Mutual information between successive reorientations

Subhaneil Lahiri
Harvard University,

May 13, 2011

Abstract

We show how mutual information can be used to describe the independence of successive reorientations

1 Reorientation sequences

sec:reoseq

As a worm navigates, it performs a sequence of turns. When turns occur sufficiently close to each other, they are grouped into a reorientation event. These reorientations have several characteristics, e.g. the types of turn of which it is composed, the difference in heading direction before and after, the duration of the run leading into it. We wish to know if the characteristics of one reorientation are independent of the characteristics of previous reorientations.

Consider a sequence of r successive reorientations. The values of a particular characteristic of these reorientations is an r-tuple of random variables: $(X_1, \ldots, X_r)$. We are asking whether or not $P(X_1, \ldots, X_r) = P(X_1) \ldots P(X_r)$. We will discuss some measures of independence in §2.

These probability distributions can be estimated from the frequencies in a sample sequence (described schematically in fig.1). However, as a consequence of finite sample size, this process has both systematic and random errors. We will look at several methods for removing systematic errors in §3 and one method for estimating random errors in appendix A.

In principle we should look at all the characteristics together. However, increasing the dimensionality of the data at each reorientation increases the number of bins used, which in turn increases the number of samples needed. Therefore, we will have to be satisfied with looking at each characteristic separately.

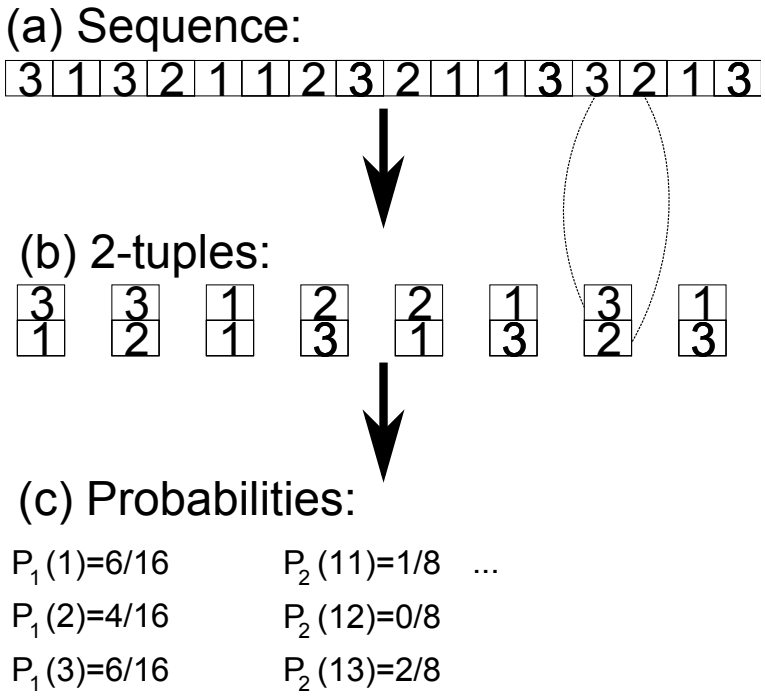

Figure 1: Schematic calculation of probabilities. (a) Sample sequence of reorientations. (b) Grouping sequence into r-tuples ($r = 2$). (c) Estimating probabilities for individual members and for r-tuples from relative frequencies in sample.

fig:schemati

2 Entropy and mutual information

sec:entropy

The **entropy** of a probability distribution is a measure of the lack of information we have about a random variable:

$$H(X) = \langle -\log P(X) \rangle \,. \tag{1}$$ eq:ent

It takes its minimum value of 0 when X can only take one value. It takes its maximum value of $\log n$ when X has a uniform distribution over n possibilities.

With several random variables, we can define a joint entropy from their joint probability distribution:

$$H(X_1, \ldots, X_r) = \langle -\log P(X_1, \ldots, X_r) \rangle \,. \tag{2}$$ eq:jointent

It satisfies the bounds

$$\max_i H(X_i) \leq H(X_1, \ldots, X_r) \leq \sum_{i=1}^{r} H(X_i). \tag{3}$$ eq:entbounds

The lower bound is saturated when one of the variables is enough to determine the others. The upper bound is saturated when the X_i are independent:

$$P(X1, \ldots, X_r) = \prod_{i=1}^{r} P(X_i) \quad \Longrightarrow \quad H(X_1, \ldots, X_r) = \sum_{i=1}^{r} H(X_i). \tag{4}$$ eq:indent

We can define the following measure of (lack of) independence:

$$I(X_1, \ldots, X_r) = \sum_{i=1}^{r} H(X_i) - H(X_1, \ldots, X_r) = \left\langle \log \frac{P(X_1, \ldots, X_r)}{P(X_1) \ldots P(X_r)} \right\rangle . \tag{5}$$ eq:mutinf

In the case $r = 2$, this is the **mutual information** between X_1 and X_2. For $r > 2$, there are many different generalisations of mutual information. This one is called the **total correlation** [1], or multiinformation. It has the properties:

- it vanishes if and only if the random variables are independent
- otherwise, it is positive.
- it is bounded from above by $\sum_{i=1}^{r} H(X_i) - \max_i H(X_i)$.

In our cases, the random variables, X_i, all have the same distribution, so the total correlation satisfies the bounds

$$0 \leq I_r(X_1, \ldots, X_r) \leq (r-1)H(X). \tag{6}$$ eq:mutinfbou

We can define a normalised total correlation:

$$C_r = \frac{I_r}{(r-1)H_1}, \qquad 0 \leq C_r \leq 1. \tag{7}$$ eq:normmutin

The lower bound corresponds to complete independence. The upper bound corresponds to complete redundancy.

3 Correcting systematic errors

sec:syscorr

We will compute the quantities defined in the previous section by estimating the probability distributions, $P_1(X_i)$ and $P_r(X_1, \ldots, X_r)$ from the relative frequencies in sample sequences of reorientations. As all the X_i have the same distribution, we will estimate $P_1(X)$ from the pooled data, rather that estimating the $P_1(X_i)$ separately. When looking at reorientation types, we will restrict attention to the 5 most common types. When looking at angles and run durations, the data has to be binned. We will follow the approach of [2] and place the bins on quantiles of the data, preserving the coordinate invariance of the mutual information. In both cases, we will use 5 bins.

The lower bounds on mutual information, (6) and (7), lead to systematic errors which tend to bias these estimates upwards. There are several methods for estimating this bias.

One approach involves expending the errors in one over the sample size, see [3] and estimating the leading order correction. One can also estimate the random errors using this approach. Our estimators are slightly different to those used there, the appropriate versions of the estimates are calculated in Appendix A.

As the bias estimates are independent of the actual probability distribution, depending only on sample size and number of bins, we can estimate the bias by computing the mutual information for a completely random sequence in the same way. As the true value is zero, the result of this computation is an estimate of the bias. Furthermore, if the probabilities of the individual elements of the sequence are the same as in the data, this can be regarded as a Monte-Carlo simulation of the null hypothesis – that the individual elements of the sequence are independent of each other. Thus, we can compute a p-value by seeing where the original result ranks amongst the results of the simulation.

We will do this using the non-parametric bootstrap. This is conceptually similar to the common alternative procedure of shuffling the sequence. Shuffling can be thought of as resampling without replacement, whereas the non-parametric bootstrap is resampling with replacement. They both involve removing any information in the sequence without changing the probabilities of the individual elements.

The direct method of [4] consists of varying the sample size and extrapolating to infinity. This can also be used to check that the number of samples is large enough compared to the number of bins. With this method, it is difficult to compute error bars and p-values, as the process is slow.

We show three examples of these methods in fig.2, one where they agreed really well, one where the agreement was not too bad and one where it was terrible. In the last case, it even showed the wrong trend: the bias is supposed to be positive and decreasing with increasing sample size. This probably indicates that the sample size is too small and we were not in the asymptotic regime where the results of appendix A can be trusted.

We can see how much this discrepancy varied with sample size in fig.3. This can be used as a guide for when N is large enough to trust the results.

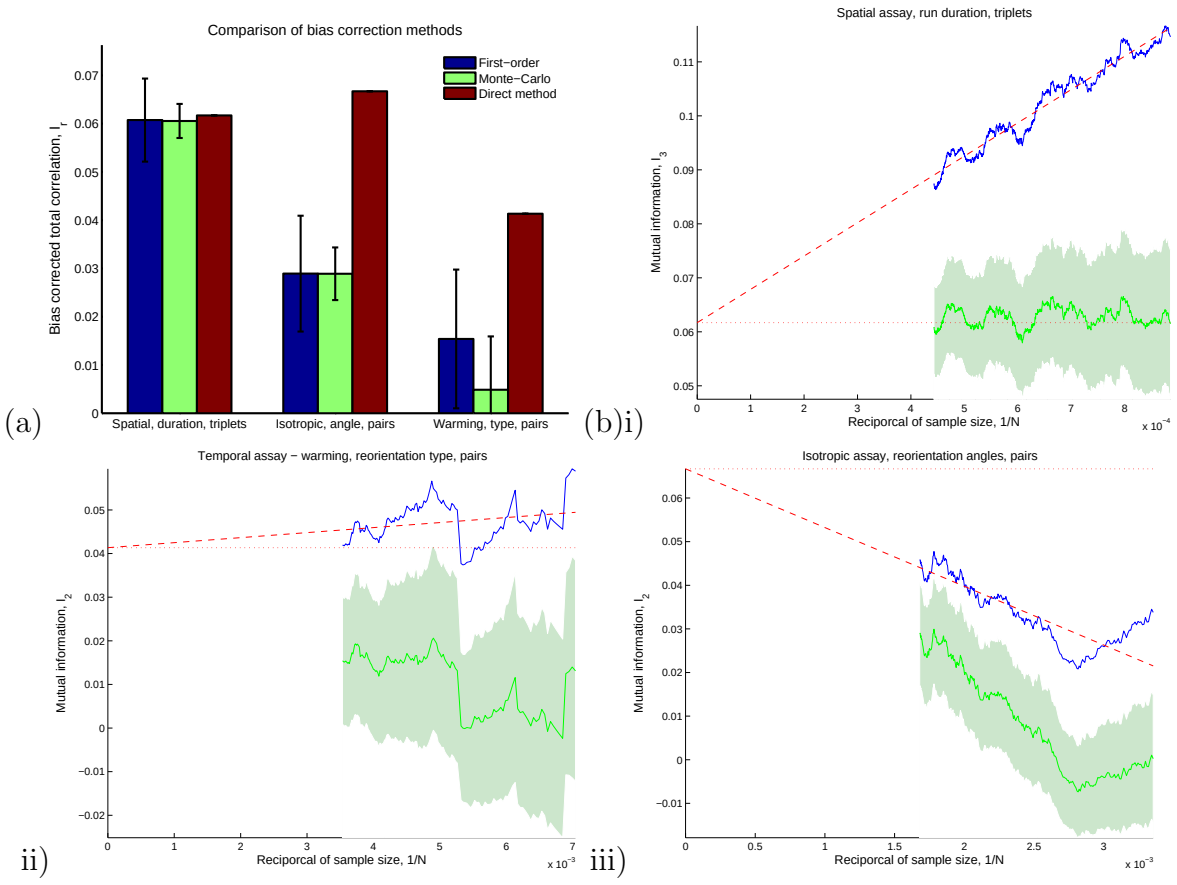

Figure 2: Comparison of different methods for removing bias. (a) Three examples of the three unbiased estimates of mutual information/total correlation. First-order refers to the methods of appendix A. Monte-Carlo refers to subtracting the mean of 1000 nonparametric bootstrap simulations, the error-bars are the standard deviation of the bootstrap simulations. (b) Illustration of the direct method, Blue line is the uncorrected estimates with different sample sizes, red dashed line shows the extrapolation to infinite sample size and red dotted line indicates the result of this extrapolation. Green line shows the unbiased estimator using the first-order method of appendix A for comparison, darker green shading is ± one standard error. i)-iii) the three cases shown in (a).

fig:biascomp

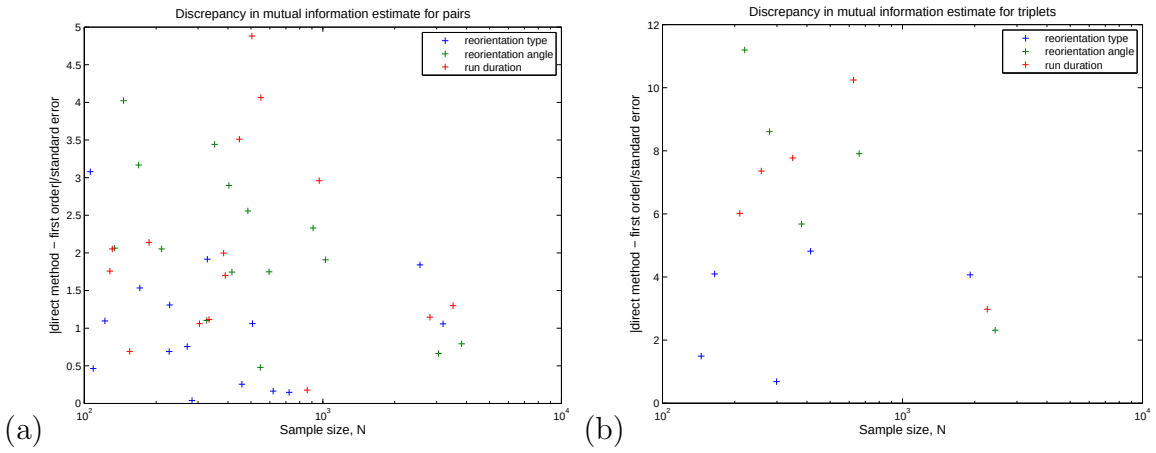

Figure 3: Effect of sample size in discrepancy between bias removal methods; size of discrepancy between unbiased estimates from direct method and first order method divided by first order standard error for (a) pairs (maximum of 25 bins) (b) triplets (maximum of 125 bins).

fig:discrep

4 Results

sec:results

The results of this analysis are presented in fig.4. The analysis was performed for an isotropic assay (where the temperature was uniform and constant), a spatial assay (where the temperature was constant in time but varied linearly in space from $18 - 23°\mathrm{C}$ over 9 cm) and a temporal assay (where the temperature was constant in space but varied sinusoidally in time from $?-?°\mathrm{C}$ with a period of 10 min). For the temporal assay, the analysis was also performed restricted to periods of warming ($\frac{\mathrm{d}T}{\mathrm{d}t} > 10^{-4}°\mathrm{C/s}$) and cooling ($\frac{\mathrm{d}T}{\mathrm{d}t} < -10^{-4}°\mathrm{C/s}$).

We looked at the cases $r = 2$ (pairs of consecutive reorientations) and $r = 3$ (triplets of consecutive reorientations). We also restricted attention to pairs of reorientations whose starts were separated by less than 20 s as well as pairs separated by more than 20 s.

Note that when we impose a restriction on the separation of reorientations, it is no longer true that the distributions of the first and second run durations in a pair are identical. This was not dealt with properly, so the corresponding bars should be ignored in fig.4.

Appendices

A Bias and standard error

sec:stderr

We will follow the approach of [3]. Our situation is slightly different from that one. As all the X_i have the same distribution, we will estimate $P(X)$ from the pooled data, rather that estimating the $P(X_i)$ separately. This means that our estimates may not satisfy the

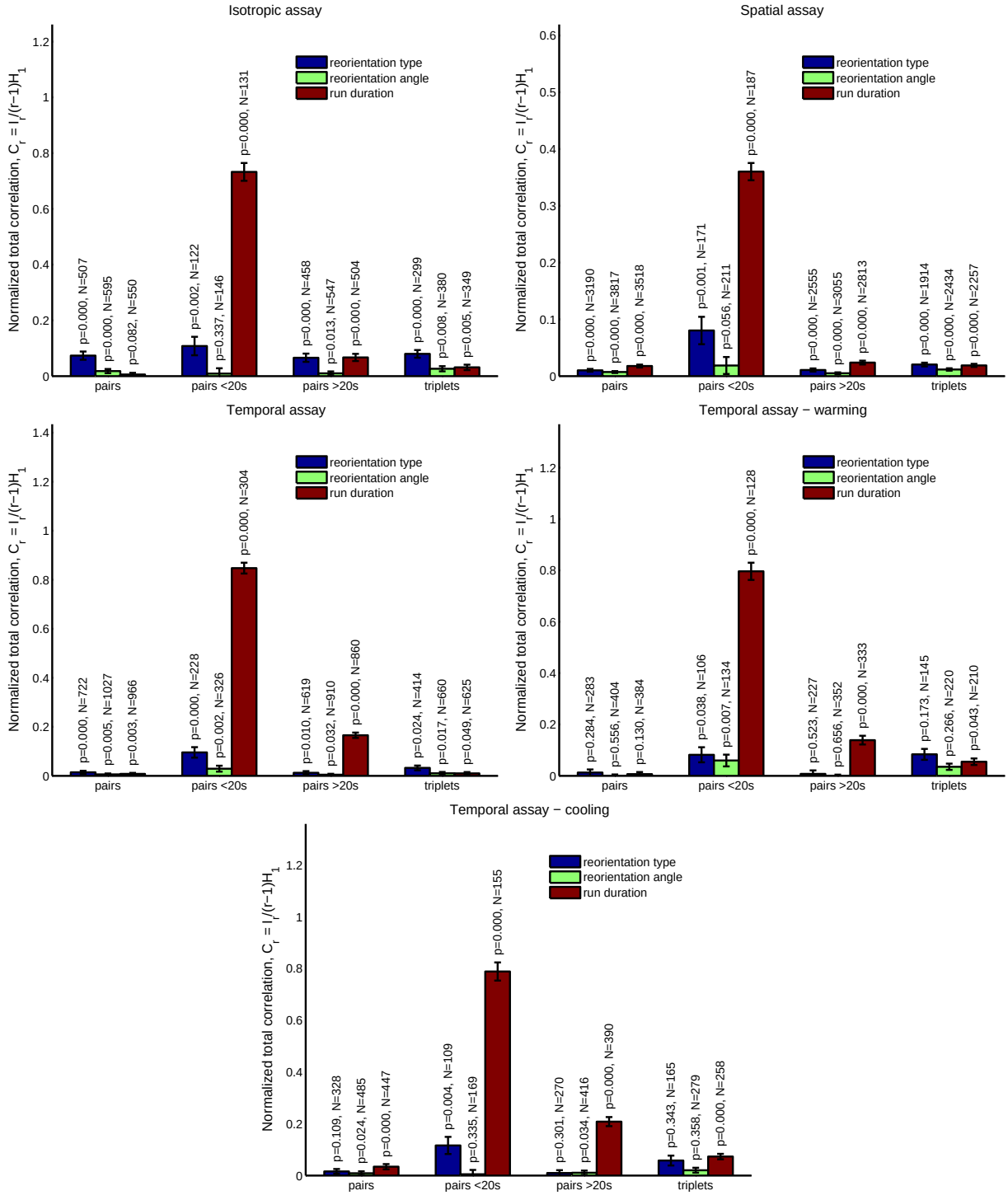

Figure 4: Normalised total correlation/mutual information for (a) isotropic assay, (b) spatial assay, (c) temporal assay, (d) temporal assay restricted to times when $\frac{\mathrm{d}T}{\mathrm{d}t} > 10^{-4}$°C/s, (e) temporal assay restricted to times when $\frac{\mathrm{d}T}{\mathrm{d}t} < -10^{-4}$°C/s. Values have bias subtracted using the methods of appendix A. Error bars are one standard error, computed using the methods of appendix A. P-values computed using 1000 nonparametric bootstrap resamples under the null hypothesis that successive reorientations are independent, therefore the third decimal place is untrustworthy. N is the number of r-tuples in the sample.

fig:results

bounds, such as (6).

Let $p_{i_1\ldots i_r}$ denote the probability $P_r(X_1 = x_{i_1}, \ldots, X_r = x_{i_r})$ and $n_{i_1\ldots i_r}$ denote the number of corresponding r-tuples in the sample. We can estimate $p_{i_1\ldots i_r}$ with

$$q_{i_1\ldots i_r} = \frac{n_{i_1\ldots i_r}}{N}, \qquad N = \sum_{j_1\ldots j i_r} n_{j_1\ldots j_r}. \tag{8}$$ eq:tupprob

We can then estimate $p_j = P_1(X = x_j)$ with

$$q_j = \sum_{i_1\ldots i_r} \left(\frac{q_{i_1\ldots i_r}}{r} \sum_{a=1}^{r} \delta_{j,i_a} \right). \tag{9}$$ eq:singprob

From now on, we will use A to denote the estimate of $A(p)$ with p replaced by q and $\widehat{A} = A - \mathrm{Bias}(A)$, where A is one of (H_1, H_r, I_r, C_r).

Our bias estimates are essentially the same as those of [3], except that the number of samples for H_1 is rN rather than N:

$$B_1 = \mathrm{Bias}(H_1) = -\frac{\#b_1}{2rN}, \qquad B_r = \mathrm{Bias}(H_r) = -\frac{\#b_r}{2N}, \tag{10}$$ eq:biasH

where $\#b_r$ is the number of non-empty bins (e.g. in fig.1, $\#b_1 = 3$, $\#b_2 = 6$). The bias estimate for I and C follow in the usual way.

We can estimate the standard errors with

$$\mathrm{Var}(A) \approx \sum_{i_1\ldots i_r} \left(\frac{\partial A}{\partial n_{i_1\ldots i_r}} \right)^2 \mathrm{Var}(n_{i_1\ldots i_r}), \qquad \mathrm{Var}(n_{i_1\ldots i_r}) \approx N q_{i_1\ldots i_r}(1 - q_{i_1\ldots i_r}). \tag{11}$$ eq:stderr

where the first formula is valid provided each term is small(we'll see later that they are proportional to $1/N$) and the corrections to the last formula are lower order in N.

We find that

$$\begin{aligned} \frac{\partial q_{i_1\ldots i_r}}{\partial n_{j_1\ldots j_r}} &= \frac{(\prod_a \delta_{i_a,j_a}) - q_{i_1\ldots i_r}}{N}, & \frac{\partial B_r}{\partial n_{j_1\ldots j_r}} &= -\frac{B_r}{N}, \\ \frac{\partial q_i}{\partial n_{j_1\ldots j_r}} &= \frac{\frac{1}{r}\left(\sum_a \delta_{i,j_a}\right) - q_i}{N}, & \frac{\partial B_1}{\partial n_{j_1\ldots j_r}} &= -\frac{B_1}{N}, \end{aligned} \tag{12}$$ eq:dqbydn

which leads to

$$\begin{aligned} \frac{\partial H_r}{\partial n_{j_1\ldots j_r}} &= -\frac{\log q_{i_1\ldots i_r} + H_r}{N}, & \frac{\partial I_r}{\partial n_{j_1\ldots j_r}} &= -\frac{\left(\sum_a \log q_{j_a}\right) - \log q_{i_1\ldots i_r} + I_r}{N}, \\ \frac{\partial H_1}{\partial n_{j_1\ldots j_r}} &= -\frac{\frac{1}{r}\left(\sum_a \log q_{j_a}\right) + H_1}{N}, & \frac{\partial C_r}{\partial n_{j_1\ldots j_r}} &= \frac{\log q_{i_1\ldots i_r} H_1 - \frac{1}{r}\left(\sum_a \log q_{j_a}\right) H_r}{(r-1)N(H_1)^2}. \end{aligned} \tag{13}$$ eq:dhbydn

All of these formulae are equally true if you put hats on every capital letter (except N).

References

258.1661265 [1] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM J. Res. Dev.* **4** (January, 1960) 66–82. http://dx.doi.org/10.1147/rd.41.0066.

.....2017S [2] N. Slonim, G. S. Atwal, G. Tkacik, and W. Bialek, "Estimating mutual information and multi–information in large networks," arXiv:cs/0502017.

.125..285R [3] M. S. Roulston, "Estimating the errors on measured entropy and mutual information," *Physica D Nonlinear Phenomena* **125** (Jan., 1999) 285–294, SAO/NASA ADS:1999PhyD..125..285R.

..80..197S [4] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek, "Entropy and Information in Neural Spike Trains," *Physical Review Letters* **80** (Jan., 1998) 197–200.