Posteriors for sufficiency hypotheses and maximum-entropy

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

Draft of 7 April 2019 (first drafted 16 February 2019)

Assessing the probability of a hypothesis of sufficiency from the observation of a sample.

1 Hypotheses about sufficient statistics

We have a population of N neurons whose activities we imagine to have time-binned into T bins and binarized. Denote their total population activity at time bin t by $S_t \in \{0, 1, \dots, N\}$, their total activity at an unspecified bin by S, and the time sequence of total activities by $\boldsymbol{S} \coloneqq (S_{t_1}, S_{t_2}, \dots, S_{t_T})$.

We have recorded the activities of a sample of n neurons from the population above. Denote the total activity of the sample at time bin t by $s_t \in \{0, 1, \dots, n\}$, at an unspecified bin by s, and the time sequence by $\boldsymbol{s} \coloneqq (s_{t_1}, s_{t_2}, \dots, s_{t_T})$.

We don't know how these sampled neurons were chosen from the full population. This fact leads, for each time bin, to the following degree of belief about the activity of the sample if we knew the activity of the full population (Porta Mana et al. 2015 § 2.3; 2018 § 2):

$$\mathrm{p}(s \mid S, I) = \binom{n}{s}\binom{N-n}{S-s}\binom{N}{S}^{-1} \eqqcolon G_{sS}, \tag{1}$$

namely, a hypergeometryc distribution.

Here and in the following I denotes the proposition stating our background information.

Now suppose that we knew the total activities $\boldsymbol{S}$ of the *full* population at some T time bins $\{t\}$, and we wanted to infer the total activities $\boldsymbol{S}'$ at T' *different* time bins $\{t'\}$:

$$\mathrm{p}(\boldsymbol{S}' \mid \boldsymbol{S}, I). \tag{2}$$

We want to consider the hypotheses that *only a specific set of statistics* about our data $\boldsymbol{S}$ are *relevant* for our inference about $\boldsymbol{S}'$; that is, they

are a sufficient statistics. Any aspect of the data not contained in those statistics would be irrelevant for our inference. This inferential property could be the result of biological properties of the population.

Let's assume that there are R such statistics (besides T, which is always part of a set of sufficient statistics). Each statistic is the sum over time of a specific function of the total activity S. We can arrange these functions in an R-by-$(N+1)$ matrix $\boldsymbol{C} \coloneqq (C_{rS})$, where C_{rS} is the value of the function for the rth statistic when the total activity is S. The R sufficient statistics for the data $\boldsymbol{S}$ would thus be

$$\overline{C}_r \coloneqq \frac{1}{T} \sum_t C_{rS_t}, \quad r \in \{1, \dots, R\}. \tag{3}$$

Our goal is to quantify our uncertainty about these hypotheses of sufficient statistics, given the activity data from a sample of neurons. It's important to note that the hypotheses we must consider are not discrete or of a yes-or-no type: they form a continuum. This is because we have a continuum of degrees of relevance. Consider for example two statistics $\overline{C}_1$ and $\overline{C}_2$ from the bins $\{t\}$. Our degrees of belief about the activities at bins $\{t'\}$ are

$$\mathrm{p}(\boldsymbol{S}' \mid \overline{C}_1, \overline{C}_2, I). \tag{4}$$

It may happen that lack of knowledge about $\overline{C}_2$ doesn't change our degree of belief:

$$\mathrm{p}(\boldsymbol{S}' \mid \overline{C}_1, I) = \mathrm{p}(\boldsymbol{S}' \mid \overline{C}_1, \overline{C}_2, I), \tag{5}$$

in which case $\overline{C}_2$ is irrelevant. It may also happen that our degree of belief is changed but in a negligible way, for all values of $\boldsymbol{S}'$ and $\overline{C}_1$:

$$\mathrm{p}(\boldsymbol{S}' \mid \overline{C}_1, I) \approx \mathrm{p}(\boldsymbol{S}' \mid \overline{C}_1, \overline{C}_2, I), \tag{6}$$

so that $\overline{C}_2$ could be dropped in practice. We can imagine larger and larger changes to the point where dropping $\overline{C}_2$ would lead to drastically different degrees of belief. The question of the relevance of $\overline{C}_2$ is therefore not dichotomous. We will thus deal with a continuum of hypotheses, each representing a degree of relevance of some statistics. We shall shortly see how to mathematically represent this continuum of hypotheses.

2 The Koopman-Pitman theorem

How does a hypothesis about a sufficient statistic affect our degrees of belief? The answer comes from the Koopman-Pitman theorem (Koopman 1936; Pitman 1936; see also Darmois 1935; Barankin et al. 1963; Denny 1967; Hipp 1974; Lauritzen 1974a; 1984; 1988; for the discrete version: Fraser 1963; Andersen 1970), which says that the degree of belief (2) has a very specific mathematical expression if only some statistics of $\boldsymbol{S}$ relevant. The main statement of the theorem is this: if R sufficient statistics are given by functions C_{rS}, then for any number of time bins T

$$\mathrm{p}(\boldsymbol{S} \mid I) = \int \mathrm{d}\boldsymbol{\lambda}\; \mathrm{p}(\boldsymbol{\lambda} \mid I) \prod_t \mathrm{p}(S_t \mid \boldsymbol{\lambda}, I) \tag{7a}$$

with

$$\mathrm{p}(S \mid \boldsymbol{\lambda}, I) \coloneqq \frac{g_S}{Z(\boldsymbol{\lambda})} \exp\bigl(\textstyle\sum_r \lambda_r C_{rS}\bigr), \tag{7b}$$

$$\boldsymbol{\lambda} \coloneqq (\lambda_1, \dots, \lambda_R) \in \mathbf{R}^R, \qquad Z(\boldsymbol{\lambda}) \coloneqq \sum_S g_S \;\exp\bigl(\textstyle\sum_r \lambda_r C_{rS}\bigr), \tag{7c}$$

and g a positive function of S.

Some important remarks about the Pitman-Koopman formula (7):

a. A hypothesis that only stated what the sufficient statistics are would not determine the density $\mathrm{p}(\boldsymbol{\lambda} \mid I)$ or the function g in the formula above. The hypotheses we are going to compare thus contain additional information besides sufficiency.

b. The formula can be interpreted this way: our degree of belief about $\boldsymbol{S}$ is given by the degree of belief we would have if we knew the values of the sufficient statistics $\overline{\boldsymbol{C}}$ for an unlimited number of time bins, mixed over our uncertainty about the values themselves:

$$\mathrm{p}(\boldsymbol{S} \mid I) = \int \mathrm{d}\overline{\boldsymbol{C}}\; \mathrm{p}(\overline{\boldsymbol{C}} \mid I)\, \mathrm{p}(\boldsymbol{S} \mid \overline{\boldsymbol{C}}, I). \tag{8}$$

c. Formula (7) is obtained from (8) by a one-to-one reparametrization:

$$\overline{C}_r(\boldsymbol{\lambda}) = \sum_S C_{rS} \frac{g_S}{Z(\boldsymbol{\lambda})} \exp\bigl(\textstyle\sum_r \lambda_r C_{rS}\bigr) \equiv \partial_{\lambda_r} \ln Z(\boldsymbol{\lambda}), \tag{9}$$

$$\mathrm{p}(\overline{\boldsymbol{C}} \mid I)\, \mathrm{d}\overline{\boldsymbol{C}} = \mathrm{p}(\boldsymbol{\lambda} \mid I)\, \mathrm{d}\boldsymbol{\lambda}, \tag{10}$$

$$\mathrm{p}[S \mid \overline{\boldsymbol{C}}(\boldsymbol{\lambda}), I] = \mathrm{p}(S \mid \boldsymbol{\lambda}, I) \equiv \frac{g_S}{Z(\boldsymbol{\lambda})} \;\exp\bigl(\textstyle\sum_r \lambda_r C_{rS}\bigr). \tag{11}$$

Equation (9) cannot be solved explicitly for $\boldsymbol{\lambda}$ in terms of $\overline{\boldsymbol{C}}$ except for very simple cases. The parametrization in terms of $\boldsymbol{\lambda}$ has several special properties:

c.1. The quantities (λ_r) can assume any values independently of one another, whereas the limit statistics $(\overline{C}_r)$ have interdependent ranges.

c.2. The expression for $\mathrm{p}(\boldsymbol{S} \mid \overline{\boldsymbol{C}}, I)$ can be written explicitly in terms of $\boldsymbol{\lambda}$, but not in terms of $\overline{\boldsymbol{C}}$.

c.3. If some λ_r vanishes then the corresponding statistic is *irrelevant* – the corresponding term indeed disappears from the exponential in formula (7).

Remark c. suggests that the absolute value of each parameter, $|\lambda_r|$, might be used to quantify the degree of relevance of the corresponding statistic, zero meaning complete irrelevance. These absolute values can therefore be used as parameters for our continuum of hypotheses about the sufficient statistics. The degree of belief we ultimately want to quantify is thus

$$\mathrm{p}\bigl(|\boldsymbol{\lambda}| \bigm| \text{data}, I\bigr) \tag{12}$$

and its marginals for the various $|\lambda_r|$.

3 Conditionalization on full-population data

First let us recapitulate the expression for

$$\mathrm{p}(\boldsymbol{\lambda} \mid \boldsymbol{S}, I) \propto \mathrm{p}(\boldsymbol{\lambda} \mid I)\, \mathrm{p}(\boldsymbol{S} \mid \boldsymbol{\lambda}, I). \tag{13}$$

Introduce the relative frequencies $\boldsymbol{F} \coloneqq (F_S)$ with which the activity S appears during the T time bins, given by

$$F_S = \frac{1}{T} \sum_t \delta(S_t - S). \tag{14}$$

From formula (7), bringing several terms within the exponential, we can rewrite, using the frequencies,

$$\begin{aligned} \mathrm{p}(\boldsymbol{S} \mid \boldsymbol{\lambda}, I) &= \prod_t \mathrm{p}(S_t \mid \boldsymbol{\lambda}, I) \\ &= \prod_S \bigl[\mathrm{p}(S \mid \boldsymbol{\lambda}, I)\bigr]^{TF_S} \\ &= \exp\Bigl(-T\bigl\{H\bigl[\boldsymbol{F}, \mathrm{p}(S \mid \boldsymbol{\lambda}, I)\bigr] - H(\boldsymbol{F})\bigr\}\Bigr), \end{aligned} \tag{15}$$

where $H(\boldsymbol{a}, \boldsymbol{b})$ is the relative entropy of $\boldsymbol{a}$ with respect to $\boldsymbol{b}$ and $H(\boldsymbol{a})$ is the Shannon entropy of $\boldsymbol{a}$.

The last expression shows that: (a) our degree of belief about the sequence $\boldsymbol{S}$ depends only on the frequencies $\boldsymbol{F}$ of that sequence; (b) if $\mathrm{p}(\boldsymbol{\lambda} \mid I)$ is constant, the mode of the density (13) for $\boldsymbol{\lambda}$ can be calculated by minimizing with respect to $\boldsymbol{\lambda}$ the relative entropy between $\boldsymbol{F}$ and $\mathrm{p}(S \mid \boldsymbol{\lambda}, I)$; (c) owing to the mathematical form (7b) of $\mathrm{p}(S \mid \boldsymbol{\lambda}, I)$, minimization of the relative entropy leads to the classical maximum-entropy equations

$$\mathrm{E}(C_{rS} \mid \boldsymbol{\lambda}, I) \equiv \partial_{\lambda_r} \ln Z(\boldsymbol{\lambda}) = \overline{\boldsymbol{C}}_r \equiv \textstyle\sum_S C_{rS} F_S. \tag{16}$$

4 Conditionalization on sample data

The data conditional on which we want to calculate the degree of belief (12) do not involve the full population, though, but only a sample of it: they are some statistics

$$\hat{\boldsymbol{c}} \coloneqq \frac{1}{T} \sum_t \boldsymbol{c}_{s_t} \equiv \sum_s \boldsymbol{c}_s f_s \tag{17}$$

of a sequence $\boldsymbol{s}$ of recorded sample activities with relative frequencies $\boldsymbol{f}$. The statistic $\boldsymbol{c}_s$ can be vector-valued.

We use Bayes's theorem:

$$\mathrm{p}(\boldsymbol{\lambda} \mid \hat{\boldsymbol{c}}, I) \propto \mathrm{p}(\boldsymbol{\lambda} \mid I)\, \mathrm{p}(\hat{\boldsymbol{c}} \mid \boldsymbol{\lambda}, I) \tag{18}$$

and focus our attention on the last term. It can be written as the sum of our degrees of belief for the frequencies satisfying the statistics:

$$\mathrm{p}(\hat{\boldsymbol{c}} \mid \boldsymbol{\lambda}, I) = \sum_f \delta\bigl(\textstyle\sum_s \boldsymbol{c}_s f_s = \hat{\boldsymbol{c}}\bigr)\, \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{\lambda}, I). \tag{19}$$

To calculate our degree of belief about the frequencies, we note that all sequences $\boldsymbol{s}$ having frequencies $\boldsymbol{f}$ have the same degree of belief under

the hypothesis of sufficiency *for the full population* and assuming the long-run statistics are known – that is, with $\boldsymbol{\lambda}$ given:

$$\mathrm{p}(\boldsymbol{s} \mid \boldsymbol{\lambda}, I) = \prod_t \mathrm{p}(s_t \mid \boldsymbol{\lambda}, I) = \prod_s \mathrm{p}(s \mid \boldsymbol{\lambda}, I)^{T f_s}. \tag{20}$$

The first, joint degree of belief factorizes because the degree of belief for each s_t is independent of the other t if the corresponding S_t is known, and the joint degree of belief for all the S_t factorizes if $\boldsymbol{\lambda}$ is known.

Our degree of belief $\mathrm{p}(s \mid \boldsymbol{\lambda}, I)$ about the sample activity at any time, given $\boldsymbol{\lambda}$, can be obtained by marginalizing the corresponding S:

$$\mathrm{p}(s \mid \boldsymbol{\lambda}, I) = \sum_S \mathrm{p}(s \mid S, \boldsymbol{\lambda}, I)\, \mathrm{p}(S \mid \boldsymbol{\lambda}, I) = \sum_S G_{sS}\, \mathrm{p}(S \mid \boldsymbol{\lambda}, I), \tag{21}$$

since the first degree of belief in the sum is independent on $\boldsymbol{\lambda}$; the second is given by eq. (7b).

Our degree of belief $\mathrm{p}(\boldsymbol{f} \mid \boldsymbol{\lambda}, I)$ about the relative frequencies of sample activities is then proportional to the degree of belief above, times the number of sequences having those frequencies:

$$\begin{aligned} \mathrm{p}(\boldsymbol{f} \mid \boldsymbol{\lambda}, I) &= \binom{T}{T\,\boldsymbol{f}}\, \mathrm{p}(\boldsymbol{s} \mid \boldsymbol{\lambda}, I) = \binom{T}{T\,\boldsymbol{f}} \prod_s \Bigl[\textstyle\sum_S G_{sS}\, \mathrm{p}(S \mid \boldsymbol{\lambda}, I)\Bigr]^{T f_s} \\ &\approx \exp\Bigl\{T\, H\Bigl[\boldsymbol{f}, \bigl(\textstyle\sum_S G_{sS}\, \mathrm{p}(S \mid \boldsymbol{\lambda}, I)\bigr)_s\Bigr]\Bigr\} \end{aligned} \tag{22}$$

The last expression valid for enough large T.

Combining eqs (19) and (22) we obtain

$$\mathrm{p}(\hat{\boldsymbol{c}} \mid \boldsymbol{\lambda}, I) \approx \sum_{\boldsymbol{f}} \delta\bigl(\textstyle\sum_s \boldsymbol{c}_s f_s = \hat{\boldsymbol{c}}\bigr)\, \exp\Bigl\{T\, H\Bigl[\boldsymbol{f}, \bigl(\textstyle\sum_S G_{sS}\, \mathrm{p}(S \mid \boldsymbol{\lambda}, I)\bigr)_s\Bigr]\Bigr\}. \tag{23}$$

When T is very large compared to the number of possible values of s, the expression above becomes negligible for all $\boldsymbol{\lambda}$ such that $\sum_S G_{sS}\ \mathrm{p}(S \mid \boldsymbol{\lambda}, I)$ equal some set of frequencies f_s that satisfy the constraints (17). It can therefore be approximated as

$$\begin{aligned} \mathrm{p}(\hat{\boldsymbol{c}} \mid \boldsymbol{\lambda}, I) \approx \delta\bigl[\textstyle\sum_s \boldsymbol{c}_s \sum_S G_{sS}\, \mathrm{p}(S \mid \boldsymbol{\lambda}, I) = \hat{\boldsymbol{c}}\bigr] \equiv \\ \delta\Bigl[\sum_S \textstyle\sum_s \boldsymbol{c}_s\, G_{sS}\, \dfrac{g_S}{Z(\boldsymbol{\lambda})}\, \exp(\boldsymbol{\lambda} \boldsymbol{C}_S) = \hat{\boldsymbol{c}}\Bigr]. \end{aligned} \tag{24}$$

This equation differ from the constraint equation obtained with the usual maximum-entropy method: since

$$\sum_s \boldsymbol{c}_s \, G_{sS} \neq \boldsymbol{C}_S, \tag{25}$$

we are not seeking $\boldsymbol{\lambda}$ for which the expectation of *its conjugate statistics* $\boldsymbol{C}_S$ satisfies a constraint, but for which the expectation of a different statistics satisfies a constraint. The only exception is when

$$\sum_s \boldsymbol{c}_s \, G_{sS} = \boldsymbol{C}_S, \tag{26}$$

which happens, for example, for the factorial moments.

***

As in the previous section, if the density $\mathrm{p}(\boldsymbol{\lambda} \mid I)$ for $\boldsymbol{\lambda}$ is uniform then from eq. (18) the mode of $\mathrm{p}(\boldsymbol{\lambda} \mid \boldsymbol{s}, I)$ can be obtained by maximizing $\mathrm{p}(\boldsymbol{f} \mid \boldsymbol{\lambda}, I)$, which amount to maximizing the relative entropy between $\boldsymbol{f}$ and $\bigl(\sum_S G_{sS} \, \mathrm{p}(S \mid \boldsymbol{\lambda}, I)\bigr)_s$ with respect to $\boldsymbol{\lambda}$. We have the maximum of the latter for the $\boldsymbol{\lambda}$ satisfying

$$\sum_S G_{sS} \, \frac{g_S}{Z(\boldsymbol{\lambda})} \, \exp\bigl(\textstyle\sum_r \lambda_r C_{rS}\bigr) = f_s \quad \forall s. \tag{27}$$

minimizing the relative entropy between $\boldsymbol{f}$ and $\mathrm{p}(s \mid \boldsymbol{\lambda}, I)$. In the present case, however, the expression for the latter probability, eq. (28b), does not lead to the classical maximum-entropy equations. We find instead

$$\partial_{\lambda_r} \ln Z(\boldsymbol{\lambda}) = \hat{c}_r(\boldsymbol{\lambda}) \tag{28a}$$

with

$$\hat{c}_r(\boldsymbol{\lambda}) \coloneqq \textstyle\sum_s f_s \; c_{rs}(\boldsymbol{\lambda}), \qquad c_{rs}(\boldsymbol{\lambda}) \coloneqq \dfrac{\sum_S C_{rS} \; G_{sS} \; \mathrm{p}(S \mid \boldsymbol{\lambda}, I)}{\sum_S G_{sS} \; \mathrm{p}(S \mid \boldsymbol{\lambda}, I)}. \tag{28b}$$

We can calculate it by marginalizing with respect to all possible *sequences* of activities $\boldsymbol{S}$ for the full population, using the hypergeometric distribution (1) and the Koopman-Pitman formula (7):

$$\mathrm{p}(\boldsymbol{s} \mid I) = \sum_S \mathrm{p}(\boldsymbol{s} \mid \boldsymbol{S}, I) \, \mathrm{p}(\boldsymbol{S} \mid I) = \int \mathrm{d}\boldsymbol{\lambda} \; \mathrm{p}(\boldsymbol{\lambda} \mid I) \; \prod_t \mathrm{p}(s_t \mid \boldsymbol{\lambda}, I) \tag{29a}$$

with

$$\mathrm{p}(s \mid \boldsymbol{\lambda}, I) \coloneqq \sum_S G_{sS} \, \frac{g_S}{Z(\boldsymbol{\lambda})} \exp\bigl(\textstyle\sum_r \lambda_r C_{rS}\bigr). \tag{29b}$$

This formula shows that our degree of belief about the sample activities does *not* have a sufficient statistics. This fact is mathematically similar to

what happens in statistical mechanics: if our uncertainty about a system of particles is expressed by a Gibbs distribution, then our uncertainty about a subsystem won't generally be of a Gibbsian type (Maes et al. 1999).

We can now calculate the density for $\boldsymbol{\lambda}$ given the sequence of sample activities $\boldsymbol{s}$ using Bayes's theorem. We can introduce the relative frequencies $\boldsymbol{f}$ for the sample activities and proceed as in § 3, obtaining

$$\begin{aligned}\mathrm{p}(\boldsymbol{s} \mid \boldsymbol{\lambda}, I) &= \prod_t \mathrm{p}(s_t \mid \boldsymbol{\lambda}, I) \\ &= \prod_s \bigl[\mathrm{p}(s \mid \boldsymbol{\lambda}, I)\bigr]^{T f_s} \\ &= \exp\Bigl(-T\bigl\{H\bigl[\boldsymbol{f}, \mathrm{p}(s \mid \boldsymbol{\lambda}, I)\bigr] - H(\boldsymbol{f})\bigr\}\Bigr).\end{aligned} \tag{30}$$

As in the previous section, if the density $\mathrm{p}(\boldsymbol{\lambda} \mid I)$ for $\boldsymbol{\lambda}$ is uniform then the mode of $\mathrm{p}(\boldsymbol{\lambda} \mid \boldsymbol{s}, I)$ can be obtained by minimizing the relative entropy between $\boldsymbol{f}$ and $\mathrm{p}(s \mid \boldsymbol{\lambda}, I)$. In the present case, however, the expression for the latter probability, eq. (28b), does not lead to the classical maximum-entropy equations. We find instead

$$\partial_{\lambda_r} \ln Z(\boldsymbol{\lambda}) = \hat{c}_r(\boldsymbol{\lambda}) \tag{31a}$$

with

$$\hat{c}_r(\boldsymbol{\lambda}) \coloneqq \textstyle\sum_s f_s \; c_{rs}(\boldsymbol{\lambda}), \qquad c_{rs}(\boldsymbol{\lambda}) \coloneqq \dfrac{\sum_S C_{rS} \; G_{sS} \; \mathrm{p}(S \mid \boldsymbol{\lambda}, I)}{\sum_S G_{sS} \; \mathrm{p}(S \mid \boldsymbol{\lambda}, I)}. \tag{31b}$$

or more explicitly

$$\frac{\sum_S C_{rS} \; g_S \exp\bigl(\sum_r \lambda_r C_{rS}\bigr)}{\sum_S g_S \exp\bigl(\sum_r \lambda_r C_{rS}\bigr)} = \sum_s f_s \; \frac{\sum_S C_{rS} \; G_{sS} \; g_S \exp\bigl(\sum_r \lambda_r C_{rS}\bigr)}{\sum_S G_{sS} \; g_S \exp\bigl(\sum_r \lambda_r C_{rS}\bigr)}. \tag{32}$$

The expressions (30) define r functions $c_{rs}(\boldsymbol{\lambda})$ of the sample activity s *and of* $\boldsymbol{\lambda}$ that have a role analogous to the statistic C_{rS} of the full-population activity S, in the sense that

$$\begin{aligned}\mathrm{E}[c_{rs}(\boldsymbol{\lambda}) \mid \boldsymbol{\lambda}, I] &= \mathrm{E}(C_{rS} \mid \boldsymbol{\lambda}, I), \\ \sum_s c_{rs}(\boldsymbol{\lambda}) \; \mathrm{p}(s \mid \boldsymbol{\lambda}, I) &= \sum_S C_{rS} \; \mathrm{p}(S \mid \boldsymbol{\lambda}, I),\end{aligned} \tag{33}$$

as can be verified by substitution and the definition (1) of G_{sS}.

The mode $\boldsymbol{\lambda}$ is given by

$$\arg\inf_{\boldsymbol{\lambda}}\Big\{\ln Z(\boldsymbol{\lambda}) - \sum_s f_s \;\ln\big[\textstyle\sum_S G_{sS}\; g_S\;\exp\big(\sum_r \lambda_r C_{rS}\big)\big]\Big\}. \tag{34}$$

Compare this with the standard maximum-entropy case (Mead et al. 1984), which can be written as

$$\arg\inf_{\boldsymbol{\lambda}}\Big[\ln Z(\boldsymbol{\lambda}) - \sum_S F_S \;\ln\exp\big(\textstyle\sum_r \lambda_r C_{rS}\big)\Big]. \tag{35}$$

Important differences from the 'dilemma' paper:

Consider the following statistics:

$$C_{rS} \coloneqq \binom{S}{r}\Big/\binom{N}{r} \qquad r \in 1, 2, \dots. \tag{36}$$

In particular, $C_{1S} = S/N$ and C_{rS} is the number of r-tuples of simultaneously active neurons divided by the number of possible ones. Analogous statistics $\binom{s}{r}/\binom{n}{r}$ can be considered for the sample.

These statistics have this special property (Porta Mana et al. 2015; 2018):

$$\sum_s \binom{s}{r}\Big/\binom{n}{r}\, p(s) = \sum_S \binom{S}{r}\Big/\binom{N}{r}\, p(S), \tag{37}$$

no matter what the degrees of belief about full-population and sample might be, provided that they are related by $p(s) = \sum_S G_{sS}\; p(S)$.

Note, however, that given $p(S)$ and any statistic C_{rS} it is always possible to create a function c_{rS} of the sample activity that satisfies

$$\sum_s c_{rs}\; p(s) = \sum_S C_{rS}\; p(S) \tag{38}$$

namely,

$$c_{rs} \coloneqq \frac{\sum_S C_{rS}\; G_{sS}\; p(S)}{\sum_S G_{sS}\; p(S)}. \tag{39}$$

But this function *depends on the specific p(S)*.

Now consider the minimization of the relative entropy in eq. (29), corresponding to eqs (30), for statistics given by (35). Remember that the assumption of sufficiency asymptotically gives zero probability to observing relative frequencies of full population and sample that lie outside a particular R-dimensional set.

We have two possibilities when T is very large:

1. If the frequencies $\boldsymbol{f}$ observed for the sample are among those admitted by the sufficiency hypothesis, then there is a $\boldsymbol{\lambda}^*$ for which

$$f_s \approx \sum_S G_{sS}\ \mathrm{p}(S \mid \boldsymbol{\lambda}^*, I). \tag{40}$$

   Moreover, for that $\boldsymbol{\lambda}$ we also have

$$\sum_s f_s \binom{s}{r} / \binom{n}{r} \approx \sum_s f_s\ c_{rs}(\boldsymbol{\lambda}^*). \tag{41}$$

   and the result of the relative-entropy minimization is equivalent to finding the maximum-entropy distribution for the full population with expectations for the statistics (35) constrained to equal the empirical ones observed in the sample.

2. If the frequencies $\boldsymbol{f}$ observed for the sample are *not* among those admitted by the sufficiency hypothesis, then the minimization of the relative entropy yields a distribution for the full-population that does *not* satisfy those constraints.

Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

Andersen, E. B. (1970): *Sufficiency and exponential families for discrete sample spaces*. J. Am. Stat. Assoc. **65**[331], 1248–1255.

Barankin, E. W., Maitra, A. P. (1963): *Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics*. Sankhyā A **25**[3], 217–244.

Barndorff-Nielsen, O. E., Blæsild, P., Schou, G., eds. (1974): *Proceedings of Conference on Foundational Questions in Statistical Inference: Aarhus, May 7–12, 1973*. (University of Aarhus, Aarhus).

Barndorff-Nielsen, O. E., Dawid, A. P., Diaconis, P., Johansen, S., Lauritzen, S. L. (1984): *Discussion of Steffen Lauritzen's paper ['Extreme Point Models in Statistics']*. Scand. J. Statist. **11**[2], 83–91. See Lauritzen (1984).

Darmois, G. (1935): *Sur les lois de probabilité à estimation exhaustive*. Comptes rendus hebdomadaires des séances de l'Académie des sciences **200**, 1265–1266.

Denny, J. L. (1967): *Sufficient conditions for a family of probabilities to be exponential*. Proc. Natl. Acad. Sci. (USA) **57**[5], 1184–1187.

Fraser, D. A. S. (1963): *On sufficiency and the exponential family*. J. Roy. Stat. Soc. B **25**[1], 115–123.

Hipp, C. (1974): *Sufficient statistics and exponential families*. Ann. Stat. **2**[6], 1283–1292.

Koopman, B. O. (1936): *On distributions admitting a sufficient statistic*. Trans. Am. Math. Soc. **39**[3], 399–409.

Lauritzen, S. L. (1974a): *Sufficiency, prediction and extreme models*. In: Barndorff-Nielsen, Blæsild, Schou (1974), 249–269. With discussion. Repr. without discussion in Lauritzen (1974b).

— (1974b): *Sufficiency, prediction and extreme models*. Scand. J. Statist. **1**[3], 128–134.

— (1984): *Extreme point models in statistics*. Scand. J. Statist. **11**[2], 65–83. See also discussion and reply in Barndorff-Nielsen, Dawid, Diaconis, Johansen, Lauritzen (1984).

— (1988): *Extremal Families and Systems of Sufficient Statistics*. (Springer, Berlin). First publ. 1982.

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

Mead, L. R., Papanicolaou, N. (1984): *Maximum entropy in the problem of moments*. J. Math. Phys. **25**[8], 2404–2417. http://bayes.wustl.edu/Manual/MeadPapanicolaou.pdf.

Pitman, E. J. G. (1936): *Sufficient statistics and intrinsic accuracy*. Math. Proc. Camb. Phil. Soc. **32**[4], 567–579.

Porta Mana, P. G. L., Rostami, V., Torre, E., Roudi, Y. (2018): *Maximum-entropy and representative samples of neuronal activity: a dilemma*. Open Science Framework doi:10.17605/osf.io/uz29n, bioRxiv doi:10.1101/329193, arXiv:1805.09084.

Porta Mana, P. G. L., Torre, E., Rostami, V. (2015): *Inferences from a network to a subnetwork and vice versa under an assumption of symmetry*. bioRxiv doi:10.1101/034199.