states (maesetal1999). Mathematically this impossibility comes from the Pitman-Koopman theorem (koopman1936; pitman1936; darmois1935) connecting sufficient statistics and the exponential family of distributions, and translates into the general impossibility of solving a system of n independent equations in m unknowns with more equations than unknowns, $n > m$ (portamanaetal2015).

This fact is important for our analysis. If, say, means and pairwise correlations seem informationally sufficient for a particular sample from a brain area, then they may well not be sufficient for the full network of

neurons constituting that area, and vice versa. So if what interests us is ‘cooperativity’ or ‘interaction’ of a brain area, it is unreliable to use a maximum-entropy distribution constructed only for the sample. The approach presented here avoids this problem, because the maximum-entropy method is applied to obtain the distribution of the full network, not of the sample alone.

Let us illustrate the remarks above with our first data set.

We measure the difference $\Delta(M'', M')$ in informational sufficiency between a set of moments, say $M'' := \{1, \dots, m''\}$, and another, say $M' := \{1, \dots, m'\}$, as follows:

- (i) from each maximum-entropy distributions $P_{\text{me}}(A | M)$ for the full network, built from each set of constraints $M = M', M''$, calculate the marginal distribution for the sample:

$$p(a | M) = \sum_A G_{aA} P_{\text{me}}(A | M), \quad M = M', M''; \quad (4)$$

- (ii) calculate the relative entropies of the measured frequency distribution f with respect to each sample marginal, and multiply them by the number of time bins T :

$$T H[f; p(a | M)] := T \sum_a f_a \ln \frac{f_a}{p(a | M)}, \quad M = M', M''; \quad (5)$$

- (iii) take their difference:

$$\begin{aligned} \Delta(M'', M') &:= T H[f; p(a | M')] - T H[f; p(a | M'')] \equiv \\ &T \sum_a f_a \ln \frac{\sum_A G_{aA} P_{\text{me}}(A | M'')}{\sum_A G_{aA} P_{\text{me}}(A | M')}. \end{aligned} \quad (6)$$

The measure $\Delta(M'', M')$ so defined is positive if M'' is ‘more informationally sufficient’ than M' .

Why is this a natural measure? Because, if we assign equal pre-data probabilities to the two hypotheses M'' and M' , then the exponential of $\Delta(M'', M')$ is approximately equal to the ratio of their probabilities conditional on the data f – their Bayes factor:

$$\Delta(M'', M') \approx \ln[p(M'' | f)/p(M' | f)]. \quad (7)$$

So the exponential of $\Delta(M'', M')$ tells us how much more probable M'' is than M' , in view of the data f . We prove this in appendix**

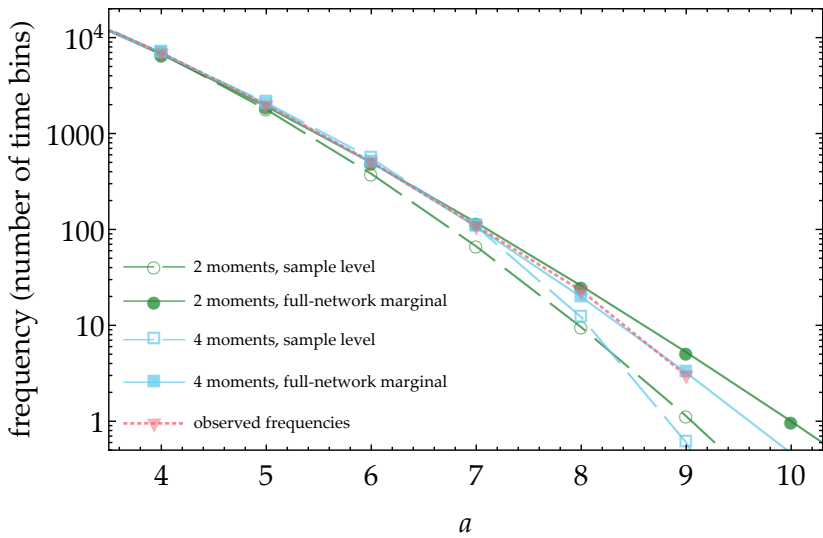
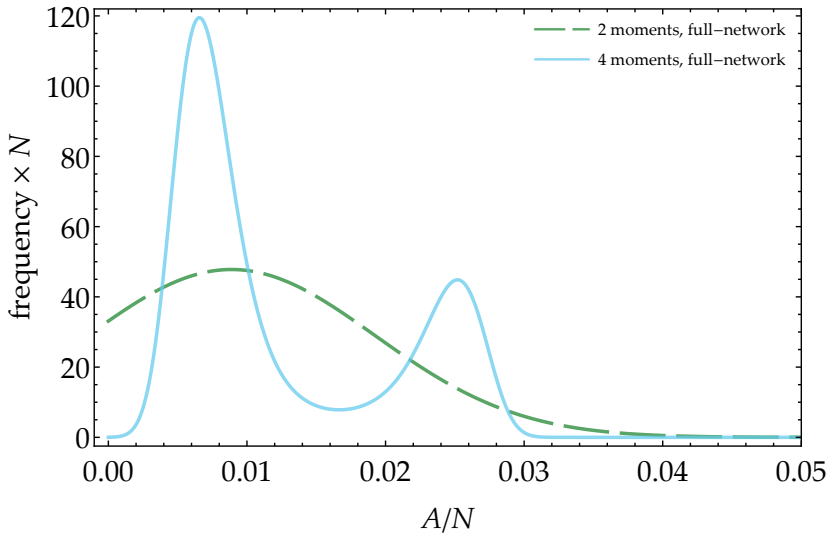


Figure 3 ***

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 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 deceptively 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. How large should or can N be in our formulae?

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:

$$p(a) = \binom{n}{a} \binom{N-n}{A-a} \binom{N}{A}^{-1} p(A). \quad (8)$$

This is a relation of ‘drawing without replacement’ (jaynes1994_r2003ross1976_r2010feller1950_r1968). 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 discussed in the next section, even if the exact number were known, the maximum-entropy distribution ought 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:

$$p(A \mid N \text{ unknown}) = \sum_N P_{\text{me}}(A \mid N) p(N). \quad (9)$$

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 (**neumann2007**; **rodriguez1991**; **skilling1998**; **catichaetal2004**; **portamana2017**):

$$p(F | I) \propto \exp[-L H(F; R)] \approx \left(\prod_{L F_0, \dots, L F_N}^L \right) \prod_A R_A^{L F_A} \quad (10)$$

where $H(F; R) := \sum_A F_A \ln(F_A/R_A)$ is the relative entropy or discrimination information (**kullback1987**; **jaynes1963**; **hobson1969**;

hobsonetal1973), 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 (14), 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

This work is financially supported by the Kavli Foundation and the Centre of Excellence scheme of the Research Council of Norway (Yasser Roudi group).

PGLPM thanks Mari, Miri, & Emma for continuous encouragement and affection; Buster Keaton and Saitama for filling life with awe and inspiration; the developers and maintainers of L^AT_EX, Emacs, AUC_TE_X, Open Science Framework, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible.

Let's ask how much more probable is the sufficiency of one set with respect to the other, conditional on our data f :

$$p(M'' | f)/p(M' | f). \quad (11)$$

Now, the probability of observing activity a in the sample at any time bin is the sample marginal of the maximum-entropy distribution for the full network, owing to the excheangeability assumption implicit in the maximum-entropy method:

$$p(a_t | M) = \sum_A G_{a_t A} P_{\text{me}}(A | M). \quad (12)$$

The probability of observing one sequence (a_t) with frequencies f is therefore

$$\prod_{t=1}^T p(a_t | M) \equiv \prod_{a=0}^n p(a | M)^{T f_a} \equiv \prod_{a=0}^n \left[\sum_A G_{aA} P_{\text{me}}(A | M) \right]^{T f_a}. \quad (13)$$

The probability of observing the frequencies f is obtained multiplying this by their multiplicity factor, the multinomial coefficient

$$\binom{T}{Tf} := \frac{T!}{\prod_a (T f_a)!} \approx \prod_a f_a^{-T f_a}, \quad (14)$$

the last expression coming from Stirling's approximation ([csiszaretal2004b](#)). If we assign equal pre-data probabilities to the two hypotheses M' and M'' , each probability in the ratio (7) then becomes, by Bayes's theorem,

$$p(M | f) \propto p(f | M) \times \text{const} \propto \binom{T}{Tf} \prod_a \left[\sum_A G_{aA} P_{\text{me}}(A | M) \right]^{T f_a} \approx \prod_a f_a^{-T f_a} \times \prod_a \left[\sum_A G_{aA} P_{\text{me}}(A | M) \right]^{T f_a}. \quad (15)$$

The logarithm of the probability above is easily seen to be the number of bins T multiplied by relative entropy between the frequency distribution f and the sample marginal of the maximum-entropy distribution.

Thus, the difference (6) is the logarithm of the probability ratio (7). The exponential of the difference (6) tells us how much more probable is the set M'' to be sufficient than the set M' .