Maximum-entropy and representative samples of neuronal activity: a dilemma

P.G.L. Porta Mana
piero.mana@ntnu.no

V. Rostami
vrostami@uni-koeln.de

E. Torre
torre@ibk.baug.ethz.ch

Y. Roudi
yasser.roudi@ntnu.no

22 May 2018

Abstract

The present work shows that the maximum-entropy method can be applied to a sample of neuronal recordings along two different routes: (1) apply to the sample; or (2) apply to a larger, unsampled neuronal population from which the sample is drawn, and then marginalize to the sample. These two routes give inequivalent results. The second route can be further generalized to the case where the size of the larger population is unknown. Which route should be chosen? Some arguments are presented in favour of the second. This work also presents and discusses probability formulae that relate states of knowledge about a population and its samples, and that may be useful for sampling problems in neuroscience.

1 Introduction: maximum-entropy and recordings of neuronal activity

Suppose that 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, Perkel, Dayhoff [1] 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. [2]. 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 method [3–7]. 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 cortex [8–24], and it has been subjected to mathematical and conceptual scrutiny [25–31].

'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 [cf. 32; 33]. Let the variable $s_i(t) \in \{0, 1\}$ denote the activity of the ith sampled neuron at time bin t. Collectively denote the n activities with $\boldsymbol{s}(t) := (s_1(t), \dots, s_n(t))$. The population-averaged activity at that bin is $\bar{\boldsymbol{s}}(t) := \sum_i s_i(t)/n$. If we count the number of distinct pairs of

active neurons at that bin we combinatorially find $\binom{n\bar{s}(t)}{2} \equiv n\bar{s}(t)\,[n\bar{s}(t) - 1]/2$. There can be at most $\binom{n}{2}$ simultaneously active pairs, so the population-averaged pair activity is $\overline{ss}(t) := \binom{n}{2}^{-1}\binom{n\bar{s}(t)}{2}$. With some combinatorics we see that the population-averaged activity of m-tuples of neurons is

$$\underbrace{\overline{s\cdots s}}_{m\text{ terms}}(t) = \binom{n}{m}^{-1}\binom{n\bar{s}(t)}{m}. \tag{1}$$

For brevity let us agree to simply call 'activity' the average $\bar{s}$, 'pair-activity' the average $\overline{ss}$, and so on.

2. Construct a sequence of relative-maximum-entropy distributions for the activity $\bar{s}$, using this sequence of constraints:

   - the time average of the activity: $\widehat{\bar{s}} := \sum_t \bar{s}(t)/T$;
   - the time averages of the activity and of the pair-activity $\widehat{\overline{ss}} := \sum_t \overline{ss}(t)/T$;
   - ...
   - the time averages of the activity, of the pair-activity, and so on, up to the k-activity.

   Call the resulting distributions $p_1(\bar{s}), p_2(\bar{s}), \dots, p_k(\bar{s})$. 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, $f(\bar{s})$, counted from the time bins.

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 information [34; 4; 35; 5]. 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 $m \geqslant 2$, 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 population 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 literature [27]. This problems extends also to more complex versions of the method, possibly except version that use 'hidden' neurons [36–39]. The problem is that the recorded neurons are a *sample* from a larger, unrecorded population, 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 rightfully be 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 millimetres [cf. 40]. This recording is a sample of the activity of the neuronal population under the recording device; which can amount to tens of thousands of neurons [41]. 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 population – at the very least that 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 population. Probability theory tells us how to make inferences from a sample to the larger population 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 population, we can also apply the maximum-entropy method to the larger (unrecorded) population. The constraints are the same: the time averages of the sampled data, since they constitute representative data about the larger population as well. The method thus yields a probability distribution for the larger population, 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 population, and show that its results differ from the application at the sample level. We also consider the case where the size of the larger population is unknown.

To apply the maximum-entropy method to the larger, unsampled population, it is necessary to use probability relations relevant to sampling [42; 43 parts I, VI; 44 ch. 8; 45 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 standards [46–50] but for the use of the comma ',' to denote logical conjunction. Probability notation follows Jaynes [45]. By 'probability' we mean a degree of belief which 'would be agreed by all rational men if there were any rational men' [51].

2 Probability relations between population and sample

We have already introduced the notation for the sample neurons. We introduce an analogous notation for the ν neurons constituting the larger population, but using the corresponding Greek letters: $\sigma_\iota(t)$ is the activity of the ιth neuron at time bin t, $\overline{\boldsymbol{\sigma}}(t) \coloneqq \sum_\iota \sigma_\iota(t)/\nu$ is the activity at that bin averaged over the larger population, and so on.

The probability relations between sample and larger population 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.

Probabilities refer to statements about the quantities we observe. We use the standard notation:

$$\begin{aligned} &\text{`}\Sigma_\iota = \sigma_\iota\text{' stands for 'the activity of the } \iota\text{th neuron is } \sigma_\iota\text{'},\\ &\text{`}\overline{\boldsymbol{\Sigma}} = \overline{\boldsymbol{\sigma}}\text{' stands for 'the (population-averaged) activity of the neurons is } \overline{\boldsymbol{\sigma}}\text{'},\\ &\text{`}S_i = s_i\text{' stands for 'the activity of the } i\text{th sample neuron is } s_i\text{'}, \end{aligned} \tag{2}$$

and similarly for other quantities.

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

$$\mathrm{P}(\Sigma_1 = \sigma_1, \Sigma_2 = \sigma_2, \dots, \Sigma_\nu = \sigma_\nu \mid K) \quad \text{or} \quad \mathrm{P}(\boldsymbol{\Sigma} = \boldsymbol{\sigma} \mid K), \qquad \boldsymbol{\sigma} \in \{0,1\}^\nu. \tag{3}$$

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

$$\mathrm{P}(S_1 = s_1, S_2 = s_2, \dots, S_n = s_n \mid K) \quad \text{or} \quad \mathrm{P}(\boldsymbol{S} = \boldsymbol{s} \mid K), \qquad \boldsymbol{s} \in \{0,1\}^n. \tag{4}$$

The theory of statistical sampling is covered in many excellent texts, for example Ghosh & Meeden [42] or Freedman, Pisani, & Purves [43 parts I, VI]; summaries can be found in Gelman et al. [44 ch. 8] and Jaynes [45 ch. 3].

We need to make an initial probability assignment for the state of the full population 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 population. Our state of knowledge and initial probability assignment should reflect that samples are somehow representative of the whole population.

In this state of knowledge, denoted I, we know that the neurons in the population 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* [52; 42 ch. 1]. We stress that this probability assignment is just an expression of the symmetry of our *ignorance* about the state of the population, not an expression of some biologic 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 population-average:

$$\mathrm{P}(\boldsymbol{\Sigma} = \boldsymbol{\sigma} \mid I) \equiv \sum_{\bar{\boldsymbol{\sigma}}} \mathrm{P}(\boldsymbol{\Sigma} = \boldsymbol{\sigma} \mid \bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\sigma}}, I)\, \mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\sigma}} \mid I) = \binom{\nu}{\nu\bar{\boldsymbol{\sigma}}}^{-1} \mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\sigma}} \mid I). \tag{5}$$

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 $\binom{\nu}{\nu\bar{\boldsymbol{\sigma}}}$ activity vectors with $\nu\bar{\boldsymbol{\sigma}}$ 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 [53], Kendall [54], Ericson [55], Diaconis & Freedman [56; 57], Heath & Sudderth [58].

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

$$\mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}} \mid \bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\sigma}}, I) = \binom{n}{n\bar{\boldsymbol{s}}}\binom{\nu - n}{\nu\bar{\boldsymbol{\sigma}} - n\bar{\boldsymbol{s}}}\binom{\nu}{\nu\bar{\boldsymbol{\sigma}}}^{-1} =: \Pi(\bar{\boldsymbol{s}} \mid \bar{\boldsymbol{\sigma}}), \tag{6}$$

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 problems [45 ch. 3; 59 § 4.8.3; 60 § II.6].

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

$$\mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}} \mid I) = \sum_{\bar{\boldsymbol{\sigma}}} \mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}} \mid \bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\sigma}}, I)\, \mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\sigma}} \mid I) = \sum_{\bar{\boldsymbol{\sigma}}} \Pi(\bar{\boldsymbol{s}} \mid \bar{\boldsymbol{\sigma}})\, \mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\sigma}} \mid I). \tag{7}$$

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 population-averaged activity:

$$\mathrm{P}(\boldsymbol{S} = \boldsymbol{s} \mid I) = \binom{n}{n\bar{\boldsymbol{s}}}^{-1} \mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}} \mid I). \tag{8}$$

This is intuitively clear: our initial symmetric ignorance should also apply to the sample. The distribution for the sample (7) indeed satisfies the same representation theorem (5) as the distribution for the full population.

The conditional probability $\mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}} \mid \bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\sigma}}, I) \equiv \Pi(\bar{\boldsymbol{s}} \mid \bar{\boldsymbol{\sigma}})$, besides relating the distributions for the population and sample activities via marginalization, also allows us to express the expectation value of any function of the sample activity, $g(\bar{\boldsymbol{s}})$, in terms of the distribution for the full population, as follows:

$$\mathrm{E}(g \mid I) \equiv \sum_{\bar{\boldsymbol{s}}} g(\bar{\boldsymbol{s}})\, \mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}} \mid I) = \sum_{\bar{\boldsymbol{s}}} g(\bar{\boldsymbol{s}}) \sum_{\bar{\boldsymbol{\sigma}}} \Pi(\bar{\boldsymbol{s}} \mid \bar{\boldsymbol{\sigma}})\, \mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\sigma}} \mid I) = \sum_{\bar{\boldsymbol{\sigma}}} \Big[\sum_{\bar{\boldsymbol{s}}} g(\bar{\boldsymbol{s}})\, \Pi(\bar{\boldsymbol{s}} \mid \bar{\boldsymbol{\sigma}})\Big]\, \mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\sigma}} \mid I), \tag{9}$$

where the second step uses eq. (7). The last expression shows that the expectation of the function $g(\bar{s})$ is equal to the expectation of the function $g^*(\bar{\sigma}) := \sum_{\bar{s}} g(\bar{s})\, \Pi(\bar{s}|\, \bar{\sigma})$.

The final expression in eq. (9) is important for our maximum-entropy application: the requirement that the function g, defined for the sample, have a value c obtained from observed data, *translates into a linear constraint for the distribution of the full population*:

$$c = \mathrm{E}(g|\, I) \equiv \sum_{\bar{\sigma}} \Bigl[\sum_{\bar{s}} g(\bar{s}) \Pi(\bar{s}|\, \bar{\sigma})\Bigr] \mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\sigma}|\, I). \tag{10}$$

In particular, when the function g is the m-activity of the sample, $g(\bar{s}) = \overline{s \dots s} \equiv \binom{n\bar{s}}{m} / \binom{n}{m}$, we find

$$\mathrm{E}(\underbrace{\overline{s \cdots s}}_{m \text{ factors}}|\, I) \equiv \sum_{\bar{s}} \binom{n}{m}^{-1} \binom{n\bar{s}}{m} \mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I) = \binom{\nu}{m}^{-1} \sum_{\bar{\sigma}} \binom{\nu\bar{\sigma}}{m} \mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\sigma}|\, I) \equiv \mathrm{E}(\underbrace{\overline{\sigma \cdots \sigma}}_{m \text{ factors}}|\, I), \tag{11}$$

that is, *the expected values of the m-activities of the sample and of the full population are equal*. The proof of the middle equality uses the expression for the mth factorial moment of the hypergeometric distribution and can be found in Potts [61]. Similar relations can be found for the raw moments $\mathrm{E}(\bar{s}^m)$ and $\mathrm{E}(\bar{\sigma}^m)$, which can be written in terms of the product expectations using eq. (1).

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 population to have the same value.

These expectation equalities between sample and full population 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 $\Pi(\bar{s}|\, \bar{\sigma})$ plays an important role in this formalization. A look at its plot, fig. 1, reveals that it is a sort of 'fuzzy identity transformation', or fuzzy Kronecker delta, between the $\bar{\sigma}$-space $\{0, \dots, \nu\}$ and $\bar{s}$-space $\{0, \dots, n\}$. From eq. (8) we thus have that

$$\mathrm{P}(\bar{\boldsymbol{S}} = a|\, I) \approx \mathrm{P}(\bar{\boldsymbol{\Sigma}} = a|\, I), \qquad \mathrm{E}[g(\bar{s})|\, I] \approx \mathrm{E}[g(\bar{\sigma})|\, I], \tag{12}$$

where g 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 population*

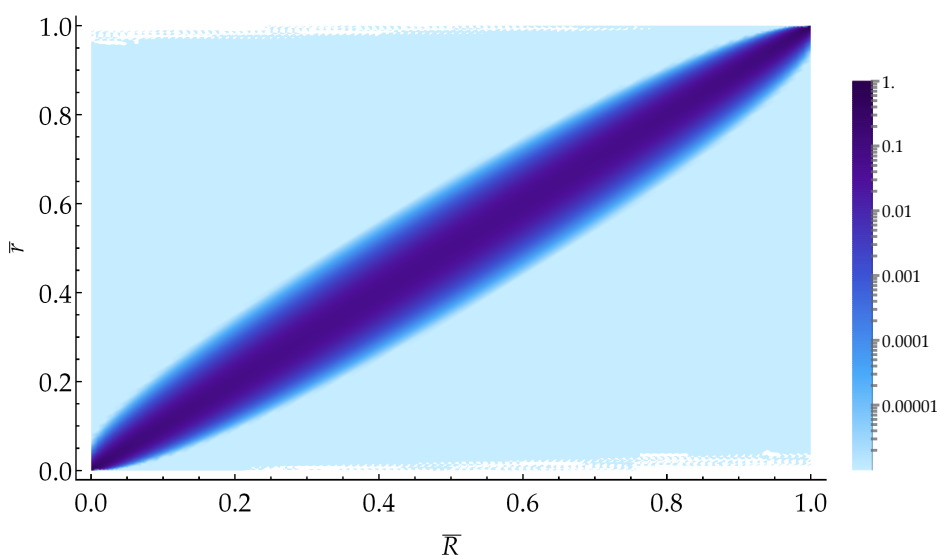

Figure 1: Log-density plot of the hypergeometric distribution $\Pi(\bar{s}|\, \bar{\sigma}) := \binom{n}{n\bar{s}} \binom{\nu-n}{\nu\bar{\sigma}-n\bar{s}} \binom{\nu}{\nu\bar{\sigma}}^{-1}$ for $\nu = 5000$, $n = 200$. (Band artifacts may appear in the colourbar depending on your PDF viewer.)

and about other samples, and vice versa. When $n = \nu$, $\Pi(\bar{s}|\,\bar{\sigma})$ becomes the identity matrix and the approximate equalities above become exact – of course, since we have sampled the full population.

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 population average $\bar{\sigma}$ is bimodal, for example, the bimodality can be lost in the distribution for the sample average $\bar{s}$, owing to the coarsening effect of $\Pi(\bar{s}|\,\bar{\sigma})$.

3 Maximum-entropy: sample level vs full-population 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 population. Let us use such constraints with the maximum-entropy method. Suppose that we want to constrain m functions of the sample activity, vectorially written $\boldsymbol{g} := (g_1, \dots, g_m)$, to m values $\boldsymbol{c} := (c_1, \dots, c_m)$. These functions are typically k-activities $\overline{\boldsymbol{s}\dots\boldsymbol{s}}$, and the values are typically the time averages of the observed sample, as discussed in § 1: $\boldsymbol{c} = \sum_t \boldsymbol{g}[\bar{s}(t)]/T$.

Let us apply the relative-maximum-entropy method [6; 7] directly to sampled neurons; denote this approach by I_{s}. Then we apply the method to the full population of neurons, most of which are unsampled; denote this approach by I_{p}.

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

$$\mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I_{\mathrm{s}}) = \frac{1}{z(\boldsymbol{l})}\binom{n}{n\bar{s}}\exp[\boldsymbol{l}^{\intercal}\boldsymbol{g}(\bar{s})] \tag{13}$$

where $z(\boldsymbol{l})$ is a normalization constant. The binomial in front of the exponential appears because we must account for the multiplicity by which the population-average activity $\bar{s}$ can be realized: $\bar{s} = 0$ can be realized in only one way (all neurons inactive), $\bar{s} = 1/n$ 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 mechanics [62 ch. 16]. The m Lagrange multipliers $\boldsymbol{l} := (l_1, \dots, l_m)$ must satisfy the m constraint equations

$$c_k = \mathrm{E}(g_k|\, I_{\mathrm{s}}) \equiv \frac{1}{z(\boldsymbol{l})}\sum_{\bar{s}} g_k(\bar{s})\binom{n}{n\bar{s}}\exp[\boldsymbol{l}^{\intercal}\boldsymbol{g}(\bar{s})], \qquad k = 1, \dots, m. \tag{14}$$

Applied to the full population, using the constraint expression (10) derived in the previous section, the method yields the distribution for the full-population activity

$$\mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\sigma}|\, I_{\mathrm{p}}) = \frac{1}{\zeta(\boldsymbol{\lambda})}\binom{\nu}{\nu\bar{\sigma}}\exp\Big[\boldsymbol{\lambda}^{\intercal}\sum_{\bar{s}}\boldsymbol{g}(\bar{s})\Pi(\bar{s}|\,\bar{\sigma})\Big]. \tag{15}$$

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

$$c_k = \mathrm{E}(g_k|\, I_{\mathrm{p}}) \equiv \frac{1}{\zeta(\boldsymbol{\lambda})}\sum_{\bar{s}}\sum_{\bar{\sigma}} g_k(\bar{s})\Pi(\bar{s}|\,\bar{\sigma})\binom{\nu}{\nu\bar{\sigma}}\exp\Big[\boldsymbol{\lambda}^{\intercal}\sum_{\bar{s}}\boldsymbol{g}(\bar{s})\Pi(\bar{s}|\,\bar{\sigma})\Big], \qquad k = 1, \dots, m. \tag{16}$$

We obtain the distribution for the sample activity by marginalization, using eq. (8):

$$\mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I_{\mathrm{p}}) = \frac{1}{\zeta(\boldsymbol{\lambda})}\sum_{\bar{\sigma}}\Pi(\bar{s}|\,\bar{\sigma})\binom{\nu}{\nu\bar{\sigma}}\exp\Big[\boldsymbol{\lambda}^{\intercal}\sum_{\bar{s}}\boldsymbol{g}(\bar{s})\Pi(\bar{s}|\,\bar{\sigma})\Big]. \tag{17}$$

The distributions for the sample activity, eqs (17) and (13), 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 $\boldsymbol{\lambda}$ and $\boldsymbol{l}$ to satisfy the constraint equations (16) and (14), and in addition also the n equations $\mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I_{\mathrm{p}}) = \mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I_{\mathrm{s}})$, $\bar{s} = 1/n, \dots, 1$ (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-population level are inequivalent*. They lead to numerically different distributions for the sample activity $\boldsymbol{s}$.

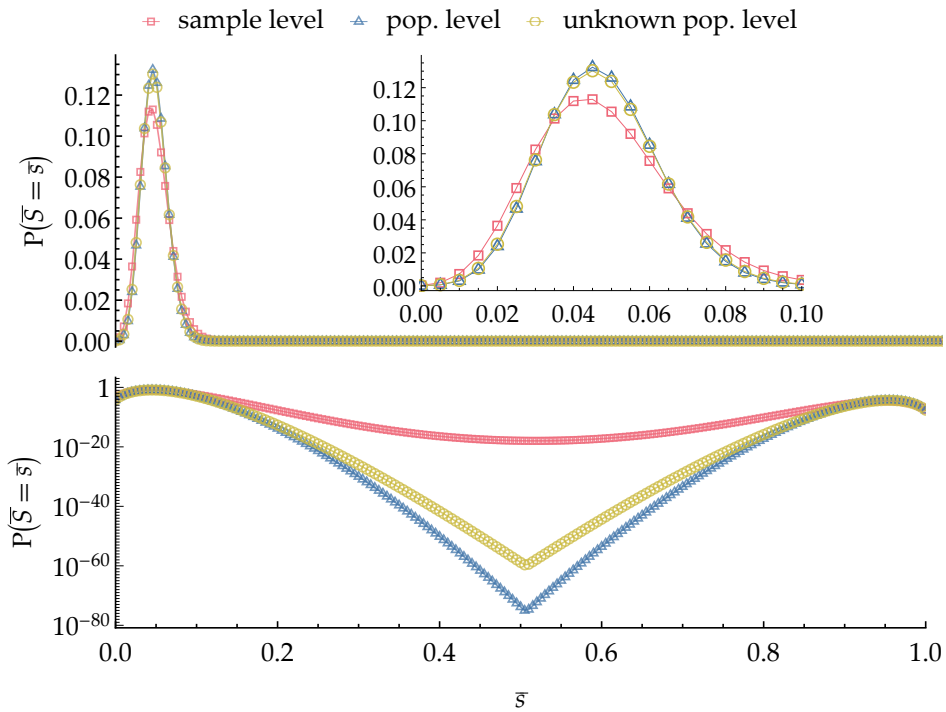

Figure 2: Linear and log-plots of $\mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}})$ for a sample of $n = 200$ and constraints as in eq. (18), constructed by: **red squares:** maximum-entropy at the sample level, eq. (13); **blue triangles:** maximum-entropy at the population level, eq. (17) with $\nu = 10\,000$, followed by sample marginalization; **yellow circles:** maximum-entropy at the population level with unknown population size, eq. (19), according to the distribution (20) for the population.

The distribution obtained at the sample level will show different features from the one obtained at the population level, like displaced or additional modes or particular tail behaviour. We show an example of this discrepancy in fig. 2, for $\nu = 10\,000$, $n = 200$, and the two constraints

$$\mathrm{E}(\bar{\boldsymbol{s}}) = 0.0478, \qquad \mathrm{E}(\bar{\boldsymbol{s}}\bar{\boldsymbol{s}}) = 0.00257, \tag{18}$$

which come from the actual recording of circa 200 neurons from macaque motor cortex [31]. The distribution obtained at the population 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 ν of the larger population to be known. This is rarely the case, however. We usually are uncertain about ν and can only guess its order of magnitude. In such a state of knowledge I_{u} our ignorance about the possible value of ν is expressed by a probability distribution $\mathrm{P}(N = \nu \mid I_{\mathrm{u}}) = h(\nu)$, and the marginal distribution for the sample activity (17) is modified, by the law of total probability, to

$$\mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}} \mid I_{\mathrm{u}}) = \sum_{\nu} \mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}} \mid N = \nu, I_{\mathrm{u}})\, \mathrm{P}(N = \nu \mid I_{\mathrm{u}}) = \sum_{\nu} \Bigg\{ \frac{1}{\zeta(\boldsymbol{\lambda}_\nu)} \sum_{\bar{\boldsymbol{\sigma}}} \Pi_\nu(\bar{\boldsymbol{s}} \mid \bar{\boldsymbol{\sigma}}) \binom{\nu}{\nu\bar{\boldsymbol{\sigma}}} \exp\Big[\boldsymbol{\lambda}_\nu{}^{\intercal} \sum_{\bar{\boldsymbol{s}}} \boldsymbol{g}(\bar{\boldsymbol{s}}) \Pi_\nu(\bar{\boldsymbol{s}} \mid \bar{\boldsymbol{\sigma}})\Big] \Bigg\} h(\nu), \tag{19}$$

where the Lagrange multipliers $\boldsymbol{\lambda}_\nu$ and the summation range for $\bar{\boldsymbol{\sigma}}$ depend on ν.

As a proof of concept, fig. 2 also shows such a distribution (yellow circles) for the same constraints as above, and a probability distribution for ν inspired by Jeffreys [63 § 4.8]:

$$h(\nu) \propto 1/\nu, \qquad \nu \in \{1\,000,\ 2\,000,\ \ldots,\ 10\,000\}. \tag{20}$$

4 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 population from which the data are sampled, eqs (15)–(17). This application leads to results that differ from the standard application which only considers the sample in isolation, eqs (13)–(14). We gave a numerical example of this difference. We have also shown how to extend the new application when the size of the larger population is unknown, eq. (19).

The latter formula, in particular, shows that the standard way of applying maximum-entropy implicitly assumes that *no* larger population exists beyond the recorded sample of neurons. One could in fact object to the application at the population level, and say that the traditional way of applying maximum-entropy, eq. (13), yields different results because it does not make assumptions about the size ν of a possibly existing larger population. Such a state of uncertainty, however, is correctly formalized according to the laws of probability by introducing a probability distribution for ν, and is expressed by eq. (19). This expression cannot generally be equal to (13) unless the distribution for ν gives unit probability to $\nu = n$; that is, unless the sample *is* the full population, and no larger population 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 population 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 population levels appears in the use of Boltzmann machines with hidden units [64], 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 state [65]. 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 state. The two distributions differ (even though the final *physical* state is obviously exactly the same [66 § 4]), 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-population 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' case [12; 13; 27], 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 neurons [36–39]?

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 population and its samples, presented in § 2. This kind of formulae is essential in neuroscience, where we try to understand properties of extended brain regions from partial observations. The formulae presented here reflect a simple, symmetric state of ignorance. More work is needed [cf. 67] to extend these formulae to account for finer knowledge of the cerebral cortex and its network properties.

Acknowledgements

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

References

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

[2] G. P. Moore, D. H. Perkel, J. P. Segundo: *Statistical analysis and functional interpretation of neuronal spike data*. Annu. Rev. Physiol. **28** (1966), 493–522.

[3] E. T. Jaynes: *Information theory and statistical mechanics*. Phys. Rev. **106**[4] (1957), 620–630. http://bayes.wustl.edu/etj/node1.html, see also ref. [68].

[4] E. T. Jaynes: *Information theory and statistical mechanics*. In: Ford [69] (1963), 181–218. Repr. in ref. [70], ch. 4, 39–76. http://bayes.wustl.edu/etj/node1.html.

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

[6] D. S. Sivia: *Data Analysis: A Bayesian Tutorial*, 2nd ed. Oxford University Press, Oxford (2006). Written with J. Skilling. First publ. 1996.

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

[8] D. J. C. MacKay: *Maximum entropy connections: neural networks*. In: Grandy, Schick [71] (1991), 237–244.

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

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

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

[12] E. Schneidman, M. J. Berry II, R. Segev, W. Bialek: *Weak pairwise correlations imply strongly correlated network states in a neural population*. Nature **440**[7087] (2006), 1007–1012. http://www.weizmann.ac.il/neurobiology/labs/schneidman/The_Schneidman_Lab/Publications.html.

[13] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, E. J. Chichilnisky: *The structure of multi-neuron firing patterns in primate retina*. J. Neurosci. **26**[32] (2006), 8254–8266. See also correction ref. [72].

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

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

[16] G. Tkačik, E. Schneidman, M. J. Berry II, W. Bialek: *Spin glass models for a network of real neurons*. (2009). arXiv:0912.5409.

[17] S. Gerwinn, P. Berens, M. Bethge: *A joint maximum-entropy model for binary neural population patterns and continuous signals*. Adv. Neural Information Processing Systems (NIPS proceedings) **22** (2009), 620–628.

[18] J. H. Macke, M. Opper, M. Bethge: *Common input explains higher-order correlations and entropy in a simple model of neural population activity*. Phys. Rev. Lett. **106**[20] (2011), 208102.

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

[20] E. Ganmor, R. Segev, E. Schneidman: *Sparse low-order interaction network underlies a highly correlated and learnable neural population code*. Proc. Natl. Acad. Sci. (USA) **108**[23] (2011), 9679–9684. http://www.weizmann.ac.il/neurobiology/labs/schneidman/The_Schneidman_Lab/Publications.html.

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

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

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

[24] H. Shimazaki, K. Sadeghi, T. Ishikawa, Y. Ikegaya, T. Toyoizumi: *Simultaneous silence organizes structured higher-order interactions in neural populations*. Sci. Rep. **5** (2015), 9821.

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

[26] Y. Roudi, E. Aurell, J. A. Hertz: *Statistical physics of pairwise probability models*. Front. Comput. Neurosci. **3** (2009), 22.

[27] Y. Roudi, S. Nirenberg, P. E. Latham: *Pairwise maximum entropy models for studying large biological systems: when they can work and when they can't*. PLoS Comput. Biol. **5**[5] (2009), e1000380.

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

[29] A. K. Barreiro, E. T. Shea-Brown, F. M. Rieke, J. Gjorgjieva: *When are microcircuits well-modeled by maximum entropy methods?* BMC Neurosci. **11**[Suppl. 1] (2010), P65.

[30] J. H. Macke, I. Murray, P. E. Latham: *Estimation bias in maximum entropy models*. Entropy **15**[8] (2013), 3109–3129. http://www.gatsby.ucl.ac.uk/~pel/papers/maxentbias.pdf.

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

[32] E. R. Caianiello: *Outline of a theory of thought-processes and thinking machines*. J. Theor. Biol. **1**[2] (1961), 204–235.

[33] E. R. Caianiello: *Neuronic equations revisited and completely solved*. In: Palm, Aertsen [73] (1986), 147–160.

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

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

[36] P. Smolensky: *Information processing in dynamical systems: foundations of harmony theory*. In: Rumelhart, McClelland, PDP Research Group [74] (1986), ch. 6, 194–281.

[37] J. E. Kulkarni, L. Paninski: *Common-input models for multiple neural spike-train data*. Netw. **18**[4] (2007), 375–407.

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

[39] B. Dunn, C. Battistin: *The appropriateness of ignorance in the inverse kinetic Ising model*. J. Phys. A **50**[12] (2017), 124002.

[40] A. Berényi, Z. Somogyvári, A. J. Nagy, L. Roux, J. D. Long, S. Fujisawa, E. Stark, A. Leonardo et al.: *Large-scale, high-density (up to 512 channels) recording of local circuits in behaving animals*. J. Neurophysiol. **111**[5] (2014), 1132–1149. http://www.buzsakilab.com/content/PDFs/Berenyi2013.pdf.

[41] M. Abeles: *Corticonics: Neural circuits of the cerebral cortex*. Cambridge University Press, Cambridge (1991).

[42] M. Ghosh, G. Meeden: *Bayesian Methods for Finite Population Sampling*, reprint. Springer, Dordrecht (1997).

[43] D. Freedman, R. Pisani, R. Purves: *Statistics*, 4th ed. Norton, London (2007). First publ. 1978.

[44] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin: *Bayesian Data Analysis*, 3rd ed. Chapman & Hall/CRC, Boca Raton, USA (2014). First publ. 1995.

[45] E. T. Jaynes: *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge (2003). Ed. by G. Larry Bretthorst. First publ. 1994. https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html, http://omega.albany.edu:8008/JaynesBook.html.

[46] ISO: *Quantities and units*, 3rd ed. International Organization for Standardization. Geneva (1993).

[47] IEEE: *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. New York (1993).

[48] *Guide for the Use of the International System of Units (SI): NIST special publication 811, 1995 edition*. National Institute of Standards and Technology. Washington, D.C. (1995). http://physics.nist.gov/cuu/Uncertainty/bibliography.html.

[49] ISO: *ISO 3534-1:2006: Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability*. International Organization for Standardization. Geneva (2006).

[50] ISO: *ISO 3534-2:2006: Statistics – Vocabulary and symbols – Part 2: Applied statistics*. International Organization for Standardization. Geneva (2006).

[51] I. J. Good: *How to estimate probabilities*. J. Inst. Maths. Applics **2**[4] (1966), 364–383.

[52] W. A. Ericson: *Subjective Bayesian models in sampling finite populations*. J. Roy. Stat. Soc. B **31**[2] (1969), 195–224. http://www.stat.cmu.edu/~brian/905-2008/papers/Ericson-JRSSB-1969.pdf. See also discussion ref. [75].

[53] B. de Finetti: *La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista*. In: de Finetti [76] (1959), 1–115. Transl. in ref. [77], ch. 9, pp. 147–227.

[54] D. G. Kendall: *On finite and infinite sequences of exchangeable events*. Studia Sci. Math. Hung. **2** (1967), 319–327.

[55] W. A. Ericson: *A Bayesian approach to two-stage sampling*. Tech. rep. AFFDL-TR-75-145. University of Michigan, Ann Arbor, USA (1976). http://hdl.handle.net/2027.42/4819.

[56] P. Diaconis: *Finite forms of de Finetti's theorem on exchangeability*. Synthese **36**^{2} (1977), 271–281. http://statweb.stanford.edu/~cgates/PERSI/year.html.

[57] P. Diaconis, D. Freedman: *Finite exchangeable sequences*. Ann. Prob. **8**^{4} (1980), 745–764.

[58] D. Heath, W. Sudderth: *De Finetti's theorem on exchangeable variables*. American Statistician **30**^{4} (1976), 188–189.

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

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

[61] R. B. Potts: *Note on the factorial moments of standard distributions*. Aust. J. Phys. **6**^{4} (1953), 498–499.

[62] H. B. Callen: *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed. Wiley, New York (1985). First publ. 1960.

[63] H. Jeffreys: *Theory of Probability*, 3rd ed. Oxford University Press, London (2003). First publ. 1939.

[64] N. Le Roux, Y. Bengio: *Representational power of restricted Boltzmann machines and deep belief networks*. Neural Comp. **20**^{6} (2008), 1631–1649.

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

[66] E. T. Jaynes: *Inferential scattering*. (1993). http://bayes.wustl.edu/etj/node1.html. Extensively rewritten version of a paper first publ. 1985 in ref. [78], pp. 377–398.

[67] A. Levina, V. Priesemann: *Subsampling scaling*. Nat. Comm. **8** (2017), 15140.

[68] E. T. Jaynes: *Information theory and statistical mechanics. II*. Phys. Rev. **108**^{2} (1957), 171–190. http://bayes.wustl.edu/etj/node1.html, see also ref. [3].

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

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

[71] W. T. Grandy Jr., L. H. Schick, eds.: *Maximum Entropy and Bayesian Methods: Laramie, Wyoming, 1990*. Kluwer, Dordrecht (1991).

[72] J. Shlens, G. D. Field, J. L. Gauthier, M. I. Grivich, D. Petrusca, A. Sher, A. M. Litke, E. J. Chichilnisky: *Correction, the structure of multi-neuron firing patterns in primate retina*. J. Neurosci. **28**^{5} (2008), 1246. See ref. [13].

[73] G. Palm, A. Aertsen, eds.: *Brain Theory*. Springer, Berlin (1986).

[74] D. E. Rumelhart, J. L. McClelland, PDP Research Group, eds.: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations*, 12th printing. MIT Press, Cambridge, USA (1999).

[75] M. R. Sampford, A. Scott, M. Stone, D. V. Lindley, T. M. F. Smith, D. F. Kerridge, V. P. Godambe, L. Kish et al.: *Discussion on professor Ericson's paper*. J. Roy. Stat. Soc. B **31**^{2} (1969), 224–233. http://www.stat.cmu.edu/~brian/905-2008/papers/Ericson-JRSSB-1969.pdf. See ref. [79].

[76] B. de Finetti, ed.: *Induzione e statistica*, reprint. Springer, Berlin (2011). First publ. 1959.

[77] B. de Finetti: *Probability, Induction and Statistics: The art of guessing*. Wiley, London (1972).

[78] C. R. Smith, W. T. Grandy Jr., eds.: *Maximum-Entropy and Bayesian Methods in Inverse Problems*. D. Reidel, Dordrecht (1985).

[79] W. A. Ericson: *A note on the posterior mean of a population mean*. J. Roy. Stat. Soc. B **31**^{2} (1969), 332–334.