# Reasoned inference of long-run mutual information
## (Bayesian theory for dummies) [draft]

P.G.L. Porta Mana <pgl@portamana.org>
C. Battistin <claudia.battistin@ntnu.no>
S. Gonzalo Cogno <soledad.g.cogno@ntnu.no>
(or any permutation thereof)

31 March 2019; updated 28 January 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 introduced the definition of Mutual Information between stimulus and responses (some measure/parametrization of neural activity), as a function of the long-run conditional frequencies of the responses on the stimulus. Unfortunately, for any biological neural system, the long-run frequencies are unknown, while what we have at our disposal is typically a limited[1] sample drawn from these unknown long-run frequencies. In order to compute the mutual information between stimulus and response, we hence want to use our data to make a guess of the long-run frequencies. How can we tell which long-run frequencies would most likely have generated our data? Let's think at the problem in reverse: if we knew the long-run frequencies then the likelihood $P(\hat{s}|\mathbf{f_s})$ would tell us how likely is our sample. It follows that for in the case of unknown long-run frequencies, the likelihood of *candidate* long-run frequencies for our data can instruct us on how likely they might have generated the sample. How to choose the candidate long-run frequencies though? And, should be the likelihood the only

[1] limited is a vague word. In this work we consider a sample limited if the number of datapoints is of the same order of magnitude of the possible combinations of stimulus and response value.

metric for appraising long-run frequencies? One might be tempted to chose as a candidate only the long-run frequencies that maximize the likelihood for our data, namely the sample frequencies. Let's see why this problematic for a limited sample:

1. By choosing the maximum likelihood estimator of the long-run frequencies, one disregards completely other long-run frequencies which might have only a slightly smaller likelihood. Intuitively this is not a big issue, if the mutual info between stimulus and response for the disregarded long-run frequencies is similar to the mutual information associated to the maximum likelihood (sample) long-run frequencies. This might not always be the case, see the continuation of Example 1 in sec. 1.1.
2. By choosing the sample frequencies as the only candidate long-run frequencies for our data one approaches the sample in a completely agnostic fashion. This means that the assumption made by this choice is that the researcher doesn't have any kind of pre-sample knowledge about what the biologically plausible candidate long-run frequencies are, which is: every long-run frequency is attained as equally likely until the sample is collected. This is often not true for at least two reasons:
    a. the exactly same neural system might have been probed before, providing us some evidence in favor of some candidate long-run frequencies;
    b. biological constraints on the long-run frequencies for the system under investigation might be well established and be reported in the literature.
3. By choosing the maximum-likelihood frequencies and computing the mutual information from them, one typically disregards the actual value of the likelihood. This value might be instead exploited to express our degree of belief (and uncertainty) on our mutual information estimate, as inherited from the degree of belief on the long-run frequencies that we chose (in this case the sample ones).
4. mutual info bias?

In the next section we will see how a logical approach to the estimate of the long-run frequencies can address these issues regarding the mutual information.

### 1.1 Example 1 continued. Mutual Infromation of high likelihood long-run frequencies.

Consider now the sample introduced in sec. **??**, for which we are aiming at an estimate of the mutual information between the spike count (r) of this neuron and the north/south head direction (s). The mutual information is a function of the long-run frequencies and the likelihood of long-run frequencies tells us how probably the sample was generated from them. Given the long-run frequencies $\mathbf{f_s} = \{f_s(r)\}_{s,r}$ their likelihood for our sample $\hat{s}$ is:

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

where $\hat{f}_s(r)$ are the sample frequencies shown in Fig. **??**. The probability distribution in Eq. (1) is the categorical one and it assumes that the probability of response $r$ given that the stimulus is $s$ is $f_s(r)$ at each time step independently.

In Fig.1.1 three examples of long-run frequencies with high likelihood (log-likelihood within 10% of maximum log-likelihood) for the sample $\hat{s}$ in sec. **??** are displayed. 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 at least 50% apart). Therefore the examples in Fig.1.1 suggest that the sample mutual information (bottom of Fig.1.1) alone - for limited sample sizes - might be poorly representative of the spectra of mutual information values corresponding to high-likelihood long-run frequencies. 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?

## 2 Posterior of long-run frequencies

In section 1 and Example 1.1 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:

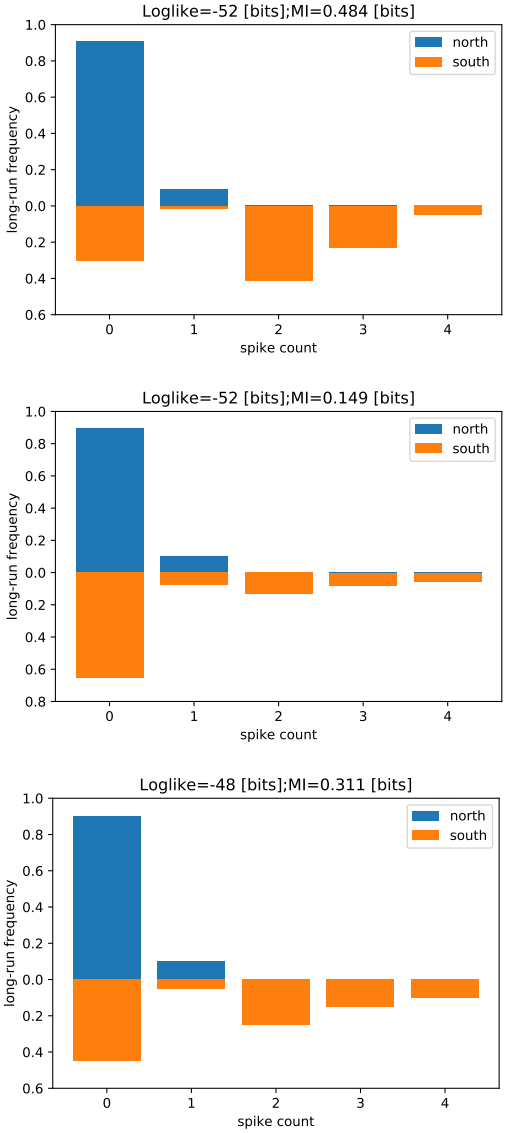


Figure 1 High likelihood long-run frequencies for the data in sec. **??**. Examples of high likelihood long-run frequencies corresponding to high MI (top), low MI (center). Maximum likelihood long-run frequencies (bottom).

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

$$P(\mathbf{f_s}^{(c)}|\hat{s}) = \frac{P(\hat{s}|\mathbf{f_s}^{(c)}) * P(\mathbf{f_s}^{(c)})}{P(\hat{s})} \tag{2}$$

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.

**3**

**4**