

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

P.G.L. Porta Mana

<pgl@portamana.org>

V. Rostami

<vrostami@uni-koeln.de>


Y. Roudi

<yasser.roudi@ntnu.no>


E. Torre

<torre@ibk.baug.ethz.ch>

Draft of 15 November 2019 (first drafted 4 November 2015)

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

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


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



1 Introduction: a simple model for questions about large neuronal networks

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

Answering this kind of questions always engages an element of uncertainty. We can't say 'the answer is such and such', but can at best assign a degree of reasonable belief – that is, probability – to every

possible answer. The assessment of this distribution of belief involves experimental data, such as recordings of neuronal activity from specific brain areas, and pre-data knowledge about biological conditions and mechanisms. Our pre-data degrees of belief are often simplified to be mathematically more tractable, and are therefore called ‘models’.

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

 [Yasser:] In the present work we aim to address the issue of building probability distributions over a large network using recording from subnetworks by considering the distribution of total population activity as an example. [In the present work we propose such an intermediate model. It answers this question: ] How much can the *total* activity of a large neuronal network have been, given the observation of the activity of a very small sample thereof? This model addresses a larger brain area, avoiding the assumption of isolation of the sample; and by focusing on the total activity, rather than the activity of individual neurons, it remains simple and numerically tractable.

The model we propose is based on the straightforward combination of the maximum-entropy method and basic sampling relations from the probability calculus, discussed in § 2. The maximum-entropy or minimum-relative-entropy method (Jaynes 1957a; much clearer in Jaynes 1963; Sivia 2006; Hobson et al. 1973; Jaynes 1996a; Grandy 1980) has been used for different kinds of estimations of the neuronal activity of various brain areas and about other phenomena of importance to the

neurosciences, for example gene and protein interaction (for example Martignon et al. 1995; Bohte et al. 2000; Shlens et al. 2006; Schneidman et al. 2006; Tkačik et al. 2006; Macke et al. 2009; Tkačik et al. 2009; Roudi et al. 2009; Barreiro et al. 2010; Gerwinn et al. 2010; Macke et al. 2011; Ganmor et al. 2011; Cohen et al. 2011; Granot-Atedgi et al. 2013; Macke et al. 2013; Tkačik et al. 2014; Shimazaki et al. 2015; Mora et al. 2015; Lezon et al. 2006; Weigt et al. 2009).

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

We want to stress the usefulness of making a quantified guess about the activity of a larger brain area. Such a guess seems indeed to be the primary idea behind recording a sample from that area. There’s also an important methodological reason. Maximum-entropy methods have been used to assess the informational sufficiency of pairwise correlations and correlations of higher order (see for example Martignon et al. 1995; Bohte et al. 2000; Schneidman et al. 2006; Shlens et al. 2006; Barreiro et al. 2010; Ganmor et al. 2011; Granot-Atedgi et al. 2013). But when the difference in size between sample and full network is too large, correlation sufficiency for the sample implies the *lack* of correlation

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

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

We obviously don't know whether the frequency distribution obtained with our approach is the *actual* one which the activity levels of the full network had during the recording; it's only the most plausible. In high dimensions, however, the features of the most plausible distribution may *not* be typical of the majority of most plausible distributions; and the set of all possible frequency distributions, if the full network for example comprises 1 000 neurons, is a 1 000-dimensional space. In § 6 we therefore try to assess which features of the frequency distribution delivered by our method may be typical and therefore expected of the actual one. We find that general features such as the bimodality of the first data set are indeed typical. Maximum-entropy models can be considered as approximations of Bayesian models based on various assumptions of inferential sufficiency (Jaynes 1996b; Porta Mana 2017a). How do our guesses change if we modify our pre-data assumptions? We show, in the same section, that the typical features indicated by our method are robust against such changes.

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

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

2 The approach: maximum-entropy and sampling

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

The context we consider is as follows. During an experimental session we have recorded the spiking activities of n neurons for a certain amount of time. These neurons are our 'sample' or 'subnetwork'. Their spikes are binned into T time bins and binarized to $\{0, 1\}$ values in each bin. Call a_t the number of neurons that fire during time bin t : this is the *total activity* of the sample, or just 'activity' for short. Obviously $a_t \in \{0, 1, \dots, n\}$; if

$a_t = 0$, no neuron spikes during bin t ; if $a_t = n$, all spike at some point during bin t . For brevity, let's say 'at t ' for 'during time bin t '. If we divide the total activity by the population size we have the normalized total activity or population-averaged activity a/n , ranging from 0 to 1 in $1/n$ steps. From the activities $\{a_t\}$ we can count how often the activity levels $a = 0, a = 1$, and so on appeared during the recording, obtaining the distribution of measured relative frequencies $(f_a) =: f$. We can also consider the sample activity at time bins *outside* of the recorded period. Such activity is unknown to us, of course.

For many animal species, the neurons that are recorded within a brain area are not specifically chosen from among the rest, owing to several limiting factors; for example, limitations in how precisely electrodes are inserted or neurons are targeted by viruses. The set of recorded neurons may even change slightly across experimental sessions that are very far apart in time. We assume that there's an area, comprising a network of N neurons, for which we believe that any other sample of size n could have equally plausibly been recorded instead of the sample of n neurons that was actually recorded. We call this larger network the 'full network'. The total activity of these N neurons at t is A_t . The relative frequencies of the various activity levels during the recording were $(F_A) =: F$. We don't know the values A_t at each t , or the frequency distribution F . We only know for certain that $A_t \in \{0, 1, \dots, N\}$ and that $A_t \geq a_t$ for obvious reasons. For the time being we assume that we know N ; in § 5 we discuss the consequences of our lack of precise knowledge about this number.

Our questions concern general features of the total activity A of the full network during and after the recording, and across sessions under the same study conditions. For example: what was its frequency distribution during the recording? How much does this frequency distribution change across sessions? How much total activity should we expect at any time bin during a recording? The approach presented here gives probability distributions over the possible answers to these questions.

The idea behind our approach is easily summarized:

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

Let's discuss these points in detail.

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

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

The maximum-entropy distribution is thus an approximation of our belief distribution about four different quantities. Note that the four distributions above numerically differ in higher-order approximations. Second remark: the maximum-entropy method based on the Shannon entropy implicitly makes some assumptions about the probabilities for the long-run frequency distributions (Jaynes 1996b; Porta Mana 2009; 2017a). We discuss these two remarks further in § 6.

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

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

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

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

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

This moment can be interpreted as the expectation of the number of distinct m -tuples of simultaneously spiking neurons (within a bin’s time width), normalized by the number of distinct m -tuples. For example, with $m = 2$, if $a = 4$ neurons spike in a network of $n = 5$, we have $\binom{4}{2} = 6$ distinct pairs of simultaneously spiking neurons, and the total number of distinct pairs is $\binom{5}{2} = 10$. The normalized number of spiking pairs is therefore $6/10$. Note that the first m factorial moments provide the same information as the first m power moments and vice versa: they linearly determine each other because $\binom{a}{m}$ is a polynomial in a of degree m . Specifying one set is therefore equivalent to specifying the other set. But the normalized factorial moments have a convenient property for our analysis (Porta Mana et al. 2015 § 2.3; Potts 1953): *the first n normalized factorial moments for the sample and for the full network are numerically identical*:

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

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

We can therefore apply the maximum-entropy method to obtain a distribution $P_{\text{me}}(A)$ for the full network by constraining its m th factorial moment to be equal to the sample’s recorded average of $\binom{a}{m} / \binom{n}{m}$, for as

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

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

Finding the constrained maximum-entropy distribution $P_{\text{me}}(A)$ amounts to a convex optimization (Mead et al. 1984; Press et al. 2007 ch. 10; Fang et al. 1997; Boyd et al. 2009; Porta Mana 2017b) and for the numbers N, n, m considered in the present work it can be done on a modern computer without approximations of the partition functions that appear in the formalism.

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

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

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

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

3 Example application: two data sets

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

- The **first data set**, from Stensola et al. (2012 rat 14147), consists of $n = 65$ neurons (27 of which classified as grid cells) from rat Medial Entorhinal Cortex, recorded for about 20 minutes. Their spikes are binned into $T = 417\,641$ bins of 3 ms width.
- The **second data set**, from Rostami et al. (2017), consists of $n = 159$ neurons from macaque Motor Cortex, recorded for about 15 minutes. Their spikes are binned into $T = 300\,394$ bins of 3 ms width.

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

The figure shows that the distribution for the full-network is less flat than the measured frequency distribution for the sample; their difference increases with N . Most remarkably, for the first data set the distribution has two distinct low-activity modes. For the second data the distribution presents a small shoulder on the right of the mode. These features are clearly not present in the sample or in the maximum-entropy distribution at the sample level. The application of the probability calculus thus reveals interesting possible features of the full network.

Use with moment constraints

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

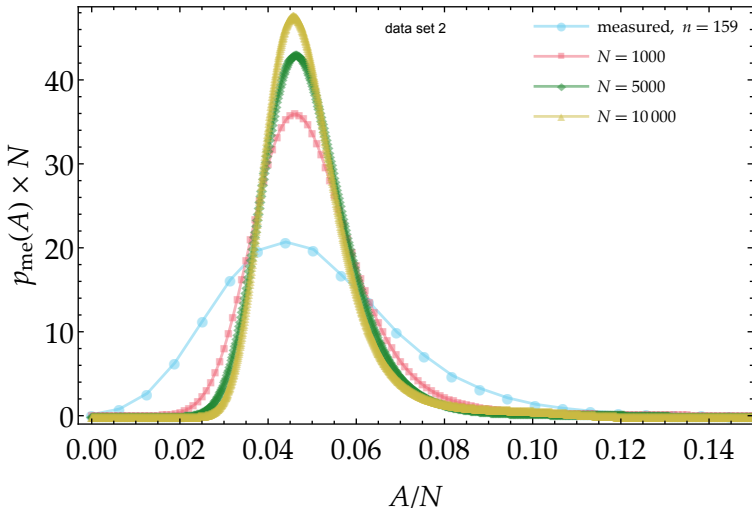
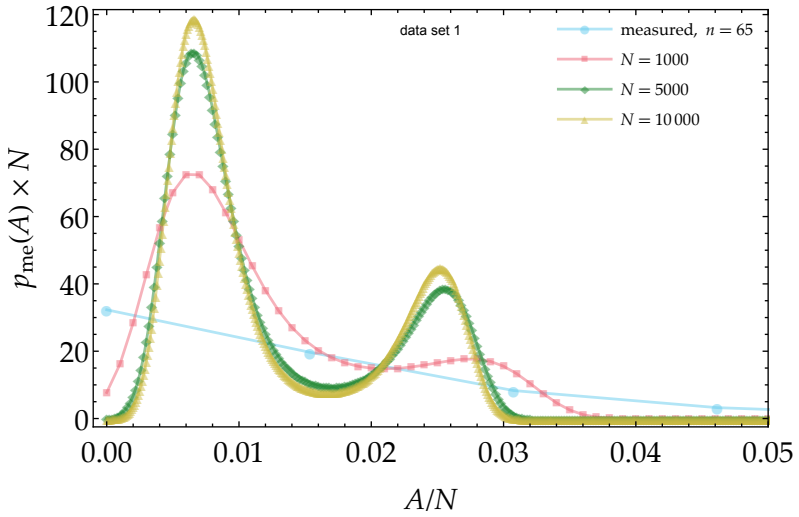


Figure 1 ***

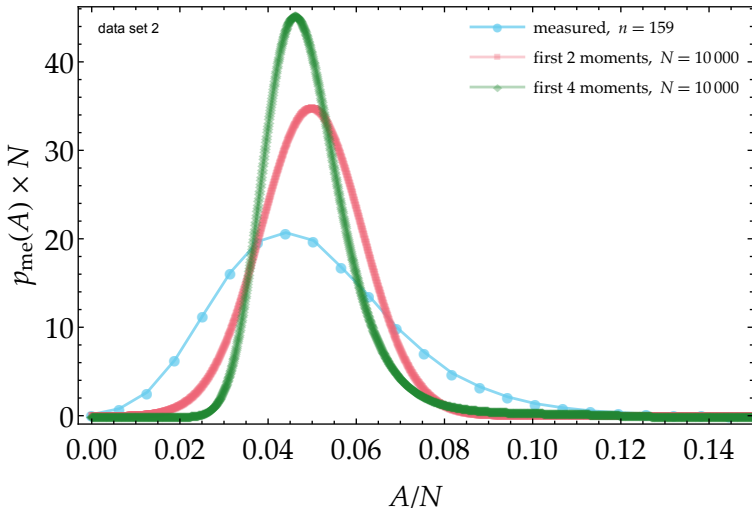
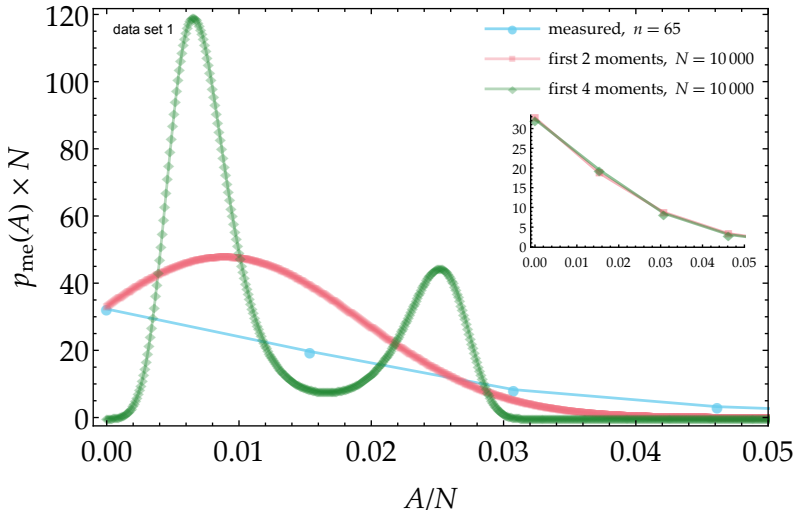


Figure 2 ***

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

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

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

As mentioned in the Introduction (see references there), in the neurosciences the maximum-entropy method has also been used as a way of quantifying the ‘cooperativity’ (Gerstein et al. 1985) or ‘interaction’ (Martignon et al. 1995; Schneidman et al. 2006; Shlens et al. 2006) or ‘synchrony’ (Bohte et al. 2000; Amari et al. 2003) of neuronal activity. In this section we discuss how our proposed application bears on this kind of quantification.

Let us translate this question into more precise terms first, because ‘cooperativity’, ‘interaction’, and similar notions are vague. A translation in terms of *informational sufficiency* seems to suit all these notions (for an intuitive understanding of sufficiency, which goes back to Fisher 1922, see e.g. Bernardo et al. 1994 § 4.5; Jaynes 2003 ch. 8, § 14.2; Cifarelli et al. 1982; Kullback et al. 1951. For technical results and connections with the notion of symmetry see e.g. Darmois 1935; Neyman 1935; Koopman 1936; Pitman 1936; Halmos et al. 1949; Bahadur 1954; Berk 1972; Lauritzen 1974a; Cifarelli et al. 1980; 1981; Diaconis et al. 1981; Lauritzen 1988; Diaconis 1992; Furmańczyk et al. 1998; Fortini et al. 2000; Nogales et al. 2000; Kallenberg 2005; Lauritzen 2007; Ay et al. 2015). The idea is as follows. We want to assess our degree of belief about the frequencies of the activities of the sample, or about the activity of the sample in a new time bin. For this assessment we should use all measured data we have. But it sometimes happens that dropping part of the data – for example, the measured third- and higher-order moments – leaves our assessment

almost unchanged: the dropped data are *informationally irrelevant* or almost so. The remaining data – for example, first and second moments – are then *informationally sufficient*.

We can approximately check how much a part of the data is irrelevant by simply dropping it from the conditional of our probability, and see if the latter changes appreciably. When our degree of belief is built with the maximum-entropy method, this is done by dropping part of the constraints and see how the resulting distribution changes. For example, we may only retain the first two moments (means and correlations) or the first four moments.

In the neurosciences this check is often done with maximum-entropy distributions constructed *for the sample*, ignoring the full network it's sampled from (for example Martignon et al. 1995; Bohte et al. 2000; Schneidman et al. 2006; Shlens et al. 2006; Barreiro et al. 2010; Ganmor et al. 2011; Granot-Atedgi et al. 2013). The method introduced in the present work, however, reveals an issue with this application at the sample level. The issue doesn't come from the method itself but from a general fact of probability theory. If a probability distribution has some sufficient statistics (Bernardo et al. 1994 § 4.5; Fortini et al. 2000), then its marginals *cannot* have the same sufficient statistics, and vice versa, except for trivial cases such as uniform distributions. This generally also holds among different marginals: if a marginal has some sufficient statistics, another won't have it. This impossibility is known in statistical mechanics: marginals of Gibbs states aren't Gibbs states (Maes et al. 1999). Mathematically this impossibility translates into a system of n independent equations in m unknowns with $n > m$ (Porta Mana et al. 2015 § 3.1; 2018 § 3).

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

Questions about informational sufficiency are therefore correctly

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

But does the application at the full-network level lead to appreciably different results from that at the sample level? after all we're interested in an approximate informational sufficiency, not in an analytically exact one. A correct answer can only be given case by case. Figure 3 gives a graphical answer to this question in the case of our first data set. The upper plot shows the maximum-entropy distributions for a full network of 10 000 neurons, constructed from two moments (---) and from four moments (—). The lower plot shows the maximum-entropy distributions for the sample, from two moments (▽) and from four moments (□), and the distributions obtained by marginalizing the full-network distribution to the sample size, from two moments (▼) and from four moments (■). The measured frequency distribution (●) is also shown. We note the following:

- The application to the full-network (upper plot) leads to two completely different distributions (even different number of modes); clearly the two sets of statistics are not even approximately equivalent for inferential purposes.
- The application at the sample level leads to deceptively similar distributions (▽ and □), which can only be distinguished on a logarithmic scale. This could lead to the erroneous conclusion that the two statistics are approximately equivalent.
- The difference between marginals from the full-network distribution and the distributions obtained at the sample level (filled vs empty markers) is larger than the difference between different statistics (triangles vs squares).
- The marginals of the application to the full network are closer to the measured frequencies than the distributions obtained at the sample level (although this closeness is not a valid criterion for their goodness).

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

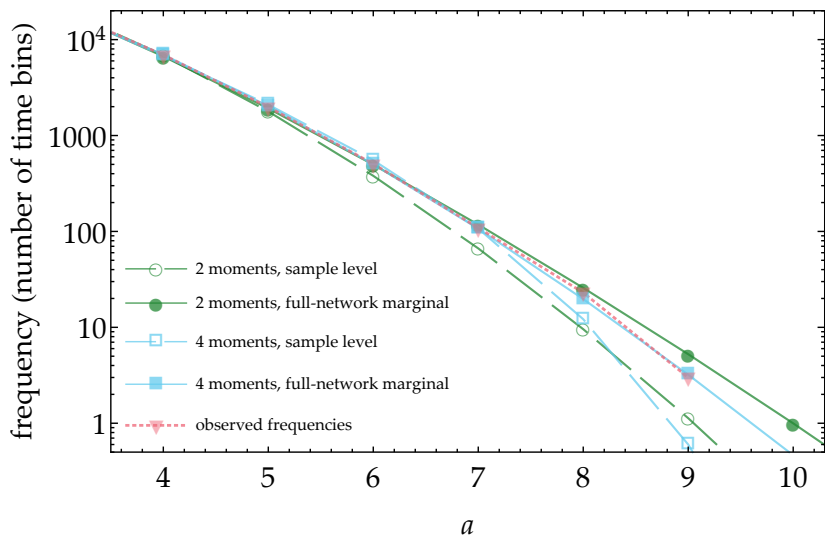
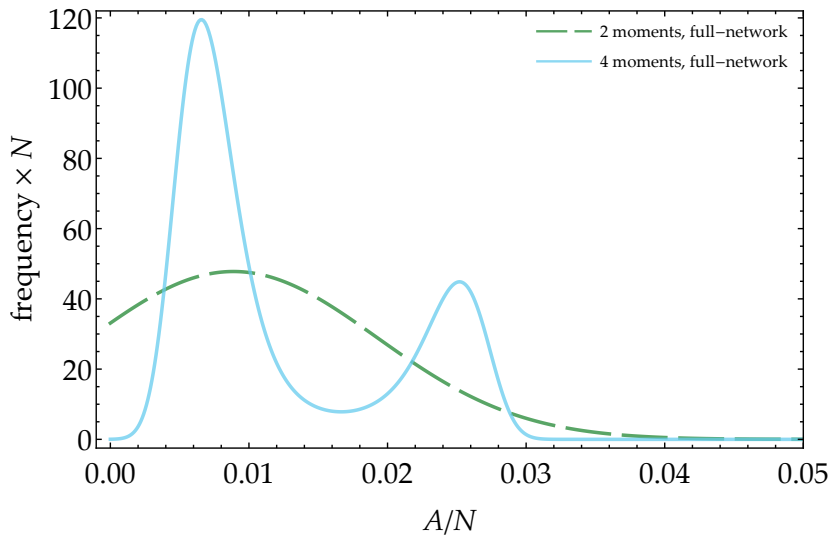


Figure 3 ***

5 The full-network size N

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

The crucial point is formula (2), on which our method is based, and which asserts the equality of the factorial moments of the full network and its sample, (or, equivalently, asserting that the power moments for the one are linear functions of the power moments for the other). This formula is only valid if our beliefs about the activity of the sample $p(a)$ and that of the full network $p(A)$ are related by a hypergeometric distribution:

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

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

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

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

$$p(A \mid N \text{ unknown}) = \sum_N P_{\text{me}}(A \mid N) p(N). \quad (5)$$

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

6 Limitations and assumptions

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

points.

The first point is that there are many possible frequency distributions which we believe, to different degrees, could be the true one that happened during the recording. The one given by our procedure is simply the one with the largest degree of belief, the mode of the belief distribution. The space of possible frequency distributions has many dimensions, however – thousands or tens of thousands. We must remember that belief distributions in high dimensions have counter-intuitive properties. For example, the mode or mean can have *atypical* features when compared with the features of most other points of the space. The mode and mean can also be very different from each other.

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

- all activities $A/N \gtrsim 5\%$ have practically zero frequencies;
- there are two regions of activity levels with high frequencies, roughly separated by a trough of lower frequencies.
- The frequencies of the region on the left ($A/N \lesssim 1.8\%$) are higher than those of the region on the right ($A/N \gtrsim 1.8\%$).

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

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

2017a):

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

where $H(F; R) := \sum_A F_A \ln(F_A/R_A)$ is the relative entropy or discrimination information (Kullback 1987; Jaynes 1963; Hobson 1969; Hobson et al. 1973), R is the reference distribution, and L a positive parameter. The approximate equality (obtained through Stirling’s approximation), where the large parentheses denote a multinomial coefficient, shows that this prior belief is proportional to the number of ways in which the distribution F can be realized in L time bins. The parameter L roughly quantifies how many time bins our data set must have to affect our initial belief. The maximum-entropy approximation is valid when L is large, but small compared to the sharpness of the constraints on F ; in our case this means $L \approx 10$, give or take an order of magnitude.

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

7 Summary and discussion

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

Thanks

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

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

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

- Cifarelli, D. M., Regazzini, E. (1982): *Some considerations about mathematical statistics teaching methodology suggested by the concept of exchangeability*. In: Koch, Spizzichino (1982), 185–205.
- Cohen, M. R., Kohn, A. (2011): *Measuring and interpreting neuronal correlations*. Nat. Neurosci. **14**⁷, 811–819. <http://marlenecohen.com/pubs/CohenKohn2011.pdf>.
- Darmois, G. (1935): *Sur les lois de probabilité à estimation exhaustive*. Comptes rendus hebdomadaires des séances de l'Académie des sciences **200**, 1265–1266.
- De Bruijn, N. G. (1961): *Asymptotic Methods in Analysis*, 2nd ed. (North-Holland, Amsterdam). First publ. 1958.
- Diaconis, P. (1992): *Sufficiency as statistical symmetry*. In: Browder (1992), 15–26. First publ. 1991 as technical report <https://statistics.stanford.edu/research/sufficiency-statistical-symmetry>.
- Diaconis, P., Freedman, D. (1981): *Partial exchangeability and sufficiency*. In: Ghosh, Roy (1981), 205–236. <http://statweb.stanford.edu/~cgates/PERSI/year.html>. Also publ. 1982 as technical report https://www.stat.berkeley.edu/~aldous/206-Exch/Papers/diaconis_freedman_PES.pdf.
- Erickson, G. J., Rychert, J. T., Smith, C. R., eds. (1998): *Maximum Entropy and Bayesian Methods*. (Springer, Dordrecht).
- Fang, S.-C., Rajasekera, J. R., Tsao, H.-S. J. (1997): *Entropy Optimization and Mathematical Programming*, reprint. (Springer, New York).
- Feller, W. (1968): *An Introduction to Probability Theory and Its Applications*. Vol. I, 3rd ed. (Wiley, New York). First publ. 1950.
- Fisher, R. A. (1922): *On the mathematical foundations of theoretical statistics*. Phil. Trans. R. Soc. Lond. A **222**, 309–368. <http://www.stats.org.uk/statistical-inference/Fisher1922.pdf>.
- Ford, K. W., ed. (1963): *Statistical Physics*. (Benjamin, New York).
- Fortini, S., Ladelli, L., Regazzini, E. (2000): *Exchangeability, predictive distributions and parametric models*. Sankhyā A **62**¹, 86–109.
- Furmańczyk, K., Niemiro, W. (1998): *Sufficiency in Bayesian models*. Applicationes Mathematicae **25**¹, 113–120.
- Ganmor, E., Segev, R., Schneidman, E. (2011): *Sparse low-order interaction network underlies a highly correlated and learnable neural population code*. Proc. Natl. Acad. Sci. (USA) **108**²³, 9679–9684. http://www.weizmann.ac.il/neurobiology/labs/schneidman/The_Schneidman_Lab/Publications.html.
- Gerstein, G. L., Perkel, D. H., Dayhoff, J. E. (1985): *Cooperative firing activity in simultaneously recorded populations of neurons: detection and measurement*. J. Neurosci. **5**⁴, 881–889.
- Gerwin, S., Macke, J. H., Bethge, M. (2010): *Bayesian inference for generalized linear models for spiking neurons*. Front. Comput. Neurosci. **4**, 12.
- Ghosh, J. K., Roy, J., eds. (1981): *Statistics: Applications and New Directions*. (Indian Statistical Institute, Calcutta).
- Grandy Jr., W. T. (1980): *Principle of maximum entropy and irreversible processes*. Phys. Rep. **62**³, 175–266.
- Granot-Atedgi, E., Tkačik, G., Segev, R., Schneidman, E. (2013): *Stimulus-dependent maximum entropy models of neural population codes*. PLoS Comput. Biol. **9**³, e1002922.
- Haken, H., ed. (1985): *Complex Systems – Operational Approaches: in Neurobiology, Physics, and Computers*. (Springer, Berlin).
- Halmos, P. R., Savage, L. J. (1949): *Application of the Radon-Nikodym theorem to the theory of sufficient statistics*. Ann. Math. Stat. **20**², 225–241.

- Hobson, A. (1969): *A new theorem of information theory*. J. Stat. Phys. **1**³, 383–391.
- Hobson, A., Cheng, B.-K. (1973): *A comparison of the Shannon and Kullback information measures*. J. Stat. Phys. **7**⁴, 301–310.
- ISO (International Organization for Standardization) (1993): *Quantities and units*, 3rd ed. International Organization for Standardization.
- (2006a): *ISO 3534-1:2006: Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability*. International Organization for Standardization.
 - (2006b): *ISO 3534-2:2006: Statistics – Vocabulary and symbols – Part 2: Applied statistics*. International Organization for Standardization.
- Jaynes, E. T. (1957a): *Information theory and statistical mechanics*. Phys. Rev. **106**⁴, 620–630. <http://bayes.wustl.edu/etj/node1.html>, see also Jaynes (1957b).
- (1957b): *Information theory and statistical mechanics. II*. Phys. Rev. **108**², 171–190. <http://bayes.wustl.edu/etj/node1.html>, see also Jaynes (1957a).
 - (1963): *Information theory and statistical mechanics*. In: Ford (1963), 181–218. Repr. in Jaynes (1989), ch. 4, 39–76. <http://bayes.wustl.edu/etj/node1.html>.
 - (1989): *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, reprint. (Kluwer, Dordrecht). Edited by R. D. Rosenkrantz. First publ. 1983.
 - (1996a): *Macroscopic prediction*. <http://bayes.wustl.edu/etj/node1.html>. First publ. in Haken (1985) pp. 254–269.
 - (1996b): *Monkeys, kangaroos, and N*. <http://bayes.wustl.edu/etj/node1.html>. First publ. 1986. (Errata: in equations (29)–(31), (33), (40), (44), (49) the commas should be replaced by gamma functions, and on p. 19 the value 0.915 should be replaced by 0.0915).
 - (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. <https://archive.org/details/XQUHIUXHIQUHIQXUIHX2>, <http://www-biba.inrialpes.fr/Jaynes/prob.html>.
- Kallenberg, O. (2005): *Probabilistic Symmetries and Invariance Principles*. (Springer, New York).
- Koch, G., Spizzichino, F., eds. (1982): *Exchangeability in Probability and Statistics*. (North-Holland, Amsterdam).
- Koopman, B. O. (1936): *On distributions admitting a sufficient statistic*. Trans. Am. Math. Soc. **39**³, 399–409.
- Kullback, S. (1987): *The Kullback-Leibler distance*. American Statistician **41**⁴, 340–341.
- Kullback, S., Leibler, R. A. (1951): *On information and sufficiency*. Ann. Math. Stat. **22**¹, 79–86.
- Lauritzen, S. L. (1974a): *Sufficiency, prediction and extreme models*. In: Barndorff-Nielsen, Blaesild, Schou (1974), 249–269. With discussion. Repr. without discussion in Lauritzen (1974b).
- (1974b): *Sufficiency, prediction and extreme models*. Scand. J. Statist. **1**³, 128–134.
 - (1988): *Extremal Families and Systems of Sufficient Statistics*. (Springer, Berlin). First publ. 1982.
 - (2007): *Sufficiency, partial exchangeability, and exponential families*. <http://www.stats.ox.ac.uk/~steffen/teaching/grad/partial.pdf>. Lecture notes.
- Lezon, T. R., Banavar, J. R., Cieplak, M., Maritan, A., Fedoroff, N. V. (2006): *Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns*. Proc. Natl. Acad. Sci. (USA) **103**⁵⁰, 19033–19038.
- Macke, J. H., Buesing, L., Cunningham, J. P., Yu, B. M., Shenoy, K. V., Sahani, M. (2011): *Empirical models of spiking in neural populations*. Adv. Neural Information Processing Systems (NIPS) **24**, 1350–1358.

- Macke, J. H., Murray, I., Latham, P. E. (2013): *Estimation bias in maximum entropy models*. Entropy **15**⁸, 3109–3129.
- Macke, J. H., 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.
- 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>.
- Mora, T., Deny, S., Marre, O. (2015): *Dynamical criticality in the collective activity of a population of retinal neurons*. Phys. Rev. Lett. **114**⁷, 078105.
- Neumann, T. (2007): *Bayesian inference featuring entropic priors*. Am. Inst. Phys. Conf. Proc. **954**, 283–292. <http://www.tilman-neumann.de/docs/BIEP.pdf>.
- Neyman, J. (1935): *Su un teorema concernente le cosiddette statistiche sufficienti*. Giorn. Ist. Ital. Att. **VI**⁴, 320–334.
- Nogales, A. G., Oyola, J. A., Pérez, P. (2000): *On conditional independence and the relationship between sufficiency and invariance under the Bayesian point of view*. Stat. Probab. Lett. **46**¹, 75–84.
- Pitman, E. J. G. (1936): *Sufficient statistics and intrinsic accuracy*. Math. Proc. Camb. Phil. Soc. **32**⁴, 567–579.
- Porta Mana, P. G. L. (2009): *On the relation between plausibility logic and the maximum-entropy principle: a numerical study*. [arXiv:0911.2197](https://arxiv.org/abs/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., Rostami, V., Torre, E., Roudi, Y. (2018): *Maximum-entropy and representative samples of neuronal activity: a dilemma*. Open Science Framework doi:10.17605/osf.io/uz29n, bioRxiv doi:10.1101/329193, arXiv:1805.09084.
- Porta Mana, P. G. L., Torre, E., Rostami, V. (2015): *Inferences from a network to a subnetwork and vice versa under an assumption of symmetry*. bioRxiv doi:10.1101/034199.
- Potts, R. B. (1953): *Note on the factorial moments of standard distributions*. Aust. J. Phys. **6**⁴, 498–499.
- Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannery, B. P. (2007): *Numerical Recipes: The Art of Scientific Computing*, 3rd ed. (Cambridge University Press, Cambridge). First publ. 1988.
- Rodríguez, C. C. (1991): *Entropic priors*. <http://omega.albany.edu:8008/>.
- Ross, S. (2010): *A First Course in Probability*, 8th ed. (Pearson, Upper Saddle River, USA). First publ. 1976.
- Rostami, V., Porta Mana, P. G. L., Grün, S., Helias, M. (2017): *Bistability, non-ergodicity, and inhibition in pairwise maximum-entropy models*. PLoS Comput. Biol. **13**¹⁰, e1005762. See

- also the slightly different version [arXiv:1605.04740](https://arxiv.org/abs/1605.04740). Data available at <https://doi.org/10.5061/dryad.n9f77>.
- Roudi, Y., Tyrcha, J., Hertz, J. (2009): *Ising model for neural data: model quality and approximate methods for extracting functional connectivity*. Phys. Rev. E **79**⁵, 051915.
- Schneidman, E., Berry II, M. J., Segev, R., Bialek, W. (2006): *Weak pairwise correlations imply strongly correlated network states in a neural population*. Nature **440**⁷⁰⁸⁷, 1007–1012. http://www.weizmann.ac.il/neurobiology/labs/schneidman/The_Schneidman_Lab/Publications.html.
- Shimazaki, H., Sadeghi, K., Ishikawa, T., Ikegaya, Y., 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>.
- 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.