

# A direct route to the mutual information of stimulus and response

P.G.L. Porta Mana

[<piero.mana@ntnu.no>](mailto:piero.mana@ntnu.no)

C. Battistin

[<claudia.battistin@gmail.com>](mailto:claudia.battistin@gmail.com)

Draft of 2 April 2019 (first drafted 31 March 2019)

This note shows a direct route to forecast the mutual information of stimulus-response long-run frequencies from sample data of any size. Using an intuitive example it is demonstrated that our pre-data forecast about the long-run frequencies is crucial when the sample is small. It turns out that estimates considered ‘biased’ in some literature should be considered seriously instead: the response is indeed very likely to be informative.


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

## 1 Bias?

The mutual information between the long-run relative frequency of a signal and that of a response is a measure of how much our uncertainty about the signal is reduced by knowledge of the response. This measure is sometimes used in neuroscience, the response being some characteristic – such as the activity or the firing rate – of a neuron or of a network of neurons.

In this note we show a direct route to forecast the long-run mutual information, given a sample of stimulus-response frequencies. The forecast consists in a probability distribution for the values of the long-run mutual information. The sample can be of any size.

The intuitive reliability of the forecast is shown by a simple numerical example. The example also shows that the forecast always depends on our pre-data forecast of the response frequencies – the more so the smaller the sample is.

Our analysis is pertinent to recent works in neuroscience (Panzeri et al. 2007)  [add others](#) that use a more topsy-turvy approach to this inference problem. These works try to correct ‘biased’ estimates that appear to predict too high values of the long-run mutual information, when the sample size is small. The example we present shows that such ‘biased’ estimates may have to be taken seriously instead: we should

indeed expect the response to be very informative. This depends, again, on our pre-data forecast for the response frequencies.

For small samples it's therefore important to be very judicious in our pre-data forecasts. We discuss some methods, used in many other fields, to make such educated assessments.

## 2 Bayes

First of all let's state what our inference is about. Given a sample of stimulus-response data we want to assess what's the most probable set of long-run<sup>1</sup> relative frequencies of the response conditional on each stimulus value, and from these assess what's the most probable value of the associated mutual information between stimulus and response. We assume that all stimuli appear with equal relative frequencies.

Let the stimulus  $s$  have two possible values  $\{-, +\}$ , and the response  $r$  ten possible values  $\{1, \dots, 10\}$ . Let the data  $D$  be a set of  $n$  stimuli  $-$  which yielded  $n$  responses  $(r_i^-)$ ,  $i \in \{1, \dots, n\}$ , and  $n$  stimuli  $+$  which yielded  $n$  responses  $(r_i^+)$ . The ten response state appeared with relative frequencies  $q^- := (q_r^-)$ ,  $r \in \{1, \dots, 10\}$ , for the stimulus  $-$ , and with relative frequencies  $q^+ := (q_r^+)$  for the stimulus  $+$ .

We can use the concrete data summarized in fig. 1, consisting in  $n = 20$  samples per stimulus.

If the long-run frequencies conditional on stimulus  $-$  are  $f^- := (f_r^-)$ , and conditional on  $+$ ,  $f^+ := (f_r^+)$ , then out of symmetry the probability of obtaining the data  $D$  is

$$p(D | f^-, f^+, K) = \prod_r \left[ (f_r^-)^{nq_r^-} (f_r^+)^{nq_r^+} \right]. \quad (1)$$

This is also the *likelihood* of the long-run frequencies in view of the sample. Their probability density is proportional to the likelihood, corrected by their initial probabilities  $p(f^-, f^+ | K)$

$$p(f^-, f^+ | D, K) \propto p(D | f^-, f^+, K) p(f^-, f^+ | K) =$$

$$p(f^-, f^+ | K) \prod_r \left[ (f_r^-)^{nq_r^-} (f_r^+)^{nq_r^+} \right]. \quad (2)$$

---

<sup>1</sup>'But this *long run* is a misleading guide to current affairs. *In the long run* we are all dead.' (Keynes 2013 § 3.I, p. 65)

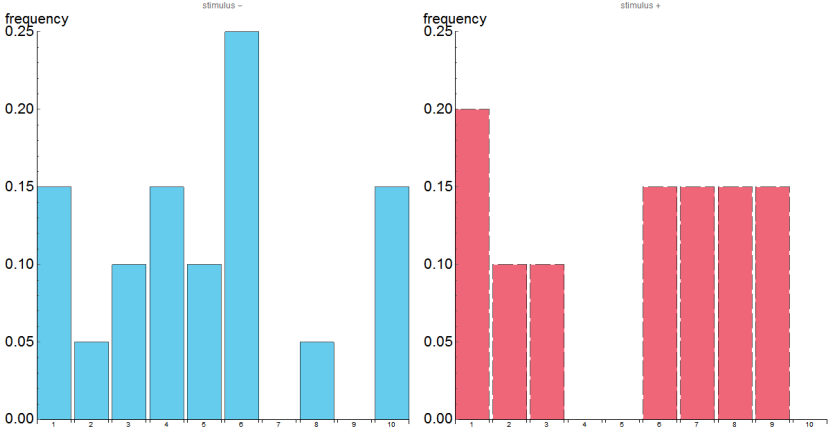


Figure 1

We can calculate this probability analytically when possible, or estimate it with Monte Carlo sampling. From such samples we estimate the probability distribution of the long-run mutual information

$$I := \sum_r \frac{1}{2} f_r^- \ln \left( \frac{\frac{1}{2} f_r^-}{\frac{1}{2} f_r^- + \frac{1}{2} f_r^+} \right) + \sum_r \frac{1}{2} f_r^+ \ln \left( \frac{\frac{1}{2} f_r^+}{\frac{1}{2} f_r^- + \frac{1}{2} f_r^+} \right). \quad (3)$$

Let's consider two possible probabilities for the long-run conditional frequencies:

## 2.1 Uniform uncertainty about the frequencies

If we initially think that equal ranges ( $\Delta f^-$ ,  $\Delta f^+$ ) of pairs of conditional frequencies are equally possible, then

$$p(f^-, f^+ | K_u) = 1. \quad (4)$$

We sample 5 000 pairs of conditional frequencies from the density (2) obtained from our example data and this pre-data probability. The resulting distribution of their associated mutual information is shown in fig. 2. It tells us that the response is likely to be informative; the most likely values of the long-range mutual information are around 0.2 bit. The next section explains intuitively why this probability distribution is correct.

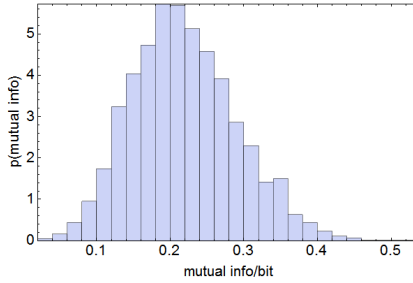


Figure 2

## 2.2 Conservative uncertainty about the frequencies

If we initially think that the long-run response frequencies conditional on the stimuli should be very similar, or if we simply want to do a conservative estimate, then our pre-data probability must be higher for pairs with similar conditional frequencies; for example [replace with combination of Dirichlet – same effect](#)

$$p(f^-, f^+ | K_c) \propto \exp \left[ \frac{\sum_r (f_r^- - f_r^+)^2}{2\sigma^2} \right]. \quad (5)$$

This density states that the two conditional frequencies should be roughly equal, but otherwise leaves a uniform uncertainty about the values of each. Smaller values of  $\sigma$  represent more conservative estimates.

A sample of 5000 pairs of conditional frequencies from the density (2) with the conservative initial density (5) [specify  \$\sigma\$](#)  leads to the distribution of mutual information of fig. 3. The estimate now says that

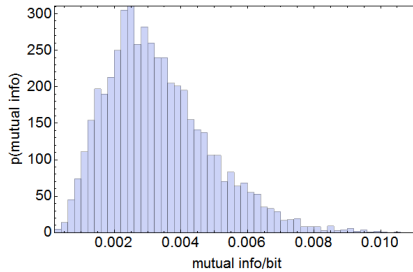


Figure 3

we should expect a negligible mutual information, with a most probable value around 100 times smaller than in the previous case.

### 3 Because

Let's try to understand why the estimate of § 2.1, fig. 2, is reliable; and to understand what happens in the conservative case of § 2.2.

If we deem all pairs of conditional frequencies equally possible, we can sample 5 000 pairs uniformly. Figure 4 shows the resulting scatter plot for first component of the two conditional frequencies, that is, the frequency of the response value 1.

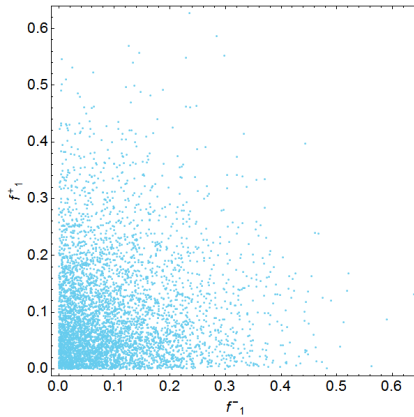


Figure 4

Let's now visualize the samples of long-run conditional frequencies in two dimensions as follows: for each sample, the horizontal coordinate is the probability that the frequencies assigns to our observed data; and the vertical coordinate is the mutual information associated with the frequencies. We obtain the scatter plot of fig. 5. All points, of three different sizes and colours, are part of the plot. The pair of two uniform conditional frequencies (1/10 probability for each response) is the largest, yellow point. This pair of long-run conditional frequencies assigns probability  $10^{-40}$  to the data and has zero mutual information.

It's clear from this scatter plot that the data strongly suggest a large estimated mutual information, for two reasons:

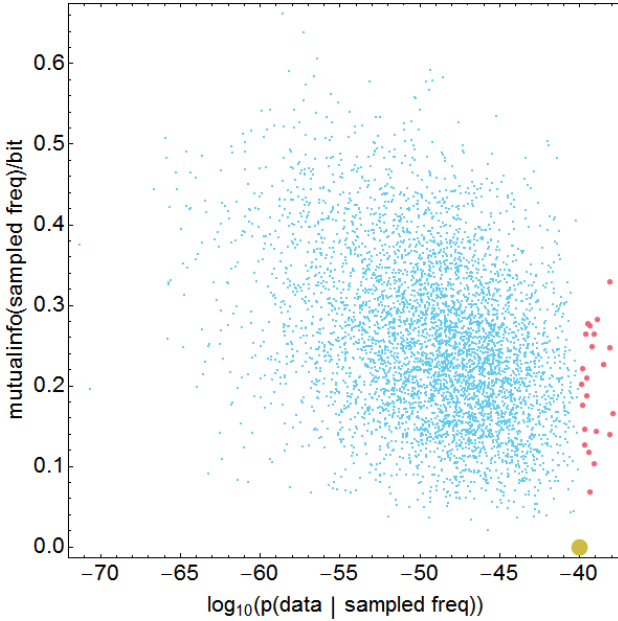


Figure 5

1. All pairs that assign very low probability to our data, say less than  $10^{-40}$ , are represented by small blue points. Each one is very unlikely to be the continuation of our data. But at the same time they constitute the overwhelming majority of possible pairs, and therefore there is a non-negligible probability that the data come from one of them. Most of them have high mutual information.
2. The pairs that assign higher probability to our data,  $10^{-40}$  or more, are represented by the larger red points and the largest yellow point. The majority of these pairs also have high mutual information. In fact, the pair of uniform conditional frequencies is an outlier: it's very unlikely that our data come from it, compared with the other possible pairs of frequencies.

The estimate of fig. 2 is therefore quite correct and reliable.

Now consider the conservative initial density (5), and sample 5 000 pairs of conditional frequencies from it. The scatter plot for the first components of the sampled frequencies is shown in fig. 6. It shows that

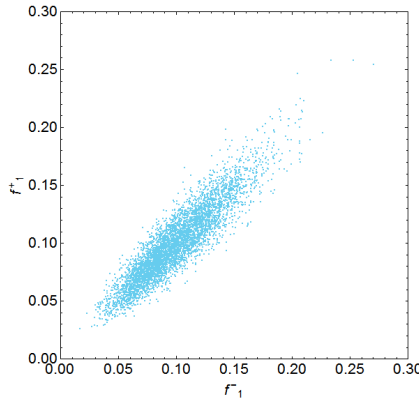


Figure 6

the two long-run frequencies are very similar to each other and close to  $1/10$ , the probability given by the uniform distribution.

Figure 7 shows the pairs of long-run frequencies plotted as in the previous case. There are now many pairs that assign roughly the same probability to the data as the uniform-distribution pair does. The majority of all pairs has a negligible mutual information. This is indeed reflected by the estimate of fig. 3.

## 4 Brief

We have calculated the probability distribution for the long-run mutual information, given our knowledge of a small sample. The calculation was conceptually straightforward and needed no corrections, even if it may require numerical sampling methods. [✚ Mention partial exchangeability.](#) We also saw (§ 3) why the distribution thus found is intuitively correct. The crucial point is the specification of the pre-data joint probability for the long-run conditional frequencies. This specification is especially important if the data are few. It's possible to specify a pre-data probability that express a conservative guess, and the resulting estimate of the mutual information is very close to zero.

This conclusion isn't surprising: the main point is the same as, for example, in the inference of diseases or some phenotypes from genetic peculiarities and vice versa. Imagine that we know the relative frequencies of two gene variants among people who have a particular

disease, and the relative frequencies among people who don't have the disease. Now we ask: given that a specific person has one of the gene variants, what's the probability that this person has the disease? The answer is that in general we don't know until we specify the incidence rate of the disease in the full population. Because even if that gene variant appears more frequently among people with the disease, the number of people having such disease may be so small that the probability still favours the person's being healthy.

The direct calculation via the probability calculus (Jaynes 2003; Sox et al. 2013; Hailperin 1996) makes any discussion about biases superfluous. We'd nevertheless like to discuss the distribution for the mutual information obtained by the sampling procedure in (Panzeri et al. 2007) for a pair of uniform conditional frequencies. Such a distribution makes sense: the long-run mutual information obtained from that small data sample should indeed be expected to be large.

The reason for this phenomenon can be explained by looking at fig. 5. We see that for our example data, the pair of uniform distributions is an outlier: there are many other pairs that have larger likelihoods under that

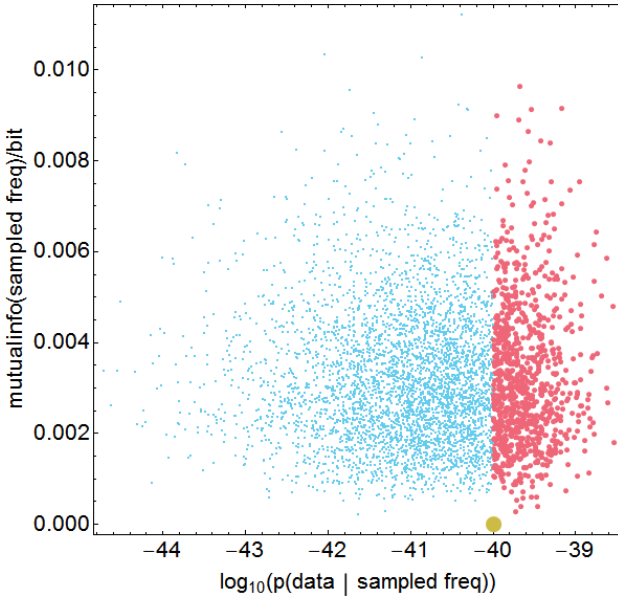



Figure 7




dataset. It turns out that for *most* small datasets generated by a pair of uniform distributions, the latter pair is an outlier; there are always many other pairs with larger likelihoods. The sampling procedure is implicitly selecting those more likely pairs and taking *their* mutual information as most likely.  The problem is that a small sample is as likely to have been generated by a uniform pair as by all possible pairs. Show the calculation about this.

The topsy-turvy point of view typical of frequentist-like analyses unfortunately leads to this kind of oversights. Our problem is not to infer data from a specific known frequency, but to infer an unknown frequency *among several possible ones* from known data – because our uncertainty about the long-run mutual information is a consequence of our uncertainty about the long-run frequency.

 Paragraph on shrinkage

## Baraka

PGLPM thanks Mari, Miri, & Emma for continuous encouragement and affection; Buster Keaton and Saitama for filling life with awe and inspiration; and the developers and maintainers of L<sup>A</sup>T<sub>E</sub>X, Emacs, AUC<sub>T</sub>E<sub>X</sub>, Open Science Framework, Python, Inkscape, Sci-Hub for making a free and unfiltered scientific exchange possible. 

## Bibliography

- (‘de X’ is listed under D, ‘van X’ under V, and so on, regardless of national conventions.)
- Hailperin, T. (1996): *Sentential Probability Logic: Origins, Development, Current Status, and Technical Applications*. (Associated University Presses, London).
- Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. <https://archive.org/details/XQUHIUXHIQUHIQXUIHX2>, <http://www-biba.inrialpes.fr/Jaynes/prob.html>.
- Keynes, J. M. (2013): *A Tract on Monetary Reform*, repr. of second ed. (Cambridge University Press, Cambridge). First publ. 1923.
- Panzeri, S., Senatore, R., Montemurro, M. A., Petersen, R. S. (2007): *Correcting for the sampling bias problem in spike train information measures*. *J. Neurophysiol.* **98**<sup>3</sup>, 1064–1072.
- Sox, H. C., Higgins, M. C., Owens, D. K. (2013): *Medical Decision Making*, 2nd ed. (Wiley, New York). First publ. 1988.