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 14 September 2020

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.

Modules:

1

Motivation for mutual info and description of the problem

2

Explain (figures, graphs) that reasoning about different samples is misleading.

3

Need to focus on long-run freqs -> they give us long-run MI. Need to focus on sample freqs -> sample MI is 1. misleading, 2. gives less info. Sufficient stats.

4

Claudia

Examples with equally likely long-run freqs – we can have very different longrun MI leading to same sample MI. Introduced by a discrete example. Multinomial probabilities

5

from the guess of the long-run freqs we guess the long-run MI -> distributions.

6

Claudia.

Weights (prior distribution) are needed besides likelihood – biological reasons or average reasons. Still from discrete example. Unavoidability of choice.

7

Soledad.

Motivation of different choices of weights/distribution

8

Soledad. Generalization to continuum -> Bayes's theorem (commented)

9

Dependence on number of samples. Why likelihood can become enough and why sample MI is good guess. Laplace approx.

10

Why a full distribution is better than a point estimate (+ variance). Possible uses in neuroscience of this distribution.

11

Discussion of bias: becomes irrelevant

12

General discussion

OLD TEXT

13 Long-run mutual info and sample mutual info

We have, say, two stimuli and ten possible neural responses to each stimulus, under specific environmental and behavioural conditions, and for a specific class of subjects. The responses could be, for example, ten different firing rates of a specific neuron, or ten different total-population activities of a brain region, to two different visual stimuli.

Imagine that a researcher gave us the results of 10^{60} stimulus-response measurements, all performed for the specified conditions and subjects, and in which the two stimuli occurred equally often. From the long-run[1] joint frequencies of the given stimulus-response pairs we could calculate the long-run mutual information between stimulus and response.

But we don't have access to such a wealth of observations, and never will. Suppose we only have actual knowledge of a sample of few, say 20, measurements for each stimulus. They are a sample from such unobservable long-run sequence. We can calculate the joint frequencies of stimulus-response pairs in this sample and the resulting sample mutual information.

Our question is: how do the mutual information of the sample and of the long-run sequence differ?

Let's explore some possibilities.

[1] "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)

Example A. Suppose the long-run frequencies of responses conditional on each stimulus are as given[2] in the top panel of fig. 1. The light-grey histogram is for the responses to the first stimulus; dark-grey for the second. In the long run all responses occur equally often for either stimulus. The long-run mutual information is 0 bit.

Now we consider a sample of 20 response measurements for each stimulus from such long-run distribution. From this sample we calculate the joint frequencies and their mutual information. The exact long-run sequence is unknown, so such a sample could yield different results. Considering all long-run sequences having the frequencies of fig. 1 as equally plausible, we plot some possible results for the sample mutual information in the bottom panel, as a scatter plot (blue dots) and with a histogram (10 000 samples, a subsample of 100 shown in the scatter plot). The long-run mutual information is indicated by the red line. The sampled mutual information is most surely always higher than the long-run one in this case. The long-run one is even outside the inner 95% interval range.

[2] as the example in Panzeri et al. 2007.

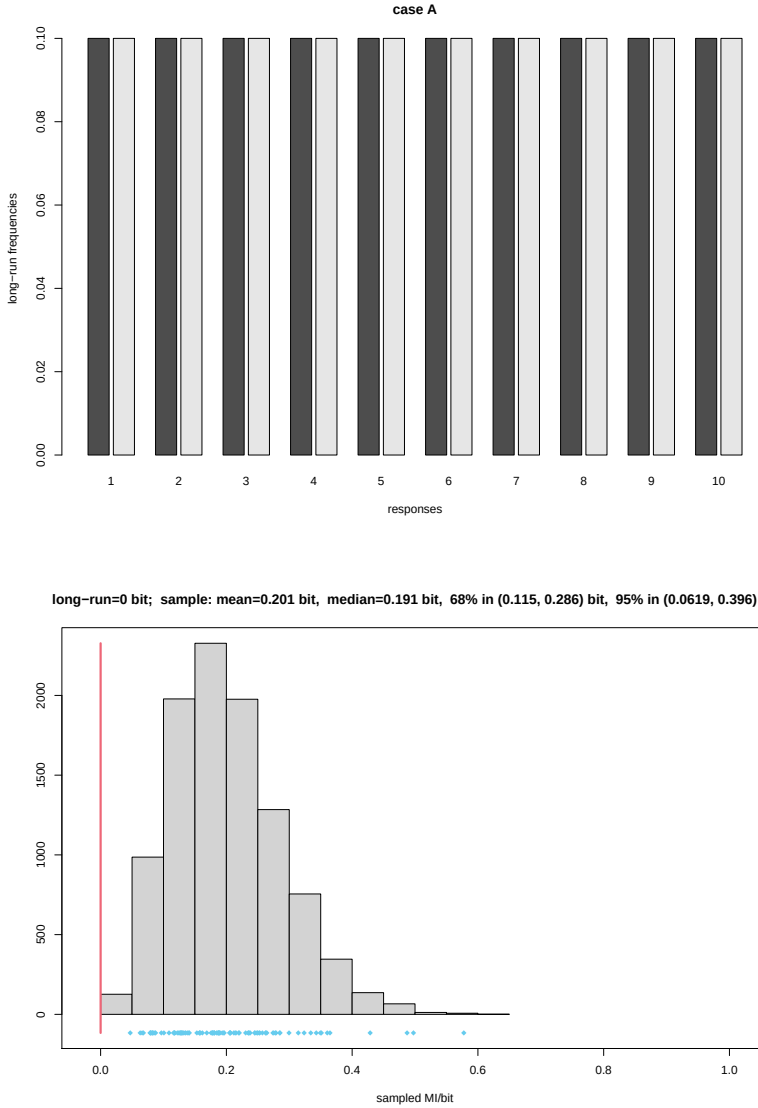
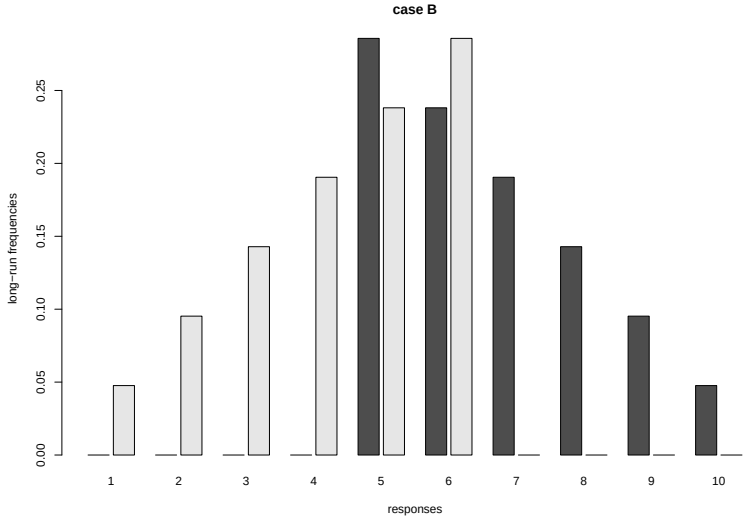


Figure 1 Example A

Example B. The long-run frequencies of responses are given in the top panel of fig. 2. The responses do not occur equally often in the long run, and the frequencies of some of them are different for the two stimuli. The long-run mutual information is 0.48 bit.

We again consider a sample of 20 response measurements for each stimulus. The histogram of the possible values of sample mutual information is shown in the bottom panel. The possible values are almost symmetrically distributed around the long-run value, which is well within their inner 68% quantile range, close to the median and mean, 0.51 bit.



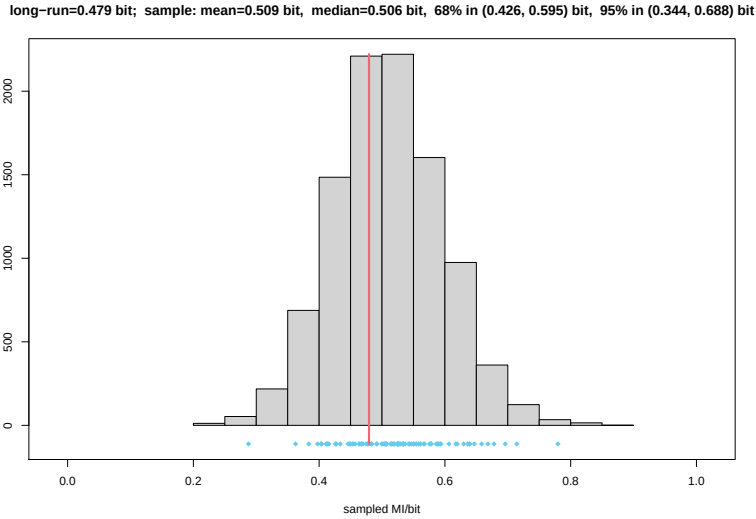


Figure 2 Example B

Example C. The long-run frequencies of responses are given in the top panel of fig. 3. There is some difference in the long-run conditional frequency distributions of the responses . The long-run mutual information is 0.38 bit.

The histogram for the possible values of the mutual information from 20 samples is shown in the bottom panel. The long-run mutual information is outside the inner 75% quantile (not reported).

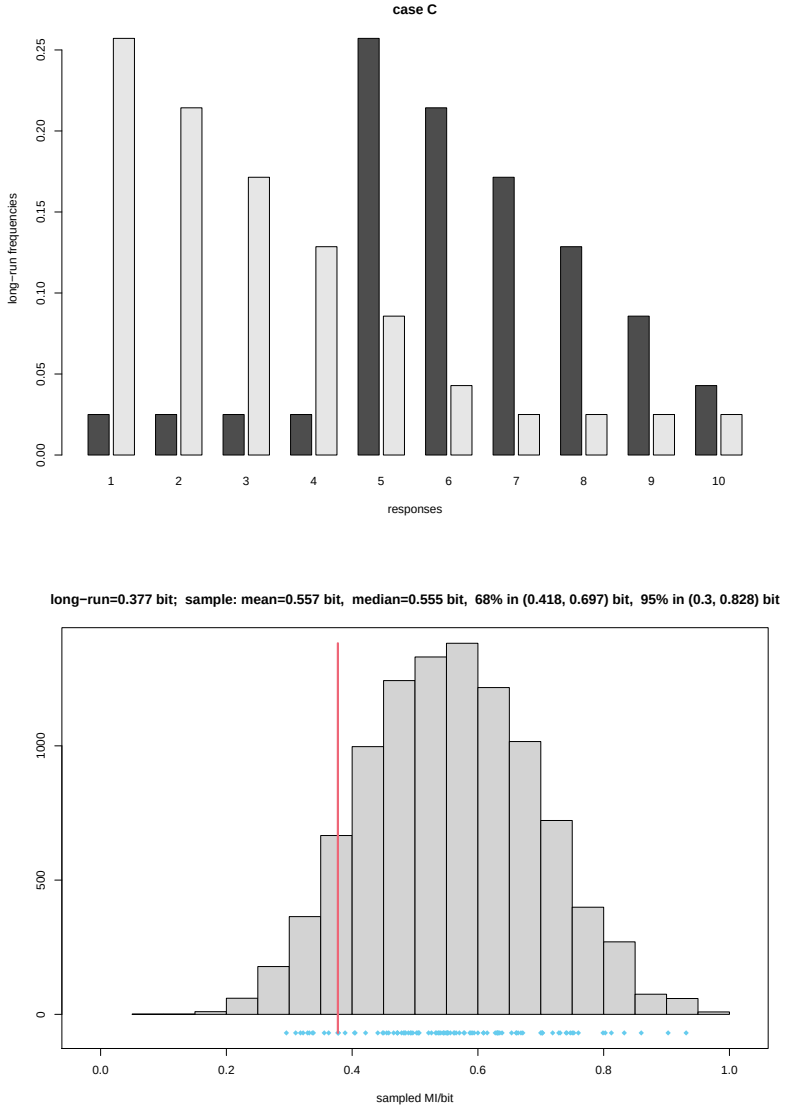


Figure 3 Example C

Example D. The long-run frequencies are shown in the top panel of fig. 4. The conditional frequency distributions of the responses have almost no overlap. The long-run mutual information is 0.92 bit.

The histogram of possible values of the sampled mutual information is shown in the bottom panel. The histogram is very skewed, with an average of 0.97 bit (the histogram's support is almost divided in distinct blocks; this is the result of the discreteness of the possible values of the frequencies from the sample and their little-overlapping supports). The long-run mutual information is within the inner 68% quantile.

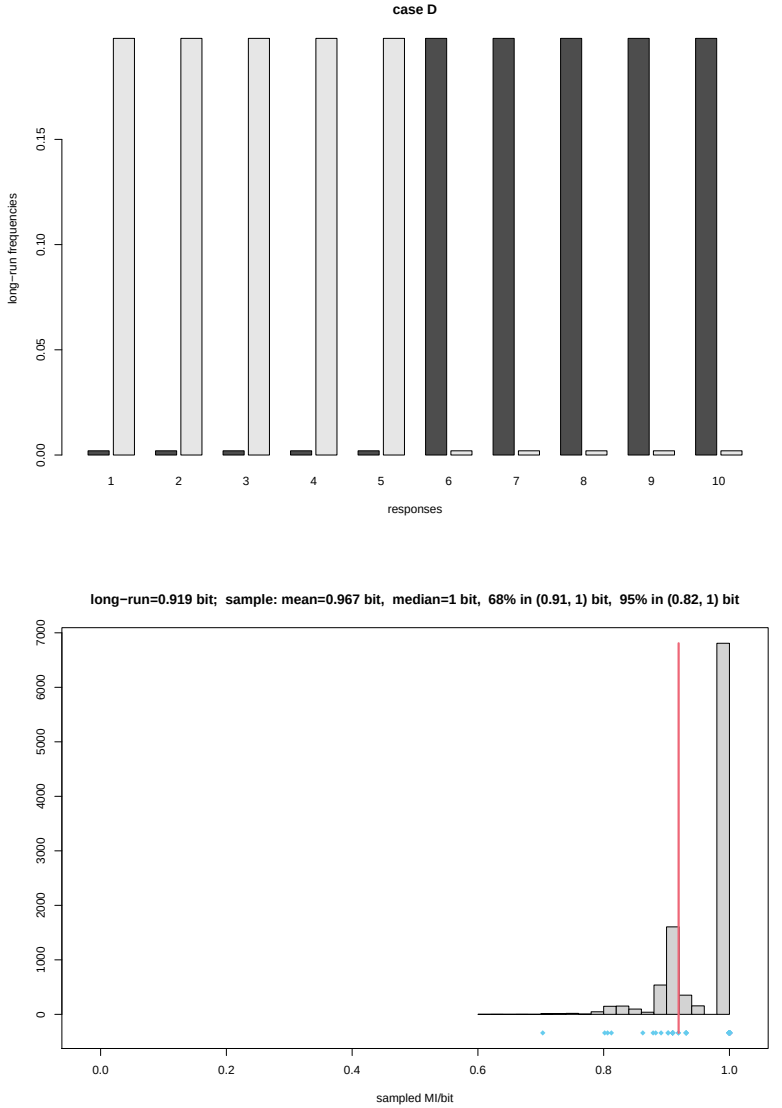


Figure 4 Example D

What conclusions can we draw from the examples above?

(i) In some cases the sample mutual information may be quite close to the long-run one; in others, very far from it.

(ii) The range and spread of the possible values of the sample mutual information are heavily dependent on the long-run response distributions.

(iii) In some cases the spread (say, the inner 95% interval) of the possible values is over a short range, compared with the maximal possible range of the long-run mutual information, $[0, 1]$ bit: for instance, 18% of the full possible range in Example D. In other cases the spread is over a very broad range compared to the maximal possible one: for instance, 53% of the full possible range in Example C.

(iv) The distribution of these possible values is in some cases very skewed.

Owing to the preceding points, it is misleading to just report the mean of the sample mutual information, even when such a mean is close to the long-run one.

Thus it is not possible to say, in general, that the long-run and the sample mutual informations have values far apart or very close. Note that even speaking of the "mean" of the sample mutual information is misleading: from a set of observations we only get *one* value of the sample mutual information, not the mean. And we cannot know what the mean is, because we don't know what the long-run conditional frequencies are – if we knew them we wouldn't need to make guesses.

Therefore, once we calculate the mutual information from a small sample we don't know whether case A, B, C, D above, or some completely different case, applies. Suppose we observe 20 trials and we find a sample mutual information of 0.5 bit. How do we know whether this should be considered a sample from the right tail of fig. 1 (so the long-run mutual info is actually lower)? or from the left side of fig. 2 (the long-run is actually higher)? or from the left of fig. 3 (long-run is lower, in this case)? or finally from the left tail of fig. 4 (long-run is higher)?

The latter remark is also important in regard of "validations by simulation". They would be validations only if the chosen long-run frequencies for the simulation were *representative of the true ones*. To check whether they are truly representative, we would need to know the true

ones. And we don't – this was indeed the problem we started from. Such "validations" are therefore logically circular and completely groundless.

14 What are the long-run frequencies?

One may say "OK, I don't know the long-run response frequency distributions, but I may check what's the relation between long-run & sample mutual informations on average among all possible such distributions". The problem with this approach is that to calculate such "average" we must specify a distribution over all possible pairs of conditional frequency distributions – let's call this a 'super-distribution' to keep it distinct from the conditional-frequency ones.

How should we choose such a superdistribution?

The answer "we choose a uniform one" has no meaning, because in a continuous space, as in the present case, there is no notion of "uniform" distribution – it depends on how we parametrize the space. And it isn't clear whether some parametrization is more natural than some other.

We could, for example, choose a superdistribution that gives equal weights to equal intervals of conditional frequencies. That is, the same weight to each hypercube $\prod_{r,s}[f(r \mid s), f(r \mid s) + \Delta]$, for fixed Δ, for all values of $f(r \mid s)$. Such a superdistribution is proportional to $\prod_{rs} \mathrm{d}f(r \mid s)$.

But if we consider the possible *sequences* of observable stimulus-response pairs, we could say that every such sequence should be given equal weight. Then the weights given to equal intervals in the frequency ranges *cannot* be uniform, because some frequencies are realized by more sequences than others. We would obtain a superdistribution proportional to $\prod_s M\{f(r \mid s)\}\, \mathrm{d}f(r \mid s)$, where M is a multinomial coefficient. will write the exact formula

But this last choice could also be debatable. If the responses for instance represent population activities or rates, we know that low responses should generally be expected more often that high responses, owing to biological reasons and out of common experience. Thus we should give more weight to sequences in which low responses occur more frequently than high responses. This would lead to yet another superdistribution on the space of frequencies. will write the exact formula

Finally, should such a superdistribution be factorizable over the frequency distributions for the two stimuli (as is the case in the three

examples above)? Owing to biological constraints there should be some similarity across such distributions, and such factorizability might be not be a sensible assumption.

We can see the importance of the choice of superdistribution over pairs of long-run conditional frequency distributions by examining some examples. Let's proceed as follows. A superdistribution is given. Sample from it a pair of long-run response-frequency distributions. From this pair calculate the long-run mutual information. Then sample 20+20 responses from the pair, and calculate the sample mutual information obtained from such sample. This way we obtain a density or scatter plot that tells us how often we should observe every pair of

$$(\text{long-run mutual info}, \text{sample mutual info})$$

under the assumption of the given superdistribution.

The results are shown in fig. 5 for the four superdistributions entertained in the discussion above.

Note that this kind of plot could be used to guess the long-run mutual information *if we only knew* the mutual information observed in the sample, under the assumption of the specific superdistribution underlying the plot. An example is given in fig. 6: if in an observation of 20 responses per stimulus we find a mutual information of 0.2 bit, then the long-run mutual information is at 68% between 0.15 bit and 0.30 bit, with a median of 0.22 bit. Such inference assumes the superdistribution underlying the bottom-right of fig. 5.

We shall see in § 16, however, that knowledge of the sample conditional frequencies provides additional data besides the mutual information that can be calculated from them. Thus our guess about the long-run mutual information should be calculated using the whole sample frequencies.

The four examples of fig. 5 show that the joint distribution of long-run & sample mutual informations can be wildly different, depending on the assumed superdistribution. Consequently, any inferences of long-run from sample and any quantifications of "bias" heavily depend on the assumed superdistribution.

We obviously can't say "let's just take the average over all possible superdistributions", because that would just lead to the analogous problem of choosing a super-superdistribution, and so on.

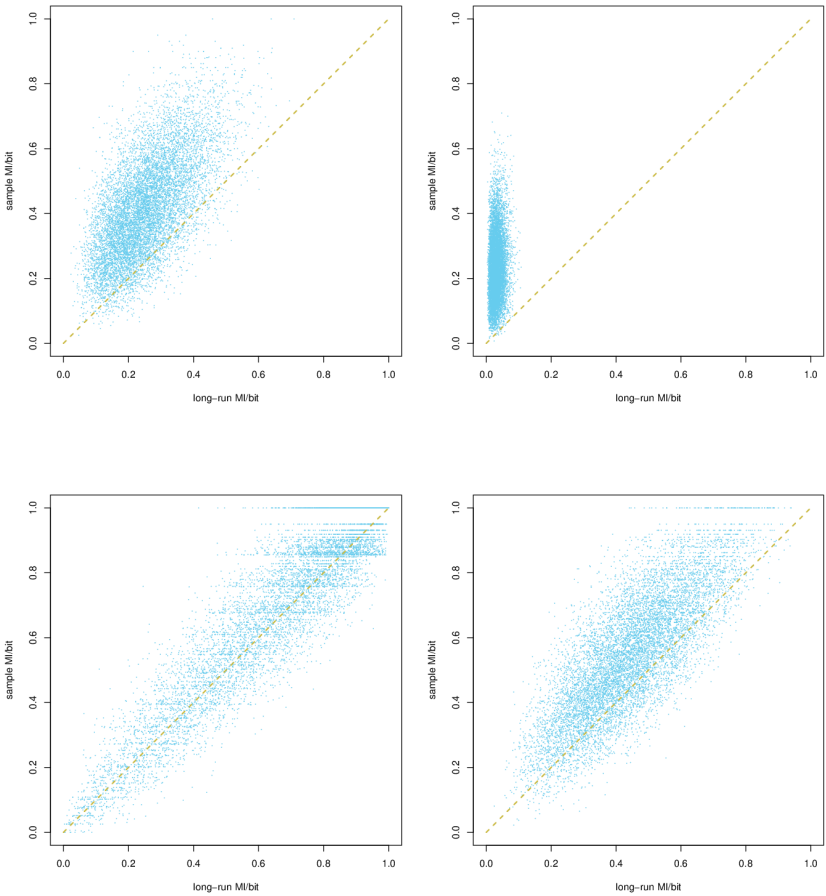


Figure 5 Top left: uniform over frequencies. Top right: more uniform over sequences. Bottom left: low responses preferred. Bottom right: not factorizable over stimuli (positive correlation; hierarchic)

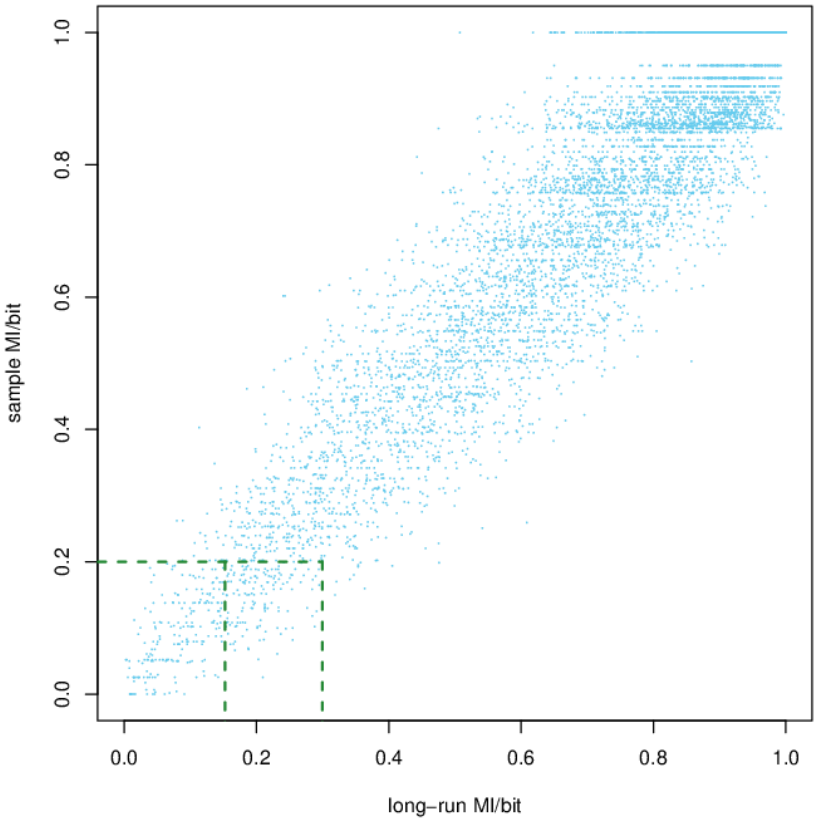


Figure 6 Example of guess for long-run mutual information from the mutual information observed in 20 samples per stimulus, under a specific superdistribution. If the sample mutual info is 0.2 bit (green dashed horizontal line), then the long-run mutual info is in this case inferred to be at 95% between $[0.097, 0.39]$ bit and at 68% between $[0.15, 0.30]$ bit (green dashed vertical lines), with a median of 0.22 bit, not far from the sample one (compare with the different conclusions in Panzeri et al. 2007 Fig. 1A).

Choosing a sort of "standard" or "default" superdistribution, to be universally used for this kind of inference[3], is not a sensible option either. The responses' frequency distributions that we can expect are extremely different depending on what kind of quantities the responses are (rates, total activities, bin totals, trial averages, and so on), on the brain region under study, and on the environmental, physiological, behavioural

[3] cf. Nemenman et al. 2004.

conditions of the stimulus-response situation. Any one-fits-all choice would simply "fit" every concrete case extremely poorly.

Not choosing a superdistribution is impossible. It should be clear from fig. 5 that any proposed algorithm or formula to infer the long-run mutual info 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.

The only possibility – it is simply forced on us – is to choose a superdistribution as best as possible, from considerations of the specific stimuli, response quantities, brain region, and so on. 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.

15 Conclusions

Speaking of estimators and biases for the long-run mutual information in small-sample situations is very misleading. The distributions involved are so broad and skewed that any estimator would be a poor guess. It is best to give one or more credible intervals for the long-run mutual information, as shown in fig. 6.

Moreover, such credible intervals or the bias of the estimator are heavily dependent on our assumptions (superdistribution) about the possible response frequencies – as should be expected in a small-sample situation. Owing to this heavy dependence, any universal formula for the credible interval or for the bias would be universally poor. Consider, for example, what would happen if any of the four assumptions underlying the plots of fig. 5 were taken as universal and used in the other three cases.

Finally, since the long-run response frequency distributions are unknown to us, and the choice of a superdistribution is heavily case-dependent, any "validation" based on simulated data[4] is logically circular – just wishful thinking. Such simulation depends on a choice either of long-run frequencies (and how do we know they are representative of the true ones?), or of a superdistribution (and is such choice appropriate for the specific case?).

[4] cf. Panzeri et al. 2007 *Comparing procedures to correct for the bias*.

(Note that the inference made in fig. 6 is simply a Bayesian posterior calculation. The prior is the superdistribution about the two long-run response frequencies.)

The long-run frequencies of the two stimuli – assumed to be 50%/50% throughout the present discussion – are also very important. Even if the conditional responses seem to yield low mutual info when stimuli occur equally often, they may yield a sufficient amount of mutual info when the stimuli don't occur equally often. In fact, this is partly a *decision* problem, which can't be judged only by checking frequencies or probabilities.

16 Using the sample frequencies for our guess

Let us use the notation

$$\boldsymbol{x}/\boldsymbol{y} \coloneqq (x_1/y_1, x_2/y_2, \dots)\,, \qquad \boldsymbol{x}\ln\boldsymbol{y} \coloneqq \textstyle\sum_i x_i \ln y_i\,, \tag{1}$$

and let

$$\mathrm{H}(\boldsymbol{x}) \coloneqq -\boldsymbol{x}\ln\boldsymbol{x}\,, \qquad \mathrm{D}(\boldsymbol{x},\boldsymbol{y}) \coloneqq \boldsymbol{x}\ln(\boldsymbol{x}/\boldsymbol{y}) \tag{2}$$

be the Shannon entropy and relative entropy for distributions $\boldsymbol{x}$, $\boldsymbol{y}$.

This is actually a vector-covector notation. A distribution is represented by a covector $\boldsymbol{x} \coloneqq (x_1\ x_2\ \dots)$. The four arithmetic operations act element-wise on vectors and covectors. The logarithm acts element-wise and maps covectors into vectors: $\ln(x_1\ x_2\ \dots) \coloneqq \left(\begin{smallmatrix}\ln x_1\\ \ln x_2\\ \dots\end{smallmatrix}\right)$. Multiplication of a covector and a vector yields a scalar.

Label the possible responses by r. Let $\boldsymbol{f}_{|\alpha} \coloneqq (f_{1|\alpha}, \dots, f_{r|\alpha}, \dots)$ and $\boldsymbol{f}_{|\beta}$ be the response frequencies observed in the sample, conditional on the two stimuli α and β. Let $\boldsymbol{F}_{|\alpha} \coloneqq (F_{r|\alpha})$ and $\boldsymbol{F}_{|\beta}$ be the long-run response frequencies, conditional on the two stimuli. If the two stimuli are assumed to occur equally often in the sample and in the long run, then the mutual information between the stimulus frequencies and response long-run frequencies is given by the specific formula

$$\begin{aligned} I(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta}) &\equiv \frac{1}{2}\boldsymbol{F}_{|\alpha}\ln\frac{2\boldsymbol{F}_{|\alpha}}{\boldsymbol{F}_{|\alpha}+\boldsymbol{F}_{|\beta}} + \frac{1}{2}\boldsymbol{F}_{|\beta}\ln\frac{2\boldsymbol{F}_{|\beta}}{\boldsymbol{F}_{|\alpha}+\boldsymbol{F}_{|\beta}} \\ &\equiv \mathrm{H}\big[\tfrac{1}{2}(\boldsymbol{F}_{|\alpha}+\boldsymbol{F}_{|\beta})\big] - \tfrac{1}{2}\mathrm{H}(\boldsymbol{F}_{|\alpha}) - \tfrac{1}{2}\mathrm{H}(\boldsymbol{F}_{|\beta})\,. \end{aligned} \tag{3}$$

The probability density that the long-run mutual information has value $I = \hat{I}$ within $\mathrm{d}\hat{I}$, given the observed conditional frequencies $\boldsymbol{f}_{|\alpha}$, $\boldsymbol{f}_{|\beta}$ in a sample of n each, and given the superdistribution

$\mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid A)\, \mathrm{d}\boldsymbol{F}_{|\alpha}\, \mathrm{d}\boldsymbol{F}_{|\beta}$ for the long-run conditional frequencies (A denotes the specific assumptions behind the superdistribution), has the following expression:

$$\mathrm{p}(\hat{I} \mid \boldsymbol{f}_{|\alpha}, \boldsymbol{f}_{|\beta}, A) \propto \iint \delta\bigl[\hat{I} - I(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta})\bigr] \times \exp\bigl[-n\, \mathrm{D}(\boldsymbol{f}_{|\alpha}, \boldsymbol{F}_{|\alpha}) - n\, \mathrm{D}(\boldsymbol{f}_{|\beta}, \boldsymbol{F}_{|\beta})\bigr] \times \mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid A)\, \mathrm{d}\boldsymbol{F}_{|\alpha}\, \mathrm{d}\boldsymbol{F}_{|\beta}\ . \quad (4)$$

Proof: By marginalization

$$\begin{aligned} &\mathrm{p}(\hat{I} \mid \boldsymbol{f}_{|\alpha}, \boldsymbol{f}_{|\beta}, A) \\ &\quad = \iint \mathrm{p}(\hat{I} \mid \boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta}, \boldsymbol{f}_{|\alpha}, \boldsymbol{f}_{|\beta}, A) \times \mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid \boldsymbol{f}_{|\alpha}, \boldsymbol{f}_{|\beta}, A)\, \mathrm{d}\boldsymbol{F}_{|\alpha}\, \mathrm{d}\boldsymbol{F}_{|\beta} \\ &\quad = \iint \delta\bigl[\hat{I} - I(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta})\bigr] \times \mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid \boldsymbol{f}_{|\alpha}, \boldsymbol{f}_{|\beta}, A)\, \mathrm{d}\boldsymbol{F}_{|\alpha}\, \mathrm{d}\boldsymbol{F}_{|\beta} \end{aligned} \quad (5)$$

where the second equality comes from the fact that the mutual information is completely determined by the long-run conditional frequencies.

The probability of observing a sequence of n responses for each stimulus, with the responses appearing with frequencies $\boldsymbol{f}_{|\alpha}$ and $\boldsymbol{f}_{|\beta}$, given the long-run frequencies $\boldsymbol{F}_{|\alpha}$, $\boldsymbol{F}_{|\beta}$, can be calculated by simple combinatorics. It's essentially the formula for drawing with replacement from two "urns" with known contents[5]. We can rewrite it with exponentials:

$$\begin{aligned} \mathrm{p}(\mathrm{seq}_\alpha, \mathrm{seq}_\beta \mid \boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta}, A) &= \left[\prod_r (F_{r|\alpha})^{n\, f_{r|\alpha}}\right] \times \left[\prod_r (F_{r|\beta})^{n\, f_{r|\beta}}\right] \\ &\equiv \exp\bigl(n\, \boldsymbol{f}_{|\alpha} \ln \boldsymbol{F}_{|\alpha} + n\, \boldsymbol{f}_{|\beta} \ln \boldsymbol{F}_{|\beta}\bigr)\ . \end{aligned} \quad (6)$$

[5] De Finetti 1938; Bernardo & Smith 2000 § 4.6.

Using the rule of conditional probability we obtain the density in $\mathrm{d}\boldsymbol{F}_{|\alpha}\,\mathrm{d}\boldsymbol{F}_{|\beta}$:

$$\begin{aligned}
&\mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid \mathrm{seq}_\alpha, \mathrm{seq}_\beta, A) \\
&\quad = \frac{\exp\bigl(n\, \boldsymbol{f}_{|\alpha} \ln \boldsymbol{F}_{|\alpha} + n\, \boldsymbol{f}_{|\beta} \ln \boldsymbol{F}_{|\beta}\bigr) \times \mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid A)}{\iint \exp\bigl(n\, \boldsymbol{f}_{|\alpha} \ln \boldsymbol{F}_{|\alpha} + n\, \boldsymbol{f}_{|\beta} \ln \boldsymbol{F}_{|\beta}\bigr) \times \mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid A)\, \mathrm{d}\boldsymbol{F}_{|\alpha}\, \mathrm{d}\boldsymbol{F}_{|\beta}} \\
&\quad = \frac{\exp\bigl[-n\, \mathrm{D}(\boldsymbol{f}_{|\alpha}, \boldsymbol{F}_{|\alpha}) - n\, \mathrm{D}(\boldsymbol{f}_{|\beta}, \boldsymbol{F}_{|\beta})\bigr] \times \mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid A)}{\iint \exp\bigl[-n\, \mathrm{D}(\boldsymbol{f}_{|\alpha}, \boldsymbol{F}_{|\alpha}) - n\, \mathrm{D}(\boldsymbol{f}_{|\beta}, \boldsymbol{F}_{|\beta})\bigr] \times \mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid A)\, \mathrm{d}\boldsymbol{F}_{|\alpha}\, \mathrm{d}\boldsymbol{F}_{|\beta}}
\end{aligned} \tag{7}$$

where the second equality is obtained multiplying numerator and denominator by $\exp[n\mathrm{H}(\boldsymbol{f}_{|\alpha}) + n\mathrm{H}(\boldsymbol{f}_{|\beta})]$.

The formula above also shows that the knowledge of the sample frequencies is equivalent to the knowledge of the specific sequences, for the purpose of guessing the long-run frequencies. That is, the sample frequencies are sufficient statistics:

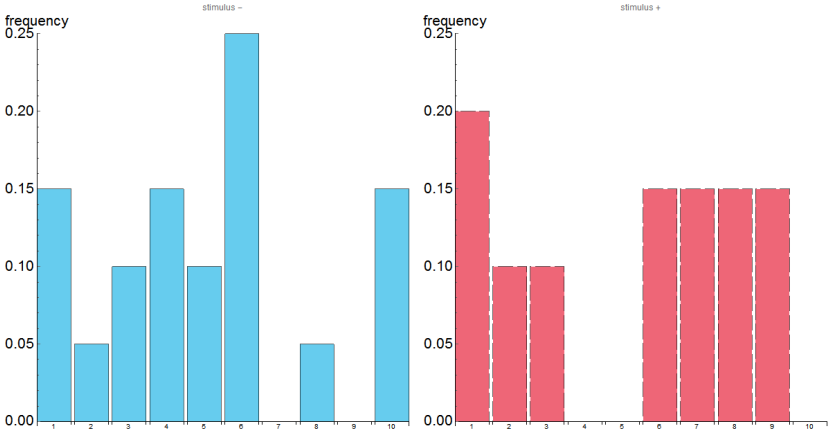
$$\mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid \mathrm{seq}_\alpha, \mathrm{seq}_\beta, A) = \mathrm{p}(\boldsymbol{F}_{|\alpha}, \boldsymbol{F}_{|\beta} \mid \boldsymbol{f}_{|\alpha}, \boldsymbol{f}_{|\beta}, A)\ . \tag{8}$$

Finally substituting formula (7) into (5) we obtain (4).

If we only knew the sample mutual-information, but not the sample frequencies, it would be necessary to marginalize the probability (4) over all possible values of the sample frequencies that have the observed mutual information (this is what figs 5–6 do). This would lead, for convexity reasons, to a broadening of the probability distribution for $\hat{I}$, indicating that we are indeed basing our inference on a reduced amount of data. In fact the sample mutual information is not a sufficient statistic.

Examination of formula (4) shows that it behaves in an intuitive way for large samples: for large n the exponential term becomes a delta centred on the observed frequencies. The sample mutual information is therefore approximately equal to the long-run mutual information. One can also do expansions in n to obtain asymptotic corrections. The exponential term also shows where maximum-entropy methods come from.

Examples of results Suppose we have observed responses with the following frequencies in a sample of 20 responses per stimulus:



This sample has a mutual information of 0.387 bit.

We want to know what mutual information we would find if we could accumulate an infinite number of samples (long-run mutual information).

As we saw in § 14, it is necessary to make an initial guess (before the sample is observed) about what the long-run frequencies could be; that is, we need to choose a superdistribution. Such choice depends on the context: the specific experiment, brain region, cell type, and all other biological details. We imagine three example cases, each of which could be appropriate in a different context.

First. We consider all long-run response-frequency distributions, for each stimulus, to be equally probable in frequency space: peaked or flat, broad or narrow, and so on (this belief is represented by a "flat" Dirichlet superdistribution with unit parameters). Our resulting guess about the long-run mutual information is given by the probability distribution at the top of fig. 7. There is a 95% probability that the long-run mutual info is between 0.093 and 0.371 bit, and a 50-50% probability that it is lower or higher than 0.214 bit.

Second. We consider long-run response-frequency distributions with one or more very sharp peaks to be more probable than broad distributions (this belief is represented here by a Dirichlet superdistribution with parameters less than unity, very close to a Jeffreys prior). Our resulting

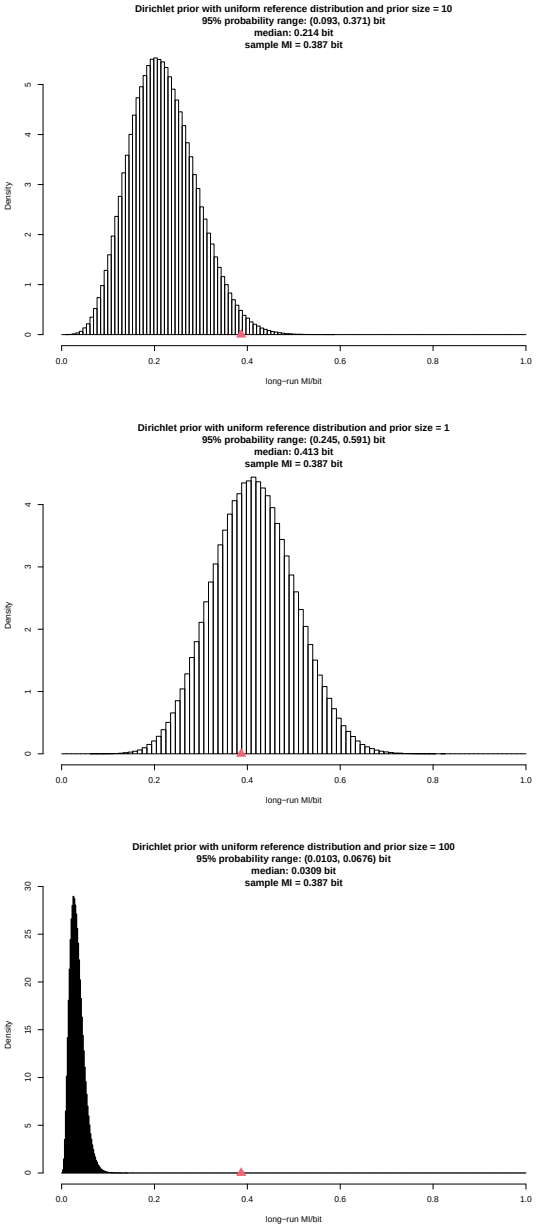


Figure 7 Top: uniform superdistribution; middle: peaked frequency distributions are more probable; bottom: uniform frequency distributions are more probable (equally probable sequences)

guess about the long-run mutual information is given by the probability distribution in the middle of fig. 7. There is a 95% probability that the long-run mutual info is between 0.245 and 0.591 bit, and a 50-50% probability that it is lower or higher than 0.413 bit.

Third. We consider long-run response-frequency distributions of almost flat shape to be more probable than peaked distributions; this is very similar to giving equal probability to all possible long-run *sequences* of responses, rather than to their frequencies (this belief is represented here by a Dirichlet superdistribution with parameters larger than unity). Our resulting guess about the long-run mutual information is given by the probability distribution at the bottom of fig. 7. There is a 95% probability that the long-run mutual info is between 0.0103 and 0.0676 bit, and a 50-50% probability that it is lower or higher than 0.0309 bit.

Note how these three different prior assumptions lead to three very different guesses about the long-run mutual information, based on the same observed sample. This strong dependence on the prior assumption is to be expected because the sample is small.

Also, there is no way to say which assumption is "true". But if we have samples from previous experiments, for which we believe that our initial guesses should be the same as in the present case, then our initial superdistribution can – and ought to – be updated with the results of those other experiments, using a hierarchic model.

Bibliography

("de X" is listed under D, "van X" under V, and so on, regardless of national conventions.)

Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, repr. (Wiley, New York). First publ. 1994.

de Finetti, B. (1938): *Sur la condition d'équivalence partielle*. In: *Colloque consacré à la théorie des probabilités. VI: Conceptions diverses*, ed. by B. de Finetti, V. Glivenko, G. Neymann (Hermann, Paris): 5–18. Transl. by P. Benacerraf and R. Jeffrey in Jeffrey (1980), pp. 193–205.

Jeffrey, R. C., ed. (1980): *Studies in inductive logic and probability. Vol. II*. (University of California Press, Berkeley).

Keynes, J. M. (2013): *A Tract on Monetary Reform*, repr. of 2nd ed. (Cambridge University Press, Cambridge). First publ. 1923.

Nemenman, I., Bialek, W., de Ruyter van Steveninck, R. (2004): *Entropy and information in neural spike trains: progress on the sampling problem*. Phys. Rev. E **69**[5], 056111.

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**[3], 1064–1072.