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

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 [**berenyietal2014**]. We bin and binarize the recorded activity into sequences of 0s (inactive) and 1s (firing) [**caianiello1961**; **caianiello1986**]. We notice that groups of two or more 1s can appear in any time bin. We may ask several questions looking at these groups. A recurring question is 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? This kind of question is very vague until we define what we mean with 'explain' or similar words.

A possible, precise definition is as follows. We want to predict, i.e. assign a probability to, the population activity in a time bin, given the population activities in the other bins:

$$\mathrm{P}(\text{activity at one time bin}\,|\,\text{activities at the remaining bins, and other info}). \tag{1}$$

We can say that the simultaneous firing of at most two neurons 'explains' the activity if only the values of the first and second empirical moments of the activities are relevant for the probability above:

$$\mathrm{P}(\text{activity at one time bin}\,|\,\text{activities at the other bins, and other info}) = \mathrm{P}(\text{activity at some time bin}\,|\,\text{1st \& 2nd empirical moments, and other info}); \tag{2}$$

in statistical terms, the first and second moments are *sufficient statistics*. Other definitions are possible, of course; the definition above has the advantage of being mathematically very precise. In fact it determines, from just the three basic laws of the probability calculus, the mathematical form of the probability for the activity of any number of bins:

$$\mathrm{P}(\text{activity at bins } 1, 2, \ldots \,|\, \text{other info}) = \int \mathrm{p}(\text{activity at bin } 1\,|\,\theta)\,\mathrm{p}(\text{activity at bin } 2\,|\,\theta)\cdots\mathrm{p}(\theta\,|\,\text{other info})\,\mathrm{d}\theta,$$

$$\text{with 'p(activity at time bin } i\,|\,\theta)\text{' having maximum-entropy form.} \tag{3}$$

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

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

Recent electrophysiological techniques [**berenyietal2014**] in fact allow experimenters to record samples comprising even a couple hundreds neurons from specific brain areas. From the observation of these samples we expect to learn something about all other neurons in the same brain area. That is, we assume that they are a 'representative sample'. We do not 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 might also entail interesting biological or functional properties of neuronal activity. The standard maximum-entropy method [**jaynes1957**; **sivia1996_r2006**; **meadetal1984**] 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 [**bohteetal2000**; **schneidmanetal2006**; **shlensetal2006**; **roudietal2009b**; **mackeetal2011**; **tkaciketal2014b**; **shimazakietal2015**]. In all these uses the maximum-entropy distribution is chosen as the 'maximally noncommittal' one [**jaynes1963**]. This adjective means little without further technical characterizations. Different works give different characterizations, but in this paper we will use the 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 show that we must face a dilemma if we want to apply the maximum-entropy method to a representative 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 the state of the sample. But our sample is representative of a larger population, in which it is biologically and functionally embedded. 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 from these two applications differ*. It goes without saying that if one is 'noncommittal', the other must be 'committal'. Which choice is most meaningful?

In the final discussion we present some arguments in favour of one choice.

In the rest of the paper we mathematically formulate this dilemma. 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 briefly discuss also two very different meaning of 'representative' that lead to similar initial probability assignments for sampling.

Our mathematical analysis pertains to 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 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 [**iso1993**; **ieee1993**; **nist1995**] but for the use of the comma ',' to denote logical conjunction. Probability notation follows Jaynes [**jaynes1994_r2003**]. By 'probability' we mean a degree of belief which 'would be agreed by all rational men if there were any rational men' [**good1966**].

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.

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

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

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 [**definetti1959b**], Kendall [**kendall1967**], Ericson [**ericson1976**], Diaconis & Freedman [**diaconis1977**; **diaconisetal1980**],

Heath & Sudderth [**heathetal1976**]. 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:

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

with

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

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

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 [**jaynes1994_r2003**; **ross1976_r2010**; **feller1950_r1968**]. 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.

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):

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

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 $\mathit{\Pi}(\bar{\boldsymbol{r}}, \overline{\boldsymbol{R}})$ 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 $\bar{\boldsymbol{x}}$:

$$p_*(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}) = \sum_{N\overline{\boldsymbol{R}}=0}^{N} \mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}}) p(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}}). \tag{13}$$

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 $\overline{\boldsymbol{X}}$:

$$\begin{aligned} f^*(\overline{\boldsymbol{X}}) &:= \sum_{n\bar{\boldsymbol{r}}=0}^{n} f(\bar{\boldsymbol{r}}) \mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{X}}), \\ \mathrm{E}[f(\bar{\boldsymbol{x}})] = \mathrm{E}[f^*(\overline{\boldsymbol{X}})] &= \sum_{n\bar{\boldsymbol{r}}=0}^{n} f(\bar{\boldsymbol{r}})\, \mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}} \mid I) = \sum_{N\overline{\boldsymbol{R}}=0}^{N} f^*(\overline{\boldsymbol{R}})\, \mathrm{P}(\overline{\boldsymbol{X}} = \overline{\boldsymbol{R}} \mid I). \end{aligned} \tag{14}$$

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

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

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 $\overline{\boldsymbol{X}}$ 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 $\mathit{\Pi}(\bar{\boldsymbol{r}} \mid \overline{\boldsymbol{R}})$.

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

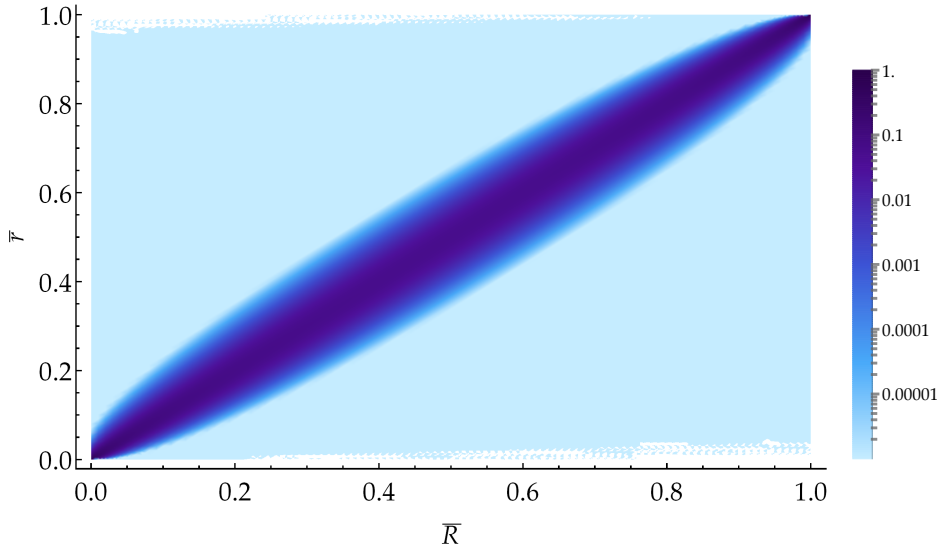

Figure 1: Log-plot of the hypergeometric distribution $\mathit{\Pi}(\bar{\boldsymbol{r}}|\,\bar{\boldsymbol{R}}) := \binom{n}{n\bar{\boldsymbol{r}}}\binom{N-n}{N\bar{\boldsymbol{R}}-n\bar{\boldsymbol{r}}}\binom{N}{N\bar{\boldsymbol{R}}}^{-1}$ for $N = 5000$, $n = 200$. (Band artifacts may appear in the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

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, $f(\bar{\boldsymbol{x}}) = \overline{\boldsymbol{x}\dots\boldsymbol{x}} \equiv \binom{n\bar{\boldsymbol{x}}}{m}/\binom{n}{m}$, formulae (14) have the useful form

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

$$\mathrm{E}(\overline{\boldsymbol{x}\cdots\boldsymbol{x}}|\,I) = \mathrm{E}(\overline{\boldsymbol{X}\cdots\boldsymbol{X}}|\,I) = \binom{n}{m}^{-1}\sum_{n\bar{\boldsymbol{r}}=0}^{n}\binom{n\bar{\boldsymbol{r}}}{m}\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}}|\,I) = \binom{N}{m}^{-1}\sum_{N\bar{\boldsymbol{R}}=0}^{N}\binom{N\bar{\boldsymbol{R}}}{m}\mathrm{P}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{R}}|\,I). \tag{16}$$

The proof uses the expression for the mth factorial moment of the hypergeometric distribution [**potts1953**]. 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 $\mathrm{P}(\bar{\boldsymbol{X}} = \bar{\boldsymbol{R}}|\,I)$ 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 (14) 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 [**sivia1996_r2006**; **meadetal1984**] at the population level, denoted by I_p, gives

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

and then by marginalization (9)

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

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

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

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 [**rostamietal2016_r2017**].

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'?

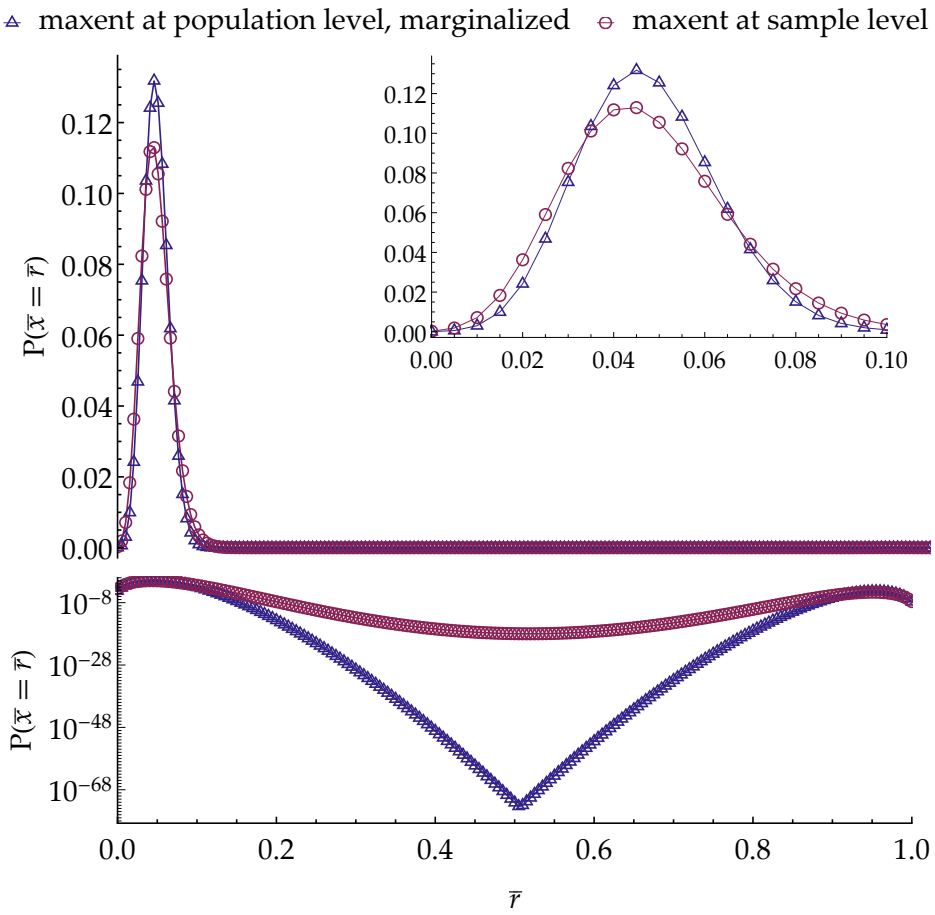

Figure 2: Linear and log-plots of $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (18), and at the sample level (red circles), eq. (20), with $N = 5000$, $n = 200$, constraints $\mathrm{E}(\bar{\boldsymbol{x}})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

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 $N = n$ 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 [**good1965**; **good1967b**] and Rissanen [**rissanen1983**] 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* [**jaynes1994_r2003**; **lindley1965b_r2008**; **bernardoetal1994_r2000**]. The probability

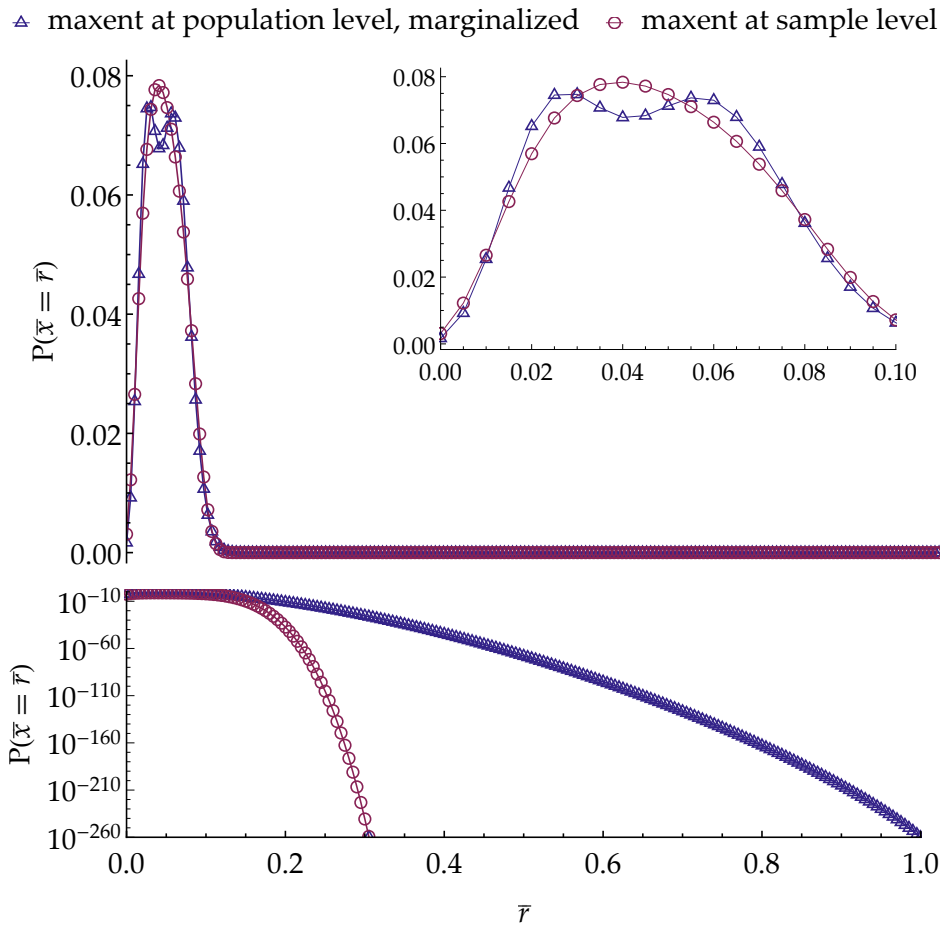

Figure 3: Linear and log-plots of $\mathrm{P}(\bar{\boldsymbol{x}} = \bar{\boldsymbol{r}})$ constructed by maximum-entropy at the population level followed by sample marginalization (blue triangles), eq. (18), and at the sample level (red circles), eq. (20), with $N = 5000$, $n = 200$, constraints $\mathrm{E}(\bar{\boldsymbol{x}})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

obtained from maximum-entropy at the population level, eq. (18), then generalizes to

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

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) 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 $N \sim 10^8$ [**leubaetal1994**]. 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 [**maesetal1999**]. 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 [**jaynes1985d_r1993**]. 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 [**schneidmanetal2006**; **shlensetal2006**; **roudietal2009b**], 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 [**tkaciketal2014b**; **shimazakietal2015**].

Acknowledgments

To be added after review.