Interpretations of maximum-entropy distributions in neuroscience

P.G.L. Porta Mana
<piero.mana@ntnu.no>
Kavli Institute for Systems Neuroscience, Trondheim

Draft of 19 February 2019 (first drafted 15 February 2019)

Possible interpretations of maximum-entropy distributions in neuroscience.
Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

And how we are going to teach well by using books written by people who don't quite *understand what they're talking about, I* cannot *understand.*
Feynman (1989 p. 267)

1 The problem

A probability distribution is our quantified uncertainty about a set of mutually exclusive and exhaustive propositions. So when we see a probability distribution it's natural to ask: *what are we uncertain about?* The answer to this question should be precise, not vague as for example 'the network state': what network? at what time? To give a meaningful answer it's useful to specify in detail what experiment would be done to answer the question with certainty.

Asking the question above is also a good way to distinguish probability distributions from other distributions. For example, if we're shown a *frequency* distribution and we ask 'what are we uncertain about?', our question makes no sense.

Together with asking what our uncertainty is about, it's also natural to ask: 1. Which initial assumptions lead to this distribution? 2. Which data lead to this distribution?

This note tries to answer these questions for probability distribution of maximum-entropy form appearing in the neuroscientific literature.

2 First interpretation

In this first interpretation the maximum-entropy distribution is actually an approximation. The context is the following. We have a recorded sequence of neural activities of N neurons, time-binned and binarized: $\bigl\{s(t) \mid t = 1, 2, \dots; s(t) \in \{0,1\}^N\bigr\}$. We wonder what activity we would measure in a *new* time bin, in the past, or in the future, or some intermediate time bin missing from the data above.

Let's explicitly describe the question, the assumptions, and the data, and then discuss them in more depth. I'd like you to pay particular attention to the assumptions: it's rare for them to be explicitly spelled out in the neuroscientific literature, thus giving the wrong impression that fewer assumptions are needed.

The question: What is the neural activity $s(t')$ in a time bin t' *outside* of the recorded ones?

The assumptions:

- A1. The order of the time bins, *including any number of new ones*, is irrelevant for our inferences.
- A2. For some biological reason, only the first R moments of observed data are *relevant* for the prediction of the activity in a new time bin; this is true for any amount of data.
- A3. Some assumption that leads to a quantified uncertainty about the first R moments of the activities of *every* number of time bins (not just the recorded ones).
- A4. Some assumption that leads to a quantified uncertainty about $s(t')$ *if we knew the first R moments* of the activities of *every* number of time bins.

Note that the second assumption implies the first.

The data: The set of recorded activities $\{s(t)\}$.

The assumptions above – denote them by A – determine a unique degree of belief for *every possible* sequence of activities $\{s(\tau)\}$, before the observation of any data. For simplicity let me consider the case $R = 3$:

$$\mathrm{p}\bigl[\{s(\tau)\} \mid A\bigr] = \int \mathrm{d}\lambda \; f(\lambda) \times \prod_{\tau} \frac{g[s(\tau)]}{Z(\lambda)} \; \exp\bigl[\lambda^i s_i(\tau) + \lambda^{ij} s_i(\tau)\, s_j(\tau) + \lambda^{ijk} s_i(\tau)\, s_j(\tau)\, s_k(\tau)\bigr] \quad \text{(1a)}$$

with

$$Z(\lambda) \coloneqq \sum_s g(s)\ \exp\bigl(\lambda^i s_i + \lambda^{ij} s_i s_j + \lambda^{ijk} s_i s_j s_k\bigr). \tag{1b}$$

Einstein's summation convention is used in the formula above, and the sums run over increasing sequences $i < j < k$. The integral is over $N + N\,(N-1)/2 + N\,(N-1)\,(N-2)/6$ parameters $\lambda \coloneqq (\lambda^i, \lambda^{ij}, \lambda^{ijk})$; in general it will be over $\binom{N}{1} + \cdots + \binom{N}{R}$ parameters.

The fact that the assumptions above determine formula (1), and vice versa, is the discrete version (Fraser 1963; Andersen 1970) of the Koopman-Pitman-Darmois theorem (Koopman 1936; Pitman 1936; Darmois 1935; see also Barankin et al. 1963; Denny 1967; Hipp 1974; Lauritzen 1974a; 1984; 1988). Assumption A2 – the assumption about a sufficient statistics – is the one leading to a mixture of exponentials, but it doesn't determine the density f and the distribution g. These are determined by assumptions A3 and A4.

From the distributions (1) we can calculate our degree of belief about any subsequence of activities conditional on any other distinct subsequence:

$$\mathrm{p}\bigl[\{s(t')\} \mid \{s(t)\}, A\bigr] = \frac{\mathrm{p}\bigl[\{s(t'), s(t)\} \mid A\bigr]}{\mathrm{p}\bigl[\{s(t)\} \mid A\bigr]}, \tag{2}$$

with numerator and denominator given by (1).

If the data $\{s(t)\}$ comprise a sequence long enough to render the product of exponentials much more peaked in ya than the density f, then the conditional probability (2) is approximated by a product of maximum-entropy distributions with Lagrange multipliers λ determined by the usual constraint equations with data $\{s(t)\}$ (Porta Mana 2017). Thus the probability distribution (2) will be approximated by a product of maximum-entropy distributions.

There are several important points about this interpretation that are rarely mentioned in the neuroscientific literature:

Which data are *relevant*? There's a natural question regarding assumptions A1 and A2: which time bins should be considered *relevant* for one another, but with their particular order still irrelevant? The choice is subjective. Note that in principle we could consider sequences of activities from different recording sessions, and even different neurons (as long as there are N of them), even from different test subjects. We may

adduce biological reasons for limiting our choices; the point is that such a choice is beyond the probability calculus. We can of course compare two different choices within the probability calculus – for example to consider only data within one recording session, or across multiple sessions – and then calculate their posterior probabilities given some data. But to do this we'd need to make an assumption of what's relevant and what's not on a more general level. The subjective choice has just been pushed somewhere else; it hasn't disappeared.

Stationarity.

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.

Feynman, R. P. (1989): *"Surely You're Joking, Mr. Feynman!": Adventures of a Curious Character*, reprint. (Bantam, New York). 'As told to Ralph Leighton', ed. by Edward Hutchings. First publ. 1985.

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.

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

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