Representative samples and maximum-entropy distributions: a dilemma

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

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

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

Y. Roudi
yasser.roudi@ntnu.no

Abstract

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

Imagine that we have recorded the firing activity of a hundred neurons from a particular brain area. This scenario is a concrete possibility thanks to recent electrophysiological techniques [1]. In 1985, Gerstein, Perkel, & Dayhoff wrote:

> the very profusion of data produced in such experiments creates serious problems of analysis and interpretation; these problems are conceptual as well as practical. The principal conceptual problems are (1) defining cooperativity or functional grouping among neurons and (2) formulating quantitative criteria for recognizing and characterizing such cooperativity. [2]

More than thirty years later these problems are still under investigation.

Some investigations in the neuroscientific literature proceed as follows. We bin in time and binarize in intensity [3; 4] the recorded activity, transforming it into a sequence of 0s (inactive) and 1s (firing) for each neuron. Groups of two or more 1s can appear in any time bin. Then we ask a question of this kind: can the presence of groups of three or more simultaneous firings be 'explained' in terms of groups of just two, or three, or…? This question is very vague until we define what we mean with 'explain' or similar words. Usually such 'explanation' has a probabilistic meaning: we wonder whether the sole knowledge about two, or three, or more simultaneous firings can yield as good predictions about the whole activity as a more complete knowledge would do. To make this question concrete we must first choose a probabilistic model from some well-defined assumptions, and then assess whether its final probabilities are 'good' according to some well-defined measure. In the following we consider only the first, model-building, step.

Maximum-entropy probability models [5–8] have been used in neuroscience with this kind of purpose, for recordings from brain areas as diverse as retina and motor cortex, at least since the 1990s [9–25]. Part of this literature focuses on the range of applicability of maximum-entropy distributions [26–31].

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

We assume that our readers are at least vaguely familiar with this kind of model. In brief, it builds a probability distribution for neuronal activity using only data about simultaneous firings of two neurons, or of three, and so on. Let us take the case of the first two moments as an example. The maximum-entropy distribution can be derived in three main ways:

1. From the maximum-entropy method [5–7]. The distribution is chosen as the one with the highest Shannon entropy among those having particular first and second moments. This distribution is deemed to be 'maximally noncommittal' [6], but this adjective means little without further technical characterization. Two characterizations are given by the next two derivations.
2. As the asymptotic distribution of a Bayesian model called 'exchangeable', with a particular prior [32–37; see also: 38 ch. VI; 39]. This distribution is the limit, when the number of time bins is very large, of

   $$\mathrm{P}(\text{activity in one time bin}\,|\,\text{1st \& 2nd empirical moments from other bins}), \tag{1}$$

   calculated with Bayes's theorem from a particular initial joint distribution in which different time bins are almost independent. This derivation produces a probability just by taking into account those two moments, but does not exclude that other information might be relevant. It can therefore be compared with a probability that uses additional information, or with empirical frequencies.
3. As the asymptotic distribution of a Bayesian model based on sufficient statistics [40; 41; 37]. This distribution is the limit, when the number of time bins is very large, of

   $$\mathrm{P}(\text{activity in one time bin}\,|\,\text{activities in other bins}) \tag{2}$$

   under the assumption that *only* the values of the first and second empirical moments of the activities are relevant for the probability above, i.e.

   $$\mathrm{P}(\text{activity in one time bin}\,|\,\text{activities in other bins}) \equiv \mathrm{P}(\text{activity in one time bin}\,|\,\text{1st \& 2nd empirical moments from other bins}); \tag{3}$$

   in other words, the assumption that the first and second moments are sufficient statistics. This limit generally does not depend on the initial joint distribution. This derivation produces a probability by explicitly assuming that only the first and second moments are informationally relevant: any other information would lead to the same conditional probability. It can therefore only be compared with empirical frequencies.

We refer our readers to the cited references for a deeper discussion of these derivations. It is remarkable that all three lead to the same distribution. We shall now see, however, that if we apply these derivations more carefully they lead to different distributions.

This kind of investigation of recorded samples of neuronal activity would have little significance if we did not expect its results to generalize beyond the particular recorded sample. In other words, we expect that the conclusions drawn with maximum-entropy methods about the recorded neurons should somehow generalize or extrapolate at least to larger areas from which those neurons are recorded. We are thus assuming that the recorded sample is a *representative sample* of some larger area. Whether such extrapolation makes biological sense is a difficult question, to be analysed case by case. For the moment let's assume that it does make biological sense, and that our sample can be assumed to be representative of a larger area.

Then a double dilemma appears in the way the maximum-entropy method is typically applied in the neuroscientific literature to recorded or artificial samples. In this note we want to discuss one part of this dilemma and offer a solution to it.

We can apply the maximum-entropy method to the sample, choosing some set of moments, and generate a probability distribution for the state of the sample. But if we consider our sample to be representative of a larger population, it makes sense to apply the maximum-entropy method to the larger population, using one of the three derivations above, and generate a probability distribution for the larger population. The distribution for the sample can then be found by marginalization. Here's the double dilemma:

- *the distribution obtained from the application at the sample level differs from those obtained from the application at the population level;*
- *at the population level, the derivations 1 and 2 lead to the same distribution, but derivation 3 leads to a different one.*

In the rest of the paper we mathematically formulate and analyse the first part of this dilemma, focusing on the derivation 1. To this purpose we also present some probability relations relevant to sampling. 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.

We consider maximum-entropy models that constrain various kinds of sample or population averages; these models are often called 'homogeneous'. The final discussion touches upon 'inhomogeneous' models as well.

The notation in this note follows ISO and ANSI standards [42–44] 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' [46].

2 Setup: population, sample, probabilities

We have a population of N 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 $(X_1, \dots, X_N) =: \boldsymbol{X}$ having fixed but unknown binary values $(R_1, \dots, R_N) =: \boldsymbol{R} \in \{0,1\}^N$. A particular sample of n neurons from this population has joint state $(x_1, \dots, x_n) =: \boldsymbol{x}$ having fixed binary values $(r_1, \dots, r_n) =: \boldsymbol{r} \in \{0,1\}^n$. We will consider various averages of the population and the sample. For this purpose we introduce a general averaging operator $\overline{\cdot}$ defined by

$$\begin{gathered}\overline{\boldsymbol{X}} \coloneqq \tfrac{1}{N}(X_1 + X_2 + \cdots + X_N), \qquad \overline{\boldsymbol{X}\boldsymbol{X}} \coloneqq \tbinom{N}{2}^{-1}(X_1X_2 + X_1X_3 + \cdots + X_{N-1}X_N), \\ \overline{\boldsymbol{X}\boldsymbol{X}\boldsymbol{X}} \coloneqq \tbinom{N}{3}^{-1}(X_1X_2X_3 + \cdots + X_{N-2}X_{N-1}X_N),\end{gathered} \tag{4}$$

and so on. These formulae say that $\overline{\boldsymbol{X}}$ is the fraction of active neurons, $\overline{\boldsymbol{X}\boldsymbol{X}}$ the fraction of simultaneously active pairs out of all $\binom{N}{2}$ pairs, $\overline{\boldsymbol{X}\boldsymbol{X}\boldsymbol{X}}$ the fraction of simultaneously active triplets, and so on. Products of states like $X_i \cdots X_j$ also have values in $\{0,1\}$; from this we can combinatorially prove that

$$\overline{\underbrace{\boldsymbol{X}\cdots\boldsymbol{X}}_{m \text{ factors}}} = \binom{N}{m}^{-1}\binom{N\overline{\boldsymbol{X}}}{m}. \tag{5}$$

Analogous formulae hold for quantities like $\boldsymbol{x}$, $\boldsymbol{R}$, $\boldsymbol{r}$.

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

$$\mathrm{P}(X_1 = R_1, X_2 = R_2, \dots, X_N = R_N \mid K) \quad \text{or} \quad \mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid K), \quad \boldsymbol{R} \in \{0,1\}^N, \tag{6}$$

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

$$\mathrm{P}(x_1 = r_1, x_2 = r_2, \dots, x_n = r_n \mid K) \quad \text{or} \quad \mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid K), \quad \boldsymbol{r} \in \{0,1\}^n. \tag{7}$$

3 Initial assumptions: the probability of representative samples

We need to make an initial probability assignment before any experimental observations are made. This initial assignment will be modified by our experimental observations. Our probability assignment should reflect that the sample is somehow 'representative' of the population. We consider here two states of knowledge that express this representativeness in different ways but lead to identical *initial* probability assignments.

In the first state of knowledge, denoted I', we know that the neurons in the population are biologically or functionally similar, for example in morphology and kind of input or output they receive or give.

Knowledge of this similarity leads us to assign a probability distribution for the population state $\boldsymbol{X}$ that is symmetric under permutations of neuron identities, or *exchangeable* as it is usually called.

DISCUSSION ABOUT THE SECOND STATE OF KNOWLEDGE IS TO BE DISCARDED

In the second state of knowledge or ignorance, denoted I'', we are completely ignorant about the physical details of the individual neurons. Our ignorance is therefore symmetric under permutations of neuron identities. This also leads to an exchangeable probability distribution for $\boldsymbol{X}$.

Let us use I to denote either of these two states of knowledge, in those probabilities that are identical for I' and I''.

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

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{R}|\, I) \equiv \mathrm{P}(\boldsymbol{X} = \boldsymbol{R}|\, \overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}, I)\, \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}|\, I) = \binom{N}{N\overline{\boldsymbol{R}}}^{-1} \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}|\, I), \tag{8}$$

the latter being the probability for the population average $\boldsymbol{X}$. A sum is only apparently missing in the central term: its summands $\overline{\boldsymbol{X}} = A$ are all zero except for $A = \overline{\boldsymbol{R}}$. Proof of this theorem and generalizations to non-binary and continuum cases are given by de Finetti [47], Kendall [48], Ericson [49], Diaconis & Freedman [50; 51], Heath & Sudderth [52]. This theorem is intuitive: owing to symmetry, we must assign equal probabilities to all states with $N\overline{\boldsymbol{R}}$ active neurons.

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

$$\mathrm{P}(\boldsymbol{x} = \boldsymbol{r}|\, I) = \binom{n}{n\bar{\boldsymbol{r}}}^{-1} \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, I), \tag{9}$$

where

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, I) = \sum_{N\overline{\boldsymbol{R}}=0}^{N} \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, \overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}, I)\, \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}|\, I), \tag{10}$$

with the conditional probability

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, \overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}, I) = \binom{n}{n\bar{\boldsymbol{r}}}\binom{N-n}{N\overline{\boldsymbol{R}} - n\bar{\boldsymbol{r}}}\binom{N}{N\overline{\boldsymbol{R}}}^{-1} =: \mathit{\Pi}(\bar{\boldsymbol{r}}|\, \overline{\boldsymbol{R}}). \tag{11}$$

The conditional probability in the last formula is a hypergeometric distribution $\mathit{\Pi}(\bar{\boldsymbol{r}}|\, \overline{\boldsymbol{R}})$, typical of 'drawing without replacement' problems. The combinatorial proof of the formulae above is in fact the same as for this class of problems [45 ch. 3; 53 § 4.8.3; 54 § II.6]. Our initial symmetric knowledge should intuitively also apply to the sample; indeed, the probability for the state of the sample (10) automatically satisfies the representation theorem (8) as well.

DISCUSSION ABOUT THE SECOND STATE OF KNOWLEDGE IS TO BE DISCARDED

How is it possible that the very different states of knowledge I' and I'' lead to the same formulae above? Their difference appears as soon as we make an experimental observation, say $X_2 = R_2 \in \{0, 1\}$ and update our initial probabilities (8):

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{R}|\, X_2 = R_2, I) \equiv \mathrm{P}(\boldsymbol{X} = \boldsymbol{R}|\, \overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}, X_2 = R_2, I)\, \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}|\, X_2 = R_2, I). \tag{12}$$

The conditional probability and the probability for the average on the right side will update in very different ways for I' and I''. The discussion of this update process in the two cases is unnecessary for the purposes of this note and outside their scope; we hope to address it in full in a future work.

The conditional probability $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, \overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}, I) \equiv \mathit{\Pi}(\bar{\boldsymbol{r}}|\, \overline{\boldsymbol{R}})$ relates the spaces of the sample average $\overline{\boldsymbol{X}} \in \{0, \dots, N\}$ and of the population average $\bar{\boldsymbol{x}} \in \{0, \dots, n\}$. It is a coarsening projector of any probability p for $\overline{\boldsymbol{X}}$ onto a marginal probability, as eq. (10) shows. It also allows us to construct a function of the population average $f^*(\overline{\boldsymbol{X}})$, given a function of the sample average $f(\bar{\boldsymbol{x}})$:

$$f^*(\overline{\boldsymbol{X}}) \coloneqq \sum_{n\bar{\boldsymbol{r}}=0}^{n} f(\bar{\boldsymbol{r}})\mathit{\Pi}(\bar{\boldsymbol{r}}|\, \overline{\boldsymbol{X}}), \tag{13}$$

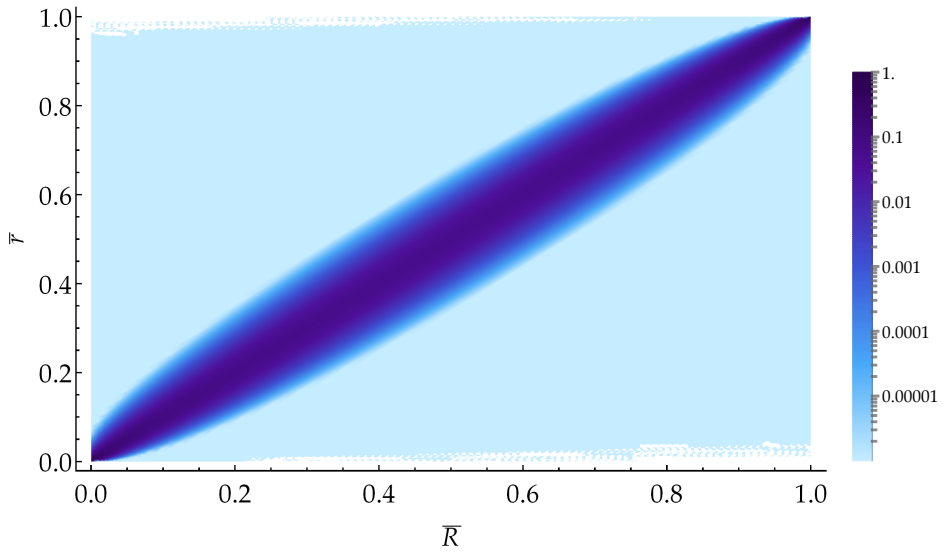

Figure 1: Log-plot of the hypergeometric distribution $\mathit{\Pi}(\bar{\boldsymbol{r}}|\,\bar{\boldsymbol{R}}) := \binom{n}{n\bar{\boldsymbol{r}}}\binom{N-n}{N\bar{\boldsymbol{R}}-n\bar{\boldsymbol{r}}}\binom{N}{N\bar{\boldsymbol{R}}}^{-1}$ for $N = 5000$, $n = 200$. (Band artifacts may appear in the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

with the property of having the same expectation value as the latter:

$$\mathrm{E}[f^*(\bar{\boldsymbol{X}})] \equiv \sum_{N\bar{\boldsymbol{R}}=0}^{N} f^*(\bar{\boldsymbol{R}})\,\mathrm{P}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{R}}|\,I) = \sum_{n\bar{\boldsymbol{r}}=0}^{n} f(\bar{\boldsymbol{r}})\,\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\,I) \equiv \mathrm{E}[f(\bar{\boldsymbol{x}})]. \tag{14}$$

A look at a plot of the hypergeometric distribution $\mathit{\Pi}(\bar{\boldsymbol{r}}|\,\bar{\boldsymbol{R}})$, fig. 1, reveals that it is a sort of 'fuzzy identity matrix' between the $\bar{\boldsymbol{X}}$-space $\{0, \dots, N\}$ and $\bar{\boldsymbol{x}}$-space $\{0, \dots, n\}$. When $n = N$ it is the identity matrix. We thus have that

$$\mathrm{P}(\bar{\boldsymbol{x}} = a) \approx \mathrm{P}(\bar{\boldsymbol{X}} = a), \qquad \mathrm{E}[f(\bar{\boldsymbol{x}})] \approx \mathrm{E}[f(\bar{\boldsymbol{X}})]. \tag{15}$$

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 $\bar{\boldsymbol{X}}$ is bimodal, for example, the bimodality can be lost in the distribution for the sample average $\bar{\boldsymbol{x}}$, owing to the coarsening effect of $\mathit{\Pi}(\bar{\boldsymbol{r}}|\,\bar{\boldsymbol{R}})$.

Yet, the approximate equalities above express the fact that *our uncertainty about the sample is representative of our uncertainty about the population and about other samples*, and vice versa. This fact comes about for very different reasons in the states of knowledge I' and I''. In I', because our sample is *physically* representative of the population and of other samples; we could write this as $\bar{\boldsymbol{x}} \approx \bar{\boldsymbol{X}}$. In I'', because we are ignorant about sample and population in similar symmetric ways; but this does *not* imply that $\bar{\boldsymbol{x}} \approx \bar{\boldsymbol{X}}$. New observations may in fact break this symmetry via eq. (12).

Note that formulae (15) say more than the limits $\mathrm{P}(\bar{\boldsymbol{x}} = a) \to \mathrm{P}(\bar{\boldsymbol{X}} = a)$ and $\mathrm{E}[f(\bar{\boldsymbol{x}})] \to \mathrm{E}[f(\bar{\boldsymbol{X}})]$, as $n \to N$, do. These limits are trivially valid because the sample becomes the full population as $n \to N$. In particular, these limits hold even in cases where the conditional probability $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\,\bar{\boldsymbol{X}} = \bar{\boldsymbol{R}})$ is not a fuzzy identity and our uncertainties about sample and about population can differ wildly.

For functions representing averaged products, $f(\bar{\boldsymbol{x}}) = \overline{\boldsymbol{x}\dots\boldsymbol{x}} \equiv \binom{n\bar{\boldsymbol{x}}}{m}/\binom{n}{m}$, the expectations at the population and at the sample levels are exactly the same, as can be seen from eq. (14):

$$\mathrm{E}(\overline{\boldsymbol{x}\cdots\boldsymbol{x}}|\,I) = \mathrm{E}(\overline{\boldsymbol{X}\cdots\boldsymbol{X}}|\,I) \tag{16}$$

The proof uses the expression for the mth factorial moment of the hypergeometric distribution [55]. Thus, in the states of knowledge I' and I'' 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 $\mathrm{E}(\bar{\boldsymbol{x}}^m)$ and $\mathrm{E}(\bar{\boldsymbol{X}}^m)$, which can be written in terms of the product expectations via eq. (5).

4 Enter maximum-entropy: dilemma

The probability formulae (8)–(11) are constraints on our initial probability assignment, but do not determine it numerically. The probability $\mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}|\, I)$ for the population average needs to be numerically specified, and by marginalization (10) it will determine that of the sample average, $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, I)$. If we numerically specify the latter, the former is not completely specified, because eq. (10) linearly constrains $N+1$ 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 $\mathrm{E}(\bar{\boldsymbol{x}}) = c_1$, $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}}) = c_2$. This is still an underdetermined problem: several distributions can have the same desired expectations, as clear from eqs (16).

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 (13) allow us to apply the method to find the probability of the population $\mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}|\, I)$, or of the sample $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, I)$. *The two applications, however, are inequivalent*. They lead to numerically different distributions for the sample average $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, I)$.

Suppose we want to constrain the sample expectations of a vector function $\boldsymbol{f} = (f_1, \dots, f_m)$ to the vector values $\boldsymbol{c} = (c_1, \dots, c_m)$, that is, $\mathrm{E}[\boldsymbol{f}(\bar{\boldsymbol{x}})] = \boldsymbol{c}$. Application of maximum-entropy [7; 8] at the population level, denoted by I_{p}, gives

$$\mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}|\, I_{\mathrm{p}}) = Z \binom{N}{N\overline{\boldsymbol{R}}} \exp\Bigl[\boldsymbol{\Lambda}^{\intercal} \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \mathit{\Pi}(\bar{\boldsymbol{r}}|\, \overline{\boldsymbol{R}})\Bigr], \tag{17}$$

and then by marginalization (9)

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, I_{\mathrm{p}}) = Z \sum_{N\overline{\boldsymbol{R}}=0}^{N} \mathit{\Pi}(\bar{\boldsymbol{r}}|\, \overline{\boldsymbol{R}}) \binom{N}{N\overline{\boldsymbol{R}}} \exp\Bigl[\boldsymbol{\Lambda}^{\intercal} \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \mathit{\Pi}(\bar{\boldsymbol{r}}|\, \overline{\boldsymbol{R}})\Bigr], \tag{18}$$

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

$$\boldsymbol{c} = Z \sum_{n\bar{\boldsymbol{r}}=0}^{n} \sum_{N\overline{\boldsymbol{R}}=0}^{N} \boldsymbol{f}(\bar{\boldsymbol{r}}) \mathit{\Pi}(\bar{\boldsymbol{r}}|\, \overline{\boldsymbol{R}}) \binom{N}{N\overline{\boldsymbol{R}}} \exp\Bigl[\boldsymbol{\Lambda}^{\intercal} \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \mathit{\Pi}(\bar{\boldsymbol{r}}|\, \overline{\boldsymbol{R}})\Bigr]. \tag{19}$$

Application of maximum-entropy at the sample level, denoted by I_{s}, gives

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, I_{\mathrm{s}}) = \zeta \binom{n}{n\bar{\boldsymbol{r}}} \exp[\boldsymbol{\lambda}^{\intercal} \boldsymbol{f}(\bar{\boldsymbol{r}})] \tag{20}$$

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

$$\boldsymbol{c} = \zeta \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \binom{n}{n\bar{\boldsymbol{r}}} \exp[\boldsymbol{\lambda}^{\intercal} \boldsymbol{f}(\bar{\boldsymbol{r}})]. \tag{21}$$

The probabilities for the sample average obtained from application at the population level (18) and at the sample level (20) should be approximately equal, by our previous observation about representativity (15) and also by the fact that they must satisfy the same expectations for $\boldsymbol{f}$.

Yet they cannot be exactly equal, because their equality would require the Lagrange multipliers $\boldsymbol{\Lambda}$ and $\boldsymbol{\lambda}$ to satisfy the constraint equations (19), (21), and also $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, I_{\mathrm{p}}) = \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\, I_{\mathrm{s}})$; 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 $\boldsymbol{f}$ 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 $N = 5000$, $n = 200$, and constraint functions of the form $\boldsymbol{f}(\bar{\boldsymbol{x}}) = (\bar{\boldsymbol{x}}, \overline{\boldsymbol{x}\boldsymbol{x}}, \dots) \equiv (\bar{\boldsymbol{x}}, \binom{n\bar{\boldsymbol{x}}}{2}/\binom{n}{2}, \dots)$, equivalent to moments constraints. The constraint values used in these examples, reported in the figure captions, have neurobiologically realistic values [31].

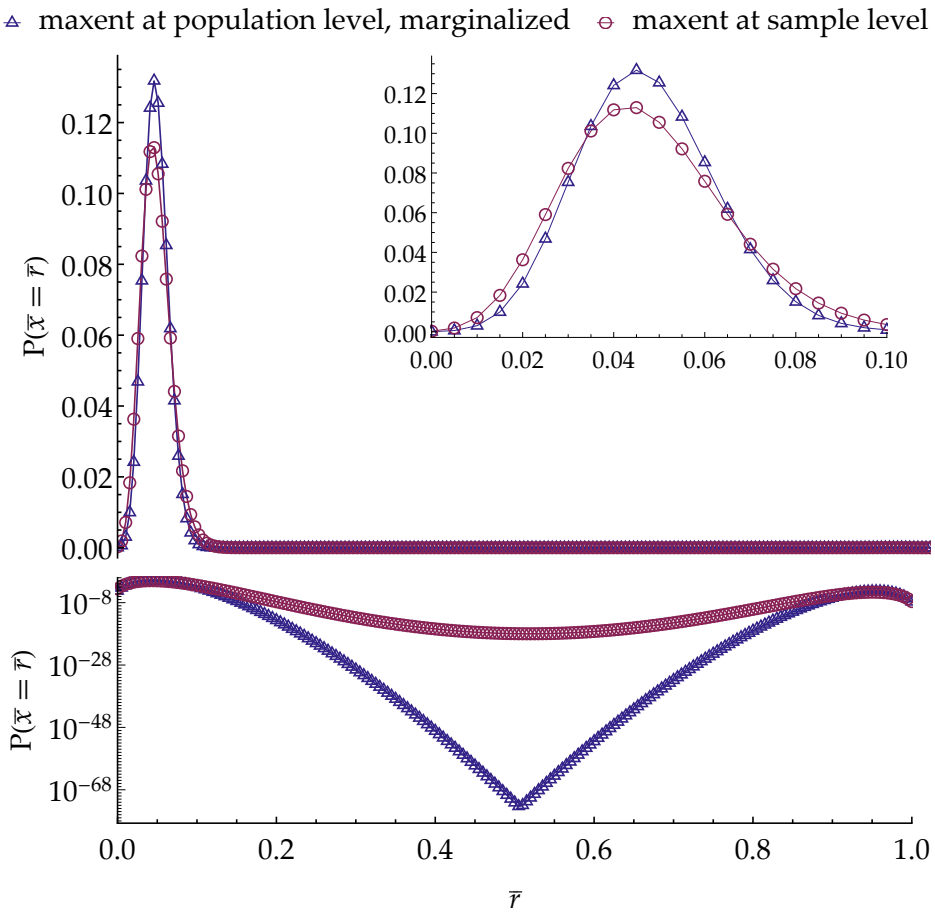

Figure 2: Linear and log-plots of $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (18), and at the sample level (red circles), eq. (20), with $N = 5000$, $n = 200$, constraints $\mathrm{E}(\bar{\boldsymbol{x}}) = 0.0478$, $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}}) = 0.00257$.

In the first example the constraint functions are $\mathrm{E}(\bar{\boldsymbol{x}})$ and $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}})$. 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 $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}})$, $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}})$. 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'?

5 Discussion

needs to be changed

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

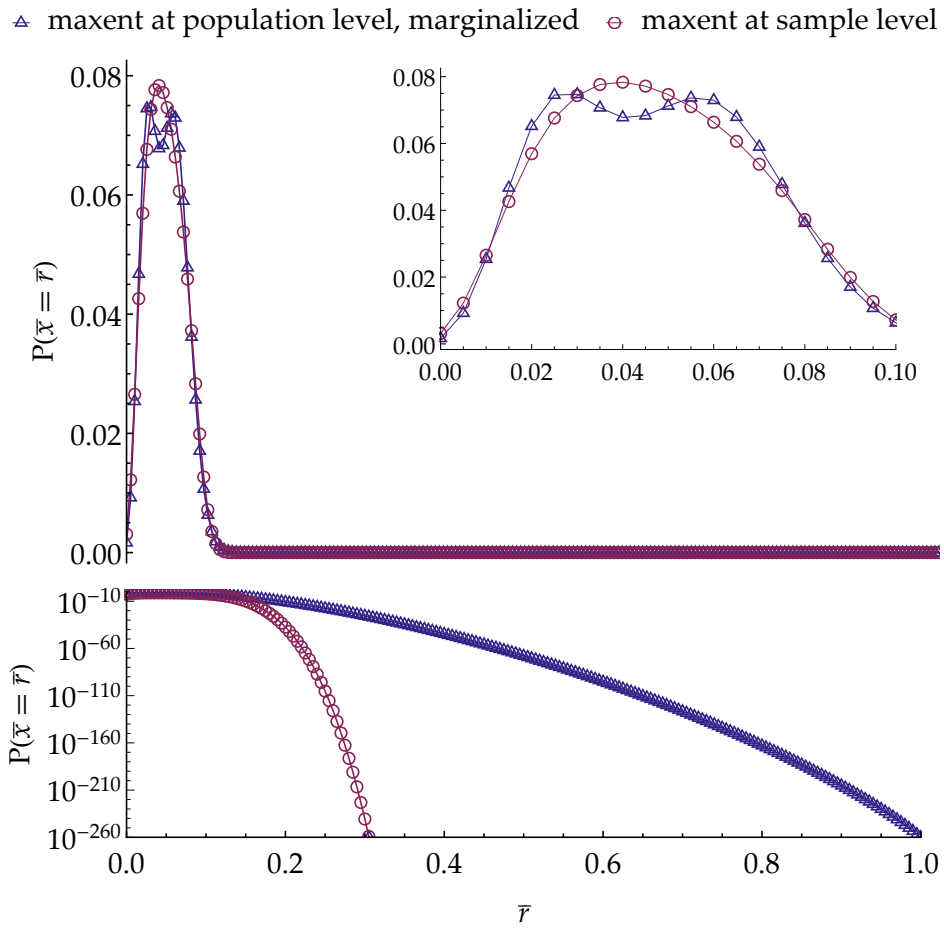

Figure 3: Linear and log-plots of $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (18), and at the sample level (red circles), eq. (20), with $N = 5000$, $n = 200$, constraints $\mathrm{E}(\bar{\boldsymbol{x}}) = 0.0478$, $\mathrm{E}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}) = 0.00257$, $\mathrm{E}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}) = 1.48 \times 10^{-4}$, $\mathrm{E}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}) = 8.81 \times 10^{-6}$.

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 $N = n$ or how large N is otherwise. Call this state of ignorance Υ. In the probability calculus this ignorance about N is expressed by assigning a probability distribution $\mathrm{P}(N|\,\Upsilon)$ that vanishes if $N < n$, since we know that $N \geqslant n$; see Good [56; 57] and Rissanen [58] for examples of such distributions over the integers. Maintaining our assumption of symmetric ignorance, probability assignments that do not assume a specific value of N are then obtained via multiplication of all N-dependent probabilities by $\mathrm{P}(N|\,\Upsilon)$ and subsequent marginalization over N. Technically speaking, N becomes a *nuisance parameter* [45; 59; 60]. The probability obtained from maximum-entropy at the population level, eq. (18), then generalizes to

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\,\Upsilon) = \sum_{N}\Bigl\{Z_N \sum_{N\bar{\boldsymbol{R}}=0}^{N} \mathit{\Pi}_N(\bar{\boldsymbol{r}}|\,\bar{\boldsymbol{R}}) \binom{N}{N\bar{\boldsymbol{R}}} \exp\Bigl[\boldsymbol{\Lambda}_N{}^{\intercal} \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}})\mathit{\Pi}_N(\bar{\boldsymbol{r}}|\,\bar{\boldsymbol{R}})\Bigr]\Bigr\}\mathrm{P}(N|\,\Upsilon), \qquad (22)$$

where N-dependencies have been made explicit. This is a formidable expression. But our question, 'is the usual maximum-entropy at the sample level (20) unassuming with regard to the existence of a larger population?', translates now into the precise mathematical question: 'are the distributions (22)

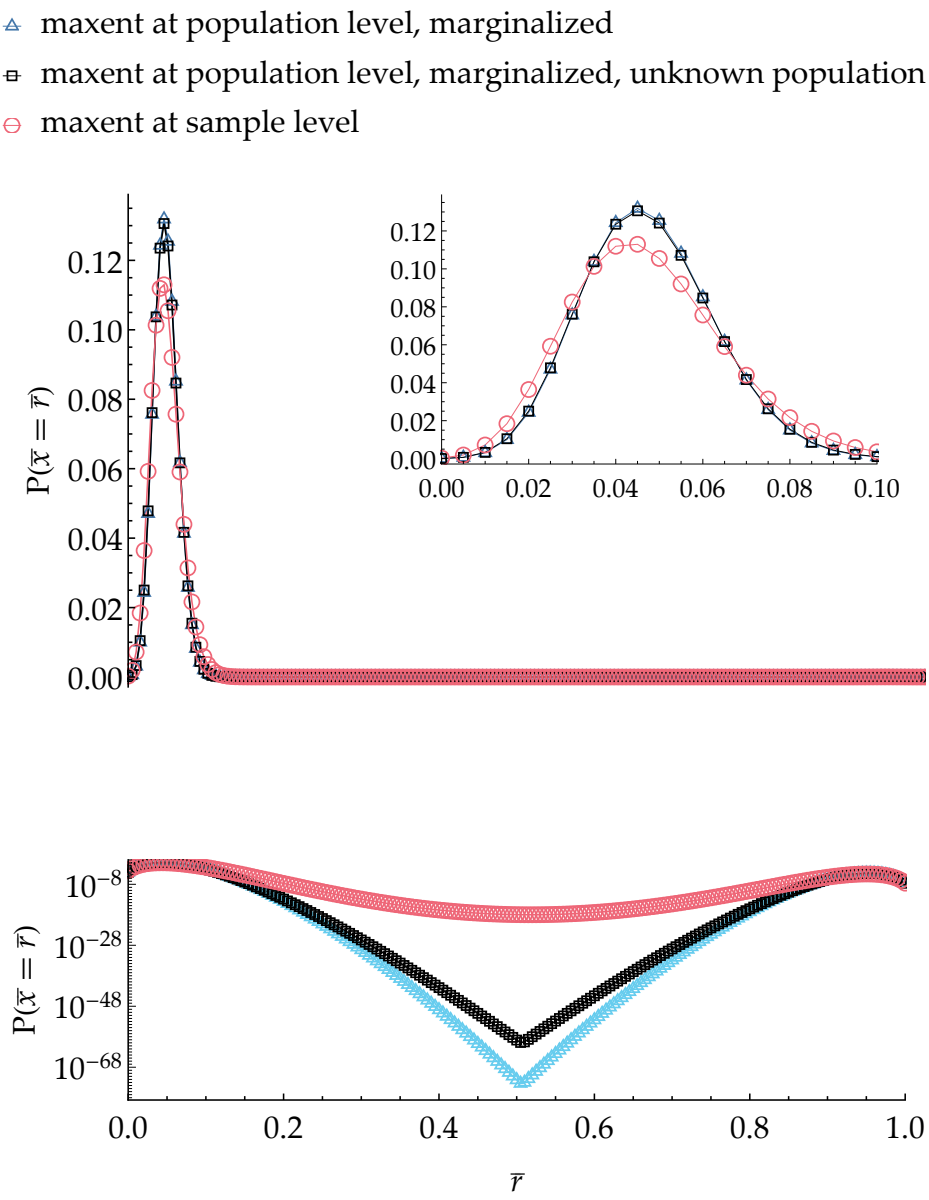

Figure 4: to be written distributions obtained using formulae (22) and the other two

and (20) equal for some choice of $\mathrm{P}(N|\,\Upsilon)$, with $\mathrm{P}(N|\,\Upsilon) \neq 0$ for $N > n$?'. We leave this mathematical problem for future work. Note, however, that this equality is satisfied if $\mathrm{P}(N = 1|\,\Upsilon) = 1$, 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 N, though. We have tacitly intended N 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 $N \sim 10^8$ [61]. For large N – unfortunately we are not yet able to translate this 'large' into a numeric order of magnitude – the final distribution becomes independent of N, 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 [62]. 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 [63 § 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, $\mathrm{E}[f(\bar{\boldsymbol{x}})]$. The corresponding models are usually called 'homogeneous' in the literature. Purely 'inhomogeneous' models have also been used [13; 14; 28], in which expectations for individual neurons or groups of neurons are constrained, for example $\mathrm{E}(x_2)$ or $\mathrm{E}(x_1 x_8 x_9)$. 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 [23; 25].

Acknowledgments

To be added after review.

References

[1] 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.

[2] 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.

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

[4] E. R. Caianiello: *Neuronic equations revisited and completely solved*. In: Palm, Aertsen [64] (1986), 147–160.

[5] 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].

[6] 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.

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

[8] 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.

[9] D. J. C. MacKay: *Maximum entropy connections: neural networks*. In: Grandy, Schick [68] (1991), 237–244.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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].

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

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

[28] 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.

[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. See ref. [70].

[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. T. Jaynes: *Monkeys, kangaroos, and N*. (1996). http://bayes.wustl.edu/etj/node1.html. First publ. 1986. (Errata: in equations (29)–(31), (33), (40), (44), (49) the commas should be replaced by gamma functions, and on p. 19 the value 0.915 should be replaced by 0.0915).

[33] J. Skilling: *Classic maximum entropy*. In: Skilling [71] (1989), 45–52.

[34] J. Skilling: *Quantified maximum entropy*. In: Fougère [72] (1990), 341–350.

[35] C. C. Rodríguez: *The metrics induced by the Kullback number*. In: Skilling [71] (1989), 415–422.

[36] C. C. Rodríguez: *Entropic priors for discrete probabilistic networks and for mixtures of Gaussians models*. Am. Inst. Phys. Conf. Proc. **617** (2002), 410–432.

[37] P. G. L. Porta Mana: *Maximum-entropy from the probability calculus: exchangeability, sufficiency*. (2017). doi:10.17605/osf.io/xdy72, arXiv:1706.02561.

[38] E. Schrödinger: *Statistical Thermodynamics: A Course of Seminar Lectures*. Cambridge University Press, London (1948). First publ. 1946.

[39] E. T. Jaynes: *Where do we stand on maximum entropy?* In: Levine, Tribus [73] (1979), 15–118. http://bayes.wustl.edu/etj/node1.html; repr. with an introduction in ref. [67], pp. 210–314.

[40] E. T. Jaynes: *Prior information in inference*. (1982). http://bayes.wustl.edu/etj/node2.html.

[41] E. T. Jaynes: *On the rationale of maximum-entropy methods*. Proc. IEEE **70**[9] (1982), 939. http://bayes.wustl.edu/etj/node1.html.

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

[43] 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).

[44] *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.

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

[47] B. de Finetti: *La probabilità e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista*. In: de Finetti [74] (1959), 1–115. Transl. in ref. [75], ch. 9, pp. 147–227.

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

[49] 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.

[50] 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.

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

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

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

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

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

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

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

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

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

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

[61] 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.

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

[63] E. T. Jaynes: *Inferential scattering*. (1993). http://bayes.wustl.edu/etj/node1.html. Extensively rewritten version of a paper first publ. 1985 in ref. [76], pp. 377–398.

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

[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. [5].

[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. [14].

[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. [29].

[71] J. Skilling, ed.: *Maximum Entropy and Bayesian Methods: Cambridge, England, 1988*. Kluwer, Dordrecht (1989).

[72] P. F. Fougère, ed.: *Maximum Entropy and Bayesian Methods: Dartmouth, U.S.A., 1989*. Kluwer, Dordrecht (1990).

[73] R. D. Levine, M. Tribus, eds.: *The Maximum Entropy Formalism: A Conference Held at the Massachusetts Institute of Technology on May 2–4, 1978*. MIT Press, Cambridge, USA (1979).

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

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

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