

Reasoned inference of long-run mutual information

(Bayesian theory for dummies) [draft]

P.G.L. Porta Mana [<pgl@portamana.org>](mailto:pgl@portamana.org)

C. Battistin

[<claudia.battistin@ntnu.no>](mailto:claudia.battistin@ntnu.no)

S. Gonzalo Cogno

[<soledad.g.cogno@ntnu.no>](mailto:soledad.g.cogno@ntnu.no)

(or any permutation thereof)

31 March 2019; updated 14 March 2021

A reasoned analysis of inference for long-run mutual information between stimuli and responses from small samples is given. The use of estimators, biased or not, is found to be inadequate for the small-sample case. Moreover, any inference or formula for bias is found to heavily depend on the specific peculiarities of the problem – the specific kind of stimuli and responses, brain region, behavioural and environmental conditions, and so on – making any one-fits-all formula universally poor.

 draft comments can be introduced with the macro `\mynote{}`

1 Likelihood of long-run frequencies

In the previous section we showed that mutual information between the stimulus s and the response r computed directly from the sample may significantly vary across samples from the same population. In reality we don't typically have access to multiple samples, but we have to rely on one single sample to make an estimate of the population mutual information. As explained in section ... the population mutual information is a function of the population frequencies, or long-run frequencies, which are unknown. Yet we can use our data to inform us on the long-run frequencies that may have generated our sample and use them to estimate the population mutual information. How do we appraise candidate long-run frequencies for our data? Given the candidate long-run frequencies $\mathbf{f}_s = \{f_s(r)\}_{s,r}$ and a sample \hat{s} , the likelihood of these long-run frequencies $P(\hat{s}|\mathbf{f}_s)$ is the probability that our sample was generated from a population with such long-run frequencies. For the dataset of spike-count responses and north/south head direction stimulus introduced in section ... the likelihood of long-run frequencies \mathbf{f}_s is:

$$P(\hat{s}|\mathbf{f}_s) = \prod_{s,r} f_s(r)^{N_{\hat{f}_s(r)}} \quad (1)$$

where $\hat{f}_s(r)$ are the sample frequencies shown in Fig. ??.

It can be easily seen that the long-run frequencies that maximize the likelihood in Eq. 1 are the sample frequencies $f_s(r) = \hat{f}_s(r)$. Thus, in order to estimate the population mutual information between head direction and spike count, it might be tempting to simply compute the mutual information from the sample frequencies being the maximum-likelihood ones, while disregarding other potential long-run frequencies. Such point estimator of the population mutual information has indeed the convenient property of being asymptotically optimal, meaning that for infinite data size, the maximum likelihood estimator converges to the population one [Ref.](#)

In real experiments the data sample size is always limited, which is why the maximum-likelihood estimator can be a poor estimator in practice. Several authors [Refs](#) have indeed pointed out that the maximum-likelihood estimate of the mutual information can be biased, because of the very nature of the mutual information being positively defined. Beyond its bias the one-shot maximum-likelihood estimation is problematic in many ways. Ultimately the main (and fixable!) issue is that it's limited to a single candidate long-run frequency, while either previous knowledge about the system, or the likelihood itself, might indicate that other long-run frequencies should be considered. In the next two subsections we illustrate this point by considering the three long-run frequencies in Fig. 1 for the sample \hat{s} , introduced in section ..., from which we want to estimate the population mutual information between head direction and spike count.

High-likelihood long-run frequencies beyond the maximum-likelihood ones

By choosing the maximum likelihood estimator of the population frequencies and computing the mutual information just from it, one disregards completely other long-run frequencies which might have only a slightly smaller likelihood.

As an example consider the three candidate long run frequencies for the sample \hat{s} in sec... displayed in Fig.1. They look pretty similar and all

have log-likelihoods which deviates less than 10% from the maximum of the log-likelihood. Although these long-run frequencies might have generated our data with similar probability (likelihood), they yield rather different values of mutual information between spike count and head direction (MIs up to 50% apart). The examples in Fig. 1 therefore suggest that the maximum-likelihood mutual information (bottom of Fig. 1) alone might be poorly representative of the spectrum of mutual information values corresponding to high-likelihood long-run frequencies. As a consequence, we argue that a logical approach to mutual information estimation must encompass a range of long-run frequencies, while taking into account their likelihood.

Maximum-likelihood estimates disregard prior information

By choosing the maximum-likelihood frequencies as the only candidate long-run frequencies for our data, one approaches the sample in a completely agnostic fashion. The assumption made by maximum-likelihood estimation is indeed that the researcher doesn't have any kind of pre-sample knowledge about the biologically plausible candidate long-run frequencies, or in other words, that the system of interest could in principle attain any population frequency with equal probability, until the sample is collected. Such an assumption is typically wrong, as:

1. **the very same neural system might have been probed before**, providing us evidence in favor of some candidate long-run frequencies over others. Suppose for example that, before collecting our sample \hat{s} in fig ... , we had recorded the activity of the same neuron, while the animal was exploring the same environment. Suppose that from this recording we estimated a population mutual information between spike count and head direction of 0.1 (an its associated uncertainty). When we now are about to collect the new sample \hat{s} in fig ..., we don't just want to disregard this previous information on the system and regard every frequency as a priori equally good candidate for generating the new sample. We would rather consider the long-run frequency at the top of Fig. 1 as a better candidate than the one at the bottom of Fig. 1, since the latter corresponds to a much higher mutual information.
2. **the neural system of interest might have been investigated before and its results reported in the literature**. For our example of spike

count vs head direction, we might know that the region we are recording from is characterized by bursty neurons with low, but different from zero, baseline firing rate. A-priori we would therefore assign a larger probability to the long-run frequency in the middle of Fig. 1 than to those at the bottom of Fig. 1, for which the neuron is always silent in one of the two head-direction conditions.

3. **biological constraints on the long-run frequencies for the system under investigation might be well established.** Assigning equal probability to every long-run frequency means that before conducting our experiment we regard long-run frequencies which are not attainable by the system of interest on equal footing with those actually attainable. In our example, where the response variable is the spike count, we know that a long-run frequency corresponding to an average firing rate larger than 500Hz violates physiological constraints on the neural activity. Intuitively we want to assign such long-run frequencies very low, if not zero, a priori probability.

Once the sample \hat{s} is collected and attributes maximum likelihood to the long-run frequency at the bottom of Fig. 1, we intuitively want to combine this new information on long-run frequencies provided by the likelihood, with the old (prior) piece of information about the system. Disregarding prior information would simply be poor practice.

In this section we explained why considering only one long-run frequency, even if it's the log-likelihood one with its nice asymptotic convergence to the truth, may be reductive for the purpose of estimating the mutual information. So if a single estimator of the mutual information based on solely the likelihood of the long-run frequencies is not a good estimator of the mutual info: how to construct a better one by encompassing a range of candidate long-run frequencies and incorporating prior knowledge on the system? We answer to this question in section ..., whereas in section ... we elaborate on how to formally express prior knowledge about the system.

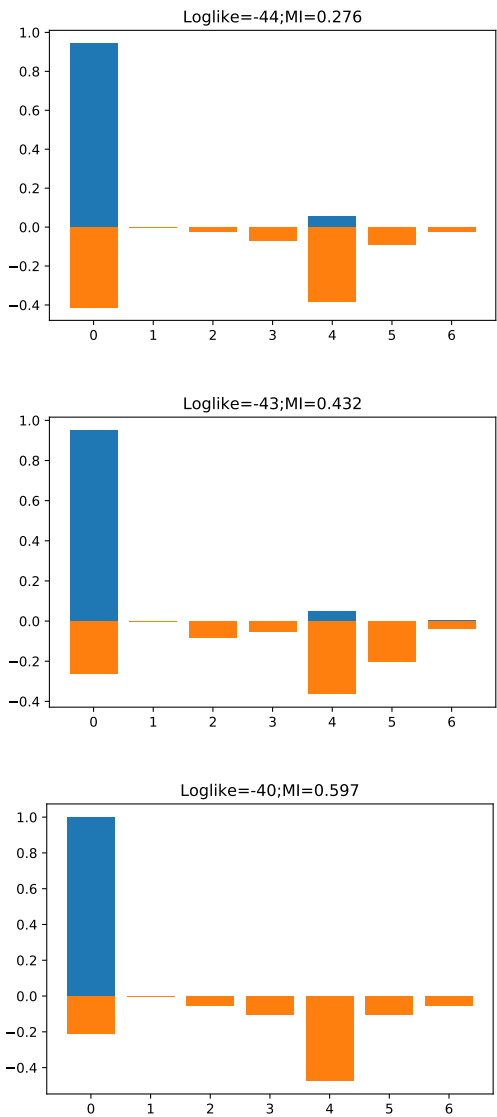


Figure 1 High likelihood long-run frequencies for the data in sec. ??. Examples of high likelihood long-run frequencies corresponding to low MI (top), medium MI (center), high MI and maximum likelihood (bottom).

2 Priors

In the previous section we calculated the likelihood of three candidate long-run frequencies. The examples presented in figure 1 exhibit similar and high likelihood values, which suggests that the data could have well been generated from those three distributions of long-run frequencies. However, we observed that those distributions lead to different values of long-run mutual information (MI). This is a direct consequence of having limited samples; the more samples are available, the more the sample frequencies will converge to the long-run frequencies and thus the sample MI will converge to the long-run MI (see section ??). When the data is scarce, however, we can tackle the uncertainty in the inference of the long-run frequencies by making use of all relevant information about the system of study available at our disposal. In the context of computing mutual information, this translates into assigning weights to the candidate long-run frequencies the data could have been sampled from. Those weights will express our prior knowledge of the data and thus will depend on characteristics of the study: for example the employed recording technique, the brain area under study, the stimulus applied, the repertoire of possible neuronal responses, etc. Based on such weights we define the *superdistribution* as the distribution of all candidate long-run frequencies. Note that this distribution lives in the space of candidate long-run frequencies. [🔧 Don't know if I should add an equation here to illustrate the superdistribution](#)

In order to illustrate what the superdistribution is, let us consider a discrete space of long-run frequencies that encompasses one, two or three spikes. We can conceptualize these distributions as lying on the surface of a triangle (see figure 2), where each dot illustrates one candidate long-run frequency, and the vertexes corresponds to firing *only* one, two or three spikes. One possible superdistribution would be the one that favours the vertexes (figure 2B). This, however, is not a biologically realistic assumption as we know cells fire stochastically and do not always fire exactly the same number of spikes per bin. Based on the assumption that the cell under study has a low firing rate, and most likely only fires one or two spikes per bin, we could instead choose a superdistribution that assigns larger weights to those histograms (figure 2C).

In the example of the NS cell figure ?? we presented three candidate long-run frequencies among all possible candidate distributions (figure 1).

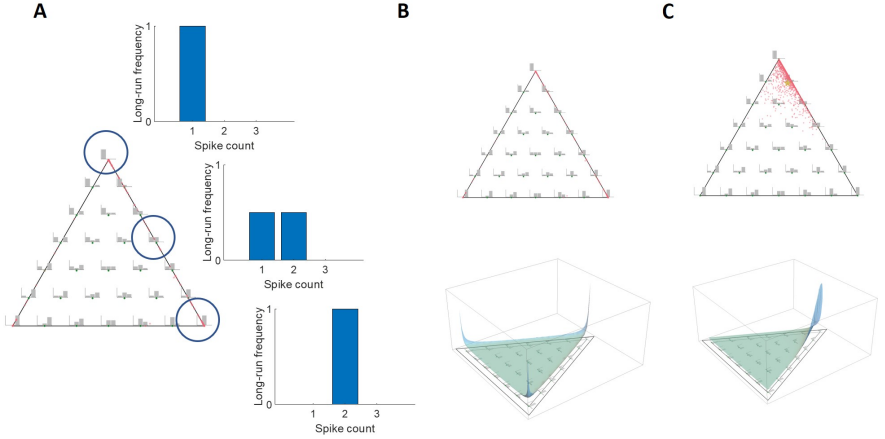



Figure 2 (DRAFT OF FIGURE) A: Schematic of the candidate long-run frequencies space. Each histogram corresponds to one candidate long-run frequency. B-C: Candidate long-run frequencies space (top) and superdistribution (bottom) for two different superdistributions. The red dots indicate the weights assigned to each long-run frequency. (check with Luca)


One possible superdistribution is the one that assigns the same weight to all candidate long-run frequencies, including the three examples. While this is a valid superdistribution (all superdistributions are) it is not biologically plausible, as it assigns equal weights to distributions that favour spiking at very high rates as well as non-spiking at all. Another possible superdistribution is the one that assigns weights different from zero to the distributions present in figure 1, and weights equal to zero to the rest. This could represent an improvement as the distributions with non-biologically plausible firing would get weights equal to zero. As for the three long-run frequencies shown in figure 1, we could assign their weights based on what we know a priori about this neuron. For example, this cell has a mean firing rate of 18.8 Hz, which favours histograms A C. If in addition we know that the cell tends to fire in bursts of two spikes, then the superdistribution should favour histogram A.

Now let us turn to more general examples that illustrate superdistributions over a continuum of candidate long-run frequencies. From now on we will use the terms *prior* and *superdistribution* interchangeably.

(i) Uniform superdistribution over frequencies: When there is no information available about the system under study the only choice left

is to use an uninformative prior, which in a discrete case consists of a uniform distribution over all candidate long-run frequencies. However, in a continuous space, this doesn't hold anymore because this will depend on the parametrization of the space. We could, for example, choose a superdistribution that gives equal weights to equal intervals of conditional frequencies $f(r | s)$. That is, the same weight to each hypercube $\prod_{r,s}[f(r | s), f(r | s) + \Delta]$, for fixed Δ , for all values of $f(r | s)$. Such a superdistribution is proportional to $\prod_{r,s} df(r | s)$.

(ii) Uniform superdistribution over sequences: Given some conditional frequencies $f(r | s)$, a *sequence* is defined as a set of n responses for each stimulus, with the responses appearing with frequencies $f(r | s)$. Therefore, instead of using equal weights over the frequencies, we could use equal weights over the sequences. In this case the weights given to equal intervals in the frequency space will not be uniform, because some frequencies are realized by more sequences than others. Then we would obtain a superdistribution proportional to $\prod_s M\{f(r | s)\} df(r | s)$, where M is a multinomial coefficient.  from Luca: will write the exact formula

(iii) Non-uniform superdistribution over sequences: Choosing a uniform prior over the sequences could be debatable from a biological point of view. If the responses for instance represent the firing rate of a population of cells, low responses should generally be expected more often than very high responses. Thus, more weight should be given to sequences in which low responses occur more frequently than high responses. This would lead to yet another superdistribution on the space of frequencies.  from Luca: will write the exact formula

(iv) Not factorizable distribution over stimuli: Should the superdistribution be factorizable over the frequency distributions? For example, for the two stimuli considered in the example above? Owing to biological constraints some similarity across the distributions should be expected. Thus such factorizability might be not be a sensible assumption.

Different superdistributions reflect different assumptions about the data. But to which extent does the choice of superdistribution affect long-run MI? We explore this by comparing the values of long-run MI obtained with the four superdistributions listed above. We proceed as follows: From a given superdistribution we sample a pair of long-run response-frequency distributions. From this pair calculate the long-run

mutual information. We then sample 20+20 responses from the pair, and calculate the sample mutual information obtained from such sample. We repeat this process ?? times. Figure 3 shows a scatter plot that tells us how often we should observe every pair of

(long-run mutual info, sample mutual info)

under the assumption of the given superdistribution, for the four superdistributions described above.

🔑 Luca wrote this: Consequently, any inferences of long-run from sample and any quantifications of “bias” heavily depend on the assumed superdistribution. – I would remove it from here and place it in the bias section

The four examples of figure 3 show that the joint distribution of long-run & sample mutual informations can be wildly different depending on the assumed superdistribution. Therefore, choosing a “default” superdistribution to be universally used for this kind of inference¹ is not a sensible option, as any one-fits-all choice would simply fit every concrete case extremely poorly. At the same time, not choosing a superdistribution is impossible, as any proposed algorithm or formula to infer the long-run MI from the sample one is explicitly or implicitly choosing a superdistribution. Hiding such a choice will just lead to the same poor inferences as a default choice. We are then left with the need to choose a superdistribution as best as possible from considerations of the specific study. Any such choice, even if based on a very cursory analysis, will always be better than any default choice that completely disregards the specific case.

3 Posterior of long-run frequencies

In section 1 and Example ?? we argued that for a limited sample estimating the MI by maximum likelihood might be a poor choice. As the mutual information, defined in ..., is a smooth function of the long-run frequency, the problem should be traced back to the estimation of the long-run frequencies from the sample. The issue of estimating the long-run frequencies can be approached logically as follows:

1. **before conducting the experiment / looking at the data** find a set (space) of candidate long-run frequencies $\{f_s^{(c)}\}_{c \in C}$ and attribute a

¹ nemenmanetal2004.

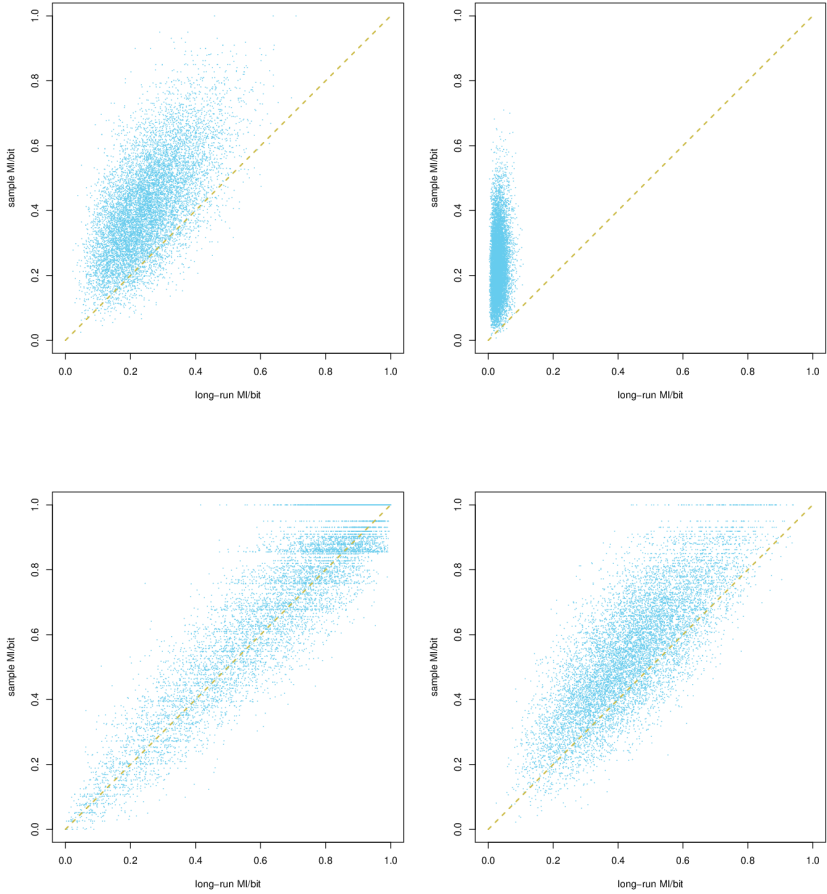


Figure 3 Top left: uniform over frequencies. Top right: more uniform over sequences. Bottom left: low responses preferred. Bottom right: not factorizable over stimuli (positive correlation; hierarchic)

probability $P(\mathbf{f}_s^{(c)})$ to each of these candidates. Set and probabilities must be chosen based of prior knowledge about the functional and physiological properties of the neural system of interest (e.g. for a single neuron average firing rate within the brain region and tuning to stimulus, if known).

2. **after conducting the experiment / looking at the data** update the probability of the candidate frequencies by means of the data \hat{s} via the likelihood:

$$P(\mathbf{f}_s^{(c)}|\hat{s}) = \frac{P(\hat{s}|\mathbf{f}_s^{(c)}) * P(\mathbf{f}_s^{(c)})}{P(\hat{s})} \quad (2)$$

where $P(\hat{s}) = \sum_c P(\hat{s}|\mathbf{f}_s^{(c)}) * P(\mathbf{f}_s^{(c)})$ is a normalization factor.

Equation (2) in probability theory is known under the name of Bayes theorem and $P(\mathbf{f}_s^{(c)}|\hat{s})$ expresses our belief into the long-run frequencies $\mathbf{f}_s^{(c)}$ after having seen the data, once we start from $P(\mathbf{f}_s^{(c)})$.

 Explain how this approach resolves all issues with maximum likelihood at least conceptually.