X facts that you maybe didn't know about the maximum-entropy method

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

Draft of 23 April 2019 (first drafted 23 April 2019)

***

Note: Dear Reader & Peer, this manuscript is being peer-reviewed by you. Thank you.

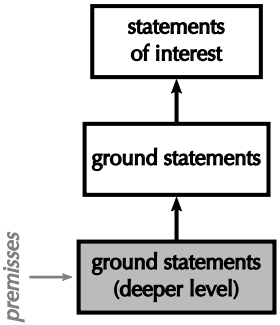
1 ***

2 Maximum-entropy method

A necessary step of every plausible inference is the assignment of some probabilistic premisses; as many as necessary to arrive at our desired probabilistic conclusions given the evidence. The necessity of this step was tersely stated by W. E. Johnson a century ago:

> Now the axioms of probability enable us to infer any probability-conclusion *only* from probability-premisses. In other words, the calculus of probability does not enable us to infer any probability-value unless we have some probabilities or probability relations *given*. Such data cannot be supplied by the mathematician. E.g. the rules of arithmetic and the axioms of the probability-calculus are utterly impotent to determine, on the supposed knowledge that the throw of a coin must yield either head or tail and cannot yield both, the probability that it will yield head or that it will yield tail. We must assume that the two co-exclusive and co-exhaustive possibilities are *equally probable*, before we can estimate the probability of either as being a half of certitude. (**johnson1924**)

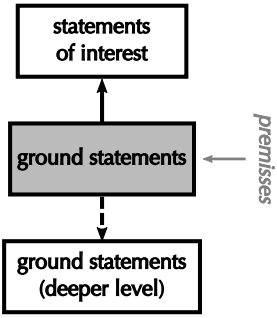
These premisses may represent either actual knowledge or just some hypotheses whose consequences we want to explore. They can be assigned at different depths of the logical analysis of our inference problem.



We can for example assign probabilistic premisses at a very deep analytical level, as

schematically illustrated on the left. They determine the probabilities of statements at more superficial levels, including the statements which we're ultimately interested in or which we can directly observe.

Or we can assign probabilistic premisses at a level of superficial or intermediate depth, as schematically illustrated on the right. In this case they determine the probabilities of statements at more superficial levels, and partly constrain the probabilities of statements at deeper levels.



The choice of the analytical depth at which to assign the premisses depends on many factors: on the difficulty of assessing our state of uncertainty at that depth, on the cost of the ensuing calculations, and especially on the general purpose of our inference. Assignments of probabilistic premisses at different levels can still turn out to be completely equivalent for the probabilities of the statements of interest.

Here's an example. Suppose that n neurons have been recorded in a sequence of time bins. We're given the time averages of the numbers of active neurons and of neuron pairs in the sequence. We don't know the exact sequence of activities or the number T of time bins, although we know that the latter is large. We want to forecast how many of the n neurons will be active in a *new* time bin.

In a deep approach to this inference problem we assign the probability for the possible value of T, and the probabilities for several hypotheses about the sequence of activities in $T+1$ or more time bins. One hypothesis could be, for instance, that the n neurons are a sample of a larger population of N neurons, and a second hypothesis could be that for this population there are some sufficient statistics, out of biological reasons. We could of course deepen this approach further, with no end, down to assigning the probabilities for the constitution and motion of the molecules of the neurons and their environment, the probabilities for the position and response of the recording electrodes, and so on. From all these probabilities and the data we could finally calculate the probability for the activity in the new time bin.

At the other end, the shallowest approach to this inference problem

is to directly assign the probabilities for the number of active neurons in the new time bin, somehow using the data we were given. (It must be remarked that such assignment partly constrains, in a backward fashion, all assignments we could do at a deeper level.)

The maximum-entropy method was proposed by Jaynes (**jaynes1957**; **jaynes1963**; **jaynes1994_r2003**; **sivia1996_r2006**; **hobsonetal1973**) as a way of assigning probability premisses at a more direct level, as in the approach just described. This point is made especially clear in **jaynes1994_r2003**, § 11.1:

> If we knew the numbers N, L, W, [some quantities relevant to our inference] then this could be solved by direct application of Bayes' theorem; without that information, we could still introduce the unknowns N, L, W as nuisance parameters and use Bayes' theorem, eliminating them at the end. [...] However, the Bayesian solution would not really address our problem; it only transfers it to the problem of assigning priors to N, L, W, leaving us back in essentially the same situation; how do we assign informative probabilities?

The idea behind this method is to translate *all* the data we're given into constraints – such as expectations – for our probability assignment, resolving the leftover freedom by choosing the constrained probability distribution having highest Shannon entropy, as a measure of its uncertainty or broadness. It can be shown that the use of this method is approximately equivalent, under specific conditions, to a probability assignment at a deeper level based on an assumption of exchangeability together with a multinomial probability distribution for long-run frequencies, or with an assumption about sufficient statistics (**jaynes1986d_r1996**; **rodriguez1989**; **rodriguez2002**; **skilling1989b**; **skilling1990bernardoetal1994_r2000diaconisetal1981**; **portamana2017**; **portamana2009**).

Sometimes this method is applied with only *part* of the available data used as constraint; such an altered approach can be a heuristic way of testing for sufficient statistics.

Owing to its directness, the maximum-entropy method is straightforward to use – and to misuse. In our previous example, for instance, it would give silly results if the number of time bins were not much larger than the possible activity values (it has thus been variously misused in the neuroscientific literature refs, especially to Tkacik).