Maximum-entropy distributions for a neuronal population from subpopulation data [draft]

P.G.L. Porta Mana
<pgl@portamana.org>
Y. Roudi
<yasser.roudi@ntnu.no>
V. Rostami
<vrostami@uni-koeln.de>
E. Torre
<torre@ibk.baug.ethz.ch>

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

Abstract must be rewritten once paper is readyThis work shows how to build a maximum-entropy probabilistic model for the total activity of a population of neurons, given only some activity data or statistics – for example, empirical moments – of a *subpopulation* thereof. This kind of model is useful because neuronal recordings are always limited to a very small sample of a population 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 population – 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 subpopulation is compared with the marginal probability distribution obtained from the maximum-entropy model applied to the full population. 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 populations

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 population. 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 population using recording from subpopulations 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 population 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[1] has been used for different kinds

[1] Jaynes 1957a; much clearer in Jaynes 1963; Sivia 2006; Hobson et al. 1973; Jaynes 1996a; Grandy 1980.

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[2].

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[3], (b) the activity of 159 neurons recorded for 15 min from a macaque's Motor Cortex[4]. For each data set the model gives the most plausible frequency distribution with which all levels of total activity of a larger population 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 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. 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 population cannot be inferred by the application of maximum-entropy *to the sample alone*. Although we don't formulate any hypotheses on why the larger population 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[5]. But when the difference in size between sample and full population is too large, correlation sufficiency for the

[2] for example Martignon et al. 1995; Bohte et al. 2000; Shlens et al. 2006; Schneidman et al. 2006; Tkačik et al. 2006; Macke et al. 2009; Tkačik et al. 2009; Roudi et al. 2009; Barreiro et al. 2010; Gerwinn et al. 2010; Macke et al. 2011; Ganmor et al. 2011; Cohen et al. 2011; Granot-Atedgi et al. 2013; Macke et al. 2013; Tkačik et al. 2014; Shimazaki et al. 2015; Mora et al. 2015; Lezon et al. 2006; Weigt et al. 2009.

[3] Stensola et al. 2012.

[4] Rostami et al. 2017.

[5] 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.

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

How large is the full population 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 population 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 population 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[6]. 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 (1993; 2006a,b) standards and Jaynes (2003) for degrees of belief. We use 'degree of belief', 'belief', and 'probability' interchangeably.

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 'subpopulation'. Their spikes are binned into T time bins and binarized to $\{0, 1\}$ values in each bin. Call

[6] Jaynes 1996b; Porta Mana 2017a.

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 $(f_a) =: \boldsymbol{f}$. 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 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. We call this larger population the 'full population'. The total activity of these N neurons at t is A_t. The relative frequencies of the various activity levels during the recording were $(F_A) =: \boldsymbol{F}$. We don't know the values A_t at each t, or the frequency distribution $\boldsymbol{F}$. We only know for certain that $A_t \in \{0, 1, \dots, N\}$ and that $A_t \geqslant a_t$ 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 population 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 a probability distribution that approximately answers all these questions.

The idea behind our approach is easily summarized:

(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 M for the total activity A of the full population.

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

We now discuss the ideas behind these steps more in detail, leaving their precise mathematical implementation and a more detailed list of references to appendix A.

Step (a), a result of sampling theory, is just an application of the probability calculus, which gives an exact linear relation between the first m moments for the full population and the first m for the sample[7]. The ones determine the others and vice versa at every time bin. This relation holds for any belief distribution $P(A_t)$ for the full-population activity and its marginal $p(a_t)$ for the sample activity at that bin.

This sampling relation is even more straightforward if instead of power moments we use *normalized factorial moments*[8]. The mth normalized factorial moment of a distribution $p(a)$ is

$$\mathrm{E}\left[\binom{a}{m}\right] \Big/ \binom{n}{m} \coloneqq \sum_{a=0}^{n} \binom{a}{m} \Big/ \binom{n}{m} \, p(a), \qquad 1 \leqslant m \leqslant n, \tag{1}$$

that is, the expected number of distinct m-tuples of simultaneously spiking 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 convenient property for our analysis: *the first n normalized factorial moments for the sample and for the full population are numerically identical*:

$$\begin{aligned} \mathrm{E}\left[\binom{a}{m}\right] \Big/ \binom{n}{m} &= \mathrm{E}\left[\binom{A}{m}\right] \Big/ \binom{N}{m} \qquad \text{or} \\ \sum_{a=0}^{n} \binom{a}{m} \Big/ \binom{n}{m} \, p(a) &= \sum_{A=0}^{N} \binom{A}{m} \Big/ \binom{N}{m} \, P(A), \qquad 1 \leqslant m \leqslant n. \end{aligned} \tag{2}$$

[7] Porta Mana et al. 2015 eqs (16).
[8] Potts 1953.

In step (b) we actually use the *minimum-relative-entropy* method[9] with respect to a uniform reference distribution. We'll still call it 'maximum-entropy' for brevity. It amounts to two 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; and judge those expectations to be equal to some measured averages, typically time averages.

In our case we don't have measured time averages of the quantity $\binom{A}{m}$ to equate to the factorial moment $\mathrm{E}\bigl[\binom{A}{m}\bigr]$ of the full-population distribution $P(A)$. But eq. (2) of step (a) comes to our rescue, because it says that the expectations for the full population are determined by those for the sample $\mathrm{E}\bigl[\binom{a}{m}\bigr]$, and we do have the time averages of the corresponding sample quantity $\binom{a}{m}$. So we can combine the two steps:

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

and apply the maximum-entropy method to obtain a distribution $P_{\mathrm{me}}(A \mid M, N)$ for the full 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

$$\underbrace{\frac{1}{T}\sum_t \binom{a_t}{m}\Big/\binom{n}{m} \equiv \sum_a \binom{a}{m}\Big/\binom{n}{m} f_a}_{\text{measured moments}} = \underbrace{\sum_A \binom{A}{m}\Big/\binom{N}{m} P(A)}_{\text{distribution moments}}, \qquad m \in M. \tag{3}$$

The result is the distribution of the form

$$\begin{aligned} P_{\mathrm{me}}(A \mid M, N) &= \frac{1}{Z(\boldsymbol{\lambda})} \exp\left[\sum_m \lambda_m \binom{A}{m}\Big/\binom{N}{m}\right] \\ \text{with} \qquad Z(\boldsymbol{\lambda}) &\coloneqq \sum_A \exp\left[\sum_m \lambda_m \binom{A}{m}\Big/\binom{N}{m}\right], \end{aligned} \tag{4}$$

where the parameters $\boldsymbol{\lambda}$ are determined by the constraints (3). It is further discussed in appendix A.

[9] Hobson et al. 1973; Csiszár 1985; Sivia 2006 § 5.2.2.

The number and degrees of the constraining moments depend on the questions and hypotheses that a researcher is exploring; for example, hypotheses about the 'cooperativity' or 'interaction' in the population activity. We discuss this kind of use in § 4. Note that the constraint equation (3) for the distribution (4) may not have exact solutions if N is strictly larger than n; and the possible discrepancy typically increases with the number of constraints. The essential reason for this discrepancy is the approximate character of two assumptions behind the maximum-entropy and of one assumption specific to our application: measured moments equal expectations (or equivalently, the number of time bins T is infinite), the order of the time bins is irrelevant, and an activity level A can be equally likely be generated by any set of A neurons in the full population. The magnitude of this discrepancy can be a signature of the (at least temporary) presence of a 'neuronal assembly'[10]. We discuss this matter in § 6.

Before applying the formula above to two concrete data sets we want to add two remarks about the maximum-entropy method that are seldom made in the neuroscientific literature. They are important for the interpretation of the results and are further discussed in § 6. First, a maximum-entropy distribution like $P_{\text{me}}(A \mid M, N)$ is the zeroth-order approximation (in the sense of Laplace's approximation)[11] of four different distributions for the full 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 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 validity of the maximum-entropy approximation decreases as the ratio N/T increases. Second, the maximum-entropy method rests on some implicit assumptions about the probabilities for the long-run

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

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

frequency distribution of activities, besides the assumptions mentioned at the end of the previous paragraph. So the often heard statement that it gives 'the maximally unbiased (or non-committal) distribution' must be taken with a grain of salt.[12].

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

3 Example application: two data sets

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

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

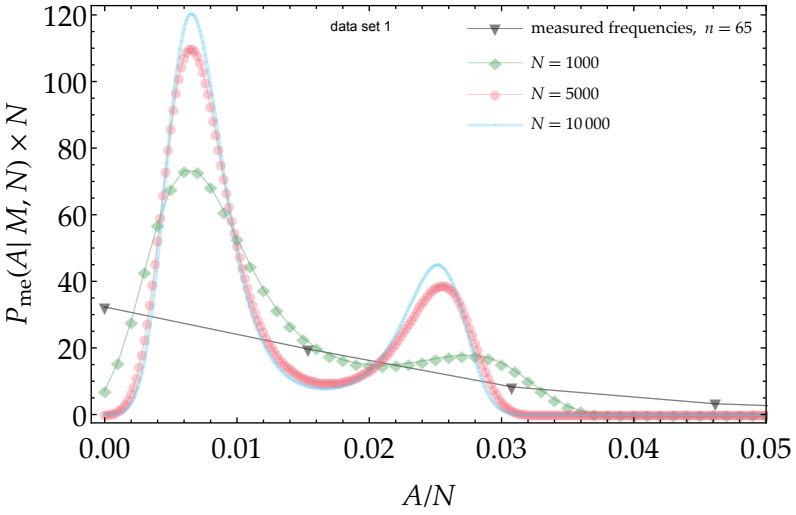
- The **second data set**, from Rostami et al. (2017 data courtesy by A. Riehle and T. Brochier), consists of $n = 159$ neurons from macaque Motor Cortex, recorded for about 15 minutes. Their spikes are binned into $T = 300\,394$ bins of 3 ms width.

The full-population maximum-entropy distribution is calculated using five moments. This number seems to provide almost as much information as the full frequency distribution of the sample (see next section). Figure 1 shows the resulting densities (distribution$\times N$) for three example values of full-population sizes: $N = 1\,000$ (green diamonds), $N = 5\,000$ (red circles), $N = 10\,000$ (blue curve). The frequency distribution of the sample activity is also shown (black triangles), and in the plot it would be indistinguishable from the maximum-entropy distribution 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 full-population is more peaked than the measured frequency distribution for the sample;

[12] Jaynes 1996b; Porta Mana 2009; 2017a move to sect.


data set 1
measured frequencies, $n = 65$
$N = 1000$
$N = 5000$
$N = 10\,000$
$P_{\mathrm{me}}(A|\,M, N) \times N$
A/N
0
20
40
60
80
100
120
0.00
0.01
0.02
0.03
0.04
0.05

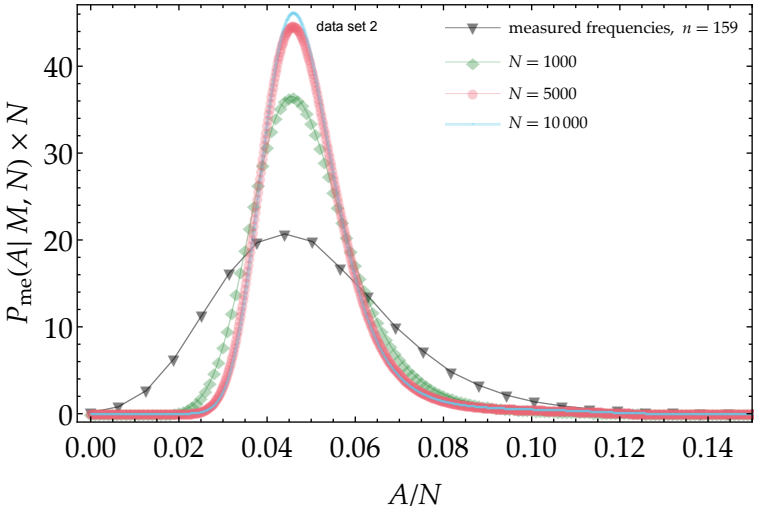

data set 2
measured frequencies, $n = 159$
$N = 1000$
$N = 5000$
$N = 10\,000$
$P_{\mathrm{me}}(A|\,M, N) \times N$
A/N
0
10
20
30
40
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14

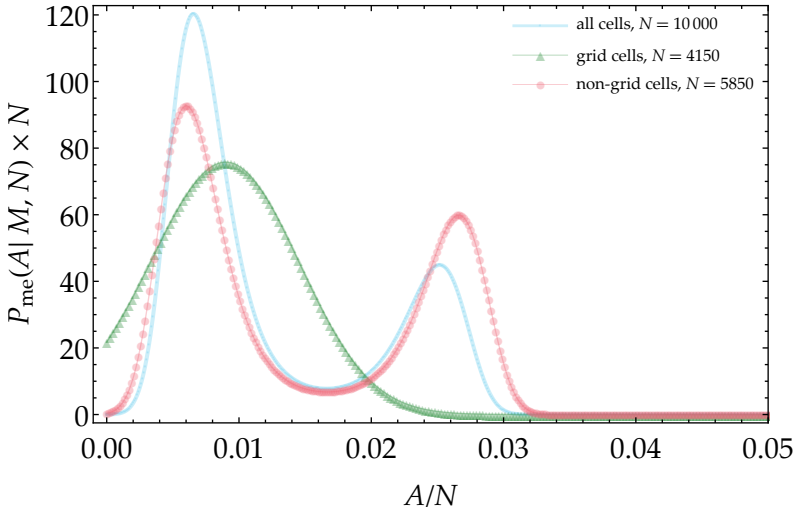
Figure 1 ***

their difference increases with N. Most remarkably, for the first data set (upper panel) the distribution for the full population has two distinct low-activity modes. For the second data set (lower panel) the distribution presents a small shoulder on the right of the mode. Such features are clearly not present in the sample frequencies or in the maximum-entropy distribution at the sample level. The application of the probability calculus thus reveals interesting possible features of the full population.

These results suggest that this maximum-entropy approach could be very useful. Here is a toy example of possible uses. 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 full-population distribution with two peaks, as in the upper panel of fig. 1, would provide some evidence for this hypothesis. Let's further imagine that we have reasons for suspecting that a specific set of the sampled neurons are of one type, and the remaining of the other type. In the case of the first data set, 27 of the 65 sampled neurons were identified as grid cells belonging to 3–4 functional modules, see Dunn et al. 2015. Could the two peaks in the distribution of fig. 1 reflect the activities of grid versus non-grid cells? We apply the method to these two sets of neurons individually, using $N = 4\,150$ for the grid set and $N = 5\,850$ for the non-grid set, to reflect their proportions in the recorded sample. The results are shown in the upper panel of fig. 2. The distribution for the larger population of grid cells (green triangles) seems to have one, broad peak, close to the first peak of the full-population distribution. The distribution for the larger population of non-grid cells (red circles) has two peaks instead, roughly at the same positions as the full-population distribution (blue curve) but closer in height. So it would seem that the population of grid cells is contributing to the first peak of the full population, but it isn't the sole contributor. We can also assess how much the distributions for the two sets are independent. If they were completely independent, the full-population distribution would be given by their convolution:

$$P_{\text{full}}(A) = \sum_{A'=0}^{A} P_{\text{grid}}(A')\, P_{\text{non-grid}}(A - A'). \tag{5}$$

However, the lower panel of figure 2 shows that such convolution (black diamonds) is different from the full-population distribution (blue curve): the two peaks of the former are closer in position and height. This


all cells, $N = 10\,000$
grid cells, $N = 4150$
non-grid cells, $N = 5850$
$P_{\mathrm{me}}(A|\,M, N) \times N$
A/N
0
20
40
60
80
100
120
0.00
0.01
0.02
0.03
0.04
0.05

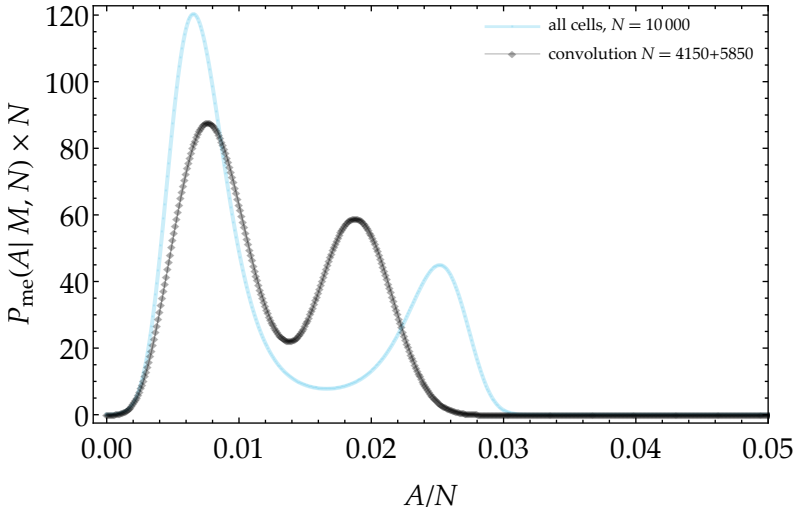

all cells, $N = 10\,000$
convolution $N = 4150+5850$
$P_{\mathrm{me}}(A|\,M, N) \times N$
A/N
0
20
40
60
80
100
120
0.00
0.01
0.02
0.03
0.04
0.05

Figure 2 ***

means that the population distributions of grid and non-grid cells are *not* independent: knowledge of the activity of either set gives us some information about the activity of the other.

The toy analysis just given should be taken not literally, but just as an illustration of the method's possible applications. The important point is that this method is computationally very cheap but can provide useful insights, even if just qualitatively.

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

To illustrate how our approach can be applied to studies of 'cooperativity' or 'interaction', fig. 3 shows the full-population distributions, for $N = 10\,000$, obtained constraining the first two moments (in red, equivalent to constraining means and correlations) and the first four moments (in blue). The frequency distribution for the sample is also shown for comparison. In both data sets the two-moment constraint leads to a distribution quite different from the that of the four-moment constraint. For the first data set in particular, one distribution is bimodal, the other unimodal. Interpreted in terms of This visual difference shows directly that the two sets of moments are not even approximately equivalent.

It is instructive to contrast this conclusion with that we would have reached by applying the maximum-entropy method *at the sample level*, as traditionally done. Such a comparison involves some important methodological caveat which we wish to discuss first.

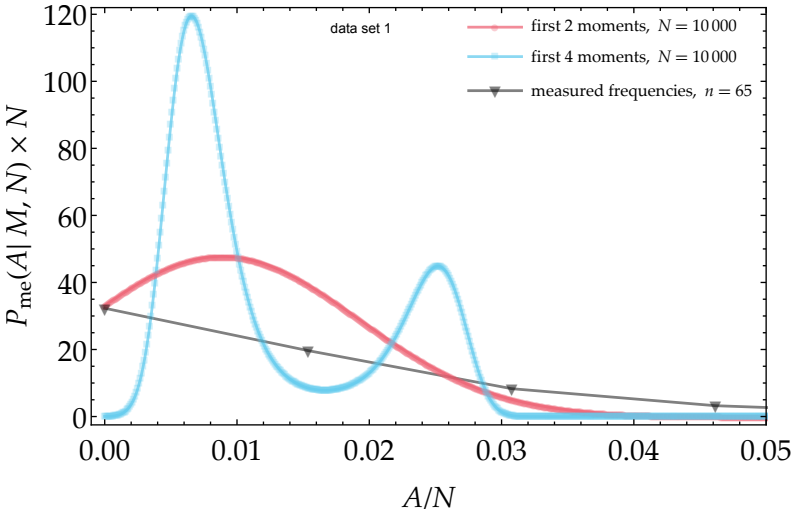
As mentioned in the Introduction (see references there), in the neurosciences the maximum-entropy method has also been used as a way of quantifying the 'cooperativity'[13] or 'interaction'[14] or 'synchrony'[15] of neuronal activity. In this section we discuss how our proposed application bears on this 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

[13] e.g. Gerstein et al. 1985.

[14] e.g. Martignon et al. 1995; Schneidman et al. 2006; Shlens et al. 2006.

[15] e.g. Bohte et al. 2000; Amari et al. 2003 – we're only citing early papers using these terms.


data set 1
first 2 moments, $N = 10\,000$
first 4 moments, $N = 10\,000$
measured frequencies, $n = 65$
$P_{\mathrm{me}}(A|\,M, N) \times N$
A/N
0
20
40
60
80
100
120
0.00
0.01
0.02
0.03
0.04
0.05

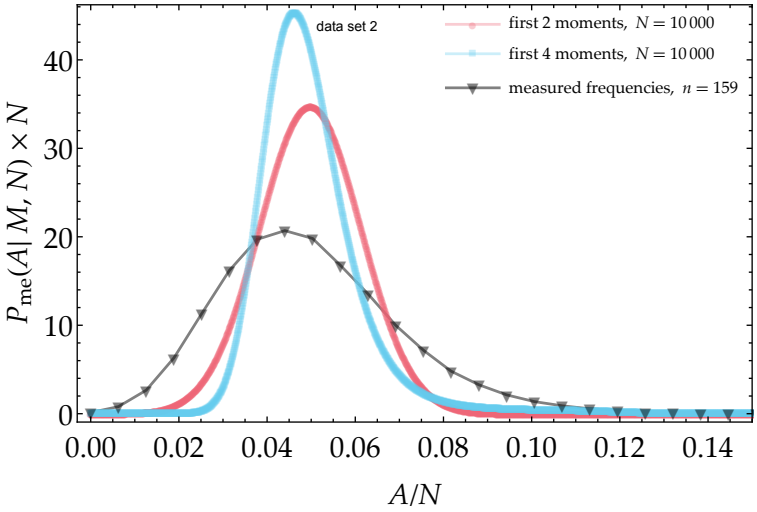

data set 2
first 2 moments, $N = 10\,000$
first 4 moments, $N = 10\,000$
measured frequencies, $n = 159$
$P_{\mathrm{me}}(A|\,M, N) \times N$
A/N
0
10
20
30
40
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14

Figure 3 ***

of *informational sufficiency*[16] 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*.[17]

There's a tight connection between informational sufficiency and maximum-entropy distributions[18]: if a probability distribution for repetitive phenomena has a sufficient statistics, then by the Pitman-Koopman theorem[19] 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,

[16] Bernardo et al. 1994 § 4.5; Jaynes 2003 ch. 8 & § 14.2; Cifarelli et al. 1982; Kullback et al. 1951; the notion goes back to Fisher 1922.

[17] For more technical results and connections with the notion of symmetry see e.g. Darmois 1935; Neyman 1935; Koopman 1936; Pitman 1936; Halmos et al. 1949; Bahadur 1954; Berk 1972; Lauritzen 1974a; 1988; 2007; Cifarelli et al. 1980; 1981; Diaconis et al. 1981; Diaconis 1992; Furmańczyk et al. 1998; Fortini et al. 2000; Nogales et al. 2000; Kallenberg 2005; Ay et al. 2015.

[18] Jaynes 1982; Bernardo et al. 1994 § 4.5.4.

[19] Koopman 1936; Pitman 1936; Darmois 1935; for later analyses and the discrete case see Hipp 1974; Andersen 1970; Denny 1967; 1972; Fraser 1963; Barankin et al. 1963; Barndorff-Nielsen 2014.

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[20]. Mathematically this impossibility comes from the Pitman-Koopman theorem mentioned above, and translates into the general impossibility of solving a system of n independent equations in m unknowns with more equations than unknowns, $n > m$[21].

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 population 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 population, 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 $M'' \coloneqq \{1, \dots, m''\}$, and another, say $M' \coloneqq \{1, \dots, m'\}$, as follows:

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

$$p(a \mid M) = \sum_A G_{aA} P_{\text{me}}(A \mid M), \qquad M = M', M''; \tag{6}$$

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

$$T\,\mathrm{H}[\boldsymbol{f}; p(a \mid M)] \coloneqq T \sum_a f_a \log \frac{f_a}{p(a \mid M)}, \qquad M = M', M''; \tag{7}$$

[20] e.g. Maes et al. 1999 and references therein.

[21] Porta Mana et al. 2015 § 3.1.

(iii) take the difference:

$$\Delta(M'', M') \coloneqq T\,\mathrm{H}[\boldsymbol{f}; p(a \mid M')] - T\,\mathrm{H}[\boldsymbol{f}; p(a \mid M'')] \equiv T \sum_a f_a \log \frac{\sum_A G_{aA} P_{\mathrm{me}}(A \mid M'')}{\sum_A G_{aA} P_{\mathrm{me}}(A \mid M')}. \quad (8)$$

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

Why is this a natural measure? Because $\Delta(M'', M')$ is equal to the log-ratio of the probabilities of the data $\boldsymbol{f}$ conditional on the hypotheses M'' and M':

$$\Delta(M'', M') = \log\bigl[\mathrm{p}(\boldsymbol{f} \mid M'')/\mathrm{p}(\boldsymbol{f} \mid M')\bigr]. \quad (9)$$

This is called their *relative weight of evidence*, the logarithm of their *relative Bayes factor*[22]. We prove this equality in appendix B. So the exponential of $\Delta(M'', M')$ tells us how much more probable the data $\boldsymbol{f}$ are conditional on M'', than conditional on 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[23].

In our case consider for example three sets of constraints M_2, M_4, M_5, consisting of the first two, four, five moments. For a full population of size $N = 10\,000$, we obtain the following differences:

$$\begin{aligned} \Delta(M_4, M_2) &= 81 \text{ nat} = 35 \text{ Hart}, \\ \Delta(M_5, M_4) &= 0.037 \text{ nat} = 0.016 \text{ Hart}, \end{aligned} \quad (10)$$

where the Hartley (Hart) denotes base-10 logarithms[24]. In words, the measured frequencies of the sample activity are 35 orders of magnitude more probable under sufficiency of the first four moments than under sufficiency of the first two only; but they are about equally probable ($10^{0.016} = 1.04$) whether we consider the first five moments to be sufficient or just the first four.

[22] Good 1950 ch. 6; 1975; 1981; 1985; 1983; Osteyee et al. 1974 § 1.4; MacKay 1992; Kass et al. 1995; see also Jeffreys 1936 p. 421; 1983 chs V, VI, A.

[23] cf. Bretthorst 2013.

[24] Iso 2009 § C.4; it was called 'ban' and used by Turing and Good in their code-breaking work at Bletchley Park: Good 1985; 1950; 1969; Jaynes 2003 § 4.2.

Compare the results above with those obtained by applying the method at the sample level only:

$$\begin{aligned} \Delta_{\text{sample level}}(M_4, M_2) &= 280 \text{ nat} = 1\,220 \text{ Hart}, \\ \Delta_{\text{sample level}}(M_5, M_4) &= 3.3 \text{ nat} = 1.4 \text{ Hart}, \end{aligned} \tag{11}$$

that is, the data are 1 220 orders of magnitude more probable conditional on four moments than conditional on two, and $10^{1.4} = 25$ times more probable conditional on five moments than on four moments.

The application at the full-population level therefore leads to different conclusions about the relative importance of the moments from the application at the sample level. It is also possible to use the measure (8) and its probabilistic meaning (9) to compare the probability of the data $\boldsymbol{f}$ conditional on the hypothesis $M_{5,N}$ of 5-moment sufficiency at the full-population level, $N = 10\,000$, versus the hypothesis $M_{5,n}$ of 5-moment sufficiency at the sample level, $N = n = 65$. We obtain

$$\Delta(M_{5,N}, M_{5,n}) = 35 \text{ nat} = 15 \text{ Hart}, \tag{12}$$

that is, the data are 15 orders of magnitude more probable under the first hypothesis than under the second.

The conclusions about 'cooperativity' or 'interaction' that we reach by applying the maximum-entropy method to the full population, as proposed here, are not only different from the application at the sample level, but also visually clearer. The plot of fig. 3, showing the two clearly different distributions constrained by two and four moments, should be compared with the plot for the distributions obtained at the sample level under the same constraints, shown in fig. 4 as dashed lines with empty red circles (○) and empty blue squares (□). The measured frequencies are shown as filled black triangles (▽). It is necessary to use a logarithmic scale to see the differences, which appear mainly in the tail.

The same plot also shows the marginals obtained from applying the method to a full population with $N = 10\,000$ and the same constraints, that is, the sample marginals of fig. 3. The two-moment distribution is represented by filled red circles (●), and the four-moment one by filled blue squares (■). The full-population distribution fits the measured frequencies (▽) very well if we constrain only four moments.

5 The full-population size N

The calculations and conclusions presented in the preceding sections depend on the size of the full population, N. The full population 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 population and those of the sample. This formula is only valid if our beliefs about the activity of the sample $p(a)$ and that of the full population $p(A)$ are related by a hypergeometric distribution:

$$p(a) = \binom{n}{a}\binom{N-n}{A-a}\binom{N}{A}^{-1} P(A), \tag{13}$$

characteristic of 'drawing without replacement'[25]. In other words, N is the size of the pool of all neurons that could just as well have been recorded by our measuring procedure. 'Just as well' indicates equal degrees of belief. Whether the equality of such beliefs is justified, or

[25] Jaynes 2003 ch. 3; Ross 2010 § 4.8.3; Feller 1968 § II.6.

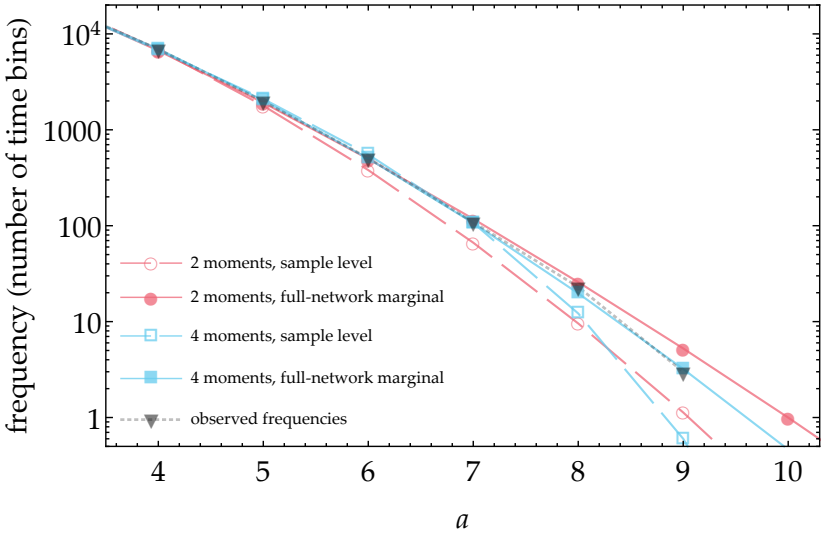


Figure 4 ***

for which sampling procedures it could be justified, is the fundamental, deep question of sampling theory, which we cannot discuss here of course[26]. 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; and similar. N can only be assessed case by case. The plots of fig. 1 suggest that its order of magnitude might 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 of course assign a probability to the possible values of N based on our background information and the frequency data $\boldsymbol{f}$. The procedure is similar to that of § 4 and appendix***. The probability for the data $\boldsymbol{f}$ conditional on a set of constraints M and N is the exponential of minus the relative entropy (7):

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

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 a simple 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 is 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 1/5 about these partial orders of magnitude.

[26] see e.g. the discussions, reviews, and references in Ericson 1969a; Smith 1976.

Thus from (16) we find

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

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

In this way we can make inferences – for example about the sufficiency of a set of moments, or about the marginal sample distribution – that take into account our uncertainty about N. We must make sure to avoid circularities, though: for example, we can't 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 whole maximum-entropy approach superfluous. We discuss this final point in the next section.

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 population 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 $\boldsymbol{F}$ is quantified by an entropic prior[27]:

$$\mathrm{p}(\boldsymbol{F} \mid I) \propto \exp[-L\ \mathrm{H}(\boldsymbol{F};\boldsymbol{R})] \approx \binom{L}{LF_0, \dots, LF_N} \prod_A {R_A}^{LF_A} \qquad (18)$$

where $\mathrm{H}(\boldsymbol{F};\boldsymbol{R}) \coloneqq \sum_A F_A \log(F_A/R_A)$ is the relative entropy or discrimination information[28], $\boldsymbol{R}$ 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 $\boldsymbol{F}$ 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 $\boldsymbol{F}$; in our case this means $L \approx 10$, give or take an order of magnitude.

[27] Neumann 2007; Rodríguez 1991; Skilling 1998; Caticha et al. 2004; Porta Mana 2017a.
[28] Kullback 1987; Jaynes 1963; Hobson 1969; Hobson et al. 1973.

We could obviously consider pre-data beliefs different from (18), for example one quantified by a Dirichlet distribution (which is equivalent to the above but with $\boldsymbol{F}$ and $\boldsymbol{R}$ switched), or a uniform distribution in $\boldsymbol{F}$-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 population 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 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 full population constrained by a set of factorial moments. For further details see Porta Mana et al. (2015).

First of all the sampling relation. We have a set of N units, A of which have some specific property, and we sample in an unknown way n of the N units. The probability that a of the n sampled units have that property is then given by the hypergeometric distribution

$$G_{aA} \coloneqq \mathrm{p}(a \mid A, n, N) = \binom{n}{na}\binom{N-n}{NA-na}\binom{N}{NA}^{-1}. \tag{19}$$

typical of 'drawing without replacement'[29]. In the following leave n, N implicit in the conditional. If we are uncertain about the number A, with belief $P(A)$, then by the theorem of total probability our belief about a is

$$p(a) = \sum_{A} G_{aA}\, P(A). \tag{20}$$

In our case the units are neurons, the set is the full population, and the property is their being active in a specific time bin.

From the definition of normalized factorial moment (1), the expression for the hypergeometric distribution (19), and the relation (20) between our beliefs about a and A, and using some combinatorial juggling[30], one can prove the equality (2) between the factorial moments of sample and full population.

For the construction of a maximum-entropy distribution from generic expectation constraints see Jaynes (1963; 2003 ch. 11). More precisely we use the minimum relative-entropy method[31] with respect to a reference distribution. 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).

[29] Jaynes 2003 ch. 3; Ross 2010 § 4.8.3; Feller 1968 § II.6.

[30] Whitworth 1965 chs I–IV; Feller 1968 ch. II; Porta Mana et al. 2015 appendix A; Potts 1953.

[31] Hobson et al. 1973; Csiszár 1985; Sivia 2006 § 5.2.2.

The result has the standard exponential-family form

$$\begin{aligned} P_{\mathrm{me}}(A) &= \frac{1}{Z(\boldsymbol{\lambda})}\, g(A)\, \exp\biggl[\sum_m \lambda_m \binom{A}{m}\binom{N}{m}^{-1}\biggr], \\ Z(\boldsymbol{\lambda}) &\coloneqq \sum_A g(A)\, \exp\biggl[\sum_m \lambda_m \binom{A}{m}\binom{N}{m}^{-1}\biggr], \end{aligned} \tag{21}$$

where $g(A)$ is the reference distribution and $\boldsymbol{\lambda} \coloneqq (\lambda_m)$ are the Lagrange multipliers, satisfying the implicit constraint equations (3):

$$\sum_A \tbinom{A}{m}\tbinom{N}{m}^{-1} \tfrac{1}{Z(\lambda)}\, g(A)\, \exp\Bigl[\sum_m \lambda_m \tbinom{A}{m}\tbinom{N}{m}^{-1}\Bigr] = \sum_a \tbinom{a}{m}\tbinom{n}{m}^{-1} f_a, \quad m \in M. \tag{22}$$

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

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 $\boldsymbol{f}$:

$$\mathrm{p}(M'' \mid \boldsymbol{f})/\mathrm{p}(M' \mid \boldsymbol{f}). \tag{23}$$

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 population, owing to the excheangeability assumption implicit in the maximum-entropy method:

$$p(a_t \mid M) = \sum_A G_{a_t A} P_{\mathrm{me}}(A \mid M). \tag{24}$$

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

$$\prod_{t=1}^{T} p(a_t \mid M) \equiv \prod_{a=0}^{n} p(a \mid M)^{T f_a} \equiv \prod_{a=0}^{n} \Bigl[\sum_A G_{aA} P_{\mathrm{me}}(A \mid M)\Bigr]^{T f_a}. \tag{25}$$

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

$$\binom{T}{T\boldsymbol{f}} \coloneqq \frac{T!}{\prod_a (T\,f_a)!} \approx \prod_a {f_a}^{-T\,f_a}, \tag{26}$$

the last expression coming from Stirling's approximation[32]. If we assign equal pre-data probabilities to the two hypotheses M' and M'', each probability in the ratio (23) then becomes, by Bayes's theorem,

$$\mathrm{p}(M \mid \boldsymbol{f}) \propto \mathrm{p}(\boldsymbol{f} \mid M) \times \mathrm{const} \propto \binom{T}{T\boldsymbol{f}} \prod_a \Big[\sum_A G_{aA} P_{\mathrm{me}}(A \mid M)\Big]^{T\,f_a} \approx$$
$$\prod_a {f_a}^{-T\,f_a} \times \prod_a \Big[\sum_A G_{aA} P_{\mathrm{me}}(A \mid M)\Big]^{T\,f_a}. \tag{27}$$

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 (8) is the logarithm of the probability ratio (23). The exponential of the difference (8) tells us how much more probable is the set M'' to be sufficient than the set M'.

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.

Ay, N., Jost, J., Lê, H. V., Schwachhöfer, L. (2015): *Information geometry and sufficient statistics*. Probab. Theory Relat. Fields **162**[1–2], 327–364.

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).

Barankin, E. W., Maitra, A. P. (1963): *Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics*. Sankhyā A **25**[3], 217–244.

Barndorff-Nielsen, O. E. (2014): *Information and Exponential Families: In Statistical Theory*, reprint. (Wiley, New York). First publ. 1978.

[32] Csiszár et al. 2004 Lemma 2.2.

Barndorff-Nielsen, O. E., Blæsild, P., Schou, G., eds. (1974): *Proceedings of Conference on Foundational Questions in Statistical Inference: Aarhus, May 7–12, 1973*. (University of Aarhus, Aarhus).

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.

Berk, R. H. (1972): *A note on sufficiency and invariance*. Ann. Math. Stat. **43**[2], 647–650.

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).

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.

Browder, F. E., ed. (1992): *Mathematics into the Twenty-first Century: 1988 Centennial Symposium August 8–12*. (American Mathematical Society, Providence, USA).

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

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

— (1982): *Some considerations about mathematical statistics teaching methodology suggested by the concept of exchangeability*. In: Koch, Spizzichino (1982), 185–205.

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

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/.

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

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

Denny, J. L. (1967): *Sufficient conditions for a family of probabilities to be exponential*. Proc. Natl. Acad. Sci. (USA) **57**[5], 1184–1187.

— (1972): *Sufficient statistics and discrete exponential families*. Ann. Math. Stat. **43**[4], 1320–1322.

Diaconis, P. (1992): *Sufficiency as statistical symmetry*. In: Browder (1992), 15–26. First publ. 1991 as technical report https://statistics.stanford.edu/research/sufficiency-statistical-symmetry.

Diaconis, P., Freedman, D. (1981): *Partial exchangeability and sufficiency*. In: Ghosh, Roy (1981), 205–236. http://statweb.stanford.edu/~cgates/PERSI/year.html. Also publ. 1982 as technical report https://www.stat.berkeley.edu/~aldous/206-Exch/Papers/diaconis_freedman_PES.pdf.

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

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

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).

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

Fisher, R. A. (1922): *On the mathematical foundations of theoretical statistics*. Phil. Trans. R. Soc. Lond. A **222**, 309–368. http://www.stats.org.uk/statistical-inference/Fisher1922.pdf.

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

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

Fraser, D. A. S. (1963): *On sufficiency and the exponential family*. J. Roy. Stat. Soc. B **25**[1], 115–123.

Furmańczyk, K., Niemiro, W. (1998): *Sufficiency in Bayesian models*. Applicationes Mathematicae **25**[1], 113–120.

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).

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

Ghosh, J. K., Roy, J., eds. (1981): *Statistics: Applications and New Directions*. (Indian Statistical Institute, Calcutta).

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

— (1969): *A subjective evaluation of Bode's law and an 'objective' test for approximate numerical rationality*. J. Am. Stat. Assoc. **64**[325], 23–49. Partly repr. in Good (1983) ch. 13.

— (1975): *Explicativity, corroboration, and the relative odds of hypotheses*. Synthese **30**[1–2], 39–73. Partly repr. in Good (1983) ch. 15.

— (1981): *Some logic and history of hypothesis testing*. In: *Philosophy in economics*. Ed. by J. C. Pitt (Reidel), 149–174. Repr. in Good (1983) ch. 14 pp. 129–148.

— (1983): *Good Thinking: The Foundations of Probability and Its Applications*. (University of Minnesota Press, Minneapolis, USA).

Good, I. J. (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. (1980): *Principle of maximum entropy and irreversible processes*. Phys. Rep. **62**[3], 175–266.

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

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

Halmos, P. R., Savage, L. J. (1949): *Application of the Radon-Nikodym theorem to the theory of sufficient statistics*. Ann. Math. Stat. **20**[2], 225–241.

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.

Hipp, C. (1974): *Sufficient statistics and exponential families*. Ann. Stat. **2**[6], 1283–1292.

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

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

ISO (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.

— (1982): *On the rationale of maximum-entropy methods*. Proc. IEEE **70**[9], 939. 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).

— (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).

Jeffreys, H. (1983): *Theory of Probability*, third ed. with corrections. (Oxford University Press, London). First publ. 1939.

Kallenberg, O. (2005): *Probabilistic Symmetries and Invariance Principles*. (Springer, New York).

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.

Koch, G., Spizzichino, F., eds. (1982): *Exchangeability in Probability and Statistics*. (North-Holland, Amsterdam).

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

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

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

Lauritzen, S. L. (1974a): *Sufficiency, prediction and extreme models*. In: Barndorff-Nielsen, Blæsild, Schou (1974), 249–269. With discussion. Repr. without discussion in Lauritzen (1974b).

— (1974b): *Sufficiency, prediction and extreme models*. Scand. J. Statist. **1**[3], 128–134.

— (1988): *Extremal Families and Systems of Sufficient Statistics*. (Springer, Berlin). First publ. 1982.

— (2007): *Sufficiency, partial exchangeability, and exponential families*. http://www.stats.ox.ac.uk/~steffen/teaching/grad/partial.pdf. Lecture notes.

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. (1992): *Bayesian interpolation*. Neural Comp. **4**[3], 415–447. http://www.inference.phy.cam.ac.uk/mackay/PhD.html.

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

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

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

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

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

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

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

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

Neyman, J. (1935): *Su un teorema concernente le cosiddette statistiche sufficienti*. Giorn. Ist. Ital. Att. **VI**[4], 320–334.

Nogales, A. G., Oyola, J. A., Pérez, P. (2000): *On conditional independence and the relationship between sufficiency and invariance under the Bayesian point of view*. Stat. Probab. Lett. **46**[1], 75–84.

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

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

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

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

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

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

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

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

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

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

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

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

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

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