

Bayesian inference for a neuronal network from subnetwork data


P.G.L. Porta Mana

Kavli Institute, Trondheim, Norway <piero.mana@ntnu.no>

Draft of 31 October 2019 (first drafted 15 May 2019)

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

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

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





1 Introduction

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

Answering this kind of questions always engages an element of uncertainty. Our answers therefore involve experimental data, such as neuronal recordings from specific brain areas, and probabilities or degrees of belief, based on prior knowledge, about biological conditions and mechanisms that cannot be experimentally ascertained. Such degrees of belief are often formulated as simplified ‘models’ to be mathematically more tractable.

Despite remarkable advances in recording technologies, the best experimental measurements of neuronal activity can still only record a very small sample of neurons compared to the numbers that constitute a functionally distinguished brain region. Many probabilistic models focus on such samples only, somehow neglecting, in their assumptions, that the recorded neurons are a sample from a larger network. This kind of isolation assumptions sometimes escape attention, being subtly hidden in the mathematics. Some probabilistic models try to take also unrecorded neurons into account, but become very complex in doing so.

In the present work we give an answer to this question: How much can the total activity of a large neuronal network be, if we have observed the activity of a very small sample thereof? We'll quantify our degrees of belief about the possible answers by combining, in a straightforward way, the maximum-entropy method and basic sampling relations of the probability calculus. Later on we'll show that our degrees of belief can be quantified by exclusively using the probability calculus. This derivation will provide a more accurate quantification, revealing that the maximum-entropy answer is only a first approximation.

We apply our approach to two concrete data sets: the activity of 65 neurons recorded from a rat's Medial Entorhinal Cortex  [ref](#), and the activity of 159 neurons recorded from a macaque's Motor Cortex  [ref](#)  [add recording length](#). In the first data set, the most interesting finding is that the most probable frequency distribution for the total activity of the full network has two very distinct modes, both at low activities, see [fig. 1](#). The analogous frequency distribution for the second data set doesn't have two modes but still presents one prominent shoulder in its one low-activity mode. Note that these guessed features of the full network aren't observed in the sample, nor can they be inferred by the application of maximum-entropy *to the sample alone*.  [Monte Carlo sampling: say something about the deviation from the most probable frequency distribution](#)

 [shall we give a two-sentence summary of the main idea here?](#)

The maximum-entropy or minimum-relative-entropy method¹ has been used for different kinds of estimations of the neuronal activity of various brain areas and about other phenomena of importance to the

¹Jaynes [1957a](#); much clearer in Jaynes [1963](#)[jaynes1985b](#); Sivia [2006](#); Hobson et al. [1973](#); Grandy [1980](#).

neurosciences, for example gene and protein interaction². This method is often used to test whether some statistics of the data, for example second-order time correlations, is sufficient for quantifying our degree of belief about some quantities of the system. It can be considered as an approximation of probabilistic models based on various assumptions of inferential sufficiency³.

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

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

✚ To be continued after structure of the rest of the article is clear. Orig intro is on p. ??

Our notation and terminology follow ISO standards⁴ and Jaynes (2003) for degrees of belief. We often simply say 'belief' for 'degree of belief'.

2 Model: maximum-entropy and sampling

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

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

²for 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. 2009c; Barreiro et al. 2010a; Gerwinn et al. 2010; Macke et al. 2011a; 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.

³Jaynes 1996; Porta Mana 2017a.

⁴ISO 1993; 2006a,b.

appeared during the recording, obtaining the distribution of measured relative frequencies (f_a) =: f . We can also consider the sample activity at time bins *outside* of the recorded range. Such activity is unknown to us, of course.

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

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

✚ Move this to intro? We want to stress the usefulness of making quantified guesses about the full-network activity. First of all, this seems to be the primary idea behind recording a sample. Second, it allows us to make comparisons across experimental sessions; such comparisons would be difficult or meaningless if made with the recorded samples,

which generally comprise non-overlapping sets of neurons and differ in size.

The idea behind our approach is easily summarized:

- (a) we build a distribution $p_{\text{me}}(A)$ for the total activity of the full network using the maximum-entropy method;
- (b) the constrained averages used in the maximum-entropy method for the full network are, in turn, determined via sampling theory from the constrained averages for the sample.

Let's discuss these points in detail.

Regarding (a), we assume that you're familiar with the maximum-entropy method. We actually use the *minimum-relative-entropy* method⁵, but call it 'maximum-entropy' for brevity. It amounts to a pair of prescriptions: choose the distribution, among those satisfying specific convex constraints, such as fixed expectations, that has minimum relative entropy with respect to a reference distribution, often taken to be the uniform one; and judge those expectations to be equal to measured averages. We add two remarks about this method that are seldom made in the literature. First, the distribution $p_{\text{me}}(A)$ given by this method is the zeroth-order approximation⁶ of four different distributions for the full network:

- the most probable *frequency* distribution for the total activity across the *recorded* bins,
- the *belief* distribution for the value of the total activity at any time bin among those *recorded*,
- the most probable *frequency* distribution for the total activity in a very long run of *new* time bins,
- the *belief* distribution for the value of the total activity at a *new* time bin.

The maximum-entropy distribution is thus an approximation of our belief distribution about four completely different quantities. Note that the four distributions above numerically differ in higher-order approximations. We discuss their differences in §***. Second, the maximum-entropy

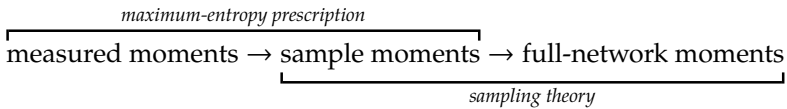
⁵Hobson et al. 1973.

⁶in the sense of Laplace's method: De Bruijn 1961 ch. 4; Tierney et al. 1986; Strawderman 2000.

method based on the Shannon entropy implicitly makes some assumptions about the probabilities for the long-run frequency distributions⁷. We discuss these assumptions in §***.

In our case, to apply the method for the total activity of the full network we need to fix some of the latter's averages, for example its moments. But we don't have any measured moments for the full network to equate the distribution moments to. Here enters point (b): the probability calculus gives an exact, linear relation between the first m moments for the full network and the first m for the sample⁸; the ones determine the others and vice versa at every time bin. This relation is a classical result of sampling theory⁹.

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



These two steps are more straightforward if instead of power moments we use *normalized factorial moments*¹⁰. The m th normalized factorial moment of a distribution $p(a)$ for the activity of the sample neurons is defined as the average

$$\sum_a \binom{na}{m} \binom{n}{m}^{-1} p(a), \quad 1 \leq m \leq n \quad (1)$$

This moment can be interpreted as the expectation of the number of distinct m -tuples of simultaneously spiking neurons (within a bin's time width), normalized by the number of distinct m -tuples. For example, with $m = 2$, if $na = 4$ neurons spike in a network of $n = 5$, we have $\binom{4}{2} = 6$ distinct pairs of simultaneously spiking neurons, and the total number of distinct pairs is $\binom{5}{2} = 10$. The normalized number of spiking pairs is therefore $6/10$. Note that the first m factorial moments linearly determine

⁷Jaynes 1996; Porta Mana 2009; 2017a.

⁸Porta Mana et al. 2015 eqs (16).

⁹Whitworth 1965 chs I-IV; Feller 1968 ch. II; Jaynes 2003 ch. 3; see also Whitworth 1897.

¹⁰Potts 1953.

the first m power moments and vice versa, because $\binom{na}{m}$ is a polynomial in a of degree m ; so fixing the ones is equivalent to fixing the others. Now we have this extremely convenient property: *the first n normalized factorial moments for a sample and for the full network are numerically identical*:

$$\sum_a \binom{na}{m} \binom{n}{m}^{-1} p(a) = \sum_A \binom{NA}{m} \binom{N}{m}^{-1} p(A), \quad 1 \leq m \leq n, \quad (2)$$

where $p(A)$ is the distribution for the full-network activity.

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

$$\frac{1}{T} \sum_t \binom{na_t}{m} \binom{n}{m}^{-1} \equiv \sum_a \binom{na}{m} \binom{n}{m}^{-1} f_a = \sum_A \binom{NA}{m} \binom{N}{m}^{-1} p_{\text{me}}(A) \quad (3)$$

for all m we wish.

The calculation amounts to a convex optimization¹¹ and for the numbers N, n, m considered in the next section it can be done on a modern computer without approximations of normalization constants or potential functions.

The number of moments used with this method depends on the questions and hypotheses that a researcher is exploring; for example, hypotheses of sufficient statistics, such as the sufficiency of pairwise correlations to quantify our degree of belief about network activity. In the present work we only want to introduce the general method without entering into biological questions of this kind.

The particular case in which n moments are constrained is especially important: it corresponds to fully constraining the marginal frequency distribution for the activity of the sample neurons. In this case, our belief about the full-network activity is based on all available measured frequency data. Note that application of the maximum-entropy method *at the sample level* is trivial and meaningless in this case – it just gives back

¹¹Mead et al. 1984; Press et al. 2007 ch. 10; Fang et al. 1997; Boyd et al. 2009; Porta Mana 2017b.

the measured frequency distribution. But application of the method *at the level of the full network* is not trivial.

Using the full frequency distribution of the sample may be a bad idea, however, because the maximum-entropy distribution may become a bad approximation of some of the four distributions described above. It is preferable to use a moderately high number of moments smaller than n . We explain this point in §***.


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

3 Application: two data sets

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

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

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

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

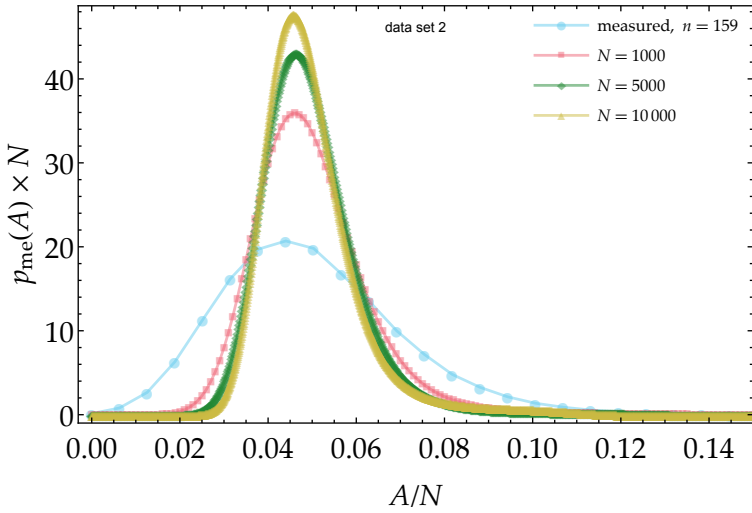
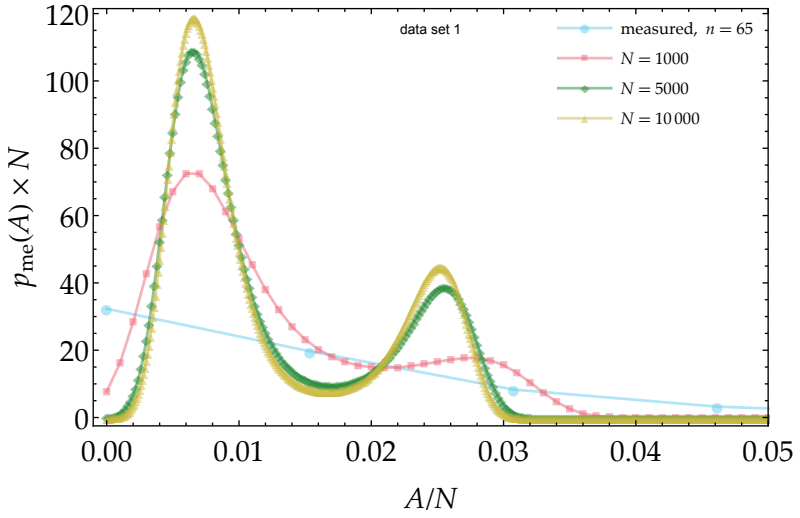


Figure 1 ***

The distribution for the full-network is sharper than the measured frequency distribution for the sample; the sharper the larger N is. Most remarkably, it has two distinct low-activity modes for the first data set. But also the second data set presents, upon closer inspection, a small shoulder on the right of the mode, suggestive of two activity regimes. We frame no hypotheses about the biological cause of these two modes (they could stem from the presence of different kinds of cells or modules). These features are clearly not present in the sample or in the maximum-entropy distribution at the sample level. The application of the probability calculus thus reveals interesting possible features of the full network.

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

4 Assumptions and corrections: beyond the maximum-entropy approximation

In § 2 we mentioned that maximum-entropy distributions come from a zeroth-order Laplace approximation of the densities obtained from the probability calculus. Zeroth-order means that we are simply considering the mode of such densities. Let's be more concrete.

Consider the question: What is the relative-frequency distribution for the different activity levels for the full-network, across the recorded bins? Call this distribution $(F_A) = F$: activity level A appeared in $T \times F(A)$ out of T time bins. We don't know F , and our beliefs about its possible values have a distribution $p(F | DI)$ conditional on data D and prior knowledge I , which we approximate with a continuous distribution

$$p(F | DI) dF. \tag{4}$$

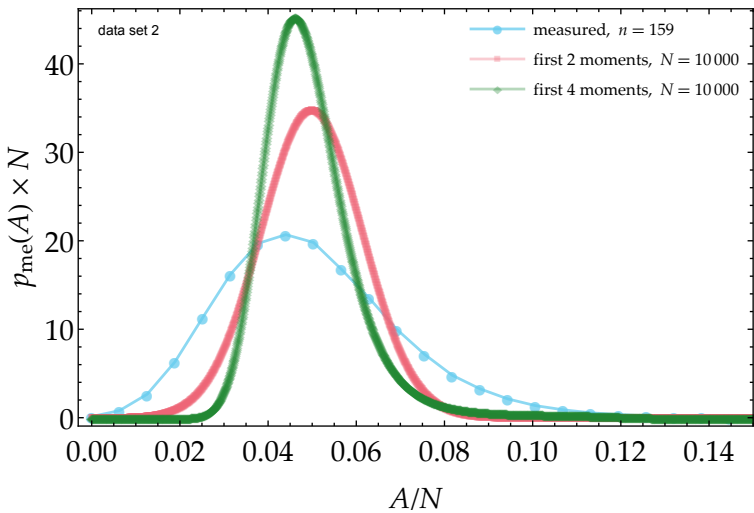
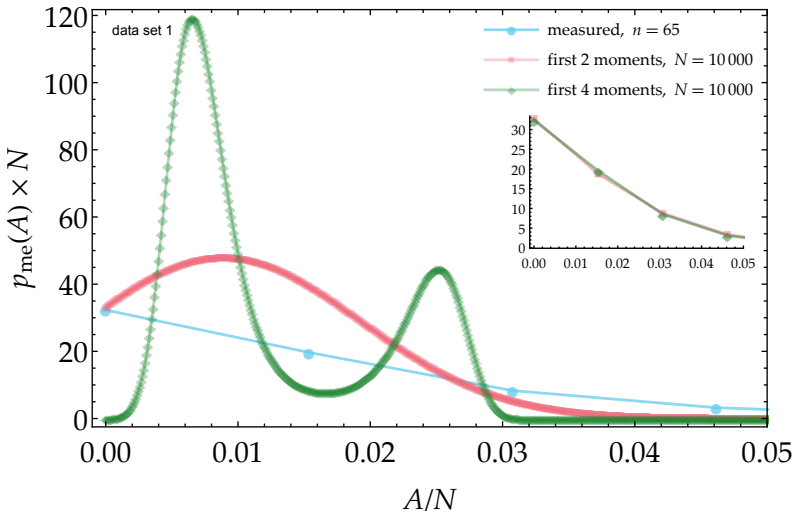


Figure 2 ***

In appendix*** we give a summary of how this distribution is derived. Note that this is a probability distribution over frequency distributions, so we must be careful with terminology to avoid confusion.

Now consider the related question: What is the activity A_t of the full network at some specific time bin t ? If we don't know the exact sequence of activities during the recording then we don't know A_t ; if we know the frequency distribution F of the activity levels, then our degree of belief about A_t is F_{A_t} , out of combinatorial reasons. But if we are also uncertain about F , with belief distribution (4), then our degree of belief about A_t by marginalization is

$$p(A_t | DI) = \int F_{A_t} p(F | DI) dF; \quad (5)$$

our belief distribution for A_t is thus simply the mean $\int F p(F | DI) dF$.

Formulae (4) and (5) show that the most probable frequency distribution F (if it's unique) and the probability distribution for A_t are generally different: the former is the mode of $p(F | DI)$, the latter its mean.

If the probability density $p(F | DI)$ is extremely peaked at some frequency distribution F^* , it can be approximately treated as a delta density,

$$p(F | DI) dF \approx \delta(F - F^*) dF, \quad (6)$$

and consequently the probability for the activity at bin t , by eq. (5), is simply the mode F^* itself:

$$p(A_t | DI) \approx F_{A_t}^*. \quad (7)$$

The maximum-entropy distribution is exactly this mode F^* , for *particular* kinds of probability densities $p(F | DI)$ which we'll discuss shortly. Formulae (6) and (7) show why the maximum-entropy distribution is an approximate answer to two different questions.

If the probability density $p(F | DI)$ is not sufficiently peaked, however, its mode and mean can be quite different. The maximum-entropy approximation can then be inadequate for two reasons. First, despite being the mode it can be a poor representative of all the possible frequency distributions around the mode. Second, it can be a bad approximation of the probability distribution (5).

Let's therefore examine the full density $p(F | DI)$ given by the probability calculus. Its derivation is given in appendix***; here we

discuss its mathematical features and examine the answers it provides. Its expression,

$$p(F | DI) \propto \exp \left[-\frac{|\mathbf{CF} - \hat{\mathbf{c}}|^2}{2\sigma^2} - \frac{|\Delta \mathbf{F}|^2}{2\delta^2} - L H(\mathbf{F}; \mathbf{R}) \right], \quad (8)$$

consists of three terms:

- The density

$$\exp[-L H(\mathbf{F}; \mathbf{R})] \quad (9)$$

expresses our pre-data beliefs about the possible frequency distributions. Here $H(\mathbf{F}; \mathbf{R}) := \sum_A F_A \ln(F_A/R_A)$ is the relative entropy or discrimination information¹², \mathbf{R} is a reference distribution, and L a positive parameter. This density expresses a belief proportional to the number of ways in which the frequency distribution \mathbf{F} can be realized if we assume that the activities appear with frequencies \mathbf{R} in the long run, as can be seen using Stirling's approximation. For \mathbf{R} we choose the uniform distribution, but any biologically sensible distribution, such as one decreasing exponentially with the activity, leads to the same final results. The parameter L roughly represents the number of data points necessary to change our initial belief, and we set it equal to 10. This density is similar to an 'entropic prior'¹³.

- The previous density is corrected by the normal density $\exp\left(-\frac{|\Delta \mathbf{F}|^2}{2\delta^2}\right)$, where the matrix Δ is the discrete equivalent of a fourth-order derivative operator. This density expresses our belief, hinging on biological arguments, that the frequency distribution should be somewhat smooth, without huge discontinuities between one activity level and the next. The fourth derivative is chosen because it still allows for the presence of modulated maxima and minima more than derivatives of lower order do. The standard deviation δ is such that the smoothness of the maximum-entropy distribution, taken as a touchstone, is within two standard deviations.

- The normal density $\exp\left(-\frac{|\mathbf{CF} - \hat{\mathbf{c}}|^2}{2\sigma^2}\right)$ is the likelihood of the frequency \mathbf{F} in view of our data. The matrix \mathbf{C} expresses the linear constraints on \mathbf{F} , and the tuple $\hat{\mathbf{c}}$ the measured values of such constraints. For example, if we consider the full frequency distribution \mathbf{f} of the sample, then \mathbf{C} effects

¹²Kullback 1987; Jaynes 1963; Hobson 1969; Hobson et al. 1973.

¹³Rodríguez 1991; Skilling 1998; Caticha et al. 2004; Porta Mana 2017a; see esp. Neumann 2007 § 3.

the marginalization to a sample of size n through a hypergeometric distribution (sampling without replacement):

$$C_{a,A} := \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1} \quad (10)$$

and $\hat{c} := f$ are the measured frequencies. If we only consider some factorial moments, instead, then

$$C_{m,A} := \binom{NA}{m} \binom{N}{m}^{-1} \quad (11)$$

as for eq. (2), and \hat{c} is the tuple of measured factorial moments. The standard deviations σ reflect our beliefs about the measurement results, which can come for example from the spike-sorting procedure. Their values are discussed in appendix***.

The post-data belief density (8) has a maximum-entropy distribution as a mode approximation when σ is small, L is large, and σL is small: the mode will be the constrained F that minimizes the relative entropy $H(F; R)$. It should be noted that a pre-data belief different from the entropic one (9) will lead to a different mode approximation. For example, a Dirichlet density (which is equivalent to (9) with F and R exchanged) would lead to a maximum-entropy distribution with *Burg's* entropy¹⁴ instead of Shannon's¹⁵.

The properties of the belief density (8) in the case of our first data set are shown in fig. 5 for $N = 1\,000$ and in fig. 6 for $N = 5\,000$. Only the first six factorial moments were considered as our data, but it turns out that they are informationally sufficient, leading to practically the same result as specifying the full frequency distribution for the sample.

The top plots in the figures show the mean $\int F p(F | DI) dF$ (—), which is also our belief about the activity at any single time bin, eq. (5), is clearly different from the maximum-entropy mode. The latter is therefore not a good approximation for the full density (8), which is quite broad, as can be seen from the ten samples (—) drawn from it. This broadness is not surprising: we're making guesses about thousands of neurons

¹⁴Burg 1975.

¹⁵Jaynes 1996; Porta Mana 2009.

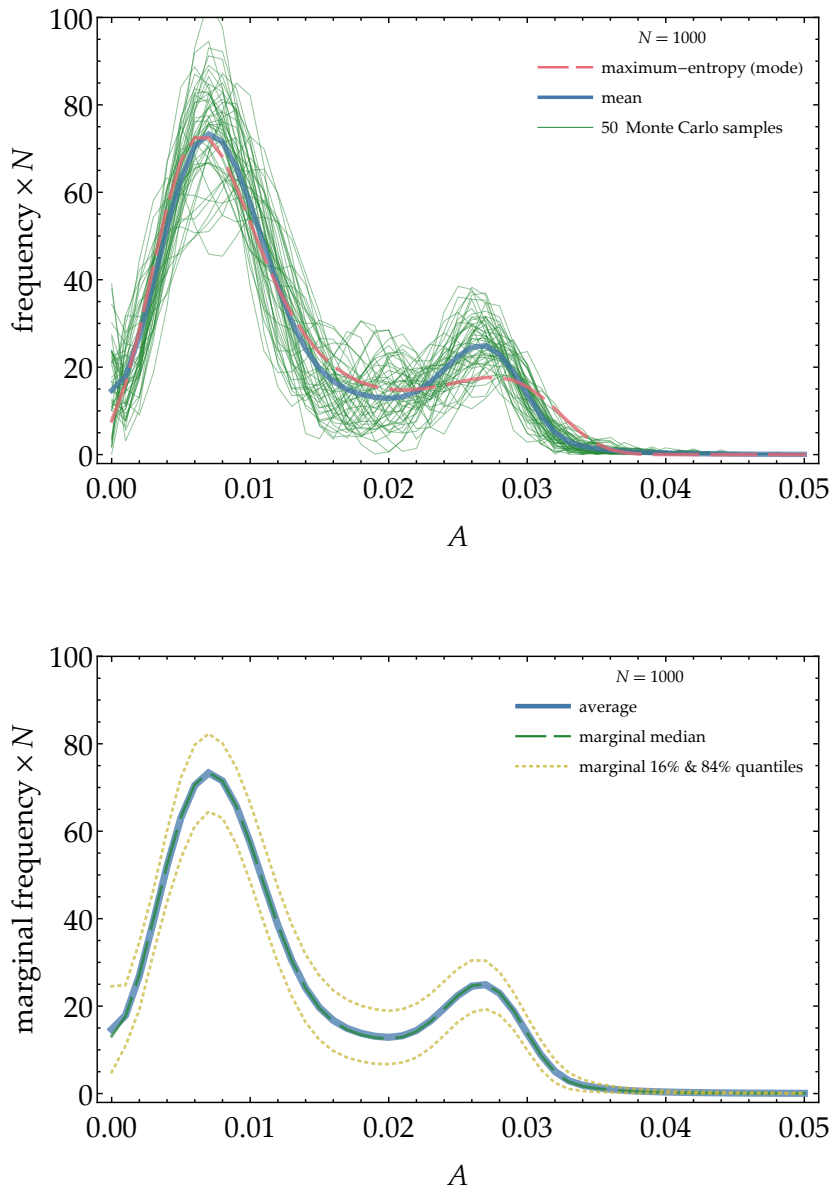


Figure 3 Entropic prior

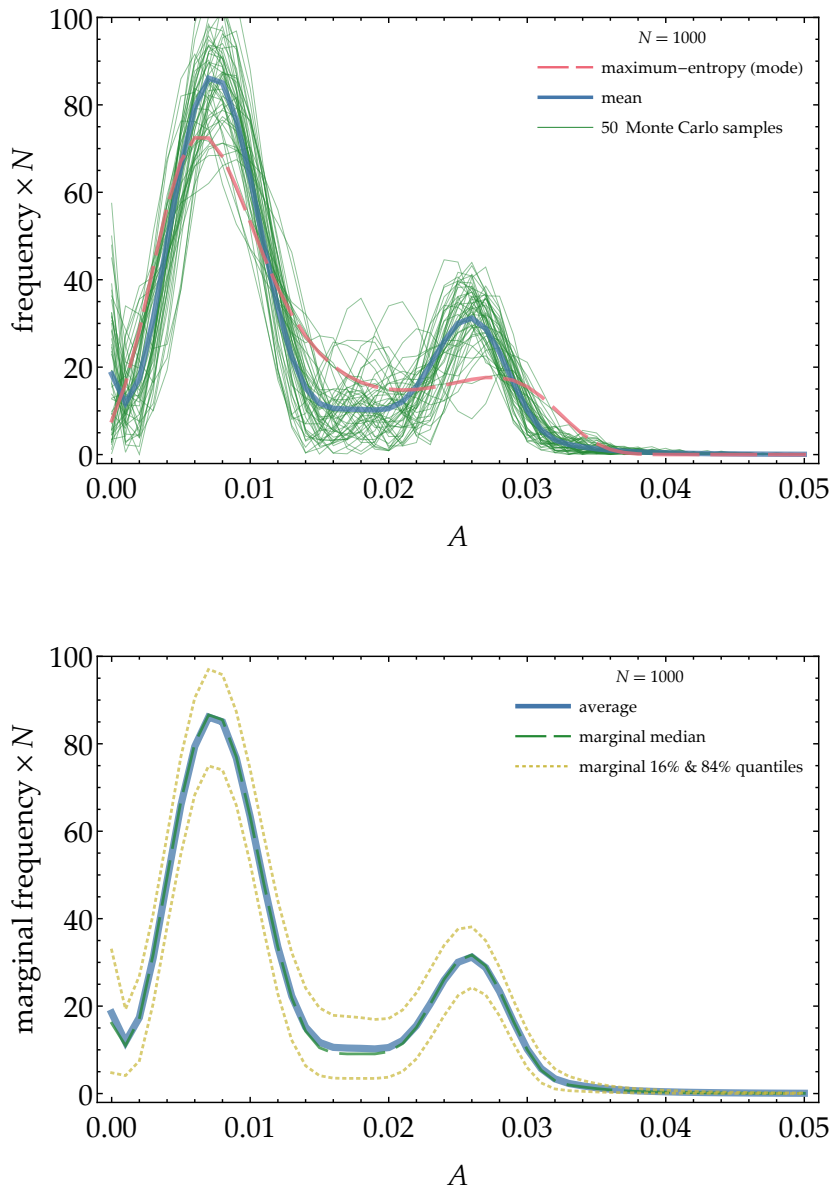


Figure 4 Prior uniform in dF

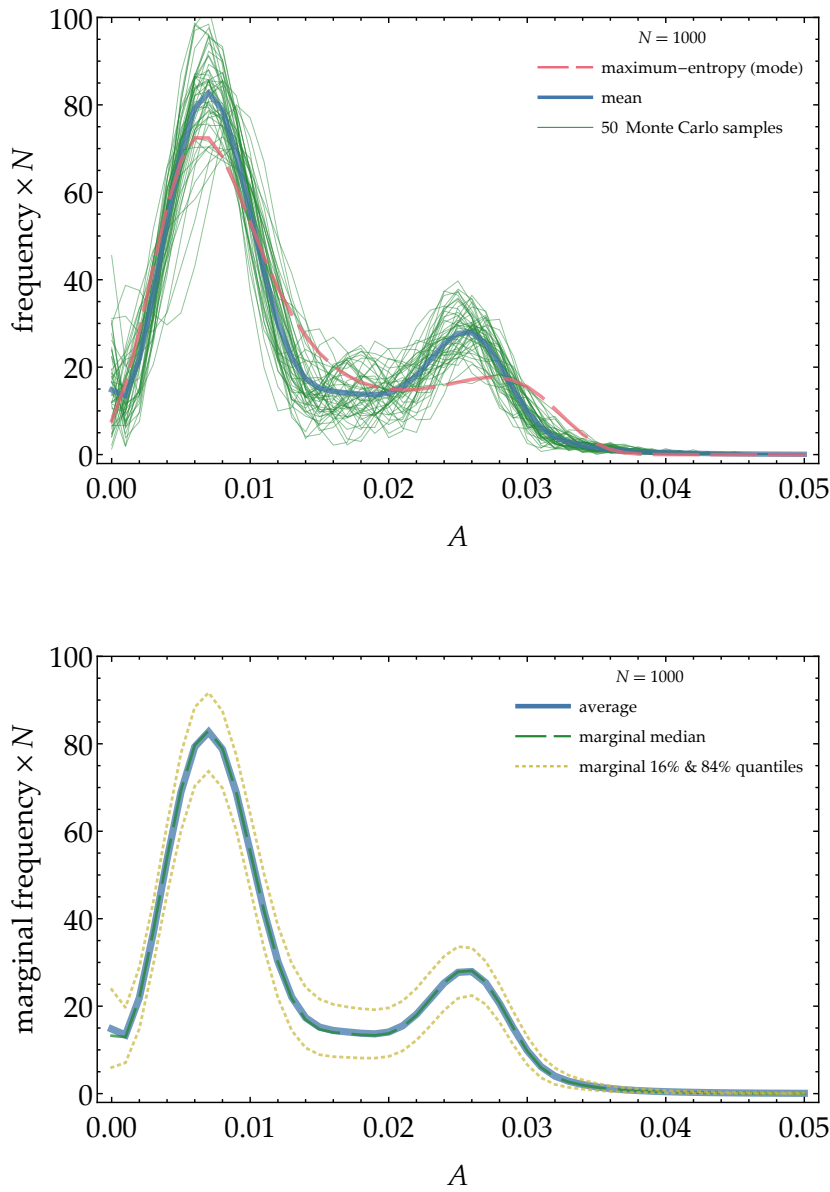


Figure 5 Dirichlet prior (entropic with reversed relative entropy)

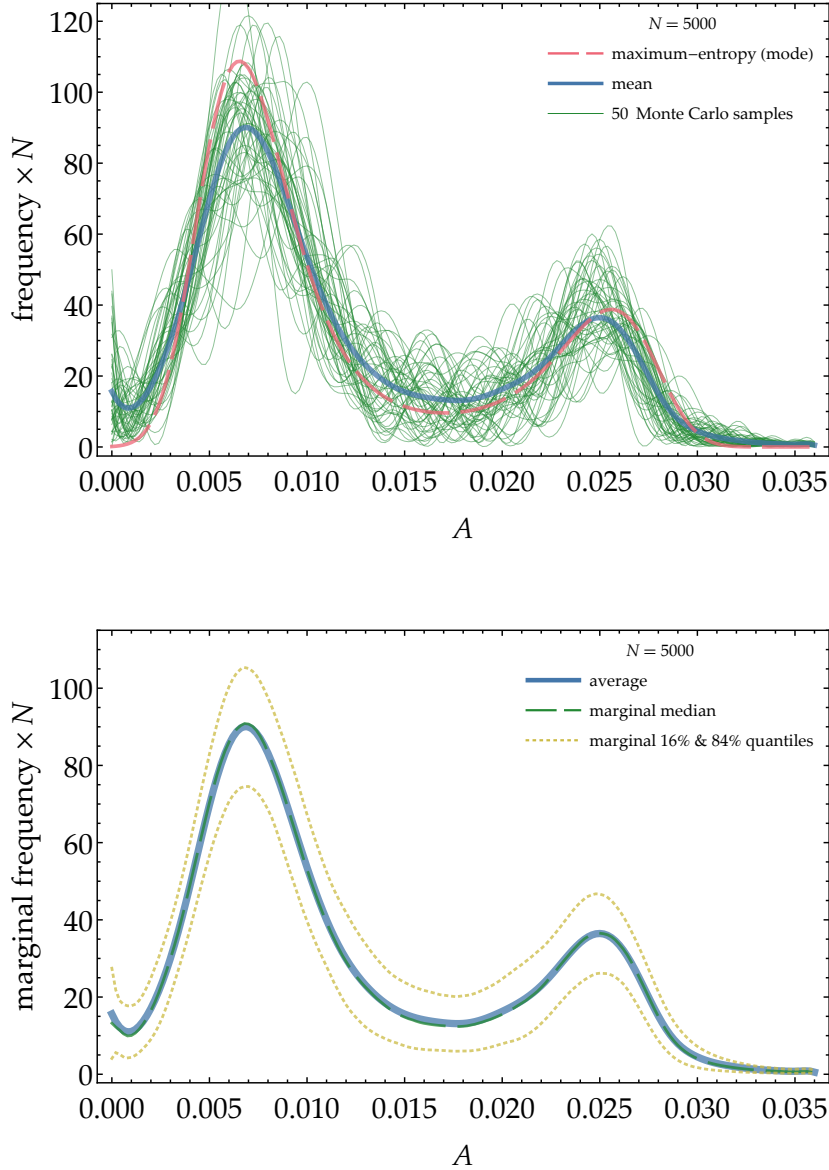


Figure 6 Entropic prior

from the observation of fewer than 100. Yet, the mean and the maximum-entropy mode are qualitatively similar. In fact, for $N = 1\,000$ the second peak of the mean is even higher than the corresponding maximum-entropy one. The samples also show that generally two high-frequency regions of activity are expected, especially for $N = 5\,000$.

The bottom plots in the figures show our belief distributions for the marginal frequency of each activity level, $p(F_A | DI)$, with the means (—), medians (— — —), and 16% and 84% quantiles (· · · corresponding to one standard deviation for a normal distribution). The coincidence of means and medians indicate that these marginals are very close to normal distributions. This happens thanks to the marginalization, which integrates out $N - 2$ quantities. Note, however, that the full distribution for F is far from normal, as indicated by the pronounced differences between mode and mean.

These results caution us against taking maximum-entropy solutions too literally. The full probabilistic analysis allows us to see more clearly the extent of our uncertainty and to make better-informed guesses.

5 Summary and discussion

 OLD TEXT

6 Mathematical development

$$\begin{aligned} E\left[\frac{s(s-1)}{n(n-1)} \mid I\right] &= E\left[\frac{S(S-1)}{N(N-1)} \mid I\right], \\ E\left[\frac{s(s-1)(s-2)}{n(n-1)(n-2)} \mid I\right] &= E\left[\frac{S(S-1)(S-2)}{N(N-1)(N-2)} \mid I\right], \end{aligned} \quad (12a)$$

and in general $E\left[\binom{s}{r} \binom{n}{r}^{-1} \mid I\right] = E\left[\binom{S}{r} \binom{N}{r}^{-1} \mid I\right], \quad r < N.$

Note that from the first r factorial moments we can calculate the first r power moments and vice versa; but the simple equalities above don't hold for the power moments¹⁶.

$$p(s \mid I) = \sum_S \binom{n}{s} \binom{N-n}{S-s} \binom{N}{S}^{-1} p(S \mid I). \quad (12b)$$

Suppose we have recorded the firing activity of a hundred neurons, sampled from a particular brain area. What are we to do with such data? Gerstein et al. (1985) posed this question very tersely (our emphasis):

The principal conceptual problems are (1) *defining cooperativity or functional grouping* among neurons and (2) *formulating quantitative criteria* for recognizing and characterizing such cooperativity.

These questions have a long history, of course; see for instance the 1966 review by Moore et al. (1966). The neuroscientific literature has offered several mathematical definitions of 'cooperativity' or 'functional grouping' and criteria to quantify it.

One such quantitative criterion relies on the maximum-entropy or relative-maximum-entropy method¹⁷. This criterion has been used in neuroscience at least since the 1990s, applied to data recorded from brain areas as diverse as retina and motor cortex¹⁸, and it has been subjected to mathematical and conceptual scrutiny¹⁹.

¹⁶Porta Mana et al. 2015 § A.

¹⁷Jaynes 1957a; 1963; Hobson et al. 1973; Sivia 2006; Mead et al. 1984.

¹⁸MacKay 1991; Martignon et al. 1995; Bohte et al. 2000; Amari et al. 2003; Schneidman et al. 2006; Shlens et al. 2006; Macke et al. 2009; Roudi et al. 2009c; Tkačik et al. 2009; Gerwinn et al. 2009; Macke et al. 2011b,a; Ganmor et al. 2011; Granot-Atedgi et al. 2013; Tkačik et al. 2014; Mora et al. 2015; Shimazaki et al. 2015.

¹⁹Tkačik et al. 2006; Roudi et al. 2009a,b; Barreiro et al. 2010a,b; Macke et al. 2013; Rostami et al. 2017.

‘Cooperativity’ can be quantified and characterized with maximum-entropy methods in several ways. The simplest way roughly proceeds along the following steps. Consider the recorded activity of a sample of n neurons.

1. The activity of each neuron, a continuous signal, is divided into T time bins and binarized in intensity, and thus transformed into a sequence of digits ‘0’s (inactive) and ‘1’s (active)²⁰.

Let the variable $a_i(t) \in \{0, 1\}$ denote the activity of the i th sampled neuron at time bin t . Collectively denote the n activities with $\mathbf{a}(t) := (a_1(t), \dots, a_n(t))$. The network-averaged activity at that bin is $a(t) := \sum_i a_i(t)/n$. If we count the number of distinct pairs of active neurons at that bin we combinatorially find $\binom{na(t)}{2} \equiv na(t)[na(t) - 1]/2$. There can be at most $\binom{n}{2}$ simultaneously active pairs, so the network-averaged pair activity is $\overline{aa}(t) := \binom{n}{2}^{-1} \binom{na(t)}{2}$. With some combinatorics we see that the network-averaged activity of m -tuples of neurons is

$$\underbrace{\overline{a \cdots a}}_{m \text{ terms}}(t) = \binom{n}{m}^{-1} \binom{na(t)}{m}. \quad (13)$$

For brevity let us agree to simply call ‘activity’ the average a , ‘pair-activity’ the average \overline{aa} , and so on.

2. Construct a sequence of relative-maximum-entropy distributions for the activity a , using this sequence of constraints:
 - the time average of the activity: $\widehat{a} := \sum_t a(t)/T$;
 - the time averages of the activity and of the pair-activity $\widehat{aa} := \sum_t \overline{aa}(t)/T$;
 - ...
 - the time averages of the activity, of the pair-activity, and so on, up to the k -activity.

Call the resulting distributions $p_1(a), p_2(a), \dots, p_k(a)$. The time-bin dependence is now absent because these distributions can be interpreted as referring to any one of the time bins t , or to a new time bin (in the future or in the past) containing new data.

We also have the empirical frequency distribution of the total activity, $f(a)$, counted from the time bins.

²⁰cf. Caianiello 1961; 1986.

3. Now compare the distributions above with one another and with the frequency distribution, using some probability-space distance like the relative entropy or discrimination information²¹. If we find, say, that such distance is very high between p_1 and f , very low between p_2 and f , and is more or less the same between all p_m and f for $m \geq 2$, then we can say that there is a ‘pairwise cooperativity’, and that any higher-order cooperativity is just a reflection or consequence of the pairwise one. The reason is that the information from higher-order simultaneous activities did not lead to appreciable changes in the distribution obtained from pair activities.

The protocol above needs to be made precise by specifying various parameters, such as the width of the time bins or the probability distance used.

We hurry to say that the description just given is just *one* way to quantify and characterize cooperativity and functional grouping, not *the only* way. It can surely be criticized from many points of view. Yet, it is quantitative and bears a more precise meaning than an undefined, vague notion of ‘cooperativity’. Two persons who apply this procedure to the same data will obtain the same numbers. Different protocols can be based on the maximum-entropy method, for instance protocols that take into account the activities or pair activities of specific neurons rather than network averages, or even protocols that take into account time dependence.

The purpose of the present work is not to assess the merits of maximum-entropy methods with respect to other methods. Its main purpose is to show that there is a problem in the way the maximum-entropy method itself, as sketched above, is applied to the activity of the recorded neurons. We believe that this problem is at the root of some quirks about this method that were pointed out in the literature²². This problems extends also to more complex versions of the method, possibly except versions that use ‘hidden’ neurons²³. The problem is that the recorded neurons are a *sample* from a larger, unrecorded network, but the maximum-entropy method as applied above is treating them as isolated from the rest of the brain. Hence, the results it provides cannot

²¹Kullback 1987; Jaynes 1963; Hobson 1969; Hobson et al. 1973.

²²Roudi et al. 2009b.

²³Smolensky 1986; Kulkarni et al. 2007; Huang 2015; Dunn et al. 2017.

be rightfully extrapolated. We will give a mathematical proof of this. Let us first analyse this issue in more detail.

Suppose that the neurons were recorded with electrodes covering an area of some square millimetres²⁴. This recording is a sample of the activity of the neuronal network under the recording device, which can amount to tens of thousands of neurons²⁵. We could even consider the recorded neurons as a sample of a brain area more extended than the recording device.

The characterization of the cooperativity of the recorded sample would have little meaning if we did not expect its results to generalize to a larger, unrecorded network – at the very least the network under the recording device. In other words, we expect that the conclusions drawn with the maximum-entropy methods about the sampled neurons should somehow extrapolate to unrecorded neurons in some larger area, from which the recorded neurons were sampled. In statistical terms we are assuming that the recorded neurons are a *representative sample* of some larger neuronal network. Probability theory tells us how to make inferences from a sample to the larger network from which it is sampled (see references below).

We can apply the maximum-entropy method to the sample, as described in the above protocol, to generate probability distributions for the activity of the sample. But, given that our sample is representative of a larger network, we can also apply the maximum-entropy method to the larger (unrecorded) network. The constraints are the same: the time averages of the sampled data, since they constitute representative data about the larger network as well. The method thus yields a probability distribution for the larger network, and the distribution for the sample is obtained by marginalization. The problem is that *the distributions obtained from these two applications differ*. Which choice is most meaningful?

In this work we develop the second way of applying the maximum-entropy method, at the level of the larger network, and show that its results differ from the application at the sample level. We also consider the case where the size of the larger network is unknown.

To apply the maximum-entropy method to the larger, unsampled network, it is necessary to use probability relations relevant to sampling²⁶.

²⁴cf. Berényi et al. 2014.

²⁵Abeles 1991.

²⁶Ghosh et al. 1997; Freedman et al. 2007 parts I, VI; Gelman et al. 2014 ch. 8; Jaynes 2003 ch. 3.

The relations we present are well-known in survey sampling and in the pedagogic problem of drawing from an urn without replacement, yet they are somewhat hard to find explicitly written in the neuroscientific literature. We present and discuss them in the next section. A minor purpose of this paper is to make these relations more widely known, because they can be useful independently of maximum-entropy methods.

The notation and terminology in the present work follow ISO and ANSI standards²⁷ but for the use of the comma ‘,’ to denote logical conjunction. Probability notation follows Jaynes²⁸. By ‘probability’ we mean a degree of belief which ‘would be agreed by all rational men if there were any rational men’²⁹.

7 Probability relations between network and sample

We have already introduced the notation for the sample neurons. We introduce an analogous notation for the N neurons constituting the larger network, but using the corresponding capital letters: $A_i(t)$ is the activity of the i th neuron at time bin t , $A(t) := \sum_i A_i(t)/N$ is the activity at that bin averaged over the larger network, and so on.

The probability relations between sample and larger network are valid at every time bin. As we mentioned above, the maximum-entropy distribution refers to any time bin or to a new bin. For these reasons we will now omit the time-bin argument ‘(t)’ from our expressions.

If K denotes our state of knowledge – the evidence and assumptions backing our probability assignments – our uncertainty about the full activity of the larger network is expressed by the joint probability distribution

$$p(A_1, A_2, \dots, A_N | K) \quad \text{or} \quad p(A | K), \quad A \in \{0, 1\}^N. \quad (14)$$

Our uncertainty about the state of the sample is likewise expressed by

$$p(a_1, a_2, \dots, a_n | K) \quad \text{or} \quad p(a | K), \quad a \in \{0, 1\}^n. \quad (15)$$

The theory of statistical sampling is covered in many excellent texts, for example Ghosh & Meeden (1997) or Freedman, Pisani, & Purves

²⁷ISO 1993; IEEE 1993; NIST 1995; ISO 2006a,b.

²⁸Jaynes 2003.

²⁹Good 1966.

(2007 parts I, VI); summaries can be found in Gelman et al. (2014 ch. 8) and Jaynes (2003 ch. 3).

We need to make an initial probability assignment for the state of the full network before any experimental observations are made. This initial assignment will be modified by our experimental observations, and these can involve just a sample of the network. Our state of knowledge and initial probability assignment should reflect that samples are somehow representative of the whole network.

In this state of knowledge, denoted I , we know that the neurons in the network are biologically or functionally similar, for example in morphology or the kind of input or output they receive or give. But we are completely ignorant about the physical details of the individual neurons. Our ignorance is therefore symmetric under permutations of neuron identities. This ignorance is represented by a probability distribution that is symmetric under permutations of neuron identities; such a distribution is usually called *finitely exchangeable*³⁰. We stress that this probability assignment is just an expression of the symmetry of our *ignorance* about the state of the network, not an expression of some biological or physical symmetry or identity of the neurons.

The *representation theorem for finite exchangeability* states that, in the state of knowledge I , the symmetric distribution for the full activity is completely determined by the distribution for its network-average:

$$p(A | I) \equiv \sum_A p(A | A, I) p(A | I) = \left(\binom{N}{NA} \right)^{-1} p(A | I). \quad (16)$$

The equivalence on the left is just an application of the law of total probability; the equality on the right is the statement of the theorem. This result is intuitive: owing to symmetry, we must assign equal probabilities to all $\binom{N}{NA}$ activity vectors with NA active neurons; the probability of each activity vector is therefore given by that of the average activity divided by the number of possible vector values. Proof of this theorem and generalizations to non-binary and continuum cases are given by de Finetti (1959), Kendall (1967), Ericson (1976), Diaconis & Freedman (1977; 1980), Heath & Sudderth (1976).

³⁰Ericson 1969; Ghosh et al. 1997 ch. 1.

Our uncertainties about the full network and the sample are connected via the conditional probability

$$p(a | A, I) = \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1} =: G_{aA}, \quad (17)$$

which is a hypergeometric distribution, typical of ‘drawing without replacement’ problems. The combinatorial proof of this expression is in fact the same as for this class of problems³¹.

Using the conditional probability above we obtain the probability for the activity of the sample:

$$p(a | I) = \sum_A p(a | A, I) p(A | I) = \sum_A G_{aA} p(A | I). \quad (18)$$

It should be proved that the probability distribution for the full activity of the sample is also symmetric and completely determined by the distribution of its network-averaged activity:

$$p(a | I) = \binom{n}{na}^{-1} p(a | I). \quad (19)$$

This is intuitively clear: our initial symmetric ignorance should also apply to the sample. The distribution for the sample (18) indeed satisfies the same representation theorem (16) as the distribution for the full network.

The conditional probability $p(a | A, I) \equiv G_{aA}$, besides relating the distributions for the network and sample activities via marginalization, also allows us to express the expectation value of any function of the sample activity, c_a , in terms of the distribution for the full network, as follows:

$$E(c | I) \equiv \sum_a c_a p(a | I) = \sum_a c_a \sum_A G_{aA} p(A | I) = \sum_A (\sum_a c_a G_{aA}) p(A | I), \quad (20)$$

where the second step uses eq. (18). The last expression shows that the expectation of the function c_a is equal to the expectation of the function $c^*(A) := \sum_a c_a G_{aA}$.

The final expression in eq. (20) is important for our maximum-entropy application: the requirement that the function c , defined for the sample,

³¹Jaynes 2003 ch. 3; Ross 2010 § 4.8.3; Feller 1968 § II.6.

have a value \hat{c} obtained from observed data, *translates into a linear constraint for the distribution of the full network*:

$$\hat{c} = E(c \mid I) \equiv \sum_A \left(\sum_a c_a G_{aA} \right) p(A \mid I). \quad (21)$$

In particular, when the function c is the m -activity of the sample, $c_a = \overline{a \dots a} \equiv \binom{na}{m} / \binom{n}{m}$, we find

$$\begin{aligned} E(\underbrace{\overline{a \dots a}}_{m \text{ factors}} \mid I) &\equiv \sum_a \binom{n}{m}^{-1} \binom{na}{m} p(a \mid I) = \\ &\quad \binom{N}{m}^{-1} \sum_A \binom{NA}{m} p(A \mid I) \equiv E(\underbrace{\overline{A \dots A}}_{m \text{ factors}} \mid I), \end{aligned} \quad (22)$$

that is, *the expected values of the m -activities of the sample and of the full network are equal*. The proof of the middle equality uses the expression for the m th factorial moment of the hypergeometric distribution and can be found in Potts (1953). Similar relations can be found for the raw moments $E(a^m)$ and $E(A^m)$, which can be written in terms of the product expectations using eq. (13).

Thus, in a maximum-entropy application, when we require the expectation of the m -activity of a sample to have a particular value, we are also requiring the expectation of the m -activity of the full network to have the same value.

These expectation equalities between sample and full network should not be surprising: we intuitively *expect* that the proportion of coloured balls sampled from an urn should be roughly equal to the proportion of coloured ball contained in the urn. The formulae in the present section formalize and mathematically express our intuition. The hypergeometric distribution G_{aA} plays an important role in this formalization. A look at its plot, fig. 7, reveals that it is a sort of ‘fuzzy identity transformation’, or fuzzy Kronecker delta, between the A -space $\{0, \dots, N\}$ and a -space $\{0, \dots, n\}$. From eq. (19) we thus have that

$$p(a = a \mid I) \approx p(A = a \mid I), \quad E[c_a \mid I] \approx E[c_A \mid I], \quad (23)$$

where c is any smooth function defined on $[0, 1]$. These approximate equalities express the intuitive fact that *our uncertainty about the sample is*

representative of our uncertainty about the network and about other samples, and vice versa. When $n = N$, G_{aA} becomes the identity matrix and the approximate equalities above become exact – of course, since we have sampled the full network.

But the approximate equalities above may miss important features of the two probability distributions. In the next section we will in fact emphasize their differences. If the distribution for the network average A is bimodal, for example, the bimodality can be lost in the distribution for the sample average a , owing to the coarsening effect of G_{aA} .

8 Maximum-entropy: sample level vs full-network level

In the previous section we have seen that observations about a sample can be used as constraints on the distribution for the activity of the full network. Let us use such constraints with the maximum-entropy method. Suppose that we want to constrain m functions of the sample activity, vectorially written $c := (c_1, \dots, c_m)$, to m values $\hat{c} := (\hat{c}_1, \dots, \hat{c}_m)$. These functions are typically k -activities $\bar{a} \dots \bar{a}$, and the values are

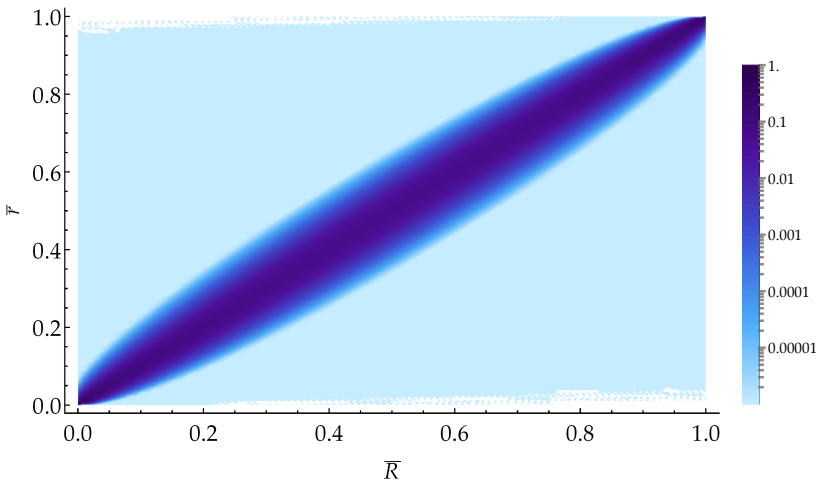


Figure 7 Log-density plot of the hypergeometric distribution $G_{aA} := \binom{n}{na} \binom{N-n}{NA-na} \binom{N}{NA}^{-1}$ for $N = 5000$, $n = 200$. (Band artifacts may appear in the colourbar depending on your PDF viewer.)

typically the time averages of the observed sample, as discussed in § 1: $\hat{c} = \sum_t c[a(t)]/T$.

Let us apply the relative-maximum-entropy method³² directly to sampled neurons; denote this approach by I_s . Then we apply the method to the full network of neurons, most of which are unsampled; denote this approach by I_p .

Applied directly to the sampled neurons, the method yields the distribution

$$p(a | I_s) = \frac{1}{z(I)} \binom{n}{na} \exp[I^\top c_a] \quad (24)$$

where $z(I)$ is a normalization constant. The binomial in front of the exponential appears because we must account for the multiplicity by which the network-average activity a can be realized: $a = 0$ can be realized in only one way (all neurons inactive), $a = 1/n$ can be realized in n ways (one active neuron out of n), and so on. This term is analogous to the ‘density of states’ in front of the Boltzmann factor in statistical mechanics³³. The m Lagrange multipliers $I := (I_1, \dots, I_m)$ must satisfy the m constraint equations

$$\hat{c} = E(c | I_s) \equiv \frac{1}{z(I)} \sum_a c_a \binom{n}{na} \exp[I^\top c_a]. \quad (25)$$

Applied to the full network, using the constraint expression (21) derived in the previous section, the method yields the distribution for the full-network activity

$$p(A | I_p) = \frac{1}{\zeta(\lambda)} \binom{N}{NA} \exp(\lambda^\top \sum_a c_a G_{aA}). \quad (26)$$

The m Lagrange multipliers $\lambda := (\lambda_1, \dots, \lambda_m)$ must satisfy the m constraint equations

$$\hat{c} = E(c | I_p) \equiv \frac{1}{\zeta(\lambda)} \sum_a \sum_A c_a G_{aA} \binom{N}{NA} \exp(\lambda^\top \sum_a c_a G_{aA}). \quad (27)$$

We obtain the distribution for the sample activity by marginalization, using eq. (19):

$$p(a | I_p) = \frac{1}{\zeta(\lambda)} \sum_A G_{aA} \binom{N}{NA} \exp(\lambda^\top \sum_a c_a G_{aA}). \quad (28)$$

³²Sivia 2006; Mead et al. 1984.

³³Callen 1985 ch. 16.

The distributions for the sample activity, eqs (28) and (24), obtained with the two approaches I_s and I_p , are different. From the discussion in the previous section we expect them to be vaguely similar; yet they cannot be exactly equal, because their equality would require the $2m$ quantities λ and I to satisfy the constraint equations (27) and (25), and in addition also the n equations $p(a | I_p) = p(a | I_s)$, $a = 1/n, \dots, 1$ (one equation is taken care of by the normalization of the distributions). We would have a set of $2m + n$ equations in $2m$ unknowns.

Hence, *the applications of maximum-entropy at the sample level and at the full-network level are inequivalent*. They lead to numerically different distributions for the sample activity a .

The distribution obtained at the sample level will show different features from the one obtained at the network level, like displaced or additional modes or particular tail behaviour. We show an example of this discrepancy in fig. ??, for $N = 10\,000$, $n = 200$, and the two constraints

$$E(a) = 0.0478, \quad E(\bar{a}\bar{a}) = 0.00257, \quad (29)$$

which come from the actual recording of circa 200 neurons from macaque motor cortex³⁴. The distribution obtained at the network level (blue triangles) has a higher and displaced mode and a quite different behaviour for activities around 0.5 than the distribution obtained at the sample level (red squares).

In our discussion we have so far assumed the size N of the larger network to be known. This is rarely the case, however. We usually are uncertain about N and can only guess its order of magnitude. In such a state of knowledge I_u our ignorance about the possible value of N is expressed by a probability distribution $p(N = N | I_u) = h(N)$, and the marginal distribution for the sample activity (28) is modified, by the law of total probability, to

$$p(a | I_u) = \sum_N p(a | N, I_u) p(N | I_u) = \sum_N \left\{ \frac{1}{\zeta(\lambda_N)} \sum_A G_{aA}^{(N)} \binom{N}{NA} \exp[\lambda_N^\top \sum_a c_a G_{aA}^{(N)}] \right\} h(N), \quad (30)$$

where the Lagrange multipliers λ_N and the summation range for A depend on N .

³⁴Rostami et al. 2017.


As a proof of concept, fig. ?? also shows such a distribution (yellow circles) for the same constraints as above, and a probability distribution for N inspired by Jeffreys³⁵:


$$h(N) \propto 1/N, \quad N \in \{1\,000, 2\,000, \dots, 10\,000\}. \quad (31)$$

9 Derivation from the probability calculus

There are three inequivalent main routes that lead to a probability distribution of the maximum-entropy form (24) or (26). The distribution carries a different interpretation under each route³⁶.

(a) One route is the choice of the distribution having the highest Shannon entropy, given only a quantitative assessment some of its properties, such as expectations. The numerical choice of the value of such properties is a (subjective) assumption.

In the two other routes the maximum-entropy distribution is obtained as an *approximation* of a distribution obtained via the probability calculus, using data coming from a set of T measurements – as in our present case.  [refs here](#) Also in this case some (subjective) assumptions are necessary: they concern our beliefs about the long-run relative frequencies of the measurement outcomes:³⁷

(b) In one case we consider all possible sets of measurement outcomes to be *roughly* equally likely; this leads to a probability for the frequencies \mathbf{v} proportional to a multinomial coefficient $\binom{L}{L\mathbf{v}}$, with L large but smaller than T . (We cannot assume the sets of measurement outcomes to be exactly equally likely, because this is equivalent to their independence: cf. Jaynes 2003 § 6.7; Porta Mana 2009 § B; 2017a § 2.) The exact expression is  [equation here](#)

³⁵Jeffreys 1983 § 4.8.

³⁶Jaynes 1979 pp. 52–55, 72–77; 1982a pp. 25–28; 1982b § I; 1996; 2003 § 11.1.

³⁷Such assumptions are always necessary at the beginning of an inference: ‘Now the axioms of probability enable us to infer any probability-conclusion *only* from probability-premisses. In other words, the calculus of probability does not enable us to infer any probability-value unless we have some probabilities or probability relations *given*. Such data cannot be supplied by the mathematician. E.g. the rules of arithmetic and the axioms of the probability-calculus are utterly impotent to determine, on the supposed knowledge that the throw of a coin must yield either head or tail and cannot yield both, the probability that it will yield head or that it will yield tail. We must assume that the two co-exclusive and co-exhaustive possibilities are *equally probable*, before we can estimate the probability of either as being a half of certitude’ (Johnson 1924 *Appendix on eduction*, § 5, p. 182).

(c) In the other case we assume that the measurements have a *sufficient statistics*: the same as appears in the exponential of the maximum-entropy distribution. The exact expression is [equation here](#)

It's important to keep in mind that the approximate equivalence of these three routes only holds under very specific assumptions – which *have physical and biological meanings and consequences*. In particular, route (c) implies that we can discard other empirical statistics of the data, if they are known; whereas route (b) requires us to specify all known empirical statistics, because using only a subset of them may lead to different results. Route (a) is also supposed to be used with all known data. Moreover, the approximate equivalence of route (a) with routes (b) and (c) *only holds if T is much larger than the possible values of the activity A* . Finally, we also obtain very different expressions depending on whether we're asking about *the activity in one of the **recorded** time bins* or about *the activity in a **new** time bin*. Works that use maximum-entropy distributions are often very vague about the latter point.

Here are the distributions for our degrees of belief about three different quantities, assuming an entropic pre-data distribution for the long-run frequencies of the full network:

The frequency distribution F of the full-network activity during the recording

$$\begin{aligned}
 P(F | f, I) &\propto \int d\mathbf{v} \sum_{\phi} \delta(\sum_A \phi_{aA} = f_a) \delta(\sum_a \phi_{aA} = F_A) \times \\
 &\quad \binom{T}{T\phi} \left[\prod_{aA} (G_{aA} v_A)^{T\phi_{aA}} \right] \exp[-L H(\mathbf{v}; \mathbf{R})] \\
 &\propto \delta(\sum_A G_{aA} F_A = f_a) \exp[-L H(F; \mathbf{R})]
 \end{aligned} \tag{32}$$

The long-run frequency distribution \mathbf{v} of the full-network activity

$$\begin{aligned}
 P(\mathbf{v} | f, I) &\propto \binom{T}{Tf} \left[\prod_a (\sum_A G_{aA} v_A)^{Tf_a} \right] \exp[-L H(\mathbf{v}; \mathbf{R})] \\
 &\propto \exp[-T H(f; \mathbf{G}\mathbf{v}) - L H(\mathbf{v}; \mathbf{R})]
 \end{aligned} \tag{33}$$

Note that if some f_a are zero, then the first exponential may be badly approximated by a delta, because the constraints lie on a facet of the simplex of frequencies $\{\mathbf{v}\}$.

The activity A' of the full-network in a *new* time bin

$$\begin{aligned} P(A' | f, I) &\propto \int d\mathbf{v} \, v_{A'} \left(\frac{T}{Tf} \right) \left[\prod_a (\sum_A G_{aA} v_A)^{Tf_a} \right] \exp[-L H(\mathbf{v}; \mathbf{R})] \\ &\approx \int d\mathbf{v} \, v_{A'} \exp[-T H(f; \mathbf{G}\mathbf{v}) - L H(\mathbf{v}; \mathbf{R})] \end{aligned} \quad (34)$$

Note that if some f_a are zero, then the first exponential may be badly approximated by a delta, because the constraints lie on a facet of the simplex of frequencies $\{\mathbf{v}\}$.

The maximum-relative-entropy distribution is, in the first two cases, an approximation of the most probable frequency distribution F or \mathbf{v} ; in the third case, an approximation of the probability distribution for A' .

10 Assumptions and limitations

Main assumptions behind this belief distribution:

We are approximating our state of knowledge with a finitely exchangeable one. In turn, this is numerically approximated by an infinitely exchangeable one for T large. But we don't really have an exchangeable belief: our degree of belief that the activity at the next time bin will differ from the activity at the present one is roughly proportional to the difference in the two subsequent activities.

The formulae say that we have equal beliefs about the underlying network states having the same activity. This isn't really our belief, for we believe there are interactions between the neurons and subnetworks thereof.

11 Discussion

The purpose of the present work was to point out and show, in a simple set-up, that the maximum-entropy method can be applied to recorded

neuronal data in a way that accounts for the larger network from which the data are sampled, eqs (26)–(28). This application leads to results that differ from the standard application which only considers the sample in isolation, eqs (24)–(25). We gave a numerical example of this difference. We have also shown how to extend the new application when the size of the larger network is unknown, eq. (30).

The latter formula, in particular, shows that the standard way of applying maximum-entropy implicitly assumes that *no* larger network exists beyond the recorded sample of neurons. One could in fact object to the application at the network level, and say that the traditional way of applying maximum-entropy, eq. (24), yields different results because it does not make assumptions about the size N of a possibly existing larger network. Such a state of uncertainty, however, is correctly formalized according to the laws of probability by introducing a probability distribution for N , and is expressed by eq. (30). This expression cannot generally be equal to (24) unless the distribution for N gives unit probability to $N = n$; that is, unless the sample *is* the full network, and no larger network exists.

The standard maximum-entropy approach therefore assumes that the recorded neurons constitute a special subnetwork, isolated from the larger network of neurons in which it is embedded, and which was also present under the recording device. This assumption is unrealistic. The maximum-entropy approach at the network level does not make such assumption and is therefore preferable. It may reveal features in a data set that were unnoticed by the standard maximum-entropy approach.

The difference in the resulting distributions between the applications at the sample and at the network levels appears in the use of Boltzmann machines with hidden units³⁸, although by a different conceptual route. It also appears in statistical mechanics: if a system is statistically described by a maximum-entropy Gibbs state, its subsystems cannot be described by a Gibbs state³⁹. A somewhat similar situation also appears in the statistical description of the final state of a non-equilibrium process starting and ending in two equilibrium states: we can describe our knowledge about the final state either by (1) a Gibbs distribution, calculated from the final equilibrium macrovariables, or (2) by the distribution obtained from the Liouville evolution of the Gibbs distribution assigned to the initial

³⁸Le Roux et al. 2008.

³⁹Maes et al. 1999.

state. The two distributions differ (even though the final *physical* state is obviously exactly the same⁴⁰), and the second allows us to make sharper predictions about the final physical state thanks to our knowledge of its preceding dynamics. In this example, though, both distributions are usually extremely sharp and practically lead to the same predictions. In neuroscientific applications, the difference in predictions of the sample vs full-network applications can instead be very relevant.

The idea of the new application leads in fact to more questions. For instance:

- Do the standard and new applications lead to different or contrasting conclusions about ‘cooperativity’, when applied to real data sets?
- How to extend the new application to the ‘inhomogeneous’ case⁴¹, in which expectations for individual neurons or groups of neurons are constrained?
- What is the mathematical relation between the new application and maximum-entropy models with hidden neurons⁴²?

Owing to space limitations we must leave a thorough investigation of these questions to future work.

Finally, we would like to point out the usefulness and importance of the probability formulae that relate our states of knowledge about a network and its samples, presented in § 7. This kind of formulae is essential in neuroscience, where we try to understand properties of extended brain areas from partial observations. The formulae presented here reflect a simple, symmetric state of ignorance. More work is needed⁴³ to extend these formulae to account for finer knowledge of the cerebral cortex and its network properties.

⁴⁰Jaynes 1993 § 4.

⁴¹Schneidman et al. 2006; Shlens et al. 2006; Roudi et al. 2009b.

⁴²Smolensky 1986; Kulkarni et al. 2007; Huang 2015; Dunn et al. 2017.

⁴³cf. Levina et al. 2017.

A Derivation of the probability distribution and of its approximations

A.1 Notation

$N = 1000$: size of the full network.

$n = 65$: size of the sample.

$T = 417\,641$: number of bins in the recording.

Bins are indexed by $t \in \{1, \dots, T\}$.

$A \in \{0, 1, \dots, N\}$: levels of total activity of the full network.

$A_t \in \{0, 1, \dots, N\}$: total activity of the full network at bin t .

$\bar{A} := (A_1, A_2, \dots, A_T)$: sequence of activities of full network during the recording.

$F_A \in \{0, 1/T, \dots, 1\}$: relative frequency of activity level A of full network during the recording. That is, activity level A appears in TF_A out of T bins: $TF_A = \sum_t \delta(A_t = A)$.

$F := (F_0, F_1, \dots, F_N)$: frequency distribution.

Note that \bar{A} completely determines F . We write the F determined by \bar{A} as $F(\bar{A})$.

$\nu_A \in [0, 1]$: long-run relative frequency of activity level A of full network during all hypothetical repetitions of the same recording in the same conditions.

$\nu := (\nu_0, \nu_1, \dots, \nu_N)$: frequency distribution.

$a \in \{0, 1, \dots, n\}$: levels of total activity of the sample.

$a_t \in \{0, 1, \dots, n\}$: total activity of the sample at bin t .

$\bar{a} := (a_1, a_2, \dots, a_T)$: sequence of activities of sample during the recording.

$f_a \in \{0, 1/T, \dots, 1\}$: relative frequency of activity level a of sample during the recording.

$f := (f_0, f_1, \dots, f_n)$: frequency distribution.

Note that \bar{a} completely determines f . We write the f determined by \bar{a} as $f(\bar{a})$.

$J_{aA} \in \{0, 1/T, \dots, 1\}$: joint relative frequency of activity levels a and A for sample and full network during the recording. That is, the pair (a, A) appears in TJ_{aA} out of T bins: $TJ_{aA} = \sum_t \delta(a_t = a) \delta(A_t = A)$.

Note that the full network can't have fewer active neurons or fewer silent neurons than its sample (for example, if $A_t = 0$ then $a_t = 0$, and if $A_t = N$ then $a_t = n$), so $J_{aA} \equiv 0$ for $a > A$ or $n - a > N - A$.

Obviously $F_A \equiv \sum_a J_{aA}$ and $f_a \equiv \sum_A J_{aA}$.

$J := (J_{00}, J_{01}, \dots, J_{nN})$: joint frequency distribution.

Note that (\bar{a}, \bar{A}) together completely determine J . We write the J determined by (\bar{a}, \bar{A}) as $J(\bar{a}, \bar{A})$.

$G_{aA} := \binom{n}{a} \binom{N-n}{A-a} \binom{N}{A}^{-1} \equiv \binom{A}{n-a} \binom{N-A}{n}^{-1}$: hypergeometric distribution.

I : background information and assumptions; includes knowledge of N and n .

A.2 Derivation

We want to know which distribution of frequencies of full-network activities occurred during the recording, given our observations of the sample. Namely, $p(F | \bar{a}, I)$. This can be obtained as the marginal of the probability distribution for the joint frequencies, $p(J | \bar{a}, I)$. If this probability distribution is represented by a set of Monte Carlo samples $\{J^{(i)}\}$, then $\{F^{(i)} \equiv (\sum_a J_{aA}^{(i)})\}$ are automatically samples of $p(F | \bar{a}, I)$.

By the theorem of total probability,

$$p(J | \bar{a}, I) = \sum_{\bar{A}} p(J | \bar{a}, \bar{A}, I) p(\bar{A} | \bar{a}, I), \quad (35)$$

the sum being over all possible sequences $\{\bar{A}\}$ of total activities.

The first factor is a singular probability distribution, because \bar{a}, \bar{A} jointly determine J :

$$\begin{aligned} p(J | \bar{a}, \bar{A}, I) &= \delta[J = J(\bar{a}, \bar{A})] \\ &\equiv \prod_{aA} \delta[T J_{aA} - \sum_t \delta(a_t = a) \delta(A_t = A)]. \end{aligned} \quad (36)$$

The second factor is derived from Bayes's theorem:

$$p(\bar{A} | \bar{a}, I) = \frac{p(\bar{a} | \bar{A}, I) p(\bar{A} | I)}{\sum_{\bar{A}} p(\bar{a} | \bar{A}, I) p(\bar{A} | I)}. \quad (37)$$

In the last formula, let's first derive $p(\bar{a} | \bar{A}, I)$. We make the following

assumption: if A_t is known, then knowledge of $a_{t'}$ or $A_{t'}$ with $t' \neq t$ is irrelevant for our inferences about a_t . (38)

It isn't a realistic assumption, but it's the one behind the maximum-entropy method in this kind of applications. From this assumption we have, using the product rule,

$$p(\bar{a} \mid \bar{A}, I) = \prod_t p(a_t \mid A_t, I), \quad (39)$$

and from simple sampling theory $p(a_t \mid A_t, I) = G_{a_t A_t}$ for each t , so that

$$p(\bar{a} \mid \bar{A}, I) = \prod_t G_{a_t A_t}. \quad (40)$$

The latter product can be rewritten by grouping together all t in which the same (a, A) pair appears: there are $T_{J_{aA}}(\bar{a}, \bar{A})$ such t . Then we consider all such pairs:

$$p(\bar{a} \mid \bar{A}, I) = \prod_{a,A} G_{aA}^{T_{J_{aA}}(\bar{a}, \bar{A})} \quad (41)$$

where $\prod_{a,A} := \prod_a \prod_A$.

Now let's decide upon $p(\bar{A} \mid I)$ in formula (37). We assume that this probability has the same value for all sequences \bar{A} having the same frequency distribution $F(\bar{A})$; that is, it functionally depends on \bar{A} only through $F(\bar{A})$. More specifically we assume that it can be hierarchically written as

$$p(\bar{A} \mid I) = \pi(F) = \int d\mathbf{v} \, q(\mathbf{v}) \prod_A v_A^{TF_A} \quad (42)$$

for some density function $q(\mathbf{v}) d\mathbf{v}$. The integral expression is simply a mathematical way to write $\pi(F)$, and doesn't need to be further interpreted (the integral disappears if we can solve it analytically). But it's also the formula for infinite exchangeability, and from this point of view \mathbf{v} can be interpreted as the long-run frequency distribution of the activities in all experiments performed in the same conditions – imagining to join together their recording times. This is the point of view underlying the maximum-entropy method.

Replacing (41) and (42) into (37) and rearranging we find

$$\begin{aligned} p(\bar{A} \mid \bar{a}, I) &= \frac{1}{Z} \int d\mathbf{v} \, q(\mathbf{v}) \prod_{a,A} [G_{aA}^{T_{J_{aA}}(\bar{a}, \bar{A})}] \prod_A v_A^{TF_A} \\ &\equiv \frac{1}{Z} \int d\mathbf{v} \, q(\mathbf{v}) \prod_{a,A} (G_{aA} v_A)^{T_{J_{aA}}(\bar{a}, \bar{A})}, \end{aligned} \quad (43)$$

where $Z := \sum_{\bar{A}} p(\bar{a} | \bar{A}, I) p(\bar{A} | I)$ is the normalization factor and the second equality comes from rewriting

$$\mathbf{v}_A^{TF_A} \equiv \mathbf{v}_A^{T \sum_a J_{aA}} \equiv \prod_a \mathbf{v}_A^{T J_{aA}}. \quad (44)$$

Note that the normalization factor is independent of \bar{A} and only depends on \bar{a} , which is given and fixed in our inference.

We can now return to our initial probability distribution (35). Replacing (36) and (43) into it and rearranging we have

$$p(J | \bar{a}, I) = \frac{1}{Z} \int d\mathbf{v} \, q(\mathbf{v}) \sum_{\bar{A}} \delta[J = J(\bar{a}, \bar{A})] \prod_{a,A} (G_{aA} \mathbf{v}_A)^{T J_{aA}(\bar{a}, \bar{A})}. \quad (45)$$

This expression can be considered as the marginal distribution of the joint density function

$$p(J, \mathbf{v} | \bar{a}, I) = \frac{1}{Z} q(\mathbf{v}) \sum_{\bar{A}} \delta[J = J(\bar{a}, \bar{A})] \prod_{a,A} (G_{aA} \mathbf{v}_A)^{T J_{aA}(\bar{a}, \bar{A})}. \quad (46)$$

If we represent this density by a set of Monte Carlo samples for (J, \mathbf{v}) , we automatically also have samples for the distribution for J , that for F , and that for \mathbf{v} .

Let's consider the sum over all possible sequences, $\sum_{\bar{A}}$, in the joint density (46). Owing to the delta in the summand, the only \bar{A} s that contribute to the sum are those for which $J(\bar{a}, \bar{A}) = J$. This also means that all terms in the sum have the same numerical value, because their values depend on \bar{A} only through $J(\bar{a}, \bar{A})$, which is fixed. We must therefore only find how many terms there are in the sum, and multiply the value of one term for the number of terms.

With \bar{a} and J fixed, we are asking how many \bar{A} s satisfy $J(\bar{a}, \bar{A}) = J$. Consider the problem from the following point of view. We have a grid of boxes with $(n + 1)$ rows and $(N + 1)$ columns, indexed by (a, A) . A sequence of T balls go into the boxes: if the t th ball goes into the (a, A) box, it means that at the t th time bin the activities of sample and full network were $a_t = a$ and $A_t = A$. In the typical combinatorial problem we would ask in how many ways we can fill the boxes, with $T J_{aA}$ balls in the box (a, A) , by throwing the T balls. The number would be given by all $T!$ possible permutations of the balls, but considering permutations

of two or more balls within the same box as equivalent, thus finding the multinomial coefficient

$$\binom{T}{TJ} \equiv \binom{T}{TJ_{00}, \dots, TJ_{nN}} := \frac{T!}{\prod_{a,A} (TJ_{aA})!}. \quad (47)$$

In our case, however, we have one constraint: \bar{a} is fixed, which means that the *row* in which each ball must fall is determined and fixed. The t th ball must perforce fall into row a_t . In considering all possible permutations we must therefore exclude those that change the rows of the balls. This means that only within-row permutations are allowed. Within each row the counting proceeds as usual, so that for row a , which has a total of $T \sum_A J_{aA} \equiv T f_a$ balls, we have

$$\binom{T f_a}{TJ_{a0}, \dots, TJ_{aN}} = \frac{(T f_a)!}{\prod_A (TJ_{aA})!}. \quad (48)$$

We therefore find that the number of terms in the sum of (46) is


$$\prod_a \binom{T f_a}{TJ_{a0}, \dots, TJ_{aN}}. \quad (49)$$

If T is large the logarithm of the multinomial coefficient can be approximated using the Shannon entropy H , which for a generic distribution (x_i) satisfies the bounds⁴⁴

$$TH(x_0, \dots, x_N) - \ln \binom{T+N}{T} \leq \ln \binom{T}{T x_0, \dots, T x_N} \leq TH(x_0, \dots, x_N). \quad (50)$$

We therefore approximate the multiplicity (49) as

$$\begin{aligned} \exp \left[T \sum_a f_a H \left(\frac{J_{a0}}{f_a}, \dots, \frac{J_{aN}}{f_a} \right) \right] &\equiv \exp \left[-T \sum_a f_a \sum_A \frac{J_{aA}}{f_a} \ln \frac{J_{aA}}{f_a} \right] \equiv \\ \exp \left[-T \sum_{a,A} J_{aA} \ln J_{aA} + T \sum_a f_a \ln f_a \right] &\equiv \exp [TH(J) - TH(f)], \end{aligned} \quad (51)$$

where the second line is obtained by combining the sums, simplifying, and using the definition of Shannon entropy.  Is it possible that this

⁴⁴Lemma 2.2 pp. 429–430 in I. Csiszár et al. (2004): *Information theory and statistics: a tutorial*. Foundations and Trends in Communications and Information Theory **1**⁴, 417–528. <http://www.renyi.hu/~csiszar/>.

approximation affects the result? – Update: the discrepancy is less than 0.1%.

The sum over \bar{A} in formula (46) can therefore be replaced by any single summand multiplied by the multiplicity (51) above. The marginal frequency distribution f is still fixed, determined by \bar{a} , so we still have the constraint $\sum_A J_{aA} = f_a$ for each a , which we compactly write $\delta(\sum J = f)$. We find

$$\begin{aligned}
 p(J, \mathbf{v} \mid \bar{a}, I) &= \frac{1}{Z} \delta(\sum J = f) q(\mathbf{v}) \exp[T H(J) - T H(f)] \prod_{a,A} (G_{aA} \mathbf{v}_A)^{T J_{aA}(\bar{a}, \bar{A})} \\
 &= \frac{1}{Z} \delta(\sum J = f) q(\mathbf{v}) \exp\{T[H(J) - H(f) + \sum_{a,A} J_{aA} \ln(G_{aA} \mathbf{v}_A)]\},
 \end{aligned} \tag{52}$$

where the second equality comes from re-expressing the product over (a, A) in exponential-logarithm form.

Now we note that $G_{aA} \mathbf{v}_A$ is a normalized distribution in (a, A) owing to the properties of the hypergeometric distribution and of \mathbf{v} . We denote it $\mathbf{G} \cdot \mathbf{v}$. We can also use the definition of relative entropy

$$H[(x_i); (y_i)] := \sum_i x_i \ln \frac{x_i}{y_i} \equiv -H[(x_i)] - \sum_i x_i \ln y_i, \tag{53}$$

to combine the first and last terms within the exponential. The logarithm of our joint density (46) can then finally be written as

$$\begin{aligned}
 \ln p(J, \mathbf{v} \mid \bar{a}, I) &= \ln \delta(\sum J = f) + \ln q(\mathbf{v}) - T H(J; \mathbf{G} \cdot \mathbf{v}) \\
 &\quad - T H(f) - \ln Z(\bar{a})
 \end{aligned}$$

(54)

The last two terms are constants: they only depend on \bar{a} and $f(\bar{a})$, which are given and fixed in our problem. They can therefore be discarded in Monte Carlo sampling. In sampling, the delta term is taken care of by restricting the sampling to the set of allowed J .

The density $q(\mathbf{v})$ remains to be specified. We use an entropic density

$$q(\mathbf{v}) \propto \exp[-L H(\mathbf{v}; \mathbf{r})] \quad (55)$$

where \mathbf{r} is a reference distribution – we take the uniform one for simplicity – and L is of order unity, say less than 10. This density is motivated by the consideration of multiplicities: a frequency distribution, like \mathbf{v} , can be realized in a number of ways equal to a multinomial coefficient. This coefficient, by formula (50), approximately equal to the exponential of the Shannon entropy of the distribution. So the density (55) keeps track of this multiplicity, but regulates its importance via the parameter L . We can also choose $L = 0$, which leads to $q(\mathbf{v})$ being uniform in \mathbf{v} .

A.3 Approximations

We consider two successive approximations of the log-density (54), based on the fact that T is large.

The relative entropy $H(J; G \cdot \mathbf{v})$ is always positive, and vanishes if only if $J = G \cdot \mathbf{v}$. This term is multiplied by $-T$ and appears in an exponential. We therefore approximate it with a delta.

This delta restricts J to the form $G \cdot \mathbf{v}$, making it completely determined by \mathbf{v} . The density (54) thus collapses onto the \mathbf{v} -space. More precisely a subspace, since J must also satisfy the constraint $\sum_a J = f$, which becomes $\sum_a G \cdot \mathbf{v} = f$, constraining \mathbf{v} .

Such a constraint is actually too strong, in the sense that no distribution \mathbf{v} can satisfy it – it leads to negative values nu_a for some a s. This problem is mitigated by considering a softer constraint between J and f , for example taking into account errors in spike sorting or just considering the first m moments of f . We express the softer constraint by replacing the delta with a normal N whose covariance matrix σ is determined from the spike-sorting error. With this approximation we reduce our problem to the density

$$\ln p(\mathbf{v} | \bar{a}, I) \propto \ln N(\sum_a G \cdot \mathbf{v} | f, \sigma) - L H(\mathbf{v}; \mathbf{r}), \quad (56)$$

where $q(\mathbf{v})$ has been replaced by the entropic density (55). We can call this the ‘maximum-entropy approximation with Bayesian correction’. It can also be represented by a set of Monte Carlo samples $\{\mathbf{v}^{(i)}\}$.

As a further approximation we can simply consider the mode of (56). This is found by maximizing the relative entropy in its expression under the moment constraints about f . The solution is the one given by the maximum-entropy method applied to the full network with subnetwork constraints.

Figure A shows comparison of the expected values (means of the Monte Carlo samples) for the density (54) (for F), the density (56) (for ν and using the first six moments of f as constraints), and the maximum-entropy solution (six moments as constraints).

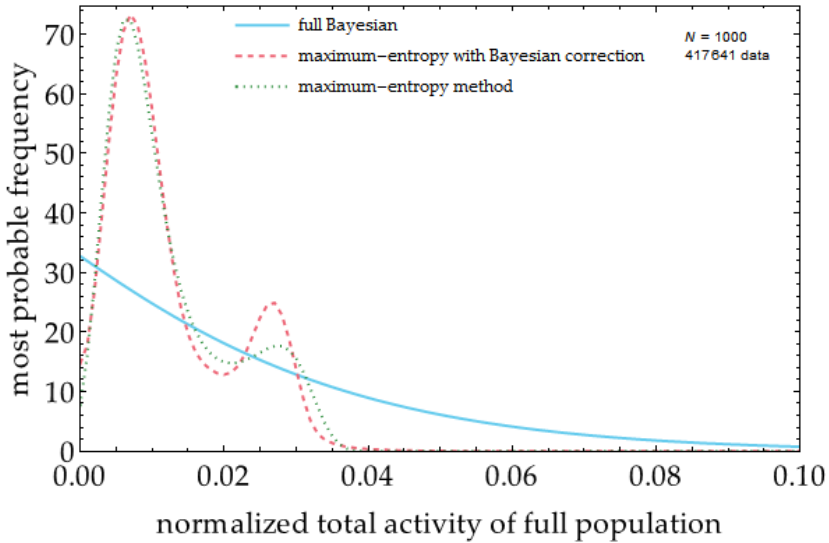


Figure 8 Comparison of (54), (56), and the maximum-entropy approximation (the vertical-axis label is wrong).

Bibliography

- (‘de X’ is listed under D, ‘van X’ under V, and so on, regardless of national conventions.)
- Abeles, M. (1991): *Corticonics: Neural circuits of the cerebral cortex*. (Cambridge University Press, Cambridge).
- 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.
- Barreiro, A. K., Gjorgjieva, J., Rieke, F. M., Shea-Brown, E. T. (2010a): *When are microcircuits well-modeled by maximum entropy methods?* [arXiv:1011.2797](https://arxiv.org/abs/1011.2797).
- Barreiro, A. K., Shea-Brown, E. T., Rieke, F. M., Gjorgjieva, J. (2010b): *When are microcircuits well-modeled by maximum entropy methods?* *BMC Neurosci.* **11**^{Suppl. 1}, P65.
- Berényi, A., Somogyvári, Z., Nagy, A. J., Roux, L., Long, J. D., Fujisawa, S., Stark, E., Leonardo, A., et al. (2014): *Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals*. *J. Neurophysiol.* **111**⁵, 1132–1149. <http://www.buzsakilab.com/content/PDFs/Berenyi2013.pdf>.
- 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.
- Burg, J. P. (1975): *Maximum entropy spectral analysis*. PhD thesis. (Stanford University, Stanford). <http://sepwww.stanford.edu/data/media/public/oldreports/sep06/>.
- Caianiello, E. R. (1961): *Outline of a theory of thought-processes and thinking machines*. *J. Theor. Biol.* **1**², 204–235.
- (1986): *Neuronic equations revisited and completely solved*. In: Palm, Aertsen (1986), 147–160.
- Callen, H. B. (1985): *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed. (Wiley, New York). First publ. 1960.
- Caticha, A., Preuss, R. (2004): *Maximum entropy and Bayesian data analysis: entropic prior distributions*. *Phys. Rev. E* **70**⁴, 046127.
- Cohen, M. R., Kohn, A. (2011): *Measuring and interpreting neuronal correlations*. *Nat. Neurosci.* **14**⁷, 811–819. <http://marlenecohen.com/pubs/CohenKohn2011.pdf>.
- 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/>.
- De Bruijn, N. G. (1961): *Asymptotic Methods in Analysis*, 2nd ed. (North-Holland, Amsterdam). First publ. 1958.
- de Finetti, B. (1959): *La probabilità e la statistica nei rapporti con l’induzione, secondo i diversi punti di vista*. In: de Finetti (2011), 1–115. Transl. in de Finetti (1972), ch. 9, pp. 147–227.
- (1972): *Probability, Induction and Statistics: The art of guessing*. (Wiley, London).
- ed. (2011): *Induzione e statistica*, reprint. (Springer, Berlin). First publ. 1959.
- Diaconis, P. (1977): *Finite forms of de Finetti’s theorem on exchangeability*. *Synthese* **36**², 271–281. <http://statweb.stanford.edu/~cgates/PERSI/year.html>.
- Diaconis, P., Freedman, D. (1980): *Finite exchangeable sequences*. *Ann. Prob.* **8**⁴, 745–764.
- Dunn, B., Battistin, C. (2017): *The appropriateness of ignorance in the inverse kinetic Ising model*. *J. Phys. A* **50**¹², 124002.
- Erickson, G. J., Rychert, J. T., Smith, C. R., eds. (1998): *Maximum Entropy and Bayesian Methods*. (Springer, Dordrecht).

- Ericson, W. A. (1969): *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/Ericks-on-JRSSB-1969.pdf>. See also discussion in Sampford, Scott, Stone, Lindley, Smith, Kerridge, Godambe, Kish, et al. (1969).
- (1976): *A Bayesian approach to two-stage sampling*. Tech. rep. AFFDL-TR-75-145. (University of Michigan, Ann Arbor, USA). <http://hdl.handle.net/2027.42/4819>
- Fang, S.-C., Rajasekera, J. R., Tsao, H.-S. J. (1997): *Entropy Optimization and Mathematical Programming*, reprint. (Springer, New York).
- Feller, W. (1968): *An Introduction to Probability Theory and Its Applications*. Vol. I, 3rd ed. (Wiley, New York). First publ. 1950.
- Ford, K. W., ed. (1963): *Statistical Physics*. (Benjamin, New York).
- Freedman, D. A., Pisani, R., Purves, R. (2007): *Statistics*, 4th ed. (Norton, London). First publ. 1978.
- 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.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B. (2014): *Bayesian Data Analysis*, 3rd ed. (Chapman & Hall/CRC, Boca Raton, USA). First publ. 1995.
- 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.
- Gerwin, S., Berens, P., Bethge, M. (2009): *A joint maximum-entropy model for binary neural population patterns and continuous signals*. Adv. Neural Information Processing Systems (NIPS) **22**, 620–628.
- Gerwin, S., Macke, J. H., Bethge, M. (2010): *Bayesian inference for generalized linear models for spiking neurons*. Front. Comput. Neurosci. **4**, 12.
- Ghosh, M., Meeden, G. (1997): *Bayesian Methods for Finite Population Sampling*, reprint. (Springer, Dordrecht).
- Good, I. J. (1966): *How to estimate probabilities*. J. Inst. Maths. Applics **2**⁴, 364–383.
- Grandy Jr., W. T. (1980): *Principle of maximum entropy and irreversible processes*. Phys. Rep. **62**³, 175–266.
- Grandy Jr., W. T., Schick, L. H., eds. (1991): *Maximum Entropy and Bayesian Methods: Laramie, Wyoming, 1990*. (Kluwer, 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.
- Heath, D., Sudderth, W. (1976): *De Finetti's theorem on exchangeable variables*. American Statistician **30**⁴, 188–189.
- 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.
- Huang, H. (2015): *Effects of hidden nodes on network structure inference*. J. Phys. A **48**³⁵, 355002.
- IEEE (1993): *ANSI/IEEE Std 260.3-1993: American National Standard: Mathematical signs and symbols for use in physical sciences and technology*. Institute of Electrical and Electronics Engineers.
- ISO (International Organization for Standardization) (1993): *Quantities and units*, 3rd ed. International Organization for Standardization.

- ISO (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.
- 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>.
- (1979): *Where do we stand on maximum entropy?* In: Levine, Tribus (1979), 15–118. <http://bayes.wustl.edu/etj/node1.html>; repr. with an introduction in Jaynes (1989), pp. 210–314.
- (1982a): *Prior information in inference*. (). <http://bayes.wustl.edu/etj/node2.html>.
- (1982b): *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.
- (1993): *Inferential scattering*. <http://bayes.wustl.edu/etj/node1.html>. First publ. 1985 in Smith, Grandy (1985) pp. 377–398.
- (1996): *Monkeys, kangaroos, and N*. <http://bayes.wustl.edu/etj/node1.html>. First publ. 1986. (Errata: in equations (29)–(31), (33), (40), (44), (49) the commas should be replaced by gamma functions, and on p. 19 the value 0.915 should be replaced by 0.0915).
- (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. <https://archive.org/details/XQUHIUXHTQUHIQXUIHX2>, <http://www-biba.inrialpes.fr/Jaynes/prob.html>.
- Jeffreys, H. (1983): *Theory of Probability*, third ed. with corrections. (Oxford University Press, London). First publ. 1939.
- Johnson, W. E. (1924): *Logic. Part III: The Logical Foundations of Science*. (Cambridge University Press, Cambridge). <https://archive.org/details/logic03john>.
- Kendall, D. G. (1967): *On finite and infinite sequences of exchangeable events*. Studia Sci. Math. Hung. **2**, 319–327.
- Kulkarni, J. E., Paninski, L. (2007): *Common-input models for multiple neural spike-train data*. Netw. **18**⁴, 375–407.
- Kullback, S. (1987): *The Kullback-Leibler distance*. American Statistician **41**⁴, 340–341.
- Le Roux, N., Bengio, Y. (2008): *Representational power of restricted Boltzmann machines and deep belief networks*. Neural Comp. **20**⁶, 1631–1649.
- Levina, A., Priesemann, V. (2017): *Subsampling scaling*. Nat. Comm. **8**, 15140.
- Levine, R. D., Tribus, M., eds. (1979): *The Maximum Entropy Formalism: A Conference Held at the Massachusetts Institute of Technology on May 2–4, 1978*. (MIT Press, Cambridge, USA).
- 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. (1991): *Maximum entropy connections: neural networks*. In: Grandy, Schick (1991), 237–244.

- Macke, J. H., Buesing, L., Cunningham, J. P., Yu, B. M., Shenoy, K. V., Sahani, M. (2011a): *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., Oppen, 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.
- (2011b): *Common input explains higher-order correlations and entropy in a simple model of neural population activity*. Phys. Rev. Lett. **106**²⁰, 208102.
- 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 Hasse, 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>.
- Moore, G. P., Perkel, D. H., Segundo, J. P. (1966): *Statistical analysis and functional interpretation of neuronal spike data*. Annu. Rev. Physiol. **28**, 493–522.
- 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>.
- NIST (National Institute of Standards and Technology) (1995): *Guide for the Use of the International System of Units (SI): NIST special publication 811, 1995 edition*. National Institute of Standards and Technology. <http://physics.nist.gov/cuu/Uncertainty/bibliography.html>.
- Palm, G., Aertsen, A., eds. (1986): *Brain Theory*. (Springer, Berlin).
- 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.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (2007): *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge). First publ. 1988.
- Rodriguez, 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](https://arxiv.org/abs/1605.04740). Data available at <https://doi.org/10.5061/dryad.n9f77>.
- Roudi, Y., Aurell, E., Hertz, J. A. (2009a): *Statistical physics of pairwise probability models*. Front. Comput. Neurosci. **3**, 22.
- Roudi, Y., Nirenberg, S., Latham, P. E. (2009b): *Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't*. PLoS Comput. Biol. **5**⁵, e1000380.
- Roudi, Y., Tyrcha, J., Hertz, J. (2009c): *Ising model for neural data: model quality and approximate methods for extracting functional connectivity*. Phys. Rev. E **79**⁵, 051915.
- Rumelhart, D. E., McClelland, J. L., PDP Research Group, eds. (1999): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Vol. 1: Foundations, 12th printing. (MIT Press, Cambridge, USA).
- 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 **ericson1969b**.
- 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., Toyozumi, 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, C. R., Grandy Jr., W. T., eds. (1985): *Maximum-Entropy and Bayesian Methods in Inverse Problems*. (Reidel, Dordrecht).
- Smolensky, P. (1986): *Information processing in dynamical systems: foundations of harmony theory*. In: Rumelhart, McClelland, PDP Research Group (1999), ch. 6, 194–281.
- 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](https://arxiv.org/abs/q-bio/0611072).
- (2009): *Spin glass models for a network of real neurons*. [arXiv:0912.5409](https://arxiv.org/abs/0912.5409).
- Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., Hwa, T. (2009): *Identification of direct residue contacts in protein-protein interaction by message passing*. Proc. Natl. Acad. Sci. (USA) **106**¹, 67–72.
- Whitworth, W. A. (1897): *DCC Exercises: Including Hints for the Solution of All the Questions in Choice and Chance*. (Deighton Bell & Co., Cambridge).
- (1965): *Choice and Chance: With One Thousand Exercises*, repr. of 5th ed. (Hafner, New York and London). First publ. 1867.