Representative random samples and maximum-entropy distributions: a dilemma

P.G.L. Porta Mana[*]
Independent researcher
pgl@portamana.org

V. Rostami[*]
Forschungszentrum Jülich INM-6
Germany
v.rostami@fz-juelich.de

E. Torre
ETH Zürich
Switzerland
torre@ibk.baug.ethz.ch

Abstract

This note has three nested purposes. The first purpose is to show that the maximum-entropy method can be applied to a representative random sample of a population, to generate its probability distribution, along two different routes. Both routes appear legitimate, but they give inequivalent results. Which route should be chosen? Some arguments are presented in favour of one. The second more general purpose, motivated by the above dilemma, is to remind readers that models like maximum-entropy may contain hidden assumptions; in this case the hidden an unnatural assumption that the sample modelled is isolated from the rest of the population. The third purpose is to promote some old but possibly forgotten probability formulae that may be useful in neuroscientific sampling contexts.

1 Introduction: maximum-entropy and sampling in neuroscience

This note is mainly addressed to neuroscientists interested in maximum-entropy methods, but we would be pleased if its discussion of the probability of "random sampling" were useful to neuroscientists that use other statistical methods to study the physical and dynamical characteristics of brain areas via neuronal recording.

Recent electrophysiological techniques [***] in fact allow experimenters to record samples of even a couple hundreds neurons from specific brain areas. These samples are usually picked out according to an unknown process, but from their observation we expect to learn something about all other neurons in the same brain area. That is, we assumed that they are a "representative random sample".

We don't elaborate on the various important purposes of such recordings here [**?*****], but stress that these sample sizes can be considered "large", because their statistical analysis requires considerable computational power.

These computational costs are one, probably not the earliest, of the many reasons why maximum-entropy methods have been introduced in neuroscience. It would be useful to somehow compress the statistical wealth of large neuronal recordings into few quantities, like sample moments for example. This compression would also entail interesting physico-biological properties of neuronal activity. The standard maximum-entropy method [1–3] accomplishes this kind of compression: it associates a unique probability distribution with few experimental quantities [***]. But this is only one of its uses. It is also used for various information-theoretic purposes or to generate reference probability distributions [***]. In all these uses the maximum-entropy distribution is chosen as the "maximally noncommittal" one [4]. This adjective means little without further technical characterizations. Different works give different characterizations, but in this paper we will use the

[*]Equally contributing first authors

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

quoted expression as an umbrella term for all of them. Our results will not depend on the specific characterization of "noncommittal".

In this note we want to show that we have to face a dilemma if we want to apply the maximum-entropy method to a representative random sample to find a maximally noncommittal distribution.

The dilemma is this. We can apply the maximum-entropy method to the sample, using a specified set of experimental constraints, and generate a probability distribution for its state. But our sample is representative of a larger population. We can apply the maximum-entropy method to the larger population, using the same constraints, and generate a probability distribution for its larger state, and then find the distribution for the sample by marginalization. Either application seems to have some commendable features. However, *the distributions obtained by these two applications differ*. It goes without saying that if one is "maximally noncommittal", the other must be somehow "committal". Which choice is most meaningful?

In the rest of the paper we mathematically formulate this dilemma. To this purpose we also present some probability relations relevant to "random sampling". These relations 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.

In the final discussion we present some arguments in favour of one choice in our dilemma, but we do not mean that to be our final answer. This note, and its dilemma, also have the more general purpose of reminding our readers the importance of asking *what is the question we're trying to answer with probability theory?* and *which assumptions do we want or need to make?*

Our mathematical analysis pertains to a neurons modelled as binary units at one specific instant or short window in time. It is possible to similarly discuss multi-state neuron models and population dynamics; but for simplicity neurons are here assumed to be in a fixed, active "1" or inactive "0" state; and evolution, change, time correlations, and similar concepts do not concern us. We consider maximum-entropy models that have constraints based on some kinds of sample or population averages; they are often called "homogeneous". The final discussion touches upon "inhomogeneous" models as well.

The notation in this note follows ISO and ANSI standards [5–7] but for the use of the comma "," to denote logical conjunction. Probability notation follows Jaynes [8]. By "probability" we mean plausibility or the degree of belief which "would be agreed by all rational men if there were any rational men" [9].

2 Setup

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 $(r_1, \dots, r_n) =: \boldsymbol{r}$. We will consider various averages of the population and the sample. For this purpose we introduce a general averaging operator $\overline{}$ defined by

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

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

$$\underbrace{\overline{\boldsymbol{X}\cdots\boldsymbol{X}}}_{m \text{ factors}} = \binom{N}{m}^{-1}\binom{N\overline{\boldsymbol{X}}}{m}. \tag{2}$$

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

$$\mathrm{P}(X_1 = R_1, X_2 = R_2, \dots, X_N = R_N \mid K) \quad \text{or} \quad \mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid K), \quad \boldsymbol{R} \in \{0,1\}^N, \tag{3}$$

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

$$\mathrm{P}(x_1 = r_1, x_2 = r_2, \dots, x_n = r_n \mid K) \quad \text{or} \quad \mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid K), \quad \boldsymbol{r} \in \{0,1\}^n. \tag{4}$$

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. We would also like our probability assignment to reflect that the sample is somehow "representative" of the population.

Let's assume that we initially know very little about the physical details of the individual neurons; their locations for example. Our initial state of knowledge or ignorance I is therefore symmetric, or "exchangeable", under their permutations. This symmetry must be reflected in our initial probability: the *representation theorem for finite exchangeability* states that it must obey

$$\mathrm{P}(\boldsymbol{X} = \boldsymbol{R} \mid I) = \binom{N}{N\overline{\boldsymbol{R}}}^{-1} \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I), \tag{5}$$

the latter being the probability for the population average $\boldsymbol{X}$. Proof of this theorem and generalizations to non-binary and continuum cases are given by de Finetti [10], Ericson [11], Diaconis [12], Heath & Sudderth [13]. 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:

$$\mathrm{P}(\boldsymbol{x} = \boldsymbol{r} \mid I) = \binom{n}{n\bar{\boldsymbol{r}}}^{-1} \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I), \tag{6}$$

with

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I) = \sum_{N\overline{\boldsymbol{R}}=0}^{N} \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid \overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}, I)\, \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I), \tag{7}$$

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid \overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}, I) = \binom{n}{n\bar{\boldsymbol{r}}}\binom{N-n}{N\overline{\boldsymbol{R}} - n\bar{\boldsymbol{r}}}\binom{N}{N\overline{\boldsymbol{R}}}^{-1} =: \mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}}). \tag{8}$$

Our initial symmetric ignorance should intuitively also apply to the sample; indeed, the probability for the state of the sample (7) automatically satisfies the representation theorem (5) as well. The conditional probability in the last formula is a hypergeometric distribution $\mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})$, typical of "drawing without replacement" problems. The combinatorial proof of the formulae above is in fact the same as for this class of problems [8 ch. 3; 14 § 4.8.3; 15 § II.6].

The conditional probability $\mathit{\Pi}(\bar{\boldsymbol{r}}, \overline{\boldsymbol{R}})$ relates the spaces of the sample average $\overline{\boldsymbol{X}} \in \{0, \dots, N\}$ and of the population average $\bar{\boldsymbol{x}} \in \{0, \dots, n\}$ in a special way. It is a coarsening projector of any probability p for $\overline{\boldsymbol{X}}$ onto a marginal probability p_* for $\bar{\boldsymbol{x}}$:

$$p_*(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}) = \sum_{N\overline{\boldsymbol{R}}=0}^{N} \mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}}) p(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}). \tag{9}$$

Conversely it also a pulls back expectations of functions f of the sample average $\bar{\boldsymbol{x}}$ to expectations of functions f^* of the population average $\overline{\boldsymbol{X}}$:

$$\begin{aligned} f^*(\overline{\boldsymbol{X}}) &:= \sum_{n\bar{\boldsymbol{r}}=0}^{n} f(\bar{\boldsymbol{r}}) \mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{X}}), \\ \mathrm{E}[f(\bar{\boldsymbol{x}})] = \mathrm{E}[f^*(\overline{\boldsymbol{X}})] &= \sum_{n\bar{\boldsymbol{r}}=0}^{n} f(\bar{\boldsymbol{r}})\, \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I) = \sum_{N\overline{\boldsymbol{R}}=0}^{N} f^*(\overline{\boldsymbol{R}})\, \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I). \end{aligned} \tag{10}$$

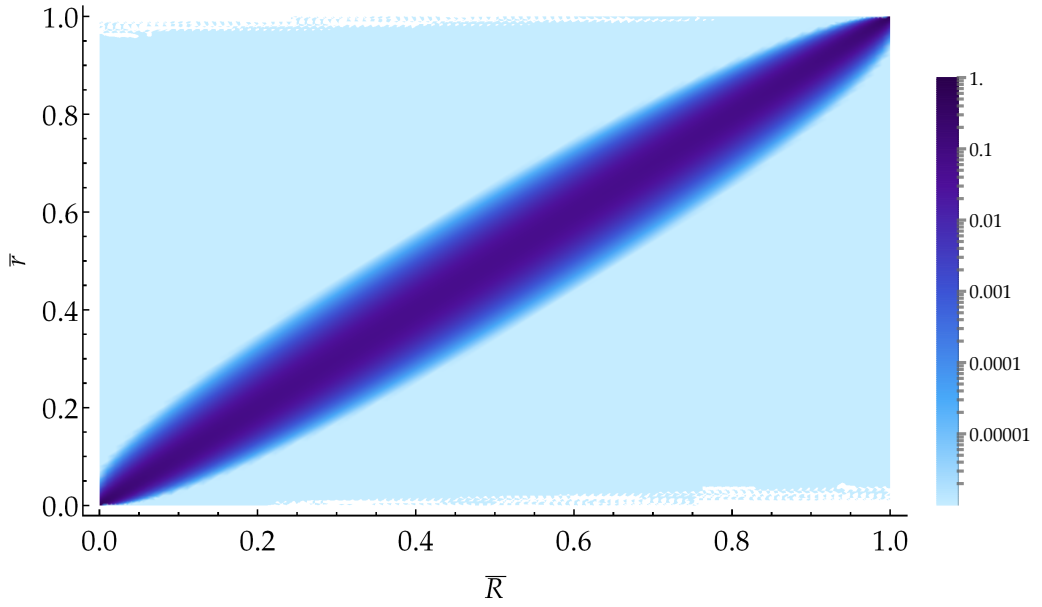

Figure 1: Log-plot of the hypergeometric distribution $\mathit{\Pi}(\bar{\boldsymbol{r}}\,|\,\bar{\boldsymbol{R}}) \coloneqq \binom{n}{n\bar{\boldsymbol{r}}}\binom{N-n}{N\bar{\boldsymbol{R}}-n\bar{\boldsymbol{r}}}\binom{N}{N\bar{\boldsymbol{R}}}^{-1}$ for $N = 5000$, $n = 200$. (Band artifacts may appear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

A look at a plot of the hypergeometric distribution $\mathit{\Pi}(\bar{\boldsymbol{r}}\,|\,\bar{\boldsymbol{R}})$, see 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

$$\mathrm{P}(\bar{\boldsymbol{x}} = a) \approx \mathrm{P}(\bar{\boldsymbol{X}} = a), \qquad \mathrm{E}[f(\bar{\boldsymbol{x}})] \approx \mathrm{E}[f(\bar{\boldsymbol{X}})]. \tag{11}$$

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 is how the idea of representativeness is translated into our probabilities. The symmetry present in our initial probabilities is not a physical property of the neuronal population or sample. It only expresses the symmetry of our initial uncertainty about them, and does not imply any sort of physical similarity between the neurons. Subsequent observations may in fact break this symmetry. Upon observation of a sample average, say $\bar{\boldsymbol{x}}\bar{\boldsymbol{x}} = a$, the updated expectations for such average in the population and in any new sample will usually be shifted towards the observed value, as follows from Bayes's theorem and the formulae above.

Note that formulae (11) say more than the limits $\mathrm{P}(\bar{\boldsymbol{x}} = a) \to \mathrm{P}(\bar{\boldsymbol{X}} = a)$, $\mathrm{E}[f(\bar{\boldsymbol{x}})] \to \mathrm{E}[f(\bar{\boldsymbol{X}})]$ as $n \to N$. 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, $f(\bar{\boldsymbol{x}}) \coloneqq \overline{\boldsymbol{x}\dots\boldsymbol{x}} = \binom{n\bar{\boldsymbol{x}}}{m}/\binom{n}{m}$, formulae (10) have the useful form

$$\Big(\underbrace{\overline{\boldsymbol{x}\cdots\boldsymbol{x}}}_{m\text{ factors}}\Big)^* = \underbrace{\overline{\boldsymbol{X}\cdots\boldsymbol{X}}}_{m\text{ factors}},$$

$$\mathrm{E}(\overline{\boldsymbol{x}\cdots\boldsymbol{x}}\,|\,I) = \mathrm{E}(\overline{\boldsymbol{X}\cdots\boldsymbol{X}}\,|\,I) = \binom{n}{m}^{-1}\sum_{n\bar{\boldsymbol{r}}=0}^{n}\binom{n\bar{\boldsymbol{r}}}{m}\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}\,|\,I) = \binom{N}{m}^{-1}\sum_{N\bar{\boldsymbol{R}}=0}^{N}\binom{N\bar{\boldsymbol{R}}}{m}\mathrm{P}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{R}}\,|\,I). \tag{12}$$

The proof uses the expression for the mth factorial moment of the hypergeometric distribution [16]. Thus, the averages of activity products *are 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 $\mathrm{E}(\overline{\boldsymbol{X}}^m)$, which can be written in terms of the product expectations via eq. (2).

4 Enter maximum-entropy: dilemma

Formulae (5)–(8) are constraints on our initial probability assignment, but do not determine it numerically. The probability $\mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I)$ for the population average needs to be numerically specified, and by (7) it will determine that of the sample average, $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I)$. If we numerically specify the latter, the former is not completely specified, because eq. (7) linearly constrains $N+1$ unknowns by only $n+1$ equations in this case.

We may want to specify the probability by enforcing the sample expectations of several functions to have specific values, for example $\mathrm{E}(\bar{\boldsymbol{x}}) = c_1$, $\mathrm{E}(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}) = c_2$. This is still an underdetermined linear problem: several distributions can have the same desired expectations, as clear from eqs (12).

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: formulae (10) allow us to apply the method to find the probability of the population $\mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I)$, or of the sample $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I)$. *The two applications, however, are inequivalent*. They lead to numerically different $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I)$.

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 at the population level, denoted by I', gives

$$\mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I') = K \binom{N}{N\overline{\boldsymbol{R}}} \exp\Big[\boldsymbol{\Lambda}^\intercal \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \Pi(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})\Big], \tag{13}$$

and then by marginalization with eq. (6)

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I') = K \sum_{N\overline{\boldsymbol{R}}=0}^{N} \Pi(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}}) \binom{N}{N\overline{\boldsymbol{R}}} \exp\Big[\boldsymbol{\Lambda}^\intercal \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \Pi(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})\Big], \tag{14}$$

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

$$\boldsymbol{c} = K \sum_{n\bar{\boldsymbol{r}}=0}^{n} \sum_{N\overline{\boldsymbol{R}}=0}^{N} \boldsymbol{f}(\bar{\boldsymbol{r}}) \Pi(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}}) \binom{N}{N\overline{\boldsymbol{R}}} \exp\Big[\boldsymbol{\Lambda}^\intercal \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \Pi(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})\Big]. \tag{15}$$

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

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I'') = \kappa \binom{n}{n\bar{\boldsymbol{r}}} \exp[\boldsymbol{\lambda}^\intercal \boldsymbol{f}(\bar{\boldsymbol{r}})] \tag{16}$$

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

$$\boldsymbol{c} = \kappa \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \binom{n}{n\bar{\boldsymbol{r}}} \exp[\boldsymbol{\lambda}^\intercal \boldsymbol{f}(\bar{\boldsymbol{r}})]. \tag{17}$$

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 (11) 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 equations (15), (17), and $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I') = \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I'')$ – that is, $2m+n$ equations in m unknowns. This could be possible, 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.

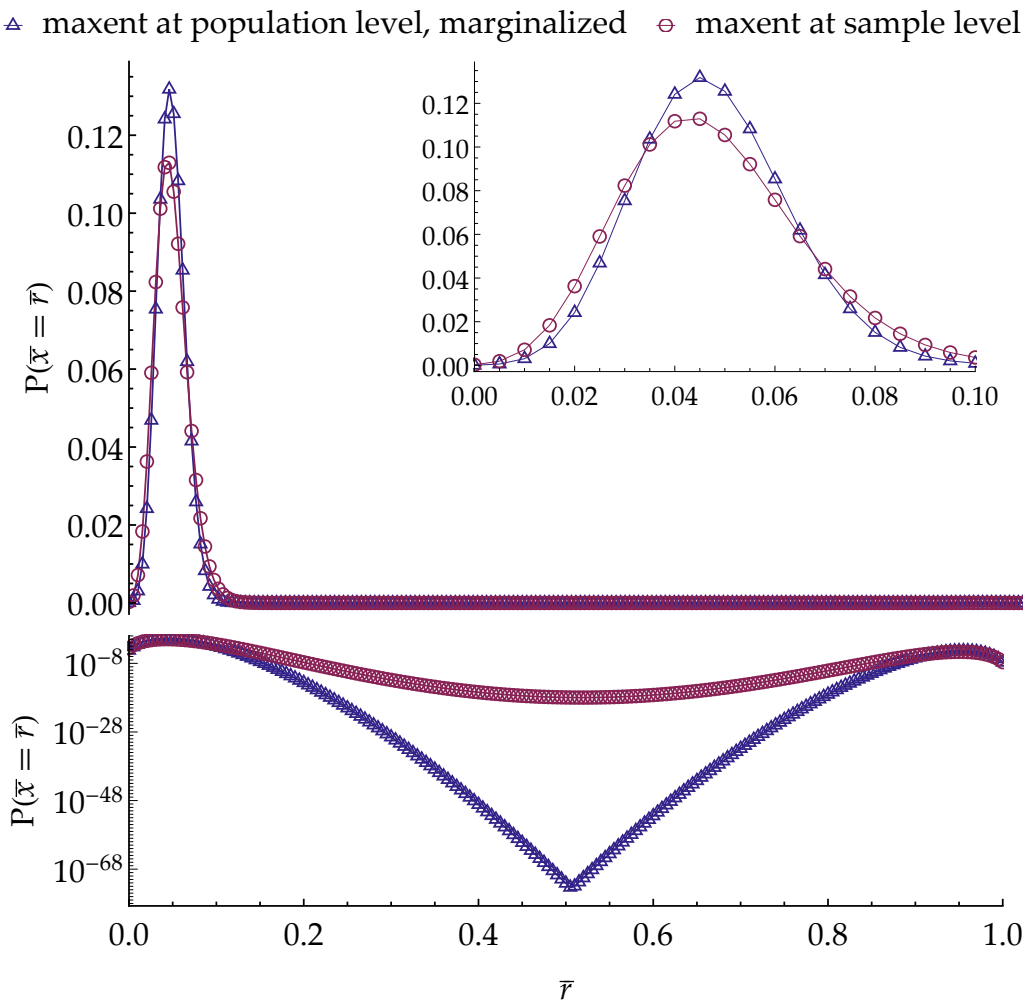

Figure 2: Linear and log-plots of $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (14), and at the sample level (red circles), eq. (16), with $N = 5000$, $n = 200$, constraints $\mathrm{E}(\bar{\boldsymbol{x}})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

We show two examples of this discrepancy in figs 2 and 3, for $N = 5000$, $n = 200$, and constraint functions of the form $\boldsymbol{f}(\bar{\boldsymbol{x}}) = (\bar{\boldsymbol{x}}, \overline{\boldsymbol{x}\boldsymbol{x}}, \dots) \equiv \left(\bar{\boldsymbol{x}}, \binom{n\bar{\boldsymbol{x}}}{2}/\binom{n}{2}, \dots\right)$. In the first example the constraints are $\mathrm{E}(\bar{\boldsymbol{x}}) = c_1$ and $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}}) = c_2$, with $c_1 = 0.0478$ and $c_2 = 0.00257$. 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 includes two additional constraints $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}}) = c_3$, $\mathrm{E}(\overline{\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}\boldsymbol{x}}) = c_4$ with $c_3 = 0.000148$, $c_4 = 8.81 \times 10^{-6}$. 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. The constraints used in these examples have neurobiologically realistic values [17].

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

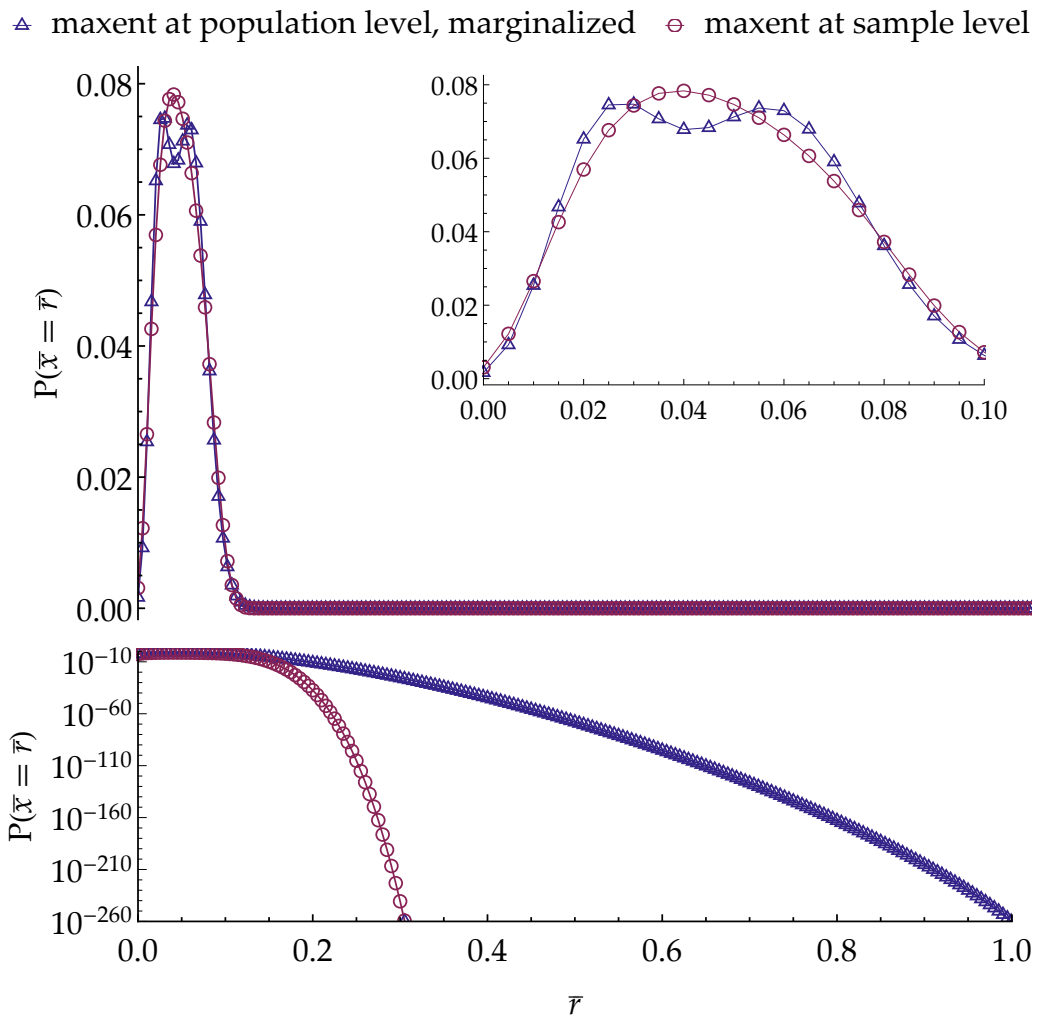

Figure 3: Linear and log-plots of $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (14), and at the sample level (red circles), eq. (16), with $N = 5000$, $n = 200$, constraints $\mathrm{E}(\bar{\boldsymbol{x}})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

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 first purpose of this note is indeed to remind ourselves that probability models require careful scrutiny, because they can rest on hidden assumptions that we don't want to make or that contradict others we are making, and we may not be aware of this. The dilemma generated by the maximum-entropy method and the assumption of representative sampling is an example.

The tricky point is this. Maximum-entropy applied at the population level and applied at the sample level give different results; they are 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 is it actually assuming that *no* larger population exists? In the latter case it would not be the correct model to use in our problem.

A perfunctory intuitive reasoning seems insufficient for clarifying this point. Let's express it via the probability calculus. Suppose we do not know whether the sample is really part of a larger population; we do not know if $N = n$ or how large is N otherwise. Call this state of ignorance Υ. From the point of view of 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 $N \geqslant n$. Maintaining our assumption of symmetric ignorance, probability assignments that do not assume a specific value of N are then obtained by multiplication of all N-dependent probabilities by $\mathrm{P}(N|\,\Upsilon)$ and subsequent marginalization over N – technically speaking, N is a *nuisance parameter* that we eliminate. The probability obtained from maximum-entropy at the population level, eq. (14) then generalizes to

$$\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\,\Upsilon) = \sum_N \left\{ K_N \sum_{N\bar{\boldsymbol{R}}=0}^{N} \mathit{\Pi}_N(\bar{\boldsymbol{r}}|\,\bar{\boldsymbol{R}}) \binom{N}{N\bar{\boldsymbol{R}}} \exp\Bigl[\boldsymbol{\Lambda}_N{}^{\intercal} \sum_{n\bar{\boldsymbol{r}}=0}^{n} \boldsymbol{f}(\bar{\boldsymbol{r}}) \mathit{\Pi}_N(\bar{\boldsymbol{r}}|\,\bar{\boldsymbol{R}})\Bigr] \right\} \mathrm{P}(N|\,\Upsilon). \tag{18}$$

This is a formidable expression. But the answer to our question, whether the usual maximum-entropy at the sample level (16) does not assume anything about a larger population, translates into the precise mathematical question: are the distributions (18) and (16) equal, for some choice of $\mathrm{P}(N|\,\Upsilon)$? We leave this mathematical problem for future work.

In a neuroscientific context we find this assumption very natural, and therefore the maximum-entropy method at the population level preferable. After all, also *physical* neuronal-network models usually include some sort of external input to the neurons, mimicking their embedding in a larger network.

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 [18]. 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 [19 § 4]. The difference in the two descriptions appears 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 cases considered in this note the difference in predictions may be relevant, and the ones made by maximum-entropy at the population level are more informed.

Importance of the formulae relating sample and population.

Discussion of case when individual $X_i \cdots X_j$ are constrained. Likely breakdown of maximum-entropy in this case [20]. Hints at full Bayesian treatment of which maximum-entropy is a limit? [21 ***]

Acknowledgments

To be added after review.

References

[1] 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. [22].

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

[3] 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.

[4] E. T. Jaynes: *Information theory and statistical mechanics*. In: Ford [23] (1963), 181–218. Repr. in ref. [24 ch. 4, pp. 39–76]; http://bayes.wustl.edu/etj/node1.html.

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

[6] *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).

[7] *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.

[8] E. T. Jaynes: *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge (2003). Ed. by G. Larry Bretthorst; http://omega.albany.edu:8008/JaynesBook.html, http://omega.albany.edu:8008/JaynesBookPdf.html, http://www-biba.inrialpes.fr/Jaynes/prob.html; first publ. 1994.

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

[10] B. de Finetti: *La probabilitá e la statistica nei rapporti con l'induzione, secondo i diversi punti di vista*. In: de Finetti [25] (1959), 1–115. Transl. as ref. [26].

[11] 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.

[12] 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.

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

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

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

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

[17] V. Rostami, P. G. L. Porta Mana, M. Helias: *Pairwise maximum-entropy models and their Glauber dynamics: bimodality, bistability, non-ergodicity problems, and their elimination via inhibition*. (2016). arXiv:1605.04740. Accepted for publ. in PloS CB.

[18] C. Maes, F. Redig, A. Van Moffaert: *The restriction of the Ising model to a layer*. J. Stat. Phys. **96**[1] (1999), 69–107. arXiv:math/9810094.

[19] E. T. Jaynes: *Inferential scattering*. (1993). http://bayes.wustl.edu/etj/node1.html; extensively rewritten version of a paper first publ. 1985 in ref. [27], pp. 377–398.

[20] P. G. L. Porta Mana: *On the relation between plausibility logic and the maximum-entropy principle: a numerical study*. (2009). arXiv:0911.2197. Also presented as invited talk at the 31st International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering 'MaxEnt 2011', Waterloo, Canada.

[21] D. J. C. MacKay: *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge (2003). http://www.inference.phy.cam.ac.uk/mackay/itila/; first publ. 1995.

[22] 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. [1].

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

[24] 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.

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

[26] B. de Finetti: *Probability, statistics and induction: their relationship according to the various points of view*. In: de Finetti [28] (1972), p. 9, 147–227. Transl. of ref. [10].

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

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