Representative samples and maximum-entropy distributions: a dilemma

P.G.L. Porta Mana
piero.mana@ntnu.org

V. Rostami
v.rostami@fz-juelich.de

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

Y. Roudi
yasser.roudi@ntnu.no

Abstract

to be rewritten This note shows that the maximum-entropy method can be applied to a representative sample from a neuronal population along two different routes: (1) apply to the sample; or (2) apply to the population and marginalize to the sample. These two routes give inequivalent results. Which route should be chosen? Some arguments are presented in favour of the second. The note also touches upon probability formulae of representative sampling and discusses their possible meanings, a discussion that may be useful for sampling problems in neuroscience.

1 Introduction: maximum-entropy and sampling in neuroscience

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 e.g. 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 [25–30].

'Cooperativity' can quantified and characterized with maximum-entropy method 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. 31; 32]. 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 $\boldsymbol{s}(t) \coloneqq (s_1(t), \dots, s_n(t))$. The population-averaged activity at that bin is $\bar{s}(t) \coloneqq \sum_i s_i(t)/n$. 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

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

can be at most $\binom{n}{2}$ simultaneously active pairs, so the population-averaged pair activity is $\overline{\bar{s}\bar{s}}(t) \coloneqq \binom{n}{2}^{-1}\binom{n\bar{s}(t)}{2}$. With some combinatorics we see that the population-averaged activity of m-tuples of neurons is

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

For brevity let us agree to simply call 'activity' the average $\bar{s}$, 'pair-activity' the average $\overline{\bar{s}\bar{s}}$, 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: $\widehat{\bar{s}} \coloneqq \sum_t \bar{s}(t)/T$;
   - the time averages of the activity and of the pair-activity $\widehat{\overline{\bar{s}\bar{s}}} \coloneqq \sum_t \overline{\bar{s}\bar{s}}(t)/T$;
   - ...
   - the time averages of the activity, of the pair-activity, etc. 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 [33; 4; 34; 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. 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. 35]. 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 [36]. 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 paper 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 [37; 38 parts I, VI; 39 ch. 8; 40 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 this note follow ISO and ANSI standards [41–45] but for the use of the comma ',' to denote logical conjunction. Probability notation follows Jaynes [40]. By 'probability' we mean a degree of belief which 'would be agreed by all rational men if there were any rational men' [46].

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, $\overline{\sigma}(t) \coloneqq \sum_\iota \sigma_\iota(t)/\nu$ is the activity at that bin averaged over the larger population, etc..

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:

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

and similarly for other quantities.

If K denotes our state of knowledge, i.e. 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 [37] or Freedman, Pisani, & Purves [38 parts I, VI]; summaries can be found in Gelman et al. [39 ch. 8] and Jaynes [40 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* [47; 37 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:

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

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 $\binom{\nu}{\nu\bar{\sigma}}$ activity vectors with $\nu\bar{\sigma}$ 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 [48], Kendall [49], Ericson [50], Diaconis & Freedman [51; 52], Heath & Sudderth [53].

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

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

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 [40 ch. 3; 54 § 4.8.3; 55 § II.6].

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

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

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:

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

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 $\mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, \bar{\boldsymbol{\Sigma}} = \bar{\sigma}, I) \equiv \mathit{\Pi}(\bar{s}|\, \bar{\sigma})$, 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, $g(\bar{s})$, in terms of the distribution for the full population, as follows:

$$\mathrm{E}(g|\, I) \equiv \sum_{\bar{s}} g(\bar{s})\, \mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I) = \sum_{\bar{s}} g(\bar{s}) \sum_{\bar{\sigma}} \mathit{\Pi}(\bar{s}|\, \bar{\sigma})\, \mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\sigma}|\, I) =$$

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

where the second step uses eq. (7). The last expression shows that the expectation of the function g, whose argument is the sample activity $\bar{s}$, is equal to the expectation of the function g^*, whose argument is the full-population activity $\overline{\sigma}$, defined as $g^*(\overline{\sigma}) := \sum_{\bar{s}} g(\bar{s})\, \mathit{\Pi}(\bar{s}|\, \overline{\sigma})$.

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*:

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

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

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

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 [56]. Similar relations can be found for the raw moments $\mathrm{E}(\bar{s}^m)$ and $\mathrm{E}(\overline{\sigma}^m)$, 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 $\mathit{\Pi}(\bar{s}|\, \overline{\sigma})$ plays an important part in this formalization. A look at its plot, fig. 1, reveals that it is a sort of 'fuzzy identity matrix', or fuzzy Kronecker delta, between the $\overline{\sigma}$-space $\{0, \dots, \nu\}$ and $\bar{s}$-space $\{0, \dots, n\}$. We thus have that

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

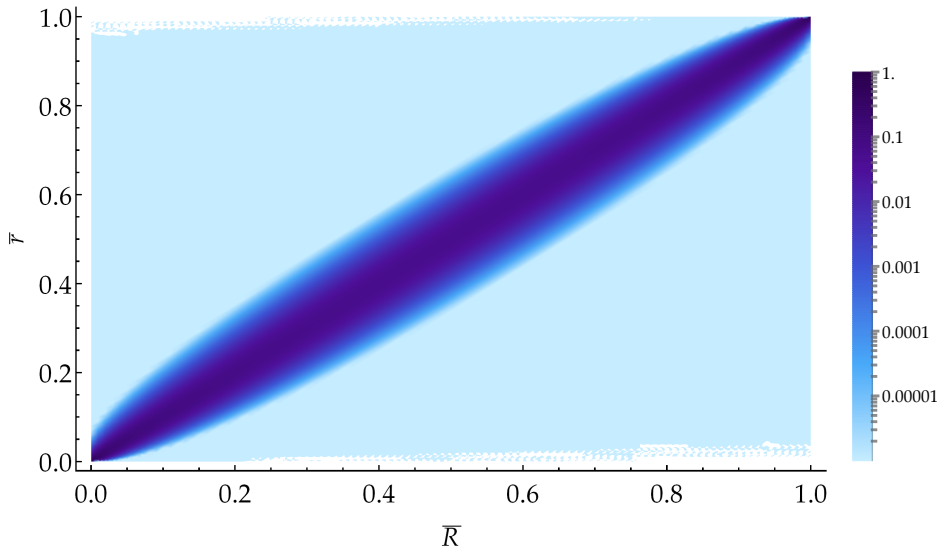

Figure 1: Log-density plot of the hypergeometric distribution $\mathit{\Pi}(\bar{s}|\, \overline{\sigma}) := \binom{n}{n\bar{s}} \binom{\nu-n}{\nu\overline{\sigma}-n\bar{s}} \binom{\nu}{\nu\overline{\sigma}}^{-1}$ for $\nu = 5000$, $n = 200$. (Band artifacts may appear in the colourbar depending on your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

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 and about other samples*, and vice versa. When $n = \nu$, $\mathit{\Pi}(\bar{s}|\,\bar{\sigma})$ 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 $\mathit{\Pi}(\bar{s}|\,\bar{\sigma})$.

3 Maximum-entropy at the population level with constraints at the sample level

The probability formulae (5)–(6) are constraints on our initial probability assignment, but do not determine it numerically. The probability $\mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\sigma}|\, I)$ for the population average needs to be numerically specified, and by marginalization (7) it will determine that of the sample average, $\mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I)$. If we numerically specify the latter, the former is not completely specified, because eq. (7) linearly constrains $\nu + 1$ unknowns by only $n + 1$ equations.

We may want to specify the probability by enforcing the sample expectations of several functions to have specific values, for example $\mathrm{E}(\bar{\boldsymbol{S}}) = c_1$, $\mathrm{E}(\overline{\boldsymbol{S}\boldsymbol{S}}) = c_2$. This is still an underdetermined problem: several distributions can have the same desired expectations, as clear from eqs (11).

The maximum-entropy method is brought into play to solve this indeterminacy. It selects one distribution, purported to be 'maximally noncommittal', among those that have the desired expectations. But here's a dilemma: the expectation formulae (9) allow us to apply the method to find the probability of the population $\mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\sigma}|\, I)$, or of the sample $\mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I)$. *The two applications, however, are inequivalent*. They lead to numerically different distributions for the sample average $\mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I)$.

Suppose we want to constrain the sample expectations of a vector function $\boldsymbol{g} = (g_1, \dots, g_m)$ to the vector values $\boldsymbol{c} = (c_1, \dots, c_m)$, that is, $\mathrm{E}[\boldsymbol{g}(\bar{\boldsymbol{S}})] = \boldsymbol{c}$. Application of maximum-entropy [6; 7] at the population level, denoted by I_p, gives

$$\mathrm{P}(\bar{\boldsymbol{\Sigma}} = \bar{\sigma}|\, I_\mathrm{p}) = Z \binom{\nu}{\nu\bar{\sigma}} \exp\Bigl[\boldsymbol{\lambda}^\intercal \sum_{n\bar{s}=0}^{n} \boldsymbol{g}(\bar{s})\mathit{\Pi}(\bar{s}|\,\bar{\sigma})\Bigr], \tag{13}$$

and then by marginalization (8)

$$\mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I_\mathrm{p}) = Z \sum_{\nu\bar{\sigma}=0}^{\nu} \mathit{\Pi}(\bar{s}|\,\bar{\sigma}) \binom{\nu}{\nu\bar{\sigma}} \exp\Bigl[\boldsymbol{\lambda}^\intercal \sum_{n\bar{s}=0}^{n} \boldsymbol{g}(\bar{s})\mathit{\Pi}(\bar{s}|\,\bar{\sigma})\Bigr], \tag{14}$$

where Z is a normalization constant and $\boldsymbol{\lambda}^\intercal = (\lambda_1, \dots, \lambda_m)^\intercal$ are Lagrange multipliers such that

$$\boldsymbol{c} = Z \sum_{n\bar{s}=0}^{n} \sum_{\nu\bar{\sigma}=0}^{\nu} \boldsymbol{g}(\bar{s})\mathit{\Pi}(\bar{s}|\,\bar{\sigma}) \binom{\nu}{\nu\bar{\sigma}} \exp\Bigl[\boldsymbol{\lambda}^\intercal \sum_{n\bar{s}=0}^{n} \boldsymbol{g}(\bar{s})\mathit{\Pi}(\bar{s}|\,\bar{\sigma})\Bigr]. \tag{15}$$

Application of maximum-entropy at the sample level, denoted by I_s, gives

$$\mathrm{P}(\bar{\boldsymbol{S}} = \bar{s}|\, I_\mathrm{s}) = \zeta \binom{n}{n\bar{s}} \exp[\boldsymbol{\lambda}^\intercal \boldsymbol{g}(\bar{s})] \tag{16}$$

where ζ is a normalization constant and $\boldsymbol{\lambda}^\intercal$ are Lagrange multipliers such that

$$\boldsymbol{c} = \zeta \sum_{n\bar{s}=0}^{n} \boldsymbol{g}(\bar{s}) \binom{n}{n\bar{s}} \exp[\boldsymbol{\lambda}^\intercal \boldsymbol{g}(\bar{s})]. \tag{17}$$

The probabilities for the sample average obtained from application at the population level (14) and at the sample level (16) should be approximately equal, by our previous observation about representativity (12) and also by the fact that they must satisfy the same expectations for $\boldsymbol{g}$.

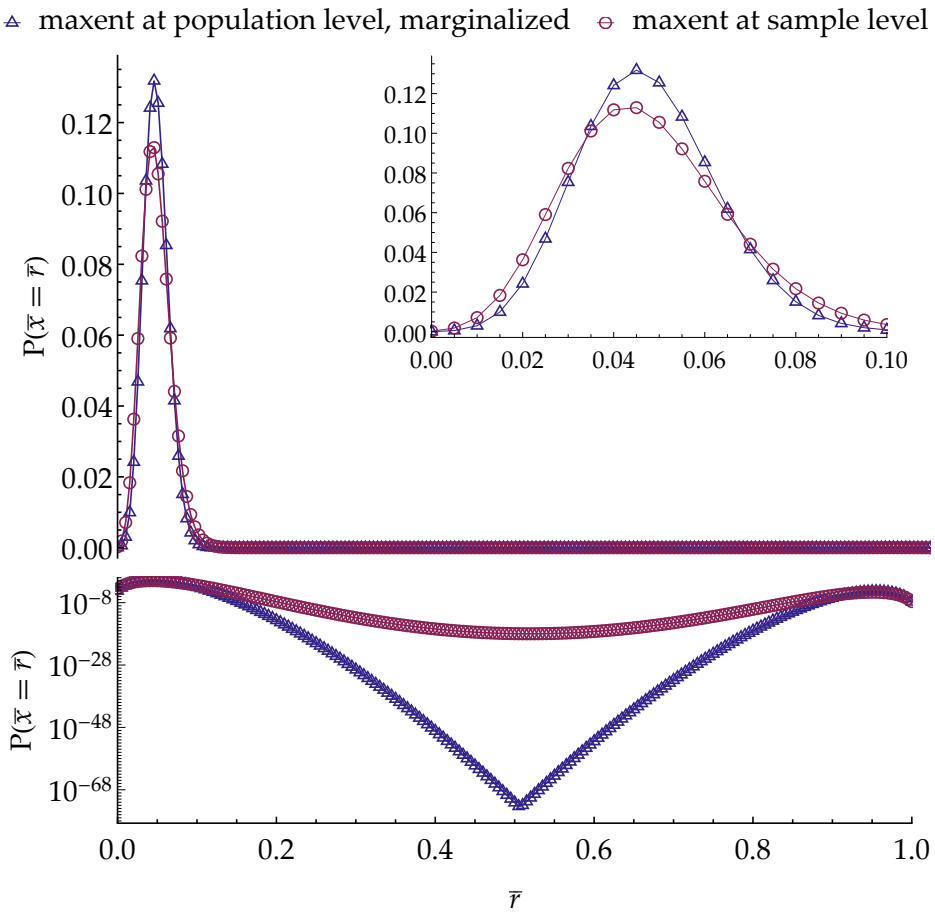

Figure 2: Linear and log-plots of $\mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (14), and at the sample level (red circles), eq. (16), with $\nu = 5000$, $n = 200$, constraints $\mathrm{E}(\bar{\boldsymbol{S}}) = 0.0478$, $\mathrm{E}(\overline{\boldsymbol{S}\boldsymbol{S}}) = 0.00257$.

Yet they cannot be exactly equal, because their equality would require the Lagrange multipliers $\boldsymbol{\Lambda}$ and $\boldsymbol{\lambda}$ to satisfy the constraint equations (15), (17), and also $\mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}} | I_{\mathrm{p}}) = \mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}} | I_{\mathrm{s}})$; that is, $2m + n$ equations (one normalization is taken care of) in $2m$ unknowns. A solution can exist, if at all, only for very special choices of constraints functions $\boldsymbol{g}$ and values $\boldsymbol{c}$.

The sample distribution obtained from maximum-entropy at the sample level will therefore likely miss important features present in the one obtained at the population level, like additional modes or particular tail behaviour. We show two examples of this discrepancy in figs 2 and 3, for $\nu = 5000$, $n = 200$, and constraint functions of the form $\boldsymbol{g}(\bar{\boldsymbol{S}}) = (\bar{\boldsymbol{S}}, \overline{\boldsymbol{S}\boldsymbol{S}}, \dots) \equiv (\bar{\boldsymbol{S}}, \binom{n\bar{\boldsymbol{S}}}{2} / \binom{n}{2}, \dots)$, equivalent to moments constraints. The constraint values used in these examples, reported in the figure captions, have neurobiologically realistic values [30].

In the first example the constraint functions are $\mathrm{E}(\bar{\boldsymbol{S}})$ and $\mathrm{E}(\overline{\boldsymbol{S}\boldsymbol{S}})$. The distribution obtained at the sample level is broader than the one obtained at the population level; the tails of the two distributions are very different.

The second example uses two additional constraint functions $\mathrm{E}(\overline{\boldsymbol{S}\boldsymbol{S}\boldsymbol{S}})$, $\mathrm{E}(\overline{\boldsymbol{S}\boldsymbol{S}\boldsymbol{S}\boldsymbol{S}})$. The distribution obtained at the population level has two modes, replaced by only one in the distribution obtained at the sample level; the tails are very different also in this case.

How should we apply the maximum-entropy method then? on the sample or on the population? Which application is 'maximally noncommittal'?

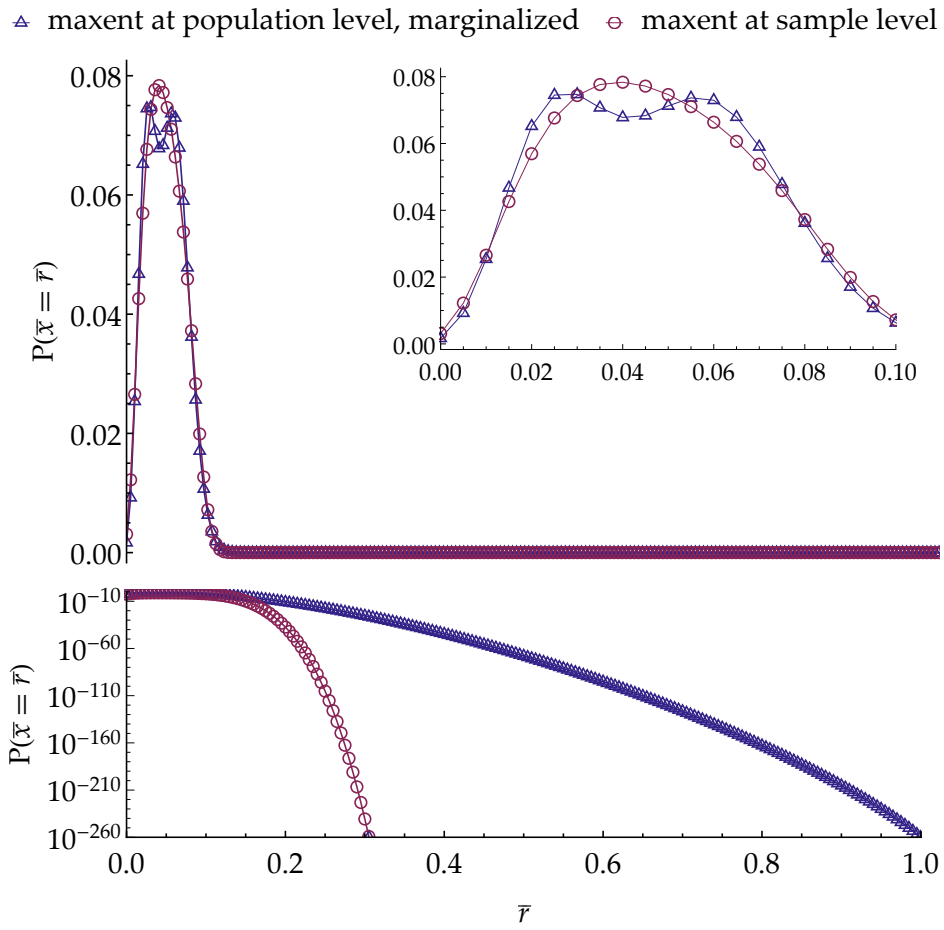

Figure 3: Linear and log-plots of $\mathrm{P}(\bar{\boldsymbol{S}} = \bar{s})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (14), and at the sample level (red circles), eq. (16), with $\nu = 5000$, $n = 200$, constraints $\mathrm{E}(\bar{\boldsymbol{S}}) = 0.0478$, $\mathrm{E}(\overline{\boldsymbol{SS}}) = 0.00257$, $\mathrm{E}(\overline{\boldsymbol{SSS}}) = 1.48 \times 10^{-4}$, $\mathrm{E}(\overline{\boldsymbol{SSSS}}) = 8.81 \times 10^{-6}$.

4 Discussion

The question that closed the preceding section cannot receive a categorical answer. An optimal answer can only be given case by case, depending on the computational power available, on which inferences we are trying to make, on which assumptions we need or want to make, and those we wish to avoid.

The tricky point is this. The maximum-entropy application at the population level and the application at the sample level give different results; they are two different statistical models. The former model clearly assumes, by construction, the existence of a larger population from which the sample is taken. What does the latter model assume in this respect? is it 'unassuming', as often claimed in the literature? or does it actually assume that *no* larger population exists? In the latter case it would not be correct to use this model in our problem.

A perfunctory intuitive reasoning seems insufficient for clarifying this point. Let's express it in the language of the probability calculus. Suppose we do not know whether the sample is really part of a larger population: we do not know whether $\nu = n$ or how large ν is otherwise. Call this state of ignorance Υ. In the probability calculus this ignorance about ν is expressed by assigning a probability distribution $\mathrm{P}(\nu|\,\Upsilon)$ that vanishes if $\nu < n$, since we know that $\nu \geqslant n$; see Good [57; 58] and Rissanen [59] for examples of such distributions over the integers. Maintaining our assumption of symmetric ignorance, probability assignments that do not assume a specific value of ν are then obtained via multiplication of all ν-dependent probabilities by $\mathrm{P}(\nu|\,\Upsilon)$ and subsequent

marginalization over ν. Technically speaking, ν becomes a *nuisance parameter* [40; 60; 61]. The probability obtained from maximum-entropy at the population level, eq. (14), then generalizes to

$$\mathrm{P}(\bar{\boldsymbol{S}} = \bar{\boldsymbol{s}}|\, \Upsilon) = \sum_{\nu} \Bigl\{ Z_\nu \sum_{\nu\bar{\boldsymbol{\sigma}}=0}^{\nu} \varPi_\nu(\bar{\boldsymbol{s}}|\, \bar{\boldsymbol{\sigma}}) \binom{\nu}{\nu\bar{\boldsymbol{\sigma}}} \exp\Bigl[\boldsymbol{\varLambda}_\nu{}^{\intercal} \sum_{n\bar{\boldsymbol{s}}=0}^{n} \boldsymbol{g}(\bar{\boldsymbol{s}}) \varPi_\nu(\bar{\boldsymbol{s}}|\, \bar{\boldsymbol{\sigma}})\Bigr]\Bigr\} \mathrm{P}(\nu|\, \Upsilon), \tag{18}$$

where ν-dependencies have been made explicit. This is a formidable expression. But our question, 'is the usual maximum-entropy at the sample level (16) unassuming with regard to the existence of a larger population?', translates now into the precise mathematical question: 'are the distributions (18) and (16) equal for some choice of $\mathrm{P}(\nu|\, \Upsilon)$, with $\mathrm{P}(\nu|\, \Upsilon) \neq 0$ for $\nu > n$?'. We leave this mathematical problem for future work. Note, however, that this equality is satisfied if $\mathrm{P}(\nu = 1|\, \Upsilon) = 1$, which means that the usual maximum-entropy model can also be interpreted as assuming that *no* larger population exists.

We find the maximum-entropy model constructed at the population level very natural and preferable. After all, physical models of neuronal networks usually include some sort of external input to the neurons as well, mimicking their embedding in a larger network. The sample distribution given by the maximum-entropy model at the population level, when used as a reference distribution for surprise analysis, may reveal features in a dataset that were unnoticed by the standard maximum-entropy model. The question remains of how to specify ν, though. We have tacitly intended ν as the size of the largest biologically or functionally homogeneous population from which our sample was recorded. It could be the amount of neurons in a functional brain area, for example the primary visual cortex, for which $\nu \sim 10^8$ [62]. For large ν – unfortunately we are not yet able to translate this 'large' into a numeric order of magnitude – the final distribution becomes independent of ν, and continuous approximations become available.

The possibility of using two different distributions is not a physical contradiction. Similar situations arise in statistical mechanics. It is known that if a system is described by a maximum-entropy Gibbs state, its subsystems need not be [63]. A dilemma quite similar to ours 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 a Gibbs distribution, or by the distribution obtained from the Liouville evolution of the Gibbs distribution assigned to the initial state. The two descriptions differ – even though the final *physical* state is obviously exactly the same [64 § 4]. The two descriptions differs because in one case we can make sharper predictions about the state thanks to our knowledge of its preceding dynamics. In this example, though, both distributions are usually immensely sharp and practically lead to the same predictions. In the neuroscientific applications considered in this note the difference in predictions may be relevant instead.

Our analysis touched only constraints of the sample average, $\mathrm{E}[g(\bar{\boldsymbol{S}})]$. The corresponding models are usually called 'homogeneous' in the literature. Purely 'inhomogeneous' models have also been used [12; 13; 27], in which expectations for individual neurons or groups of neurons are constrained, for example $\mathrm{E}(S_2)$ or $\mathrm{E}(S_1 x_8 x_9)$. A short computation shows that the maximum-entropy method with this kind of constraints gives the same result whether applied at the sample or at the population level: the states of any unconstrained neurons marginalize out. This is understandable: expressing different uncertainties about, say, neurons 2 and 5 we are breaking the symmetry of our uncertainty, which thus cannot be representative of other neurons in the sample or in the population. Inhomogeneous models, however, require enormous computational power for large sample sizes; homogeneous models therefore retain their importance. Our analysis and dilemma also persist for hybrid homogeneous-inhomogeneous models [22; 24].

Acknowledgments

To be added after review.

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. [65].

[4] E. T. Jaynes: *Information theory and statistical mechanics*. In: Ford [66] (1963), 181–218. Repr. in ref. [67], 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 [68] (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. [69].

[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, E. T. Shea-Brown, F. M. Rieke, J. Gjorgjieva: *When are microcircuits well-modeled by maximum entropy methods?* BMC Neurosci. **11**$^{\text{Suppl. 1}}$ (2010), P65. See ref. [70].

[29] 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.

[30] 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.

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

[32] E. R. Caianiello: *Neuronic equations revisited and completely solved*. In: Palm, Aertsen [71] (1986), 147–160.

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

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

[35] 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.

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

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

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

[39] 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.

[40] 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.

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

[42] 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).

[43] *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.

[44] 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).

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

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

[47] 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. [72].

[48] B. de Finetti: *La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista*. In: de Finetti [73] (1959), 1–115. Transl. in ref. [74], ch. 9, pp. 147–227.

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

[50] 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.

[51] 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.

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

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

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

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

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

[57] I. J. Good: *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, Cambridge, USA (1965).

[58] I. J. Good: *A Bayesian significance test for multinomial distributions*. J. Roy. Stat. Soc. B **29**[3] (1967), 399–431. With discussion.

[59] J. Rissanen: *A universal prior for integers and estimation by minimum description length*. Ann. Stat. **11**[2] (1983), 416–431.

[60] D. V. Lindley: *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference*, reprint. Cambridge University Press, Cambridge (2008). First publ. 1965.

[61] J.-M. Bernardo, A. F. Smith: *Bayesian Theory*, reprint. Wiley, New York (2000). First publ. 1994.

[62] G. Leuba, R. Kraftsik: *Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age*. Anat. Embryol. **190**[4] (1994), 351–366.

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

[64] E. T. Jaynes: *Inferential scattering*. (1993). http://bayes.wustl.edu/etj/node1.html. Extensively rewritten version of a paper first publ. 1985 in ref. [75], pp. 377–398.

[65] 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].

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

[67] 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.

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

[69] 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].

[70] 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. See also ref. [28].

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

[72] 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. [76].

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

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

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

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

arXiv eprints available at http://arxiv.org/.