Inferring the total activity of a large neuronal population from a small sample

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 22 December 2019 (first drafted 4 November 2015)

to be written

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

1 Introduction: a model for collective inferences about large neuronal populations

What correlations are dominant in the neuronal activity of a specific brain area? How does such activity change when external stimuli or the activity of other areas change? Does such activity range over all its mathematically possible values, or only within restricted bounds?

Answering this kind of questions always engages an element of uncertainty. We cannot say 'the answer is such and such'; at best we can assign degrees of reasonable belief – that is, probabilities – to the possible answers. The assessment of such distributions of belief involves experimental data, such as recordings of neuronal activity from specific brain areas, and pre-data knowledge and hypotheses about biological conditions and mechanisms. When we translate such pre-data beliefs into an initial probability distribution we often simplify them in more or less realistic ways, to make the analysis mathematically tractable. This is why such initial probability distributions are called 'models'; and that is how we intend this word in the present work.

The best experimental measurements of instantaneous neuronal activity use remarkable technologies[1], but 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: they neglect that the recorded neurons are a sample from a larger population. Some probabilistic models try to take unrecorded neurons into account, but they do so by describing

[1] Hong et al. 2019.

This document is designed for screen reading and two-up printing on A4 or Letter paper

each neuron individually, thus becoming very complex[2]. It would be useful to explore models that operate in between, addressing the activity of the larger brain area whence the sample comes, but *collectively*, that is, without asking about individual neuronal details. Such intermediate models would be useful at least for preliminary investigations, to help us decide which hypotheses to consider or discard for more complex and costly models, or to suggest new hypotheses.

In the present work we propose such an intermediate probabilistic model[3]. It answers this question: how much was the *total* activity of a large neuronal population, 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 can be viewed as a 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[4] to build initial probability distributions has been used in the neurosciences for different kinds of investigations about the neuronal activity of various brain areas and about other phenomena of importance, for example gene and protein interaction[5]. However, our model can also be viewed as an approximation of a more detailed model within Bayesian theory; we shall constantly keep this point of view in mind in our discussion.

We illustrate possible uses of our 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[6], (b) the activity of 159 neurons recorded for 15 min from a macaque's Motor Cortex[7]. For each data set the model gives us the most plausible frequencies of all levels of total activity of a much larger population, 1 000–10 000 neurons in size, during the recording. For example it can tell us that 60 out of 10 000 neurons were most likely active during 1% of the recording time (though not

[2] Pillow et al. 2008; Dunn et al. 2013; Tyrcha et al. 2014; Huang 2015; Battistin et al. 2017.
[3] cf. Porta Mana et al. 2015. [4] Jaynes 1957a; 2003 ch. 11; for the minimum-relative entropy method see: Jaynes 1963 § 4.b; Hobson et al. 1973 for a more rigorous derivation; Sivia 2006 for applications. [5] Martignon et al. 1995; Bohte et al. 2000; Shlens et al. 2006; Schneidman et al. 2006; Tkačik et al. 2006; Roudi et al. 2009; Barreiro et al. 2010; Shimazaki et al. 2015; Mora et al. 2015; Lezon et al. 2006; Weigt et al. 2009; for further references see Latham et al. 2013. [6] Stensola et al. 2012. [7] Rostami et al. 2017.

necessarily always the same 60), 250 neurons out of 10 000 were active during 0.4% of the recording time, and so on. The precise meaning of this frequency distribution is explained in § 2. For the two example data sets, the guessed frequency distributions for the larger population are distinctly different from those for the sample. For example, for the first data set the frequency distribution for the larger population 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. These results show that the proposed model can be used for the formulation or preliminary assessment of interesting hypotheses. We illustrate this use with toy examples in §§ 3 and 4. Note that it is not possible to guess these features of the larger population by applying the maximum-entropy method *at the level of the sample alone*.

Our model also solves a methodological problem in the use of maximum-entropy methods to assess the 'cooperativity', 'interaction', or 'synchrony' in neuronal activity, for example studying its pairwise correlations and correlations of higher order[8]. When the difference in size between a large population and a sample thereof is too large, the sufficiency of correlations of some order for the sample implies the *lack* of sufficiency of correlations of the same order for the larger population, and vice versa. Maximum-entropy applications at the sample level can therefore deceive us in questions regarding the cooperativity of the larger population. Our model addresses directly the latter instead. We explain this point more precisely in § 4.

How large is the 'larger population' considered by the model proposed here? It obviously cannot include the full brain. The size of the larger population is determined by the validity of the formulae from sampling theory, and can range from a set of neurons around the recording probe to part of a brain area, depending on some assumptions about the recording procedure and the brain region under consideration. This matter is discussed in detail in § 5.

In § 6 we discuss in detail the assumptions and approximations which define our model, from the point of view of the probability-calculus.

A summary of all points above and a discussion of the usefulness of the model is given in the final § 7.

[8] see for example Martignon et al. 1995; Bohte et al. 2000; Schneidman et al. 2006; Shlens et al. 2006; Barreiro et al. 2010; Ganmor et al. 2011; Granot-Atedgi et al. 2013.

Our notation and terminology follow ISO (1993; 2006a,b) standards and Jaynes (2003) for probability. We use 'degree of belief', 'belief', and 'probability' interchangeably.

2 Model: maximum-entropy and sampling, and a Bayesian perspective

Some context and mathematical notation first.

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'. 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, and so on. For brevity we say 'at t' for 'during time bin t'. 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 $(f_a) =: \boldsymbol{f}$. We can also consider the (unknown) sample activity at time bins *outside* of the recorded period.

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. The sample 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 population 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 what we will mean with 'larger population'. This population need not be a whole functionally or anatomically distinct region. Loosely speaking it is the area of which we believe our sample to be 'representative'[9].

The total activity of the N neurons at t is A_t. The relative frequencies of the various activity levels during the recording were $(F_A) =: \boldsymbol{F}$. The values A_t at each t and the frequency distribution $\boldsymbol{F}$ are unknown to us. We only know for certain that $A_t \in \{0, 1, \dots, N\}$, that $A_t \geqslant a_t$, and that

[9] we intend this term in the Meaning 2 of Kruskal et al. 1979a; or Meanings 8 and 9 of Kruskal et al. 1979b.

$N - A_t \geqslant n - a_t$ 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 larger population during and after the recording, and across sessions under the same study conditions. For example: what was its frequency distribution $\boldsymbol{F}$ 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 model presented here gives a probability distribution over the answers to these questions.

The idea behind our approach can be easily summarized in terms of the maximum-entropy method:

(a) Using sampling theory we determine the relation between some expected values – specifically, moments – for the total activity a of the sample and corresponding expected values for the total activity A of the larger population.

(b) Using the maximum-entropy method we build a distribution $P_{\mathrm{me}}(A \mid M, N)$ for the total activity of the larger population of N neurons, using as constraints the expected values M found in the previous step.

We now discuss these steps more in detail, but leave their precise mathematical implementation and a more detailed list of references to appendix A.

Step (a) is just an application of the probability-calculus, which gives an exact linear relation between the first m moments for the larger population and the first m for the sample[10]. The ones determine the others and vice versa at every time bin. This relation holds for any belief distribution $P(A_t)$ for the larger-population activity and its marginal $p(a_t)$ for the sample activity at that bin. These two distributions are related by

$$p(a_t) = \sum_{A_t=0}^{N} G_{a_t A_t}\, P(A_t) \qquad a_t \in \{0, \dots, n\}, \tag{1}$$

where G_{aA} is the hypergeometric distribution defined in eq. (18), characteristic of 'drawing without replacement'[11].

[10] see Porta Mana et al. 2015 for a complete discussion and proofs. Some proofs are summarized again here in appendix A. [11] Jaynes 2003 ch. 3.

This sampling relation is even more straightforward if instead of power moments we use *normalized factorial moments*[12]. The mth normalized factorial moment of a distribution $p(a)$ is

$$\sum_{a=0}^{n} \binom{a}{m} \Big/ \binom{n}{m} p(a) \qquad m \in \{1, \dots, n\}, \tag{2}$$

that is, the expected number of distinct m-tuples of simultaneously active neurons (within a time-bin's width), normalized by the maximum possible number of distinct m-tuples. Note that the first m factorial moments together provide the same information as the first m power moments together, and vice versa: they are linearly related because $\binom{a}{m}$ is a polynomial in a of degree m. So we'll just say 'first m moments' from now on. But the normalized factorial moments have a mathematically convenient property: *the first n normalized factorial moments for the sample and for the larger population are numerically identical*:[13]

$$\sum_{a=0}^{n} \binom{a}{m} \Big/ \binom{n}{m} p(a) = \sum_{A=0}^{N} \binom{A}{m} \Big/ \binom{N}{m} P(A), \qquad m \in \{1, \dots, n\}. \tag{3}$$

In step (b) we actually use the *minimum-relative-entropy* method[14] with respect to a uniform reference distribution. We shall still call it 'maximum-entropy' for brevity. It amounts to two prescriptions: first, take the distributions satisfying specific convex constraints – such as fixed expectations – and among them select the one having minimum relative entropy with respect to a reference distribution; second, judge – and therefore set – those expectations to be equal to some measured averages, typically time averages. (A judgement is required because probability expectations and empirical averages differ numerically in general; a misjudgement of their approximate equality can lead to 'silly answers', to quote MacKay[15].)

In our case we don't know the time averages of the quantity $\binom{A}{m}$, so we cannot directly equate them to the factorial moment $\mathrm{E}\big[\binom{A}{m}\big]$ of the larger-population distribution $P(A)$. But eq. (3) of step (a) comes to our rescue, because it says that the expectation for the larger population are determined by that for the sample $\mathrm{E}\big[\binom{a}{m}\big]$, and we do have the time

[12] Potts 1953. [13] Porta Mana et al. 2015 Appendix. [14] Hobson et al. 1973; Csiszár 1985; Sivia 2006 § 5.2.2. [15] MacKay 2003 p. 308.

average of its corresponding sample quantity $\binom{a}{m}$. So we can combine the two steps:

$$\overbrace{\text{measured averages} \to \text{sample moments}}^{\textit{maximum-entropy prescription}} \to \underbrace{\text{sample moments} \to \text{larger-population moments}}_{\textit{sampling theory}}$$

We obtain a distribution $P_{\mathrm{me}}(A \mid M, N)$ for the larger population of N neurons by constraining some of its factorial moments, for example $M = \{1, \dots, m'\}$ with $m' \leqslant n$, to be equal to the sample's recorded averages. In formulae, the constraints on $P(A)$ are

$$\underbrace{\frac{1}{T}\sum_t \binom{a_t}{m}\Big/\binom{n}{m} \equiv \sum_a \binom{a}{m}\Big/\binom{n}{m} f_a}_{\text{measured averages}} \quad = \quad \underbrace{\sum_A \binom{A}{m}\Big/\binom{N}{m} P(A)}_{\text{distribution moments}} \qquad m \in M. \tag{4}$$

The result is the distribution of the form

$$\begin{aligned} P_{\mathrm{me}}(A \mid M, N) &= \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[\sum_m \lambda_m \binom{A}{m}\Big/\binom{N}{m}\right] \\ \text{with} \qquad Z(\boldsymbol{\lambda}) &:= \sum_A \exp\left[\sum_m \lambda_m \binom{A}{m}\Big/\binom{N}{m}\right], \end{aligned} \tag{5}$$

where the m' parameters $\boldsymbol{\lambda} := (\lambda_m)$ are determined by the constraints (4) and can be found by convex extremization, discussed further in appendix A. Note that for $N > n$ the constraint equation (4) may never be exactly satisfied by the distribution (5) for any $\boldsymbol{\lambda}$. In this case the extremization minimizes the discrepancy between the left and right sides of (4). This discrepancy comes from an approximation implicit in the maximum-entropy method (namely that the number of time bins T is infinite) combined with one assumption specific to our application: that an activity level A can equally likely be generated by any set of A neurons in the larger population. The magnitude of this discrepancy can be a signature of the presence of neuronal 'assemblies' or 'clusters'[16]. We discuss this matter in § 6.

One important question remains about the distribution P_{me} of eq. (5): *what* is it a distribution of? – Of probability? Of relative frequency? The

[16] Gerstner et al. 2014 ch. 12; Hebb 2002.

answer is related to an important relation between maximum-entropy distributions and the probability-calculus, and also to our alternative point of view of the present model.

A maximum-entropy distribution like $P_{\mathrm{me}}(A \mid M, N)$ is equivalent to the zeroth-order Laplace approximation[17] of four distinct distributions for the larger population, which differ numerically from one another in higher-order approximations:

(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 the *recorded* ones,

(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.

Their mathematical relationship is presented more explicitly in appendix C. The distribution (5) can therefore be interpreted in each of the four ways above, in this approximate sense. In the present case the validity of the approximation decreases as the ratio nN/T increases. The most robust interpretation is (i), and that is how we prefer to interpret the distribution P_{me}: as *the most probable relative-frequency distribution* of the activity A. See § 6 for further discussion about assumptions and approximations.

3 Example application: two data sets

We apply the model 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 from rat Medial Entorhinal Cortex, simultaneously 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 supplementary material), consists of $n = 159$ neurons from macaque Motor Cortex, simultaneously recorded for about 15 minutes. Their spikes are binned into $T = 300\,394$ bins of 3 ms width.

[17] De Bruijn 1961 ch. 4; Tierney et al. 1986; Strawderman 2000.

The maximum-entropy distribution for the larger population is calculated using five moments, because this number provides almost as much information as the full frequency distribution of the sample, as discussed in § 4. Figure 1 shows the resulting distributions for three example values of larger-population sizes: $N = 1\,000$ (green diamonds), $N = 5\,000$ (red circles), $N = 10\,000$ (blue curve). The frequency density of the sample activity is also shown (black triangles), and in the plot it would be indistinguishable from the maximum-entropy density for $N = n$, that is, applied at the sample level. We discuss the case of unknown N in § 5.

The most expensive calculation, for $N = 10\,000$, takes less than 15 minutes on a laptop with two 2 GHz cores. In all cases the moments were recovered with relative errors smaller than 10^{-12}.

The figure shows that the distribution for the larger-population is more peaked than the measured frequency distribution for the sample; their difference increases with N. Most remarkably, for the first data set, fig. 1(a), the distribution for the larger population has two distinct low-activity modes. For the second data set, fig. 1(b), the distribution presents a small shoulder on the right of its mode. Such features are clearly not present in the sample frequencies or in the maximum-entropy distribution at the sample level. The model thus suggests interesting features of the larger population.

Here is a toy example of possible uses of this model, based on the first data set. We could be interested in the hypothesis that two distinct cell types or assemblies be present in the region where the recording was made. Finding a larger-population distribution with two peaks, as in fig. 1(a), would provide some evidence for this hypothesis. We might have reasons for suspecting that a specific set of the sampled neurons is of the first type, and the remaining of the second type. In the case of the first data set, 27 of the 65 sampled cells were identified as grid cells belonging to 3–4 functional modules, the remaining 38 'non-grid' cells being interneurons and unidentified cells[18]. Could the two peaks in the distribution of fig. 1 reflect the activities of grid versus non-grid cells? We apply the model to these two sets of neurons individually, using $N_{\mathrm{g}} = 4\,150$ for the grid set and $N_{\mathrm{ng}} = 5\,850$ for the non-grid set, to reflect their proportions (27/65 and 38/65) in the recorded sample. The

[18] Stensola et al. 2012.

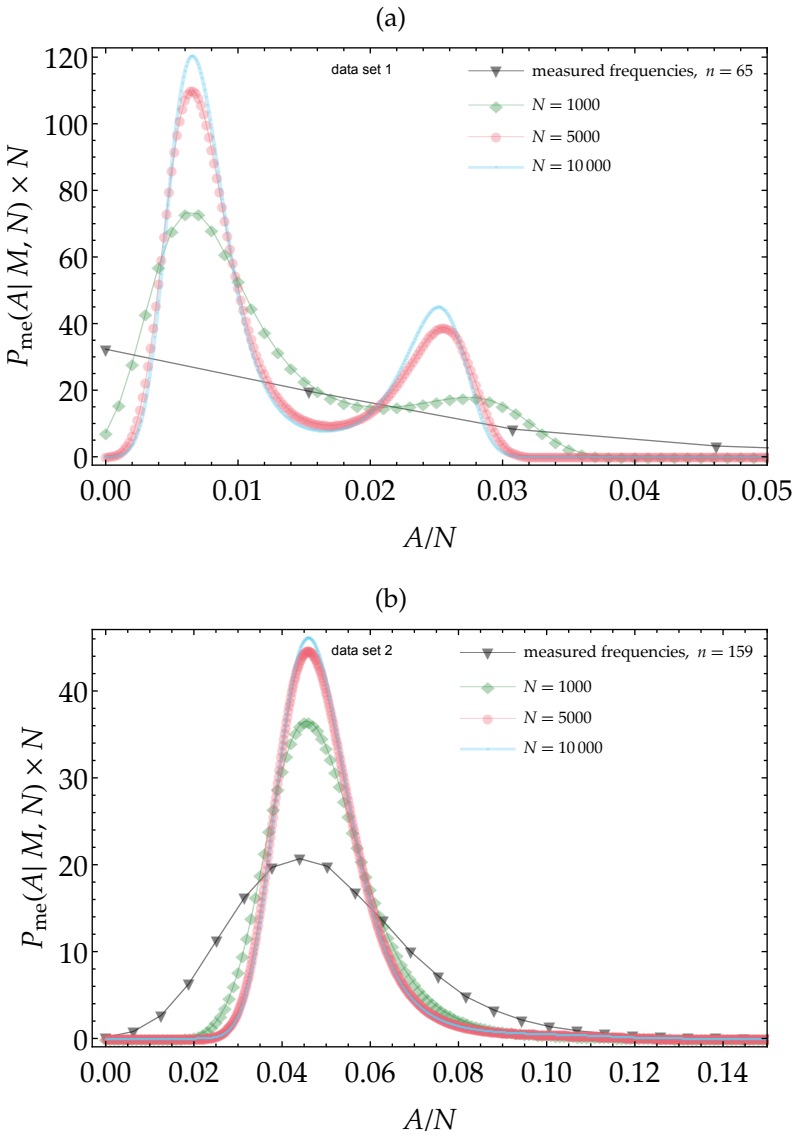

Figure 1 Maximum-entropy distributions for the total activity of the larger population, using five moments and assuming several population sizes N. In order to compare different N, the distributions are multiplied by N (obtaining densities) and the activity divided by N (obtaining a population average). The distributions for large N have very different features from the measured frequency distribution, and how their peaks become more pronounced as N increases.

results are shown in fig. 2(a). The distribution for the larger population of grid cells (green triangles) seems to have one broad peak, close to the first peak of the larger-population distribution. The distribution for the larger population of non-grid cells (red circles) has two peaks instead, roughly at the same population-averaged activities as the peaks of the larger-population distribution (blue curve) but closer in height. It would thus seem that the population of grid cells is contributing to the first mode of the larger population, but it is not its sole contributor. We can also assess if the distributions for the two sets of neurons are independent. If they were independent, the larger-population distribution would be given by their convolution:

$$P_{\text{full}}(A) = \sum_{A'} P_{\text{grid}}(A')\, P_{\text{non-grid}}(A - A') \tag{6}$$

where the index A' runs from $\max(0, A - N_{\text{ng}})$ to $\min(A, N_{\text{g}})$. But this is not the case: figure 2(b) shows that such convolution (black diamonds) is quite different from the larger-population distribution (blue curve): the two peaks of the former are closer in position and height than those of the latter. The population distributions of grid and non-grid cells are therefore *not* independent: knowledge of the activity of either set gives us some information about the activity of the other.

The toy analysis above should not be taken literally, but just as an illustration of the model's possible applications. The important point is that this model is computationally very cheap and yet it can provide useful insights, even if just qualitative ones, and even suggest new hypotheses.

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

As mentioned in the Introduction, in the neurosciences the maximum-entropy method has also been used to quantify the 'cooperativity'[19] or 'interaction'[20] or 'synchrony'[21] of neuronal activity. Most, if not all, such applications use the method *at the sample level*, that is, they find a distribution for the sample without considering that it comes from a larger population. In this section we discuss how our proposed application to

[19] e.g. Gerstein et al. 1985. [20] e.g. Martignon et al. 1995; Schneidman et al. 2006; Shlens et al. 2006. [21] Bohte et al. 2000; Amari et al. 2003 e.g.

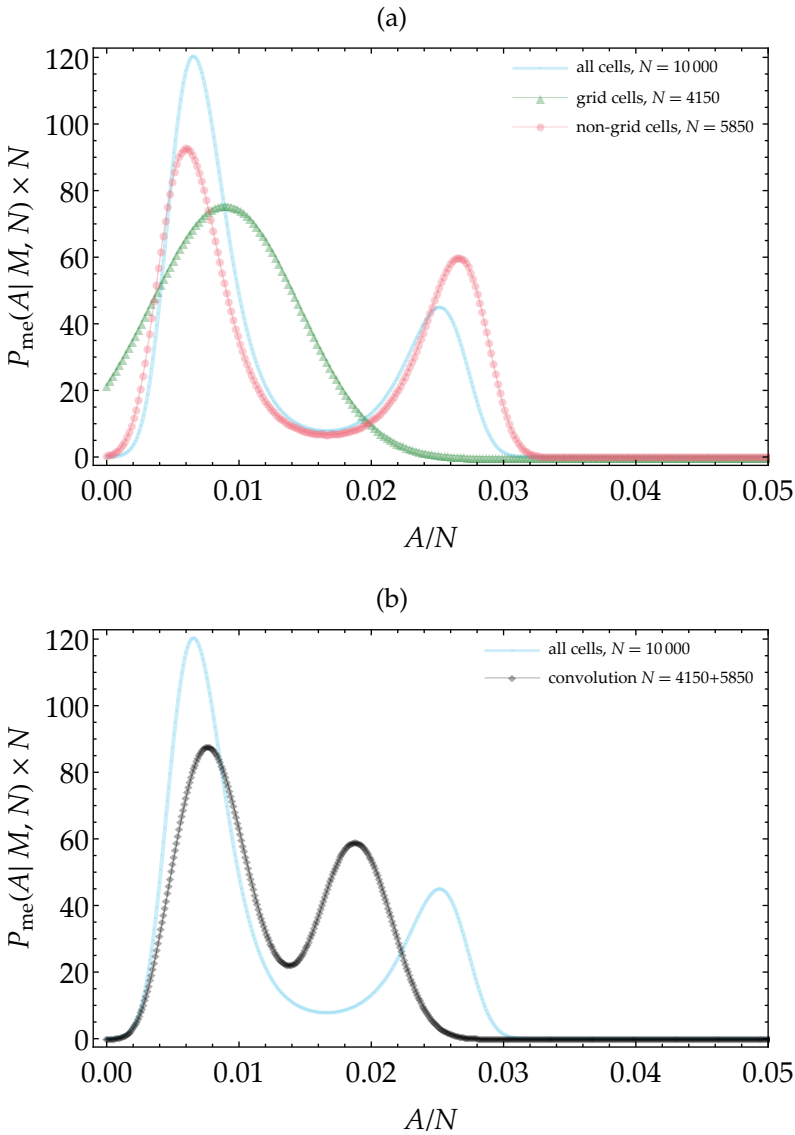

Figure 2 (a) Maximum-entropy distributions for larger subpopulations or grid and non-grid cells. (b) The convolution of the two subpopulation distributions is very different from the distribution for all cells; this means that the two subpopulation distributions are not independent.

a larger population bears on this kind of quantification and compares with the application at the sample level.

As a first step, terms like 'cooperativity', 'interaction', and similar need to be translated into a more precise, operational[22] notion. Here we propose the notion of *informational sufficiency*, because (i) it seems closely related to those terms as used in the literature; (ii) it is mathematically connected with maximum-entropy distributions; (iii) it reveals important consequences of sampling. Its meaning will now be briefly explained. For an extensive discussion of this notion, which was probably first suggested by Kullback & Leibler (1951), see Bernardo et al. (1994 § 4.5) and Jaynes (2003 ch. 8 and § 14.2).

Our belief distribution about the frequencies of the neuronal activities, or about the activity in a new time bin, is conditional on all experimental data and statistics we have collected. This distribution, however, can be such that it remains unchanged if we discard part of the data or statistics – for example, the third- and higher-order empirical moments. This means that the discarded statistics are *informationally irrelevant*: knowing them does not change our beliefs. The remaining statistics – in the example, first and second empirical moments – are *informationally sufficient*.

Now let us turn the informational sufficiencies of two different subsets of statistics into two hypotheses. We can then ask: which hypothesis is more probable, given the data? equivalently, what is the ratio of their probabilities? By translating our question about 'cooperativity' into a hypothesis-testing problem about sufficiency, we can find a well-defined answer determined by the probability-calculus.

But such a translation does more: through a mathematical theorem[23], it leads to the automatic appearance of maximum-entropy distributions: *if a probability distribution for a repetitive phenomenon has some sufficient statistics, then it is a mixture of distributions of maximum-entropy form having those statistics as constraints, and vice versa*. The most important consequence of this theorem is that, by Bayes's formula, the probability that some statistics are informationally sufficient can approximately be found by first building the maximum-entropy distribution constrained by those statistics, and then calculating the probability for the whole

[22] cf. Bridgman 1958 ch. I. [23] Andersen 1970 and references therein; this theorem was first proved for continuous probability distributions by Koopman 1936; Pitman 1936; Darmois 1935; see Bernardo et al. 1994 § 4.5.3 for further references.

data under that distribution. We shall shortly illustrate this procedure with the data sets of § 3.

Owing to the theorem above, studies about cooperativity (and similar ideas) that use maximum-entropy can be naturally reinterpreted as studies about informational sufficiency. An important fact must then be taken into account: *informational sufficiency is not preserved under sampling*. If a probability distribution has some sufficient statistics, then its sample marginals *cannot* have the same sufficient statistics, and vice versa[24]; except for trivial cases (e.g. uniform probability distributions). This impossibility is known in another guise in statistical mechanics: if a system is described by a Gibbs state, its subsystems cannot be perfectly described by Gibbs states[25].

This fact affects the interpretation of the results obtained from maximum-entropy distributions. For example, if means and pairwise correlations seem informationally sufficient for a sample from a brain area, then they are probably not sufficient for the larger population of neurons constituting that area, and vice versa. If we are interested in the cooperativity of the neurons in a brain area, it is therefore unreliable to use a maximum-entropy distribution constructed only for a sample thereof. The model presented here avoids this problem because it targets the larger population, not the sample alone.

We can compare the informational sufficiency between of a set of moments, say $M'' := \{1, \dots, m''\}$, with respect to another, say $M' := \{1, \dots, m'\}$, by introducing the quantity $\Delta(M'', M')$ defined as follows:

(i) From each maximum-entropy distribution $P_{\mathrm{me}}(A \mid M, N)$ for the larger population, eq. (5), built from each set of constraints $M = M', M''$, calculate the marginal distribution for the sample by eq. (1):

$$p(a \mid M, N) = \sum_A G_{aA} P_{\mathrm{me}}(A \mid M, N), \qquad M = M', M''. \tag{7}$$

(ii) Calculate the relative entropies of the measured frequency distribution $\boldsymbol{f}$ with respect to each sample marginal from the previous step, and multiply them by the number of time bins T:

$$T\,\mathrm{H}[\boldsymbol{f}; p(a \mid M, N)] := T \sum_a f_a \log \frac{f_a}{p(a \mid M, N)}, \qquad M = M', M''. \tag{8}$$

[24] Porta Mana et al. 2015 § 3.1. [25] e.g. Maes et al. 1999 and references therein.

(iii) Take the difference:

$$\begin{aligned}\Delta_N(M'', M') &\coloneqq T\,\mathrm{H}[\boldsymbol{f}; p(a \mid M', N)] - T\,\mathrm{H}[\boldsymbol{f}; p(a \mid M'', N)] \\ &\equiv T \sum_a f_a \log \frac{\sum_A G_{aA} P_{\mathrm{me}}(A \mid M'', N)}{\sum_A G_{aA} P_{\mathrm{me}}(A \mid M', N)}.\end{aligned} \tag{9}$$

The quantity $\Delta_N(M'', M')$ so defined is positive if M'' is more probable to be informationally sufficient than M', and negative otherwise.

We show in appendix B that $\Delta_N(M'', M')$ is approximately equal to the log-ratio of the probabilities of the data $\boldsymbol{f}$ conditional on the hypotheses M'' and M':

$$\Delta_N(M'', M') \approx \log \frac{\mathrm{p}(\boldsymbol{f} \mid M'', N)}{\mathrm{p}(\boldsymbol{f} \mid M', N)}. \tag{10}$$

This is their *relative weight of evidence*[26], the logarithm of their *relative Bayes factor*[27]. The exponential of $\Delta_N(M'', M')$ tells us how much more probable the data $\boldsymbol{f}$ are, conditional on the sufficiency of M'', than conditional on the sufficiency of M'. We can also combine this measure with pre-data probabilities for the two hypotheses to obtain the ratio of their probabilities conditional on the data[28].

Let us illustrate the quantity just defined and the previous discussion about sufficiency and sampling by using our data sets. Consider for example three sets of constraints M_2, M_4, M_5, consisting of the first two, four, five moments; and a larger population of size $N = 10\,000$. Figure 3 shows the resulting distributions for the two data sets. There is a clear visual difference between the distributions based on two (dashed red) and on four (solid blue) moments; for the first data set, panel (a), they even differ in the number of modes. On the other hand there is barely any visual difference between the distributions based on four and on five (dotted yellow) moments. From these visual differences we immediately conclude that two moments are less likely to be informationally sufficient than four; but four should be almost as likely as five. We shall see that the measure (9) reflects this swift visual judgement.

Compare these plots with those obtained applying the model at the sample level (equivalent to $N = n$) to the first data set, shown in

[26] Good 1950 ch. 6; 1985; Osteyee et al. 1974 § 1.4. [27] Jeffreys 1936 p. 421; Kass et al. 1995 and references therein. [28] cf. Bretthorst 2013.

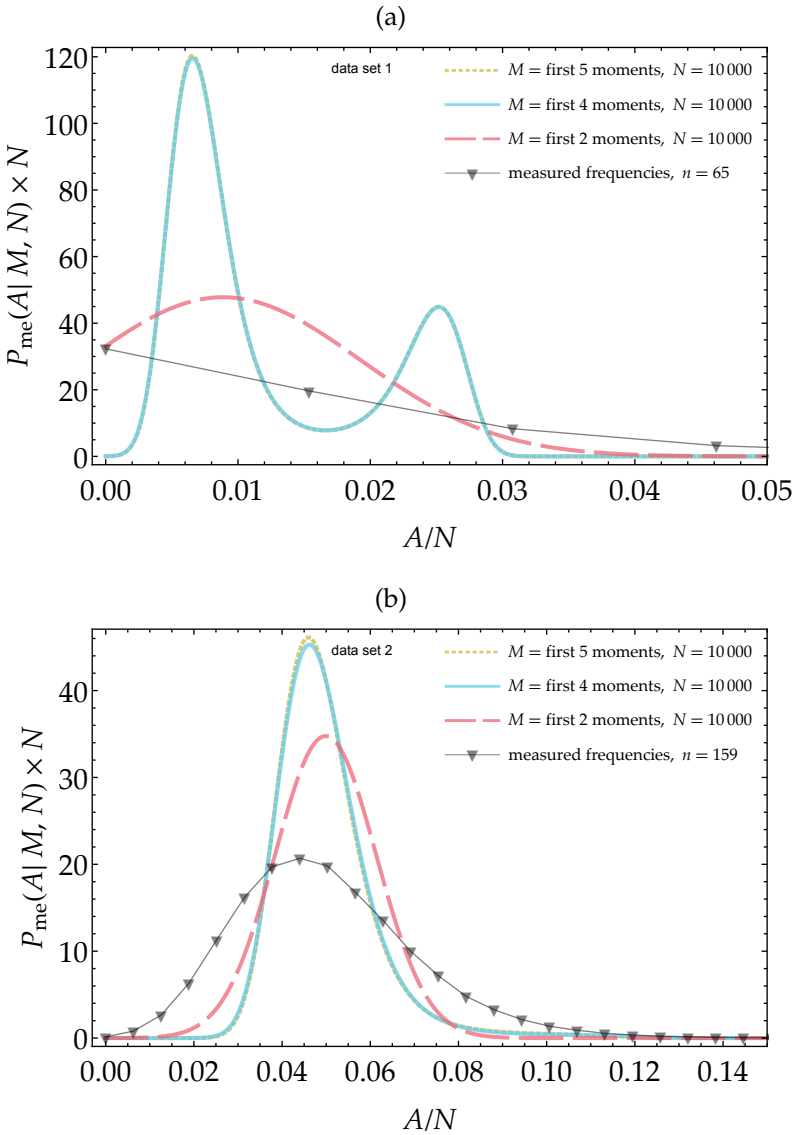

Figure 3 Maximum-entropy distributions for the total activity of the larger population, based on the first two, four, and five moments, and a larger-population size $N = 10\,000$. The two- and four-moment distributions are very different, suggesting that two moments are not informationally sufficient. The four- and five-moment distributions are almost indistinguishable, suggesting that four and five moments are equally likely to be sufficient.

fig. 4. The distributions based on two (red circles), four (blue empty squares), and five (yellow lozenges) moments can only be distinguished on a logarithmic scale, and only in their tails. The measured frequencies (black triangles) are also shown. Two moments seem slightly less likely to be sufficient than four; and four and five moment seem about equally likely. But it is difficult to visually make these comparisons.

Calculation of the weight of evidence (9) for the first data set and $N = 10\,000$, corresponding to fig. 3(a), leads to the following results:

$$\begin{aligned} \Delta_N(M_4, M_2) &= 118 \text{ bit} \equiv 35.4 \text{ Hart}, \\ \Delta_N(M_5, M_4) &= 0.0544 \text{ bit} \equiv 0.0159 \text{ Hart}, \end{aligned} \tag{11}$$

where the Hartley (Hart) denotes base-10 logarithms[29]. In words, our data are 35 orders of magnitude more probable by assuming sufficiency of M_4 than by assuming sufficiency of M_2. But the data are about as probable ($10^{0.016} = 1.04$) by assuming sufficiency of M_4 as of M_5. The qualitative visual information of fig. 3(a) was therefore reliable.

For the model at the sample level, that is, with $N = n$, the weights of evidence are

$$\begin{aligned} \Delta_n(M_4, M_2) &= 404 \text{ bit} \equiv 122 \text{ Hart}, \\ \Delta_n(M_5, M_4) &= 4.77 \text{ bit} \equiv 1.44 \text{ Hart}. \end{aligned} \tag{12}$$

So the application at the sample level says that the data are 120 orders of magnitude more probable conditional on M_4 than on M_2, and $10^{1.4} = 25$ times more probable conditional on M_5 than on M_4. These results are different from those of the application at the larger population, and they would be hard to deduce visually from the distributions of fig. 4.

It is possible to use the measure (9) and its probabilistic interpretation (10) to compare the probability of the data f conditional on the hypothesis (M_5, N) of five-moment sufficiency at the larger-population level, $N = 10\,000$, versus the hypothesis (M_5, n) of five-moment sufficiency at the sample level, $N = n = 65$. We obtain

$$\Delta[(M_5, N), (M_5, n)] = 50.9 \text{ bit} \equiv 15.3 \text{ Hart}. \tag{13}$$

This result is also shown in fig. 4. We see that the five-moment distribution constructed at the larger-population level (yellow stars) fits the

[29] ISO 2009 § C.4; called 'ban' and used by Turing and Good in their code-breaking work at Bletchley Park: see Good 1985; Jaynes 2003 § 4.2.

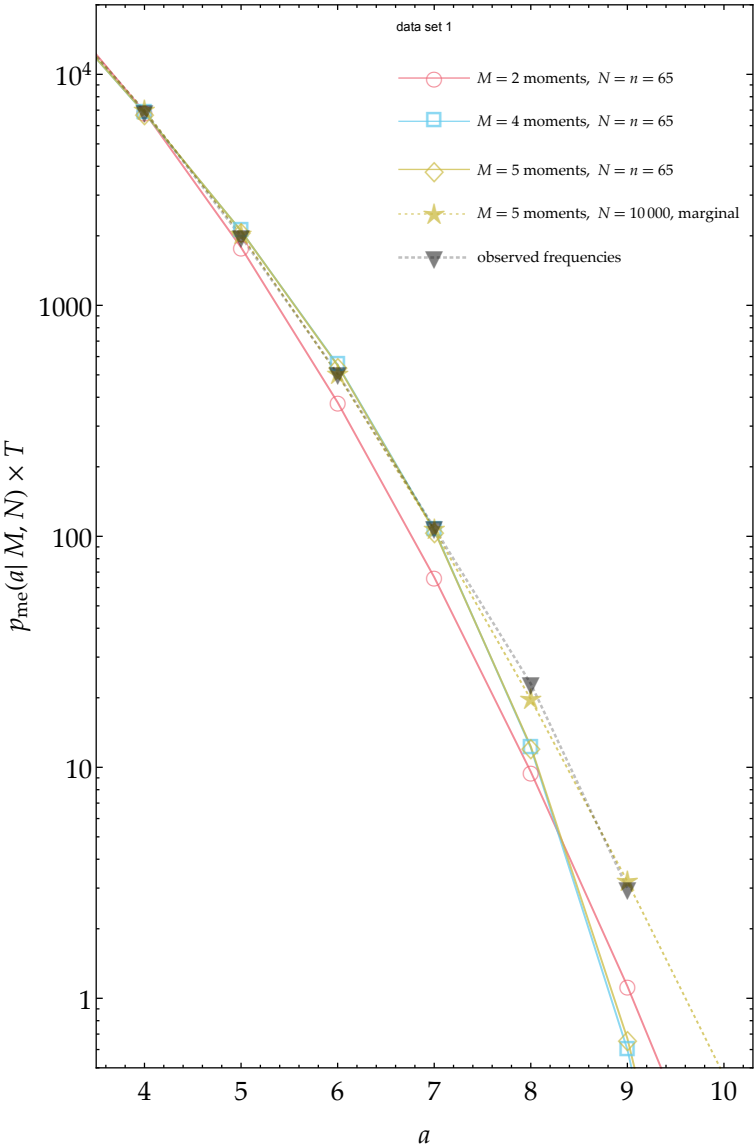

Figure 4 Maximum-entropy distributions constructed at the sample level, using the first two, four, and five moments; and sample marginal of the distribution constructed with five moments for a larger population with $N = 10\,000$ (dotted yellow in 3(a)). The distributions are multiplied by T, obtaining absolute frequencies; note that activities with frequency lower than 1 cannot be observed in T bins. The distributions at the sample level are only distinguishable on a logarithmic scale. The five-moment distribution constructed at the larger-population level fits the frequency data better than the one at the sample level.

measured frequency distribution (black triangles) more closely than the corresponding sample-level distribution (yellow lozenges).

5 The size N of the larger population

The calculations and conclusions presented in the preceding sections depend on the size N of the larger population. The larger population cannot be the whole brain, of course. How large should or can N be in our formulae?

The crucial point is our belief distribution for the activity of the sample *conditional* on the activity of the larger population. It is given by the hypergeometric distribution (18), reprinted here:

$$G_{aA} \coloneqq \mathrm{p}(a \mid A, n, N) = \binom{A}{a}\binom{N-A}{n-a}\binom{N}{n}^{-1} \equiv \binom{n}{a}\binom{N-n}{A-a}\binom{N}{A}^{-1}. \quad (18)_{\mathrm{r}}$$

This distribution leads to the equality of factorial moments (3), on which our model rests. This conditional probability, characteristic of 'drawing without replacement'[30], ensues when we believe, with the same degree, that any of the N units (balls or neurons) could be drawn or sampled. Thus, consider the pool of all neurons which we believe – with equal degree for each neuron – could have been recorded. Then N is the size of that pool. N therefore depends on factors such as: the shape, dimensions, and technical specifications of the recording probe; the inaccuracy in the insertion of the probe, leading for example to slightly different insertion points or angles; the density of neurons around the probe. But it also depends on the homogeneity of the brain region where the recording was made: if we believe that it doesn't matter whether the probe had been inserted in some other point of the same brain region, then formula (18) is appropriate. N can therefore only be assessed case by case.

Whether such equal beliefs are justified, or for which sampling procedures they can be justified, is the fundamental, deep question of sampling theory, which we cannot discuss here[31]. It is of course possible to probabilistically assess, case by case, whether this equality assumption is appropriate. But we must also be aware that such assessment in turn rests on analogous assumptions at a higher level, and so on. The

[30] e.g. Jaynes 2003 ch. 3. [31] see e.g. the discussions and references in Ericson 1969a; Smith 1976.

probability-calculus, just like the truth-calculus, cannot yield conclusions without premisses[32].

Luckily the plots of fig. 1 suggest that the order of magnitude of N ought to be enough for qualitative inferences. As discussed in the next section, even if the exact number N were known, the maximum-entropy distribution ought to be interpreted qualitatively or semi-quantitatively.

If our uncertainty about N spans several orders of magnitude we can assign a probability to the possible values of N based on our background information and the frequency data $\boldsymbol{f}$, similarly to the procedure in § 4 and appendix B. The probability for the data $\boldsymbol{f}$ conditional on a set of constraints M and size N is the exponential of the negative relative entropy (8):

$$\mathrm{p}(\boldsymbol{f} \mid M, N) = \binom{T}{T\boldsymbol{f}} \prod_a p(a \mid M, N)^{T f_a} \approx \exp\bigl\{-T\,\mathrm{H}[\boldsymbol{f}; p(a \mid M, N)]\bigr\}. \tag{14}$$

If our pre-data probability for N is $p(N)$, then by Bayes's theorem

$$\mathrm{p}(N \mid M, \boldsymbol{f}) \propto p(N)\ \exp\bigl\{-T\,\mathrm{H}[\boldsymbol{f}; p(a \mid M, N)]\bigr\}. \tag{15}$$

Here is an illustrative example with our first data set. Suppose our uncertainty spans slightly more than an order of magnitude, from $N = 1\,000$ to $N = 20\,000$. Divide this range roughly into thirds of order of magnitude, considering the values $N \in \{1\,000, 2\,000, 5\,000, 10\,000, 20\,000\}$. Assuming the set M of first five moments to be sufficient, formula (14) gives

$$\begin{gathered}\mathrm{p}(\boldsymbol{f} \mid N = 1\,000, M) = 0.00222, \qquad \mathrm{p}(\boldsymbol{f} \mid N = 2\,000, M) = 0.00704,\\ \mathrm{p}(\boldsymbol{f} \mid N = 5\,000, M) = 0.0127, \qquad \mathrm{p}(\boldsymbol{f} \mid N = 10\,000, M) = 0.0150,\\ \mathrm{p}(\boldsymbol{f} \mid N = 20\,000, M) = 0.0135.\end{gathered} \tag{16}$$

Since our uncertainty regards a scale factor, it can be represented by equal pre-data probabilities of 1/5 about these partial orders of magnitude. Thus from (16) we find

$$\begin{gathered}\mathrm{p}(N = 1\,000 \mid \boldsymbol{f}, M) = 0.044, \qquad \mathrm{p}(N = 2\,000 \mid \boldsymbol{f}, M) = 0.140,\\ \mathrm{p}(N = 5\,000 \mid \boldsymbol{f}, M) = 0.251, \qquad \mathrm{p}(N = 10\,000 \mid \boldsymbol{f}, M) = 0.298,\\ \mathrm{p}(N = 20\,000 \mid \boldsymbol{f}, M) = 0.267,\end{gathered} \tag{17}$$

[32] Hailperin 2011; Johnson 1924 p. 182.

which gives a slightly higher probability to $N = 10\,000$.

In this way we can make inferences that take into account our uncertainty about N. We must make sure to avoid circularities, though: for example, we cannot assume a set of moments M to be sufficient and assess our uncertainty about N conditional on it, and then use this uncertainty to assess the sufficiency of M. But we can approximately assess the sufficiency of a subset of moments $M' \subset M$ taking into account the uncertainty of N conditional on M.

A rigorous assessment would involve a more expensive, full Bayesian calculation; but such calculation would make the maximum-entropy method superfluous. We discuss this final point in the next section.

6 Assumptions and approximations that define the model

The possible uses of the model here presented depend on the assumptions and approximations which define it: three main assumptions and two main approximation all in all. These become clear within a full-fledged probabilistic approach. We summarize them here, leaving a more mathematical sketch to appendix C.

The first and most important assumption, typical of this kind of maximum-entropy application in the neurosciences, can be expressed in two equivalent ways: (i) the frequencies of the activity during a recording are informationally sufficient for inferences about unrecorded times; (ii) our degree of belief about the sequence of activities is invariant under permutations of the time bins. Equating time averages and expectations, eq. (4), would not make sense without this assumption. Its mathematical consequence is to relate the probability for the frequency distribution $\boldsymbol{F}$ during the recording to that of the *long-run* frequency distribution for the activities of the larger population. Note that the maximum-entropy method in statistical physics is meant to be used not with this assumption, but with averages across *repetitions* of the same experiment, because its constraints represent macroscopically *reproducible* quantities and kinematics, even in non-equilibrium[33].

The second assumption, typical of maximum-entropy in general, concerns the form of the probability distribution for the long-run frequencies ensuing from the first assumption. It is taken to be either an 'entropic

[33] Jaynes 1996a; Grandy 1988; De Roeck et al. 2006; for experiments e.g. Gunton et al. 1983.

prior'[34], which takes into account the multiplicity of long-run frequencies, or a prior for a model with sufficient statistics, which expresses the sufficiency of the moments under study[35]. This assumption is crucial for deriving the maximum-entropy based on the Shannon entropy. A different density leads to alternative maximum-entropy methods, for example based on the Burg entropy[36] if a Dirichlet density is used[37]. The entropic prior depends on a reference distribution and on a coefficient L which determines the sharpness of the prior's peak. The computations of the preceding sections were insensitive to the reference distribution in their significant digits. We used three biologically reasonable alternatives: a flat distribution, one linearly decreasing with A, and a normal one with a broad peak at low activities.

The third assumption is specific to our application: all subsets of A neurons in the larger population are equally likely to give rise to total activity A, at every time bin. A mathematically equivalent assumption is that n neurons are sampled anew at each time bin. This assumption allows us to factorize the joint probability of the two sequences of activities and apply the sampling relation (18) of appendix A at each bin.

Finally, the maximum-entropy method is typically equivalent to approximating the number T of time bins, the prior coefficient L, and the ratio T/L to infinity. These approximations lead to the identification of the four distributions listed at the end of § 2: the recorded frequency distribution becomes equal to the long-run one, and also to the probability distribution for the activity at recorded and new time bins. In our application it also leads to the equality of the hypergeometric distribution (18) and the conditional frequencies of the sample activity a given A. The goodness of this approximation decreases as the ratio between the number of possible joint states and the number of bins – roughly nN/T – increases, and also as observed quantities approach their mathematical bounds, such as some frequencies $f_a = 0$. Without these approximations it is still possible to interpret the maximum-entropy distribution (5) as the mode of the probability distribution for the recorded frequencies $\boldsymbol{F}$, that is, the most probable recorded frequency.

The first and third assumptions above are the least realistic, but they make sense as reference hypotheses for judging the informational

[34] Skilling et al. 1984; Rodríguez 1991; Neumann 2007. [35] Porta Mana 2017a. [36] Burg 1975. [37] Jaynes 1996b.

relevance of correlations of some order, as in § 4, to be compared with hypotheses about more realistic time dependences and clustering.

The $T = \infty$ approximation can break our model for specific distributions of sample frequencies, if too many moments are used. For example take $n = 2$ with frequency distribution $\boldsymbol{f} = (0.5, 0, 0.5)$ for the activities $a = 0, 1, 2$; that is, the sample is silent half of the time and completely active half of the time. The first two normalized factorial moments have both value 0.5, and together they uniquely determine $\boldsymbol{f}$. If we now take $N = 3$, it can be proved that eq. (5), constrained on those two moments, has no solution. The reason is clear from a sampling perspective: if there is zero probability of sampling one active neuron, then the larger population cannot have any active neurons, and there must be zero probability of sampling two active ones. But this reasoning holds only for each single time bin, not for the averages over time bins. The $T = \infty$ approximation identifies the two instead. Outside this approximation there are no contradictions because the conditional frequencies and conditional probabilities are kept distinct.

The two infinity approximations are at the limit of their validity in our examples. For this reason we interpret the distributions as *most probable frequency distributions*, as mentioned at the end of § 2. The space of possible frequency distributions for the larger population has N dimensions, however, and probability distributions in high dimensions have counter-intuitive properties. For example, the mode or mean can have *atypical* features when compared with the majority of other probable points of the probability space. The question then is *which features of the maximum-entropy frequency distribution are typical of the majority of plausible frequency distributions?* Preliminary studies using a full probability calculation (to be published separately) show that several semi-quantitative features of the maximum-entropy distribution are indeed typical. For the first data set, for example, most frequency distributions have two high-frequency regimes of activity, as in fig. 1(a), with roughly the same heights and almost the same locations, within a few percentages of population-averaged activity. Each of high-frequency regimes, though, may consist of two or three very close modes rather than one. The same preliminary studies also indicate that these typical features remain with prior choices in the second assumption above, for example using a Dirichlet prior instead of an entropic one.

This means that the maximum-entropy approximation presented

here can be favourably used to get an idea of what a fully probabilistic calculation could yield, but with a far lower computational cost. For example, for $N = 1\,000$ the maximum-entropy distribution can be calculated in ten minutes on a laptop, whereas the full probability calculation takes several days on a high-performance computing cluster.

7 Summary and discussion

The reason why we record neurons from a brain region is to have an idea of the activity of the other neurons in that region, similarly to survey sampling. If we thought that *precisely* the recorded ones were exceptions or completely unrelated to the others surrounding the recording probe, our recording would serve little purpose. Probabilistic models that allow us to make direct inferences about the larger population are therefore useful. But they become complex and sensitive if we ask for very detailed inferences.

We have presented a model that allows us to make direct inferences about a larger population from very small samples. Its inferences regard only the frequencies of the total activity of the larger population, but such limited inferential scope makes the model computationally very cheap.

Yet, and most important, this model leads to inferences that are not at all trivial and not easily discernible by looking at the sample. We showed this by applying the model to two real data sets.

By 'inference' we mean the quantification of a degree of belief about possible observations, based on specific assumptions; not the response of an oracle. The model here presented gives us a *forecast* based on specific assumptions discussed in § 6, and therefore has two main uses, just like any other probabilistic model. If our assumptions are only hypotheses under testing, then the subsequent confirmation or confutation of the forecast will increase or decrease our confidence in those hypotheses. If we strongly trust our assumptions, on the other hand, then we rely on the forecast and even use it in place of actual observations, especially if the latter are difficult to make. (In fact, we typically shift from the first to the second use.) In the preceding sections we have given examples of additional hypotheses that could be explored with the present model. The assumptions of the model itself (§ 6) can either be under test or relied upon, depending on the case.

Some approximations implicit in the model make its numerical results semi-quantitative with respect to an exact analysis. The results should not be off by more than a few percentages for the population-averaged activity and for the frequency density. This imprecision, however, is compensated by the extreme computational cheapness of the model, far less costly than an exact analysis. The model is therefore very convenient at least for essaying hypotheses. We hope to present an exact quantitative comparison with the full probabilistic approach in a future publication.

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 the staff of the NTNU library for their always prompt support; Mari, Miri, & Emma for continuous encouragement and affection; Buster Keaton and Saitama for filling life with awe and inspiration; the developers and maintainers of LaTeX, Emacs, AUCTeX, Open Science Framework, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible.

Appendices

A Derivation of the maximum-entropy distribution

Here is a summary derivation of the maximum-entropy distribution for the larger population constrained by a set of factorial moments.

First of all the sampling relation. We have a population of N units – neurons – A of which have some specific property – being active in a specific time bin – and we sample in an unknown way n of the N units. The probability that a of the n sampled neurons are active is then given by the hypergeometric distribution

$$G_{aA} \coloneqq \mathrm{p}(a \mid A, n, N) = \binom{A}{a}\binom{N-A}{n-a}\binom{N}{n}^{-1} \equiv \binom{n}{a}\binom{N-n}{A-a}\binom{N}{A}^{-1} \tag{18}$$

typical of 'drawing without replacement'[38]. In the following we leave n, N implicit in the conditional. If we are uncertain about the number

[38] e.g. Jaynes 2003 ch. 3.

A, with belief $P(A)$, then by the theorem of total probability our belief about a is

$$p(a) = \sum_{A} G_{aA}\, P(A). \tag{19}$$

From the definition of normalized factorial moment (2), the expression for the hypergeometric distribution (18), and the relation (19) between our beliefs about a and A, and using some combinatorial juggling[39], one can prove the equality (3) between the factorial moments of sample and larger population.

For the construction of a maximum-entropy distribution from generic expectation constraints see Jaynes (1963; 2003 ch. 11), and Hobson et al. (1973) and Sivia (2006 § 5.2.2) for the minimum-relative-entropy construction. For the solution of the extremization problem using Lagrangians and Lagrange multipliers see Mead et al. (1984) and the extensive texts by Fang et al. (1997) and Boyd et al. (2009). For a geometric understanding of the extremization and of the relation between expectations and multipliers see Porta Mana (2017b).

The result has the standard exponential-family form

$$\begin{aligned} P_{\mathrm{me}}(A) &= \frac{1}{Z(\boldsymbol{\lambda})}\, r_A\, \exp\biggl[\sum_m \lambda_m \binom{A}{m}\binom{N}{m}^{-1}\biggr], \\ Z(\boldsymbol{\lambda}) &:= \sum_A r_A\, \exp\biggl[\sum_m \lambda_m \binom{A}{m}\binom{N}{m}^{-1}\biggr], \end{aligned} \tag{20}$$

where (r_A) is the reference distribution and $\boldsymbol{\lambda} := (\lambda_m)$ are the Lagrange multipliers, satisfying the implicit constraint equations (4):

$$\sum_A \tbinom{A}{m}\tbinom{N}{m}^{-1} \tfrac{1}{Z(\boldsymbol{\lambda})}\, r_A\, \exp\biggl[\sum_m \lambda_m \tbinom{A}{m}\tbinom{N}{m}^{-1}\biggr] = \sum_a \tbinom{a}{m}\tbinom{n}{m}^{-1} f_a, \quad m \in M. \tag{21}$$

The reference distribution (r_A) represents our pre-data beliefs about the activity levels A. We know that the majority of neurons in a brain area are rarely simultaneously active within a window of some milliseconds, so we could choose a distribution with slightly higher weights on low values of A. On the other hand, considering the number of ways in which A out of N neurons can be simultaneously active would suggest

[39] Porta Mana et al. 2015 appendix A and references therein.

the multiplicity distribution proportional to $\binom{N}{A}$. It turns out that our results of §§ 3–4 are insensitive to the choice between these two possible reference distributions, or even a uniform reference distribution.

For a derivation from a Bayesian perspective see appendix C.

B Measure of informational sufficiency

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

$$\frac{p(M'' \mid \boldsymbol{f})}{p(M' \mid \boldsymbol{f})}. \tag{22}$$

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

$$p(a_t \mid M) = \sum_A G_{a_t A}\, P_{\mathrm{me}}(A \mid M). \tag{23}$$

The probability of observing one sequence (a_t) with frequencies $\boldsymbol{f}$ is therefore

$$\prod_{t=1}^{T} p(a_t \mid M) \equiv \prod_{a=0}^{n} p(a \mid M)^{T f_a} \equiv \prod_{a=0}^{n} \Bigl[\sum_A G_{aA}\, P_{\mathrm{me}}(A \mid M)\Bigr]^{T f_a}. \tag{24}$$

The probability of observing the frequencies $\boldsymbol{f}$ is obtained multiplying this by their multiplicity factor, the multinomial coefficient

$$\binom{T}{T\boldsymbol{f}} := \frac{T!}{\prod_a (T f_a)!} \approx \prod_a f_a^{-T f_a}, \tag{25}$$

the last expression coming from Stirling's approximation[40]. If we assign equal pre-data probabilities to the two hypotheses M' and M'', each probability in the ratio (22) then becomes, by Bayes's theorem,

$$\mathrm{p}(M \mid \boldsymbol{f}) \propto \mathrm{p}(\boldsymbol{f} \mid M) \times \mathrm{const} \propto \binom{T}{T\boldsymbol{f}} \prod_a \Bigl[\sum_A G_{aA} P_{\mathrm{me}}(A \mid M)\Bigr]^{T f_a} \approx$$
$$\prod_a f_a^{-T f_a} \times \prod_a \Bigl[\sum_A G_{aA} P_{\mathrm{me}}(A \mid M)\Bigr]^{T f_a}. \tag{26}$$

[40] Csiszár et al. 2004 Lemma 2.2.

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

Thus, the difference (9) is the logarithm of the probability ratio (22). The exponential of the difference (9) tells us how much more probable the hypothesis that the set M'' is sufficient than the hypothesis that the set M' is.

C Assumptions and approximations: mathematical sketch

We give here a sketch of the mathematical formulae behind the assumptions and approximations discussed in § 6. We indicate them all collectively with I.

The first assumption, of infinite exchangeability[41] of the probability distribution $\mathrm{p}[(A_1, \dots, A_T) \mid I]$ for the sequence of activities, implies the following mixture form by de Finetti's theorem[42]:

$$\mathrm{p}[(A_1, \dots, A_T) \mid I] = \int \mathrm{d}\boldsymbol{\nu}\, \mathrm{p}(\boldsymbol{\nu} \mid I) \prod_{A=0}^{N} {\nu_A}^{T F_A}, \tag{27}$$

where $\boldsymbol{\nu} \coloneqq (\nu_A)$ is the long-run frequency distribution of activities in an imaginary continuation or repetition of the recording, and $\mathrm{p}(\boldsymbol{\nu} \mid I)\, \mathrm{d}\boldsymbol{\nu}$ our belief distribution about it. The integration is over an N-dimensional simplex.

The second assumption concerns the latter distribution. Its density function has one of the following expressions:

(a) an entropic prior[43]

$$\mathrm{p}(\boldsymbol{\nu} \mid I) \propto \exp[-L\, \mathrm{H}(\boldsymbol{F}; \boldsymbol{r})] \approx \binom{L}{L\nu_0, \dots, L\nu_N} \prod_A {r_A}^{L\nu_A} \tag{28a}$$

where $\boldsymbol{r} \coloneqq (r_A)$ is a reference distribution and the expression in larger parentheses is a multinomial coefficient;

[41] Bernardo et al. 2000 ch. 4; Dawid 2013. [42] De Finetti 1930; Hewitt et al. 1955. [43] Skilling et al. 1984; Rodríguez 1991; Neumann 2007.

(*b*) a 'sufficiency prior', from the exponential family, which follows from the Pitman-Koopman theorem if some moments $\{m\}$ are sufficient statistics for our belief about $\boldsymbol{\nu}$:

$$\mathrm{p}(\boldsymbol{\nu} \mid I) \propto \int \mathrm{d}\boldsymbol{\lambda}\ \mathrm{p}(\boldsymbol{\lambda} \mid I) \prod_A \delta\left\{\boldsymbol{\nu}_A - \frac{r_A}{Z(\boldsymbol{\lambda})}\ \exp\left[\textstyle\sum_m \lambda_m \binom{A}{m} / \binom{N}{m}\right]\right\}, \tag{28b}$$

where $\boldsymbol{\lambda} := (\lambda_m)$ are Lagrange multipliers and $Z(\boldsymbol{\lambda})$ a normalization factor for the exponential distribution within the delta. This is a formal expression, the delta symbolizing that the possible domain of $\boldsymbol{\nu}$ is restricted from an N-dimensional simplex to a lower-dimensional subset thereof, parametrized by $\boldsymbol{\lambda}$. When this prior is used, all integrations over $\boldsymbol{\nu}$ are replaced by integrations over $\boldsymbol{\lambda}$.

The third assumption corresponds to the formula

$$\mathrm{p}[(a_1, \dots, a_T) \mid (A_1, \dots, A_T), I] = \prod_{t=1}^{T} \mathrm{p}(a_t \mid A_t, I) = \prod_{t=1}^{T} G_{a_t A_t} \equiv \prod_{a,A} G_{a_t A_t}{}^{T J_{aA}}, \tag{29}$$

where J_{aA} is the joint frequency distribution of the activity pairs (a, A), so that its marginals are the frequency distributions $\boldsymbol{f}$ and $\boldsymbol{F}$. As discussed in § 6 this formula follows from either of two beliefs: that all subsets of A neurons in the larger population are always equally likely to give rise to total activity A, or that a new sampling is done at every time bin.

It can be proved that the formulae above lead to the following belief distribution for the joint frequency distribution $\boldsymbol{J} := (J_{aA})$ and the long-run frequency distribution $\boldsymbol{\nu}$, conditional on $\boldsymbol{f}$:

$$\mathrm{p}(\boldsymbol{J}, \boldsymbol{\nu} \mid \boldsymbol{f}, I) \propto \mathrm{p}(\boldsymbol{\nu} \mid I) \binom{T}{T J_{00}, \dots, T J_{nN}} \prod_{a,A} (G_{aA}\ \nu_A)^{T J_{aA}} \approx \mathrm{p}(\boldsymbol{\nu} \mid I)\ \exp\left(-T \textstyle\sum_{a,A} J_{aA} \ln \dfrac{J_{aA}}{G_{aA}\ \nu_A}\right) \quad \text{with} \quad \textstyle\sum_A J_{aA} = f_a, \tag{30}$$

where $\mathrm{p}(\boldsymbol{\nu} \mid I)$ is either of the distributions (28). Note the relative entropy in the exponential. If we want to conditionalize not on $\boldsymbol{f}$ but on a set of measured averages $C(m) = \sum_a c_a(m) f_a$, $m \in M$, such as (2), we have

to replace the integration constraint $\sum_A J_{aA} = f_a$ with $\sum_{aA} c_a(m)\, J_{aA} = C(m)$.

From this belief distribution we can find the one for $F_A \equiv \sum_a J_{aA}$ and any others of interest by marginalization. Within the exponential we can recognize the relative entropy of $\boldsymbol{J}$ with respect to $(G_{aA}\ \ \nu_A)$, which is the reason why these two joint distributions must be equal if T tends to infinity.

From the formulae above we can also find the four distributions mentioned at the end of § 2. The frequency distribution (i) is the mode of $\mathrm{p}(\boldsymbol{F} \mid \boldsymbol{f}, I)$. The belief distribution (ii) is

$$\mathrm{p}(A_t \mid \boldsymbol{f}, I) = \sum_{\boldsymbol{F}} F_{A_t}\, \mathrm{p}(\boldsymbol{F} \mid \boldsymbol{f}, I), \qquad t \in \{1, \dots, T\}. \tag{31}$$

The frequency distribution (iii) is the mode of $\mathrm{p}(\boldsymbol{\nu} \mid \boldsymbol{f}, I)$. Finally, the belief distribution (iv) is

$$\mathrm{p}(A_t \mid \boldsymbol{f}, I) = \int \mathrm{d}\boldsymbol{\nu}\ \nu_{A_t}\, \mathrm{p}(\boldsymbol{\nu} \mid \boldsymbol{f}, I), \qquad t \notin \{1, \dots, T\}. \tag{32}$$

Bibliography

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

Amari, S.-i., Nakahara, H., Wu, S., Sakai, Y. (2003): *Synchronous firing and higher-order interactions in neuron pool*. Neural Comp. **15**[1], 127–142.

Andersen, E. B. (1970): *Sufficiency and exponential families for discrete sample spaces*. J. Am. Stat. Assoc. **65**[331], 1248–1255.

Bahadur, R. R. (1954): *Sufficiency and statistical decision functions*. Ann. Math. Stat. **25**[3], 423–462. See also Bahadur, Lehmann (1955).

Bahadur, R. R., Lehmann, E. L. (1955): *Two comments on "Sufficiency and statistical decision functions"*. Ann. Math. Stat. **26**[1], 139–142. See Bahadur (1954).

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.

Battistin, C., Dunn, B., Roudi, Y. (2017): *Learning with unknowns: analyzing biological data in the presence of hidden variables*. Curr. Opin. Syst. Biol. **1**, 122–128.

Bernardo, J.-M., DeGroot, M. H., Lindley, D. V., Smith, A. F. M., eds. (1985): *Bayesian Statistics 2*. (Elsevier and Valencia University Press, Amsterdam and Valencia). https://www.uv.es/~bernardo/valenciam.html.

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

— (2000): *Bayesian Theory*, reprint. (Wiley, New York). First publ. 1994.

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.

Bretthorst, G. L. (2013): *The maximum entropy method of moments and Bayesian probability theory*. Am. Inst. Phys. Conf. Proc. **1553**, 3–15. http://bayes.wustl.edu/glb/BretthorstHistograms.pdf.

Bridgman, P. W. (1958): *The Logic of Modern Physics*, eight printing. (Macmillan, New York). First publ. 1927.

Burg, J. P. (1975): *Maximum entropy spectral analysis*. PhD thesis. (Stanford University, Stanford). http://sepwww.stanford.edu/data/media/public/oldreports/sep06/.

Cifarelli, D. M., Regazzini, E. (1980): *Sul ruolo dei riassunti esaustivi ai fini della previsione in contesto bayesiano (1ª parte)*. Riv. mat. scienze econ. sociali **3**[2], 109–125. See also Cifarelli, Regazzini (1981).

— (1981): *Sul ruolo dei riassunti esaustivi ai fini della previsione in contesto bayesiano (2ª parte)*. Riv. mat. scienze econ. sociali **4**[1], 3–11. See also Cifarelli, Regazzini (1980).

Csiszár, I. (1985): *An extended maximum entropy principle and a Bayesian justification*. In: Bernardo, DeGroot, Lindley, Smith (1985), 83–98. With discussion by G. A. Barnard, E. T. Jaynes, T. Seidenfeld, W. Polasekand, and reply.

Csiszár, I., Shields, P. C. (2004): *Information theory and statistics: a tutorial*. Foundations and Trends in Communications and Information Theory **1**[4], 417–528. http://www.renyi.hu/~csiszar/.

Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A., eds. (2013): *Bayesian Theory and Applications*. (Oxford University Press, Oxford).

Darmois, G. (1935): *Sur les lois de probabilité à estimation exhaustive*. Comptes rendus hebdomadaires des séances de l'Académie des sciences **200**, 1265–1266.

Dawid, A. P. (2013): *Exchangeability and its ramifications*. In: Damien, Dellaportas, Polson, Stephens (2013), ch. 2, 19–29.

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

De Roeck, W., Maes, C., Netočný, K. (2006): *H-theorems from macroscopic autonomous equations*. J. Stat. Phys. **123**[3], 571–584. mp_arc:05-263.

de Finetti, B. (1930): *Funzione caratteristica di un fenomeno aleatorio*. Atti Accad. Lincei: Sc. Fis. Mat. Nat. **IV**[5], 86–133. http://www.brunodefinetti.it/Opere.htm.

Domb, C., Lebowitz, J. L., eds. (1983): *Phase Transitions and Critical Phenomena. Vol. 8*. (Academic Press, London).

Dunn, B., Roudi, Y. (2013): *Learning and inference in a nonequilibrium Ising model with hidden nodes*. Phys. Rev. E **87**[2], 022127.

Ericson, W. A. (1969a): *Subjective Bayesian models in sampling finite populations*. J. Roy. Stat. Soc. B **31**[2], 195–224. http://www.stat.cmu.edu/~brian/905-2008/papers/Ericson-JRSSB-1969.pdf. See also discussion in Sampford, Scott, Stone, Lindley, Smith, Kerridge, Godambe, Kish, et al. (1969).

— (1969b): *A note on the posterior mean of a population mean*. J. Roy. Stat. Soc. B **31**[2], 332–334.

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

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

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.

Gerstein, G. L., Perkel, D. H., Dayhoff, J. E. (1985): *Cooperative firing activity in simultaneously recorded populations of neurons: detection and measurement*. J. Neurosci. **5**[4], 881–889.

Gerstner, W., Kistler, W. M., Naud, R., Paninski, L. (2014): *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. (Cambridge University Press, Cambridge).

Good, I. J. (1950): *Probability and the Weighing of Evidence*. (Griffin, London).

— (1985): *Weight of evidence: a brief survey*. In: Bernardo, DeGroot, Lindley, Smith (1985), 249–270. With discussion by H. Rubin, T. Seidenfeld, and reply.

Grandy Jr., W. T. (1988): *Foundations of Statistical Mechanics. Vol. II: Nonequilibrium Phenomena*. (Reidel, Dordrecht).

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.

Gunton, J. D., San Miguel, M., Sahni, P. S. (1983): *The dynamics of first order phase transitions*. In: Domb, Lebowitz (1983), ch. 3 & addendum, 267–466, 479–482. `https://ifisc.uib-csic.es/publications/publication-detail.php?indice=2005`.

Hailperin, T. (2011): *Logic with a Probability Semantics: Including Solutions to Some Philosophical Problems*. (Lehigh University Press, Plymouth, UK).

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

Hebb, D. O. (2002): *The Organization of Behavior: A Neuropsychological Theory*, repr. (Lawrence Erlbaum, Mahwah, USA). First publ. 1949. `http://s-f-walker.org.uk/pubsebooks/pdfs/The_Organization_of_Behavior-Donald_O._Hebb.pdf`.

Hewitt, E., Savage, L. J. (1955): *Symmetric measures on Cartesian products*. Trans. Am. Math. Soc. **80**[2], 470–501.

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

Hong, G., Lieber, C. M. (2019): *Novel electrode technologies for neural recordings*. Nat. Rev. Neurosci. **20**, 330–345, 376.

Huang, H. (2015): *Effects of hidden nodes on network structure inference*. J. Phys. A **48**[35], 355002.

ISO (International Organization for Standardization) (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.

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

— (2009): *ISO 80000-1:2009: Quantities and units 1: General*. 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`.

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

— (1996a): *Macroscopic prediction*. `http://bayes.wustl.edu/etj/node1.html`. First publ. in Haken (1985) pp. 254–269.

— (1996b): *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).

Jaynes, E. T. (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.

Jeffreys, H. (1935): *Some tests of significance, treated by the theory of probability*. Proc. Cambridge Philos. Soc. **31**^{2}, 203–222. See also Jeffreys (1936).

— (1936): *Further significance tests*. Proc. Cambridge Philos. Soc. **32**^{3}, 416–445. See also Jeffreys (1935).

Johnson, W. E. (1924): *Logic. Part III: The Logical Foundations of Science*. (Cambridge University Press, Cambridge). https://archive.org/details/logic03john.

Kass, R. E., Raftery, A. E. (1995): *Bayes factors*. J. Am. Stat. Assoc. **90**^{430}, 773–795. https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf; https://www.andrew.cmu.edu/user/kk3n/simplicity/KassRaftery1995.pdf.

Koopman, B. O. (1936): *On distributions admitting a sufficient statistic*. Trans. Am. Math. Soc. **39**^{3}, 399–409.

Kruskal, W., Mosteller, F. (1979a): *Representative sampling, II: Scientific literature, excluding statistics*. Int. Stat. Rev. **47**^{2}, 111–127. See also Kruskal, Mosteller (1979c,b; 1980).

— (1979b): *Representative sampling, III: The current statistical literature*. Int. Stat. Rev. **47**^{3}, 245–265. See also Kruskal, Mosteller (1979c,a; 1980).

— (1979c): *Representative sampling, I: Non-scientific literature*. Int. Stat. Rev. **47**^{1}, 13–24. See also Kruskal, Mosteller (1979a,b; 1980).

— (1980): *Representative sampling, IV: The history of the concept in statistics, 1895–1939*. Int. Stat. Rev. **48**^{2}, 169–195. See also Kruskal, Mosteller (1979c,a,b).

Kullback, S., Leibler, R. A. (1951): *On information and sufficiency*. Ann. Math. Stat. **22**^{1}, 79–86.

Latham, P. E., Roudi, Y. (2013): *Role of correlations in population coding*. In: Quian Quiroga, Panzeri (2013), ch. 7, 121–138.

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.

MacKay, D. J. C. (2003): *Information Theory, Inference, and Learning Algorithms*. (Cambridge University Press, Cambridge). http://www.inference.phy.cam.ac.uk/mackay/itila/. First publ. 1995.

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.

Osteyee, D. B., Good, I. J. (1974): *Information, Weight of Evidence, the Singularity between Probability Measures and Signal Detection*. (Springer, Berlin).

Pillow, J. W., Latham, P. E. (2008): *Neural characterization in partially observed populations of spiking neurons*. Adv. Neural Information Processing Systems (NIPS) **20**, 1161–1168.

Pitman, E. J. G. (1936): *Sufficient statistics and intrinsic accuracy*. Math. Proc. Camb. Phil. Soc. **32**[4], 567–579.

Porta Mana, P. G. L. (2017a): *Maximum-entropy from the probability calculus: exchangeability, sufficiency*. Open Science Framework doi:10.17605/osf.io/xdy72, HAL:hal-01533985, arXiv:1706.02561.

— (2017b): *Geometry of maximum-entropy proofs: stationary points, convexity, Legendre transforms, exponential families*. Open Science Framework doi:10.17605/osf.io/vsq5n, HAL:hal-01540184, arXiv:1707.00624.

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.

Quian Quiroga, R., Panzeri, S., eds. (2013): *Principles of Neural Coding*. (CRC Press, Boca Raton, USA).

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

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.

Sampford, M. R., Scott, A., Stone, M., Lindley, D. V., Smith, T. M. F., Kerridge, D. F., Godambe, V. P., Kish, L., et al. (1969): *Discussion on professor Ericson's paper*. J. Roy. Stat. Soc. B **31**[2], 224–233. http://www.stat.cmu.edu/~brian/905-2008/papers/Ericson-JRSSB-1969.pdf. See Ericson (1969b).

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., Bryan, R. K. (1984): *Maximum entropy image reconstruction: general algorithm*. Mon. Not. Roy. Astron. Soc. **211**[1], 111–124.

Smith, T. M. F. (1976): *The foundations of survey sampling: a review*. J. Roy. Stat. Soc. A **139**[2], 183–195. See also discussion and reply in Sampford, Scott, Stone, Lindley, Smith, Kerridge, Godambe, Kish, et al. (1969).

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

Tyrcha, J., Hertz, J. (2014): *Network inference with hidden units*. Math. Biosci. Eng. **11**[1], 149–165.

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.