
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 binned in time 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 i th sampled neuron at time bin t . The average activity at that bin is therefore $\bar{S}(t) := \sum_i S_i(t)/n$. 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 average pair activity is $\overline{SS}(t) := \binom{n}{2}^{-1} \binom{n\bar{S}(t)}{2}$. With some combinatorics we see that the average activity of m -tuples of neurons is

$$\overbrace{S \cdots S}^{m \text{ terms}} = \binom{n}{m}^{-1} \binom{n\bar{S}(t)}{m}. \quad (1)$$

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 activities $\bar{S}(t)$;
 - the time averages of the activities $\bar{S}(t)$ and pair-activities $\overline{SS}(t)$;
 - ... and so on, including the time averages of $\bar{S}(t)$ up to $\overline{S \cdots S}(t)$ with k terms.

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 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 $m \geq 2$, then we can say that there is a ‘pairwise cooperativity’, and that any higher-order cooperativity is just a reflection or consequence of the pairwise one. The reason is that the information from higher-order simultaneous activities did not lead to appreciable changes in the distribution obtained from pair activities.

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

We hurry to say that the description just given is just *one* way to quantify and characterize cooperativity and functional grouping, not *the only* way. It can surely be criticized from many points of view. Yet, it is quantitative and bears a more precise meaning than an undefined, vague notion of ‘cooperativity’. Two persons who apply this procedure to the same data will obtain the same numbers. Different protocols can be based on the maximum-entropy method, for instance protocols that take into account the activities or pair activities of specific neurons rather than 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 problem extends also to more complex versions of the method, possibly except version that use ‘hidden’ neurons. The problem stems from the fact that the recorded neurons are a *sample* from a larger, unrecorded population.

Suppose that the neurons were recorded with electrodes covering an area of some square millimetres [berenyietal2014]. This recording is a sample of the activity of the neuronal population under the recording device; such population can amount to tens of thousands of neurons [abeles1991]. 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).

Here is the problem. 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, if our sample is representative of a larger population, we can apply the maximum-entropy method to the larger (unrecorded) population, using the constraints obtained from the recorded sample, and generate a probability distribution for the larger population. The distribution for the sample is then obtained by marginalization. Either application seems to have some commendable features. However, *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, 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 [35; 36 parts I, VI; 37 ch. 8; 38 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, so they may be of interest on their own.

The notation and terminology in this note follows ISO and ANSI standards [39–43] but for the use of the comma ‘,’ to denote logical conjunction. Probability notation follows Jaynes [38]. By ‘probability’ we mean a degree of belief which ‘would be agreed by all rational men if there were any rational men’ [44].

2 Notation: population, sample, probabilities

We have a population of v binary neurons. We assume that they can be distinguished, by their spike shapes for example; but other details, like their locations, are unknown. The neurons have a joint state $(\Sigma_1, \dots, \Sigma_v) =: \Sigma$ having fixed but unknown binary values $(\sigma_1, \dots, \sigma_v) =: \sigma \in \{0, 1\}^v$; the ‘ Σ ’ denote the variables, and ‘ σ ’ their values. A particular sample of n neurons from this population has joint state $(S_1, \dots, S_n) =: S$ having fixed binary values $(s_1, \dots, s_n) =: s \in \{0, 1\}^n$. We will consider various averages of the population and the sample. For this purpose we introduce a population-averaging operator $\bar{\cdot}$ defined by

$$\begin{aligned}\bar{\Sigma} &:= \frac{1}{v}(\Sigma_1 + \Sigma_2 + \dots + \Sigma_v), & \overline{\Sigma\Sigma} &:= \binom{v}{2}^{-1}(\Sigma_1\Sigma_2 + \Sigma_1\Sigma_3 + \dots + \Sigma_{v-1}\Sigma_v), \\ \overline{\Sigma\Sigma\Sigma} &:= \binom{v}{3}^{-1}(\Sigma_1\Sigma_2\Sigma_3 + \dots + \Sigma_{v-2}\Sigma_{v-1}\Sigma_v), & \text{and so on.}\end{aligned}\tag{2}$$

These formulae say that $\bar{\Sigma}$ is the fraction of active neurons, $\overline{\Sigma\Sigma}$ the fraction of simultaneously active pairs out of all $\binom{v}{2}$ pairs, $\overline{\Sigma\Sigma\Sigma}$ the fraction of simultaneously active triplets, and so on. Products of states like $\Sigma_i \dots \Sigma_j$ also have values in $\{0, 1\}$; from this we can combinatorially prove that

$$\overbrace{\Sigma \dots \Sigma}^{m \text{ factors}} = \binom{v}{m}^{-1} \binom{v\bar{\Sigma}}{m}.\tag{3}$$

Analogous formulae hold for quantities like S , σ , s .

Our uncertainty about the actual state of the population is completely expressed by the joint probability distribution

$$P(\Sigma_1 = \sigma_1, \Sigma_2 = \sigma_2, \dots, \Sigma_v = \sigma_v | K) \quad \text{or} \quad P(\Sigma = \sigma | K), \quad \sigma \in \{0, 1\}^v,\tag{4}$$

where K denotes our state of knowledge, i.e. the evidence and assumptions backing this particular probability assignment. Our uncertainty about the state of the sample is likewise expressed by

$$P(S_1 = s_1, S_2 = s_2, \dots, S_n = s_n | K) \quad \text{or} \quad P(S = s | K), \quad s \in \{0, 1\}^n.\tag{5}$$

3 Probability: relations between population and sample

The theory of statistical sampling is covered in many excellent texts, for example Ghosh & Meeden [35] or Freedman, Pisani, & Purves [36 parts I, VI]; summaries can be found in Gelman et al. [37 ch. 8] and Jaynes [38 ch. 3].

We need to make an initial probability assignment for the state of the population before any experimental observations are made. This initial assignment will be modified by our experimental observations, which can involve only a sample of the population. Our initial probability assignment should reflect that samples are somehow representative of the whole population. We consider here a state of knowledge that express this representativeness.

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* [45; 35 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 identity of the neurons.

The *representation theorem for finite exchangeability* states that the symmetric distribution in the state of knowledge I must obey

$$P(\Sigma = \sigma | I) \equiv P(\Sigma = \sigma | \bar{\Sigma} = \bar{\sigma}, I) P(\bar{\Sigma} = \bar{\sigma} | I) = \left(\frac{\nu}{\nu \bar{\sigma}} \right)^{-1} P(\bar{\Sigma} = \bar{\sigma} | I), \quad (6)$$

the latter being the probability for the population average Σ . A sum is only apparently missing in the central term: its summands $\bar{\Sigma} = A$ are all zero except for $A = \bar{\sigma}$. Proof of this theorem and generalizations to non-binary and continuum cases are given by de Finetti [46], Kendall [47], Ericson [48], Diaconis & Freedman [49; 50], Heath & Sudderth [51]. This theorem is intuitive: owing to symmetry, we must assign equal probabilities to all states with $\nu \bar{\sigma}$ active neurons.

By marginalization we obtain the probability for the state of the sample:

$$P(S = s | I) = \binom{n}{n\bar{s}}^{-1} P(\bar{S} = \bar{s} | I), \quad (7)$$

with

$$P(\bar{S} = \bar{s} | I) = \sum_{\nu \bar{\sigma}=0}^{\nu} P(\bar{S} = \bar{s} | \bar{\Sigma} = \bar{\sigma}, I) P(\bar{\Sigma} = \bar{\sigma} | I), \quad (8)$$

$$P(\bar{S} = \bar{s} | \bar{\Sigma} = \bar{\sigma}, I) = \binom{n}{n\bar{s}} \binom{\nu - n}{\nu \bar{\sigma} - n\bar{s}} \binom{\nu}{\nu \bar{\sigma}}^{-1} =: \Pi(\bar{s} | \bar{\sigma}). \quad (9)$$

The conditional probability in the last formula is a hypergeometric distribution $\Pi(\bar{s} | \bar{\sigma})$, typical of ‘drawing without replacement’ problems. The combinatorial proof of the formulae above is in fact the same as for this class of problems [38 ch. 3; 52 § 4.8.3; 53 § II.6]. Our initial symmetric ignorance should intuitively also apply to the sample; the probability for the state of the sample (8) indeed satisfies the representation theorem (6) as well.

The conditional probability $P(\bar{S} = \bar{s} | \bar{\Sigma} = \bar{\sigma}, I) \equiv \Pi(\bar{s} | \bar{\sigma})$ relates the spaces of the sample average $\bar{\Sigma} \in \{0, \dots, \nu\}$ and of the population average $\bar{S} \in \{0, \dots, n\}$. It is a coarsening projector of any probability for $\bar{\Sigma}$ onto a marginal probability for \bar{S} , as eq. (8) shows. For each function f of the sample average \bar{S} , the hypergeometric distribution $\Pi(\bar{s} | \bar{\sigma})$ also allows us to construct a function f^* of the population average $\bar{\Sigma}$

$$f^*(\bar{\Sigma}) := \sum_{n\bar{s}=0}^n f(\bar{s}) \Pi(\bar{s} | \bar{\Sigma}), \quad (10)$$

that has the same expectation as f :

$$E[f^*(\bar{\Sigma})] \equiv \sum_{\nu \bar{\sigma}=0}^{\nu} f^*(\bar{\sigma}) P(\bar{\Sigma} = \bar{\sigma} | I) = \sum_{n\bar{s}=0}^n f(\bar{s}) P(\bar{S} = \bar{s} | I) \equiv E[f(\bar{S})]. \quad (11)$$

The middle equality uses the combination of formulae (10) and (8).

A look at a plot of the hypergeometric distribution $\Pi(\bar{s} | \bar{\sigma})$, see fig. 1, reveals that it is a sort of ‘fuzzy identity matrix’ between the $\bar{\Sigma}$ -space $\{0, \dots, \nu\}$ and \bar{S} -space $\{0, \dots, n\}$. When $n = \nu$ it is the



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

identity matrix. We thus have that

$$P(\bar{S} = a) \approx P(\bar{\Sigma} = a), \quad E[f(\bar{S})] \approx E[f(\bar{\Sigma})]. \quad (12)$$

These are only very approximate equalities: they 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 $\Pi(\bar{s} | \bar{\sigma})$.

The approximate equalities above, on the other hand, express the fact that *our uncertainty about the sample is representative of our uncertainty about the population and about other samples*, and vice versa. These approximate equalities are not generally valid for other probability distributions for the population. They should be contrasted with the limits $P(\bar{S} = a) \rightarrow P(\bar{\Sigma} = a)$ and $E[f(\bar{S})] \rightarrow E[f(\bar{\Sigma})]$, as $n \rightarrow \nu$, do: these limits are trivially, universally valid because the sample becomes the full population as $n \rightarrow \nu$. In particular, these limits hold even in cases where the conditional probability $P(\bar{S} = \bar{s} | \bar{\Sigma} = \bar{\sigma})$ is not a fuzzy identity, our uncertainties about sample and about population can differ wildly, and the approximate equalities (12) do not hold.

For averaged products, $\overbrace{\bar{S} \cdots \bar{S}}^{m \text{ factors}} \equiv \binom{n\bar{S}}{m} / \binom{n}{m}$, formulae (10) yield the useful equality

$$E(\overbrace{\bar{S} \cdots \bar{S}}^{m \text{ factors}} | I) \equiv \binom{n}{m}^{-1} \sum_{n\bar{s}=0}^n \binom{n\bar{s}}{m} P(\bar{S} = \bar{s} | I) = \binom{\nu}{m}^{-1} \sum_{\nu\bar{\sigma}=0}^{\nu} \binom{\nu\bar{\sigma}}{m} P(\bar{\Sigma} = \bar{\sigma} | I) \equiv E(\overbrace{\bar{\Sigma} \cdots \bar{\Sigma}}^{m \text{ factors}} | I). \quad (13)$$

The proof of the middle equality uses the expression for the m th factorial moment of the hypergeometric distribution [54]. Thus, in the state of knowledge I the expectations of the averages of activity products *are initially the same for the sample and for the full population*. Similar relations can be found for the raw moments $E(\bar{S}^m)$ and $E(\bar{\Sigma}^m)$, which can be written in terms of the product expectations via eq. (3).

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

The probability formulae (6)–(9) are constraints on our initial probability assignment, but do not determine it numerically. The probability $P(\bar{\Sigma} = \bar{\sigma} | I)$ for the population average needs to be

numerically specified, and by marginalization (8) it will determine that of the sample average, $P(\bar{S} = \bar{s} | I)$. If we numerically specify the latter, the former is not completely specified, because eq. (8) 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 $E(\bar{S}) = c_1$, $E(\bar{S}\bar{S}) = c_2$. This is still an underdetermined problem: several distributions can have the same desired expectations, as clear from eqs (13).

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 (10) allow us to apply the method to find the probability of the population $P(\bar{S} = \bar{s} | I)$, or of the sample $P(\bar{S} = \bar{s} | I)$. *The two applications, however, are inequivalent.* They lead to numerically different distributions for the sample average $P(\bar{S} = \bar{s} | I)$.

Suppose we want to constrain the sample expectations of a vector function $\mathbf{f} = (f_1, \dots, f_m)$ to the vector values $\mathbf{c} = (c_1, \dots, c_m)$, that is, $E[\mathbf{f}(\bar{S})] = \mathbf{c}$. Application of maximum-entropy [6; 7] at the population level, denoted by I_p , gives

$$P(\bar{S} = \bar{s} | I_p) = Z \left(\frac{\nu}{\nu \bar{s}} \right) \exp \left[\boldsymbol{\Lambda}^\top \sum_{n\bar{s}=0}^n \mathbf{f}(\bar{s}) \Pi(\bar{s} | \bar{s}) \right], \quad (14)$$

and then by marginalization (7)

$$P(\bar{S} = \bar{s} | I_p) = Z \sum_{\nu \bar{s}=0}^{\nu} \Pi(\bar{s} | \bar{s}) \left(\frac{\nu}{\nu \bar{s}} \right) \exp \left[\boldsymbol{\Lambda}^\top \sum_{n\bar{s}=0}^n \mathbf{f}(\bar{s}) \Pi(\bar{s} | \bar{s}) \right], \quad (15)$$

where Z is a normalization constant and $\boldsymbol{\Lambda}^\top = (\Lambda_1, \dots, \Lambda_m)^\top$ are Lagrange multipliers such that

$$\mathbf{c} = Z \sum_{n\bar{s}=0}^n \sum_{\nu \bar{s}=0}^{\nu} \mathbf{f}(\bar{s}) \Pi(\bar{s} | \bar{s}) \left(\frac{\nu}{\nu \bar{s}} \right) \exp \left[\boldsymbol{\Lambda}^\top \sum_{n\bar{s}=0}^n \mathbf{f}(\bar{s}) \Pi(\bar{s} | \bar{s}) \right]. \quad (16)$$

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

$$P(\bar{S} = \bar{s} | I_s) = \zeta \left(\frac{n}{n\bar{s}} \right) \exp[\boldsymbol{\lambda}^\top \mathbf{f}(\bar{s})] \quad (17)$$

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

$$\mathbf{c} = \zeta \sum_{n\bar{s}=0}^n \mathbf{f}(\bar{s}) \left(\frac{n}{n\bar{s}} \right) \exp[\boldsymbol{\lambda}^\top \mathbf{f}(\bar{s})]. \quad (18)$$

The probabilities for the sample average obtained from application at the population level (15) and at the sample level (17) should be approximately equal, by our previous observation about representativity (12) and also by the fact that they must satisfy the same expectations for \mathbf{f} .

Yet they cannot be exactly equal, because their equality would require the Lagrange multipliers $\boldsymbol{\Lambda}$ and $\boldsymbol{\lambda}$ to satisfy the constraint equations (16), (18), and also $P(\bar{S} = \bar{s} | I_p) = P(\bar{S} = \bar{s} | I_s)$; 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 \mathbf{f} and values \mathbf{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 $\mathbf{f}(\bar{S}) = (\bar{S}, \bar{S}\bar{S}, \dots) \equiv (\bar{S}, \binom{n\bar{S}}{2} / \binom{n}{2}, \dots)$, 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 $E(\bar{S})$ and $E(\bar{S}\bar{S})$. 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 $E(\bar{S}\bar{S}\bar{S})$, $E(\bar{S}\bar{S}\bar{S}\bar{S})$. 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.

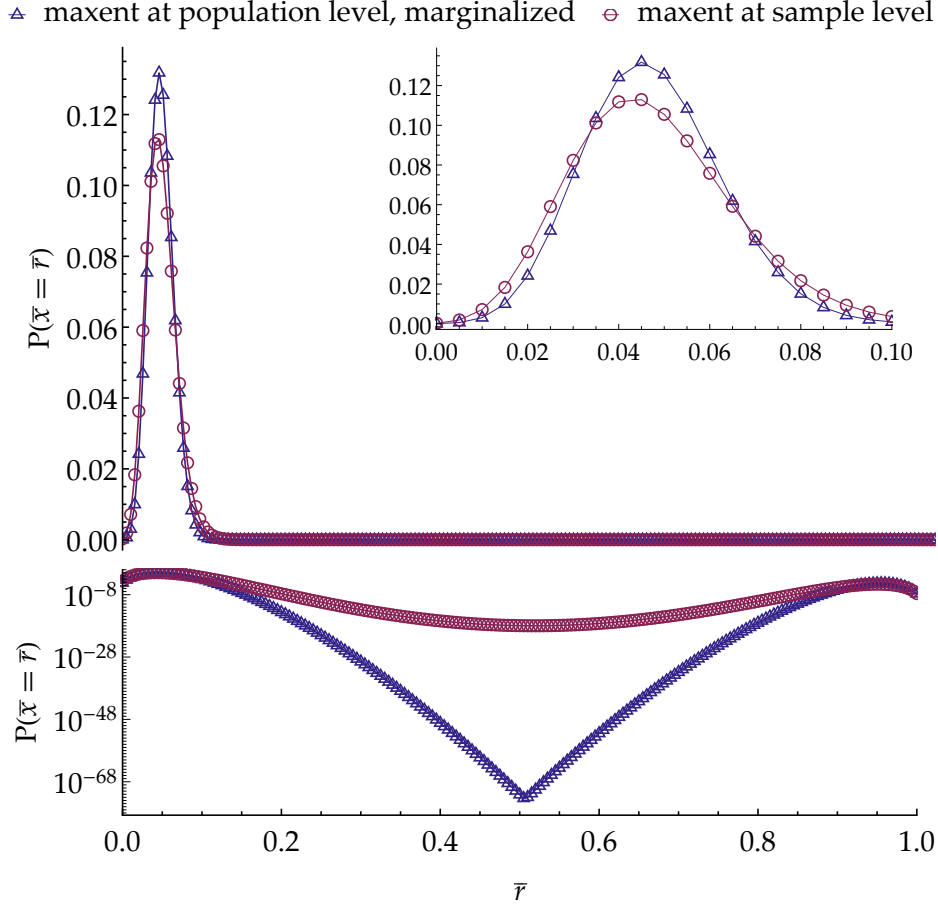


Figure 2: Linear and log-plots of $P(\bar{S} = \bar{s})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (15), and at the sample level (red circles), eq. (17), with $\nu = 5000$, $n = 200$, constraints $E(\bar{S}) = 0.0478$, $E(\bar{S}\bar{S}) = 0.00257$.

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

5 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 \mathcal{Y} . In the probability calculus this ignorance about ν is expressed by assigning a probability distribution $P(\nu | \mathcal{Y})$ that vanishes if $\nu < n$, since we know that $\nu \geq n$; see Good [55; 56] and Rissanen [57] for examples of such distributions over the integers. Maintaining our

△ maxent at population level, marginalized ○ maxent at sample level

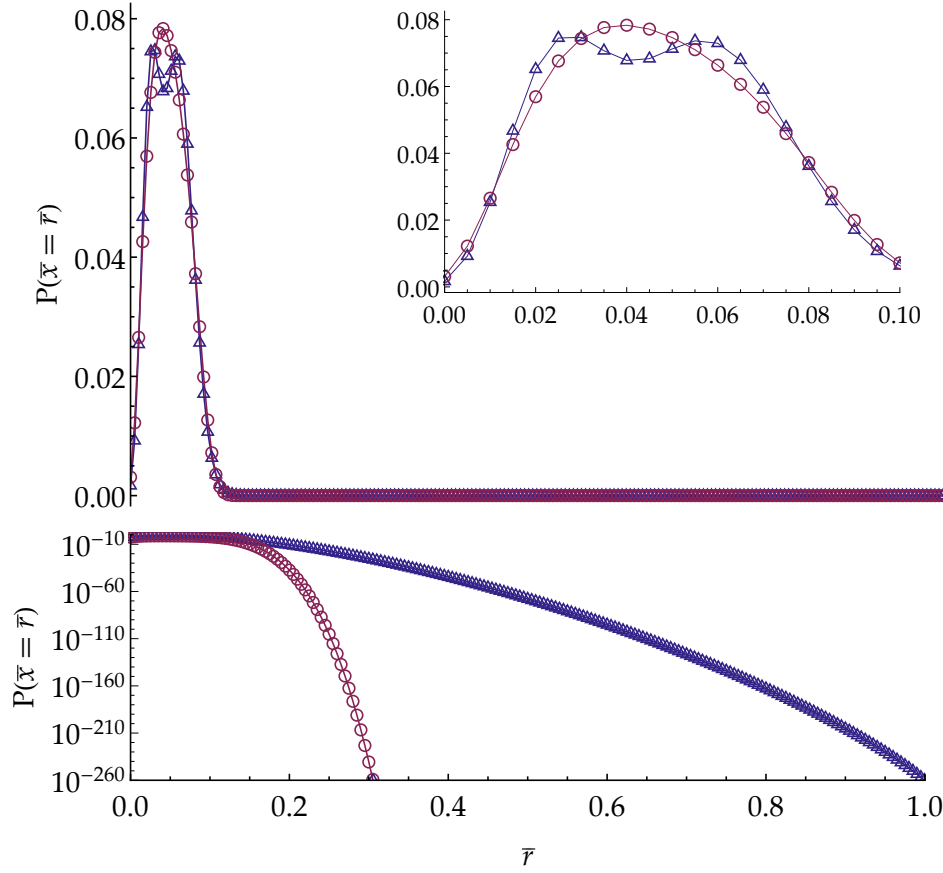


Figure 3: Linear and log-plots of $P(\bar{S} = \bar{s})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (15), and at the sample level (red circles), eq. (17), with $\nu = 5000$, $n = 200$, constraints $E(\bar{S}) = 0.0478$, $E(\bar{S}\bar{S}) = 0.00257$, $E(\bar{S}\bar{S}\bar{S}) = 1.48 \times 10^{-4}$, $E(\bar{S}\bar{S}\bar{S}\bar{S}) = 8.81 \times 10^{-6}$.

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 $P(\nu | \gamma)$ and subsequent marginalization over ν . Technically speaking, ν becomes a *nuisance parameter* [38; 58; 59]. The probability obtained from maximum-entropy at the population level, eq. (15), then generalizes to

$$P(\bar{S} = \bar{s} | \gamma) = \sum_{\nu} \left\{ Z_{\nu} \sum_{\nu \bar{\sigma}=0}^{\nu} \Pi_{\nu}(\bar{s} | \bar{\sigma}) \binom{\nu}{\nu \bar{\sigma}} \exp \left[\mathcal{A}_{\nu}^{\top} \sum_{n \bar{s}=0}^n f(\bar{s}) \Pi_{\nu}(\bar{s} | \bar{\sigma}) \right] \right\} P(\nu | \gamma), \quad (19)$$

where ν -dependencies have been made explicit. This is a formidable expression. But our question, ‘is the usual maximum-entropy at the sample level (17) unassuming with regard to the existence of a larger population?’, translates now into the precise mathematical question: ‘are the distributions (19) and (17) equal for some choice of $P(\nu | \gamma)$, with $P(\nu | \gamma) \neq 0$ for $\nu > n$?’ We leave this mathematical problem for future work. Note, however, that this equality is satisfied if $P(\nu = 1 | \gamma) = 1$, 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 $\nu \sim 10^8$ [60]. 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 [61]. 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 [62 § 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, $E[f(\bar{S})]$. 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 $E(S_2)$ or $E(S_1 x_8 x_9)$. 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. [63].
- [4] E. T. Jaynes: *Information theory and statistical mechanics*. In: Ford [64] (1963), 181–218. Repr. in ref. [65], 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 [66] (1991), 237–244.
- [9] L. Martignon, H. V. Hasse, S. Grün, A. Aerts, 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. [67].
- [14] J. H. Macke, M. Oppen, 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](https://arxiv.org/abs/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. Oppen, 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. Amodè, 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. Toyozumi: *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](https://arxiv.org/abs/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**^{Suppl. 1} (2010), P65. See ref. [68].
- [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](https://arxiv.org/abs/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 [69] (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] M. Ghosh, G. Meeden: *Bayesian Methods for Finite Population Sampling*, reprint. Springer, Dordrecht (1997).
- [36] D. Freedman, R. Pisani, R. Purves: *Statistics*, 4th ed. Norton, London (2007). First publ. 1978.
- [37] 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.

- [38] 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/XQUHIUXHIQUHIQUHX2>, <http://www-biba.inrialpes.fr/Jaynes/prob.html>, <http://omega.albany.edu:8008/JaynesBook.html>.
- [39] ISO: *Quantities and units*, 3rd ed. International Organization for Standardization. Geneva (1993).
- [40] 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).
- [41] *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>.
- [42] 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).
- [43] ISO: *ISO 3534-2:2006: Statistics – Vocabulary and symbols – Part 2: Applied statistics*. International Organization for Standardization. Geneva (2006).
- [44] I. J. Good: *How to estimate probabilities*. J. Inst. Maths. Applics **2**⁴ (1966), 364–383.
- [45] 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. [70].
- [46] B. de Finetti: *La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista*. In: de Finetti [71] (1959), 1–115. Transl. in ref. [72], ch. 9, pp. 147–227.
- [47] D. G. Kendall: *On finite and infinite sequences of exchangeable events*. Studia Sci. Math. Hung. **2** (1967), 319–327.
- [48] 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>.
- [49] 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>.
- [50] P. Diaconis, D. Freedman: *Finite exchangeable sequences*. Ann. Prob. **8**⁴ (1980), 745–764.
- [51] D. Heath, W. Sudderth: *De Finetti's theorem on exchangeable variables*. American Statistician **30**⁴ (1976), 188–189.
- [52] S. Ross: *A First Course in Probability*, 8th ed. Pearson, Upper Saddle River, USA (2010). First publ. 1976.
- [53] W. Feller: *An Introduction to Probability Theory and Its Applications. Vol. I*, 3rd ed. Wiley, New York (1968). First publ. 1950.
- [54] R. B. Potts: *Note on the factorial moments of standard distributions*. Aust. J. Phys. **6**⁴ (1953), 498–499.
- [55] I. J. Good: *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. MIT Press, Cambridge, USA (1965).
- [56] I. J. Good: *A Bayesian significance test for multinomial distributions*. J. Roy. Stat. Soc. B **29**³ (1967), 399–431. With discussion.
- [57] J. Rissanen: *A universal prior for integers and estimation by minimum description length*. Ann. Stat. **11**² (1983), 416–431.
- [58] D. V. Lindley: *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference*, reprint. Cambridge University Press, Cambridge (2008). First publ. 1965.
- [59] J.-M. Bernardo, A. F. Smith: *Bayesian Theory*, reprint. Wiley, New York (2000). First publ. 1994.
- [60] 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.
- [61] C. Maes, F. Redig, A. Van Moffaert: *The restriction of the Ising model to a layer*. J. Stat. Phys. **96**¹ (1999), 69–107.
- [62] E. T. Jaynes: *Inferential scattering*. (1993). <http://bayes.wustl.edu/etj/node1.html>. Extensively rewritten version of a paper first publ. 1985 in ref. [73], pp. 377–398.
- [63] 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].
- [64] K. W. Ford, ed.: *Statistical Physics*. Benjamin, New York (1963).
- [65] 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.
- [66] W. T. Grandy Jr., L. H. Schick, eds.: *Maximum Entropy and Bayesian Methods: Laramie, Wyoming, 1990*. Kluwer, Dordrecht (1991).

- [67] 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].
- [68] 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](https://arxiv.org/abs/1011.2797). See also ref. [28].
- [69] G. Palm, A. Aertsen, eds.: *Brain Theory*. Springer, Berlin (1986).
- [70] 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. [74].
- [71] B. de Finetti, ed.: *Induzione e statistica*, reprint. Springer, Berlin (2011). First publ. 1959.
- [72] B. de Finetti: *Probability, Induction and Statistics: The art of guessing*. Wiley, London (1972).
- [73] C. R. Smith, W. T. Grandy Jr., eds.: *Maximum-Entropy and Bayesian Methods in Inverse Problems*. D. Reidel, Dordrecht (1985).
- [74] 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/>.