Open Science Framework DOI:10.31219/osf.io/***

Guessing long-run mutual information [draft]

P.G.L. Porta Mana
<pgl@portamana.org>

C. Battistin
<>

S. Gonzalo Cogno
<@>

***; updated 12 July 2020

***

1 Long-run mutual info and sample mutual info

We have two kinds of stimuli and, say, ten kinds of neural responses to these stimuli, under specific environmental and behavioural conditions, and for a specific kind 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.

Imagine that a researcher told us the results of 10^{60} stimulus-response measurements, all performed for the specified conditions and subjects, and in which the two kinds of stimulus occurred equally often. From the long-run[1] joint frequencies of 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 kind of stimulus, from such 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)

Case A. Suppose the long-run frequencies of responses conditional on each kind of stimulus are as given in the top panel of fig. 1[2]. The light-grey histogram is for the responses to the first kind of stimulus; dark-grey for the second kind. 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 same long-run distribution) as equally plausible, we plot some possible results for the sample mutual information in the bottom panel, with a histogram. 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; it is even outside the inner 95% interval range.

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

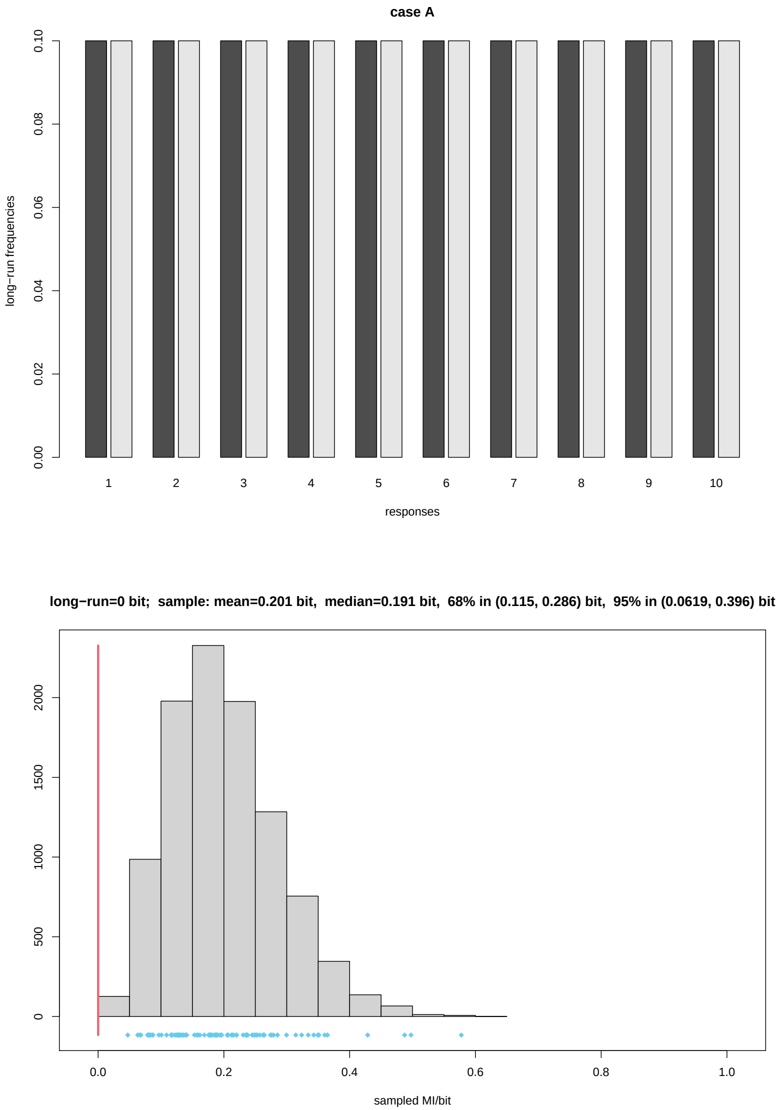
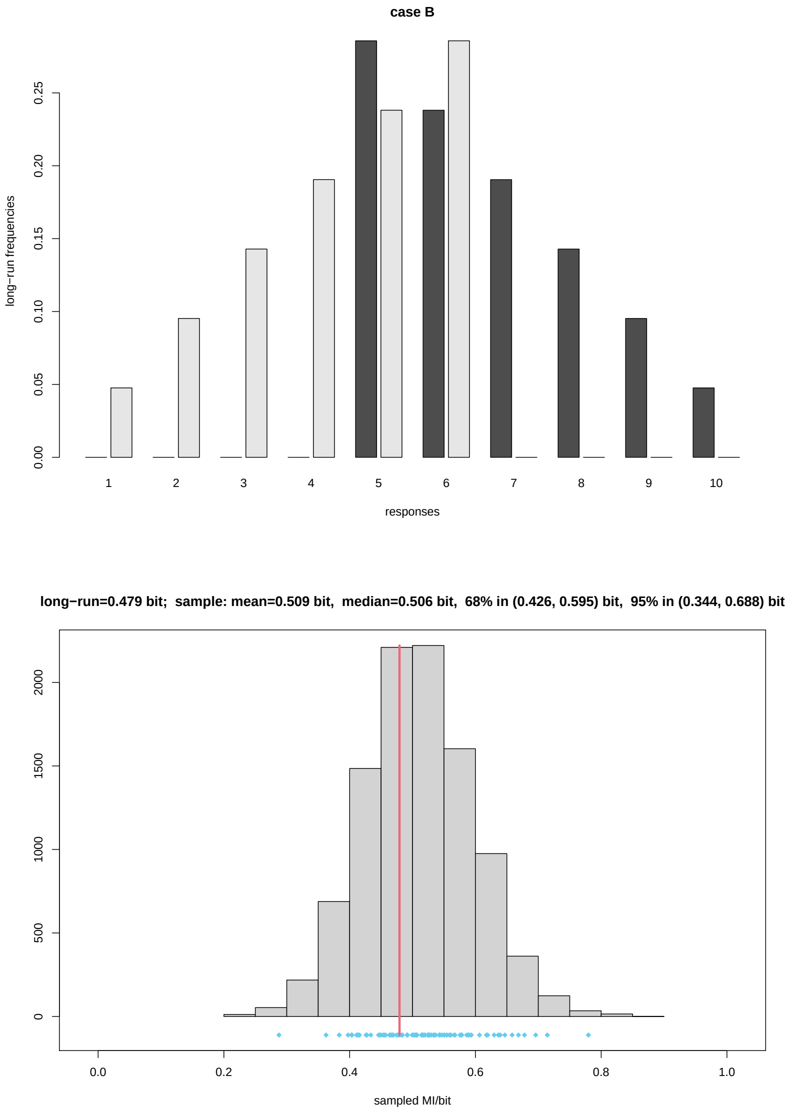


Figure 1 Case A

Case 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 kinds of stimulus. 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 sampled 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, 0.50 bit, and the mean, 0.51 bit.


case B

long−run=0.479 bit; sample: mean=0.509 bit, median=0.506 bit, 68% in (0.426, 0.595) bit, 95% in (0.344, 0.688) bit

Figure 2 Case B

Case 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).

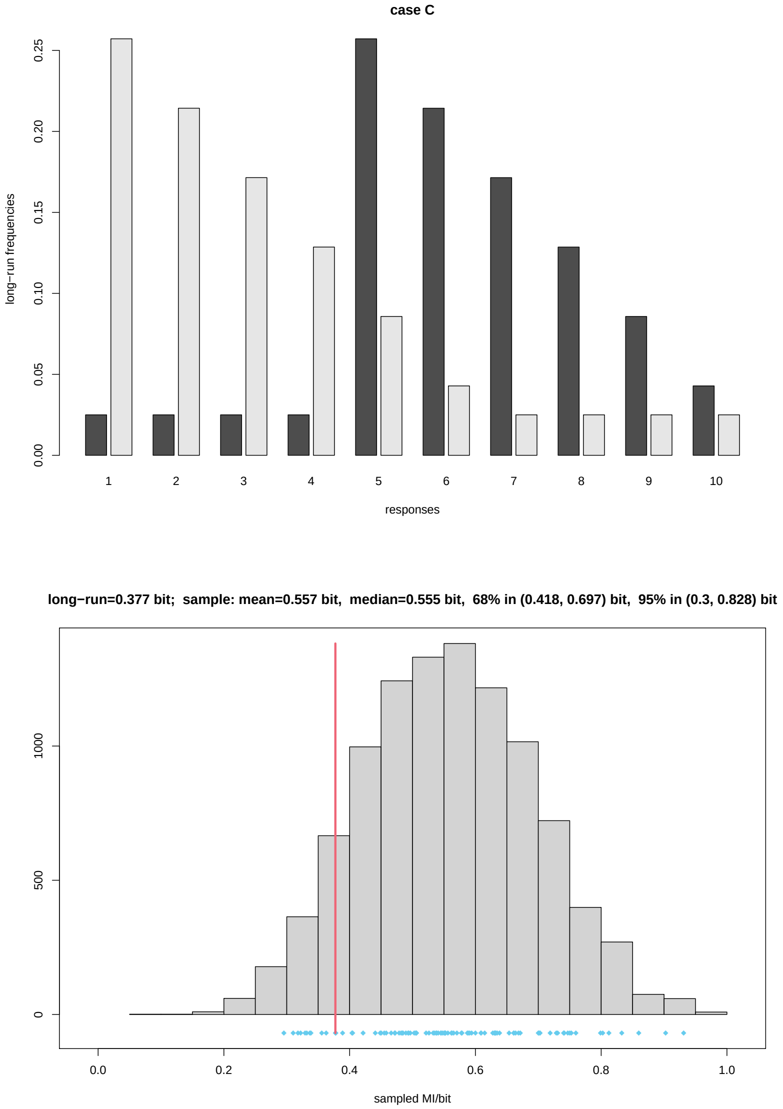


Figure 3 Case C

Case 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: they must be multiples of 1/20). The long-run mutual information is within the inner 68% quantile.

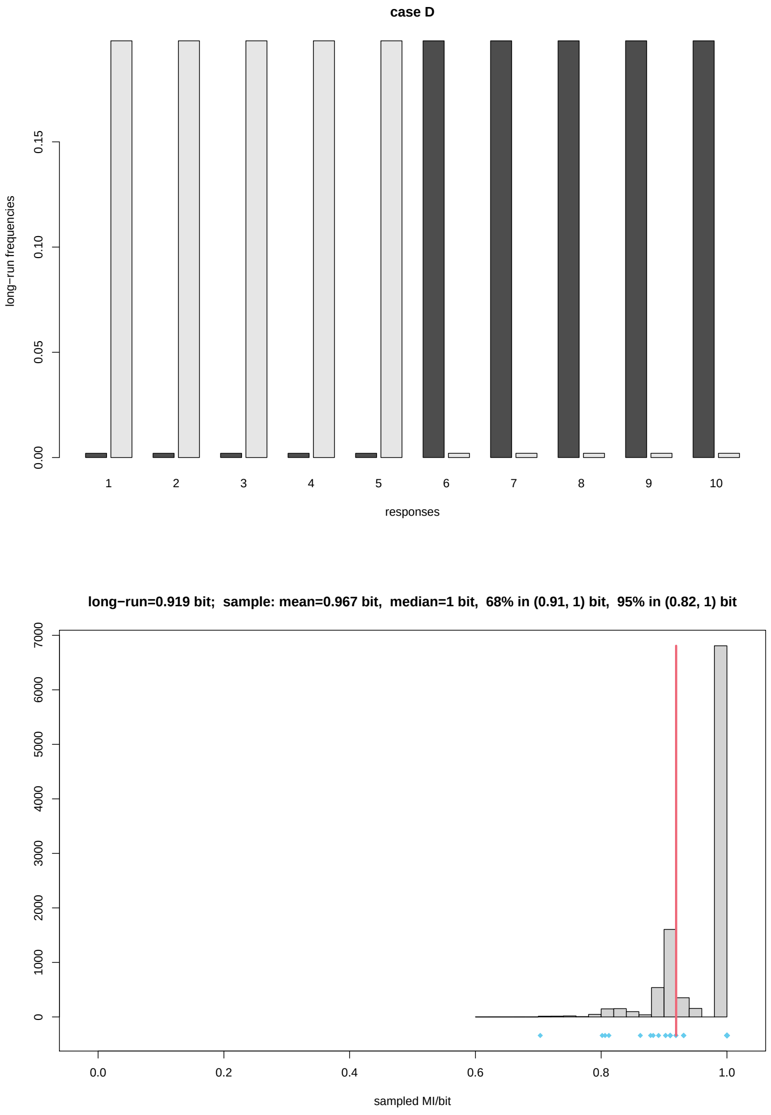


Figure 4 Case D

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 conditional response distributions.

(iii) In some cases the spread (say, 95%) is over a short range, compared with the possible range of the long-run mutual information, $[0, 1]$ bit; for example 18% for Case D. In other cases it is over a very broad range, for example 53% for Case C.

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

(v) 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 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. So, 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)?

This last remark is also important for "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 really are representative, we therefore 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.

2 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 the long-run and 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 proportional to a multinomial coefficient.

But also this last choice could be debatable. If the responses represent population activities or rates, for example, we know that low values should be expected more often that high values, 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.

Finally, should such a superdistribution be factorizable over the frequency distributions for the different stimuli (as 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. We proceed as follows: given a superdistribution, we choose from it two long-run response-frequency distributions; from these we calculate the long-run mutual information; then we sample 20 responses from each, and calculate the sample mutual information obtained from such sample of responses. This way we obtain a density or scatter plot that tells us how often we should observe every given pair of long-run & sample mutual informations – under the assumption of that specific superdistribution.

The results are shown in fig. 5 for the four kinds of superdistribution entertained in the discussion above.

Note that this kind of plot can be used to guess the long-run mutual information given the one observed in a sample, under the assumption of the specific superdistribution related to 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,

The 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 the former from the latter 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.

Choosing a "standard" superdistribution, to be universally used[3], is not a sensible option either.

3 Caveats

The long-run frequencies of the two kinds of stimulus 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

[3] cf. Nemenman et al. 2004.

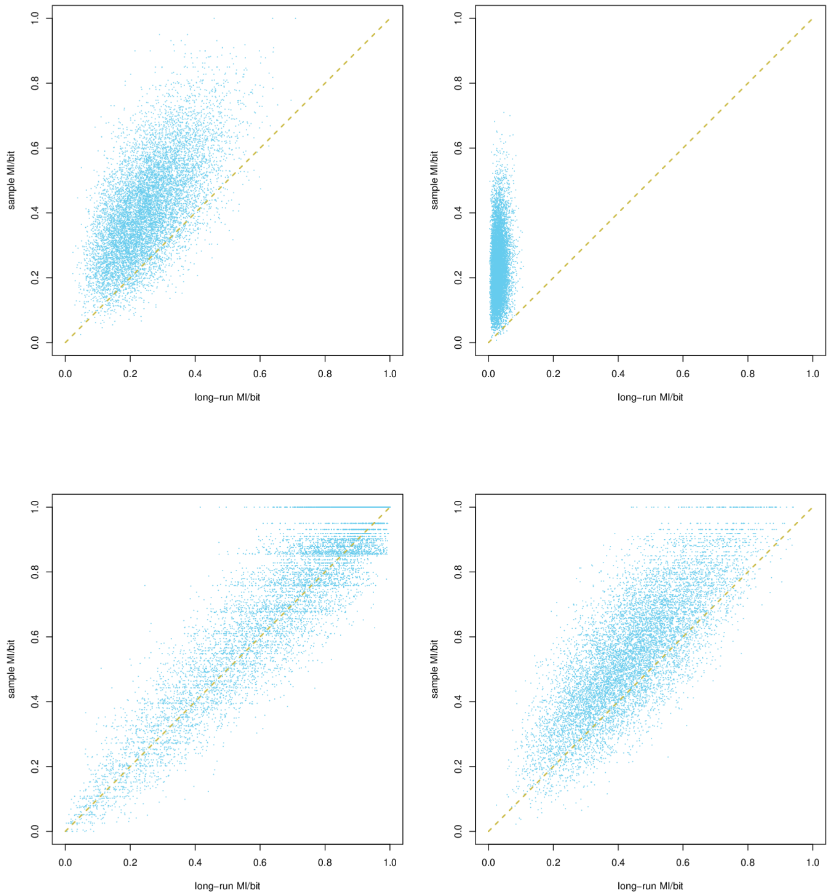


Figure 5 Top left: uniform over frequencies. Top right: more uniform over sequences. Bottom left: low responses preferred. Bottom right: not factorizable over stimuli (positive correlation; hierarchic)

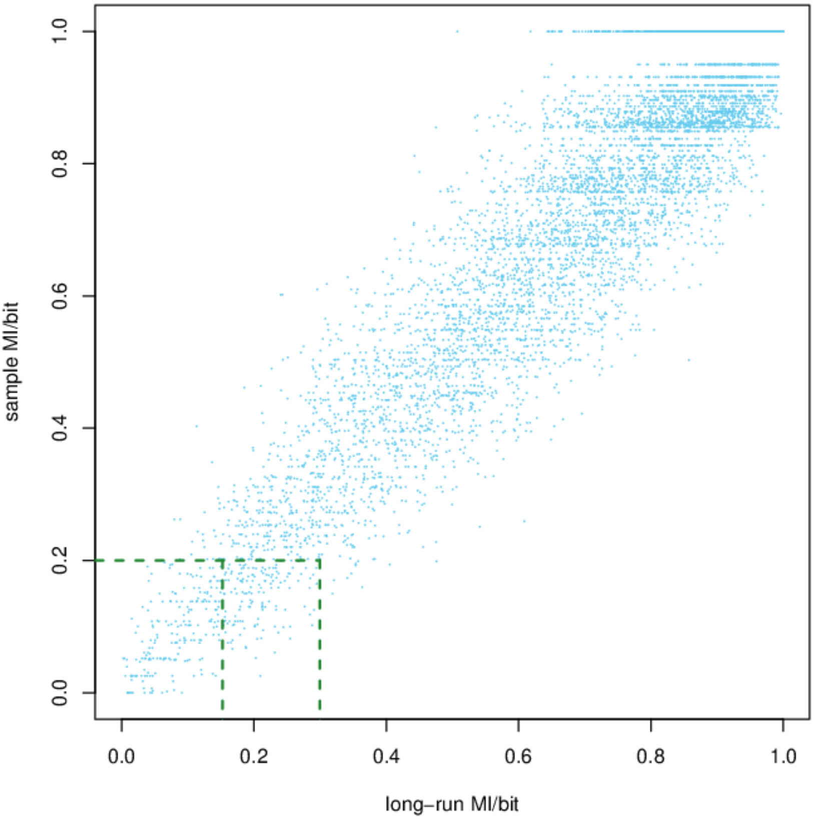


Figure 6 Example of guess for long-run mutual information from the mutual information observed in 20 samples per stimulus, under a specific superdistribution. If the sample mutual info is 0.2 bit (horizontal, yellow dashed line), then the long-run mutual info can in this case inferred to be at 95% between (0.097, 0.39) bit and at 68% between (0.15, 0.30) bit (vertical, yellow dashed lines), with a median of 0.22 bit

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.

Bibliography

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

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 $\mathbf{69}^{5}$, 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. $\mathbf{98}^{3}$, 1064–1072.