Mutual information between successive reorientations

Subhaneil Lahiri

Harvard University

July 11, 2011

Abstract

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

Contents

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. Some example tracks, consisting of runs and reorientations, are shown in fig.1. 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[1]: $(X_1, \ldots, X_r)$. 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.2). 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 §3.1 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.

Note that this method can confuse animal-to-animal variability with causation. E.g. suppose that one of the worms makes larger turns than the others. In this situation, if the first reorientation in an r-tuple is large it would make it more likely that the worm in question is the large-turning one, making it more likely that the other reorientations in the r-tuple are large. Thus the probability distributions would not show independence even in the absence of actual causation. The same thing applies to lack of stationarity. If runs tend to get longer over time, the same issue will arise.

2 Entropy and mutual information

The **entropy** of a probability distribution is a measure of the lack of information we have about a random variable [1]:

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

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

---

[1] Actually, reorientation type is not a random variable, as it is not described by a number. However, the concept of a probability distribution still makes sense, and the probability of the outcome can be considered a random variable.
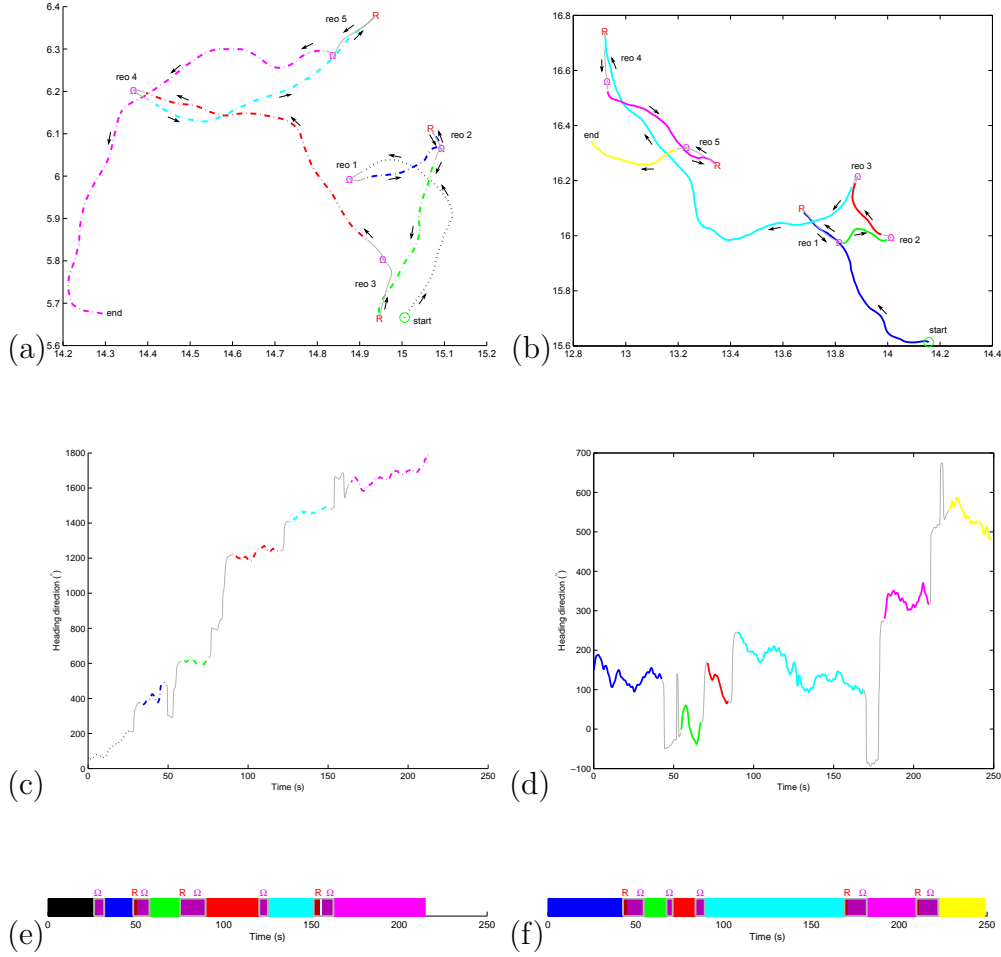
Figure 1: (a-b)Example tracks. Runs: Coloured, thin lines, reorientations: black, thin lines. Omega turns and reversals labelled as Ω and R. Grouping of turns into reorientations indicated by reo 1, reo 2, etc. (c-d) heading direction during tracks in (a-b). (e-f) Segmentation into runs and turns.
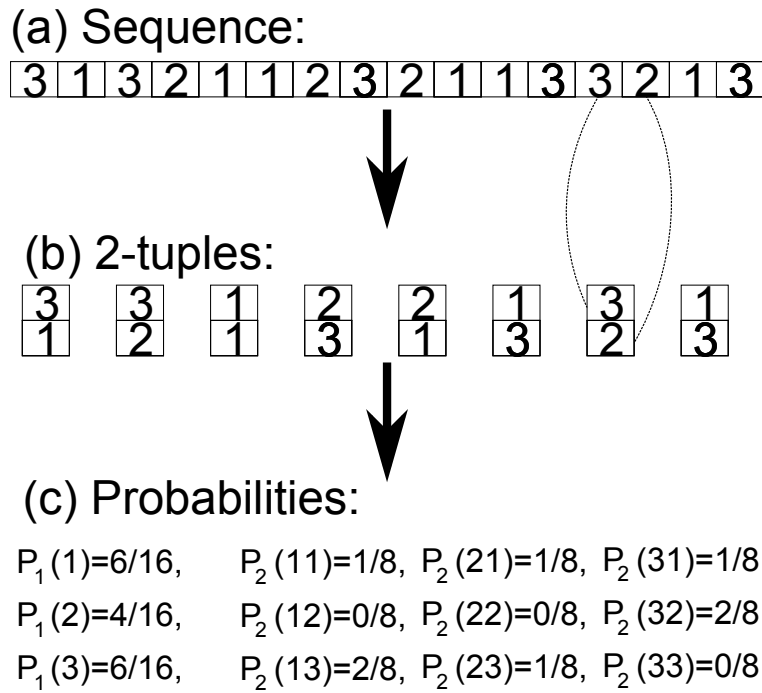
fig:tracks

3

(a) Sequence:

| 3 | 1 | 3 | 2 | 1 | 1 | 2 | 3 | 2 | 1 | 1 | 3 | 3 | 2 | 1 | 3 |

(b) 2-tuples:

$\begin{array}{c}3\\1\end{array}$ $\begin{array}{c}3\\2\end{array}$ $\begin{array}{c}1\\1\end{array}$ $\begin{array}{c}2\\3\end{array}$ $\begin{array}{c}2\\1\end{array}$ $\begin{array}{c}1\\3\end{array}$ $\begin{array}{c}3\\2\end{array}$ $\begin{array}{c}1\\3\end{array}$

(c) Probabilities:

$P_1(1)=6/16,$ $\quad P_2(11)=1/8,$ $P_2(21)=1/8,$ $P_2(31)=1/8,$

$P_1(2)=4/16,$ $\quad P_2(12)=0/8,$ $P_2(22)=0/8,$ $P_2(32)=2/8,$

$P_1(3)=6/16,$ $\quad P_2(13)=2/8,$ $P_2(23)=1/8,$ $P_2(33)=0/8.$

Figure 2: 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`

The lower bound is saturated when one of the variables is enough to determine the others, but it only makes sense for discrete distributions. The upper bound is saturated when the X_i are independent:

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

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

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

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** [2], 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 $\sum_{i=1}^{r} H(X_i) - \max_i H(X_i)$.

We can define a normalised total correlation:

$$C_r = \frac{I_r}{\sum_{i=1}^{r} H(X_i) - \max_i H(X_i)}, \qquad 0 \le C_r \le 1. \tag{6} \boxed{\texttt{eq:normmutin}}$$

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

In our cases, the random variables, X_i, all have the same marginal distribution, so $H(X_i) = H_1 \ \forall i$ and the total correlation satisfies the bounds

$$0 \le I_r(X_1, \ldots, X_r) \le (r-1)H_1. \tag{7} \boxed{\texttt{eq:mutinfbou}}$$

and the normalised version is:

$$C_r = \frac{I_r}{(r-1)H_1}, \qquad 0 \le C_r \le 1. \tag{8} \boxed{\texttt{eq:normmutin}}$$

We will use this version except where explicitly stated otherwise.

2.1 Entropy, information and bits

$\boxed{\texttt{sec:bits}}$

If we take the logarithms to base 2, we can think of entropy as a measure of "the average number of bits" of information we are missing by not knowing the value of X, and hence the number of bits of information we'd gain by measuring it. E.g. suppose X can take on two values with equal probability, then $H(X) = 1$. If the probabilities were 1 and 0,

then $H(X) =$, as we'd gain no information by measuring it due to the fact that we already know its value. If we have n independent two state systems of the first type, we'd have an entropy of n due to (4). On the other hand, if the first bit determined the remaining $n-1$ bits, we'd have an entropy of 1.

There is another way of thinking of mutual information in terms of **conditional entropy**:

$$H(Y|X) = \langle -\log P(Y|X) \rangle, \tag{9}$$

where the average is over all values of X and Y. This can be thought of as the number of bits of missing information due to not knowing Y that remain after measuring X.

Then the mutual information can be written as

$$I(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y), \tag{10}$$

i.e. it is the number of bits of information about Y that we gain by measuring X (and vice-versa).

When we have three random variables, we can write

$$
\begin{aligned}
I_3 &= \left\langle \log \frac{P(X_1, X_2, X_3)}{P(X_1)P(X_2)P(X_3)} \right\rangle \\
&= \left\langle \log \frac{P(X_2|X_1)}{P(X_2)} \right\rangle + \left\langle \log \frac{P(X_3|X_1, X_2)}{P(X_3)} \right\rangle \\
&= [H(X_2) - H(X_2|X_1)] + [H(X_3) - H(X_3|X_1, X_2)].
\end{aligned}
\tag{11}
$$

In general, we can write

$$
\begin{aligned}
I_r &= \left\langle \log \frac{P(X_1, X_2, \ldots, X_r)}{P(X_1)P(X_2)\ldots P(X_r)} \right\rangle \\
&= \left\langle \log \frac{P(X_2|X_1)}{P(X_2)} \right\rangle + \ldots + \left\langle \log \frac{P(X_r|X_1, \ldots, X_{r-1})}{P(X_r)} \right\rangle \\
&= [H(X_2) - H(X_2|X_1)] + \ldots + [H(X_r) - H(X_r|X_1, \ldots, X_{r-1})] \\
&= \sum_{i=2}^{r} [H(X_i) - H(X_i|X_1, \ldots, X_{i-1})].
\end{aligned}
\tag{12}
$$

i.e. its the sum of number of bits of information about each X_i that we gain by measuring the previous ones.

2.2 Markov processes

If the process is Markovian, i.e. one reorientation/run is only affected by the previous reorientation/run and not any earlier ones, then $P(X_3|X_1, X_2)$ is independent of X_1, and is the same function as $P(X_2|X_1)$. Therefore

$$H(X_3|X_1, X_2) = H(X_2|X_1).$$

$$I_3 = 2I_2, \qquad C_3 = \frac{I_3}{2H_1} = C_2. \tag{13}$$

6

Similar arguments apply for longer r-tuples:

$$I_r = (r-1)I_2, \qquad C_r = \frac{I_r}{(r-1)H_1} = C_2. \qquad (14) \quad \boxed{\texttt{eq:markovinf}}$$

One can look for non-Markovian behaviour by looking for differences in C_r for different r.

2.3   Effect of animal-to-animal variability

In this section we will look at the effect that animal-to-animal variability can have on mutual information in the case of binary output, i.e. the characteristics of each reorientation are reduced to two possible outcomes which we will call $\{+, -\}$. In specific cases these two outcomes will be the two most important reorientation types (omega or reversal-omega) and the reorientation direction (right or left). We will assume that animal-to-animal variability is the only effect and that there is no memory

Label the individual worms with $i = 1 \ldots N$. Then, for a single reorientation

$$P(i) = q_i, \qquad P(+|i) = p_i, \qquad P(-|i) = 1 - p_i. \qquad (15) \quad \boxed{\texttt{eq:varisingl}}$$

The prior probability that a reorientation was from a particular animal, q_i, will be proportional to that animal's reorientation rate, which may not be the same for all animals. We also introduce the notation

$$\bar{p} = \sum_i q_i p_i, \qquad \sigma_p = \sqrt{\sum_i q_i (p_i - \bar{p})^2}. \qquad (16) \quad \boxed{\texttt{eq:varinot}}$$

So that

$$P(+) = \bar{p}, \qquad P(-) = 1 - \bar{p}. \qquad (17) \quad \boxed{\texttt{eq:varimarg}}$$

Then, using Bayes' theorem

$$P(i|+) = \frac{P(+|i)P(i)}{P(+)} = \frac{q_i p_i}{\bar{p}}, \qquad P(i|-) = \frac{P(-|i)P(i)}{P(-)} = \frac{q_i(1-p_i)}{1-\bar{p}}. \qquad (18) \quad \boxed{\texttt{eq:singlereo}}$$

For pairs of successive reorientations, we use the notation

$$P(\text{next} = a | \text{prev} = b) = P(a|b), \qquad P(\text{prev} = b \cap \text{next} = a) = P(b, a). \qquad (19) \quad \boxed{\texttt{eq:varipairn}}$$

We can calculate conditional probabilities for next reorientation, given previous reorientation

$$
\begin{aligned}
P(+|+) &= \sum_i P(+|i)P(i|+) = \bar{p} + \frac{\sigma_p^2}{\bar{p}}, \\
P(+|-) &= \sum_i P(+|i)P(i|-) = \bar{p} - \frac{\sigma_p^2}{1-\bar{p}}, \\
P(-|+) &= \sum_i P(-|i)P(i|+) = 1 - \bar{p} - \frac{\sigma_p^2}{\bar{p}}, \\
P(-|-) &= \sum_i P(-|i)P(i|-) = 1 - \bar{p} + \frac{\sigma_p^2}{1-\bar{p}},
\end{aligned}
\qquad (20) \quad \boxed{\texttt{eq:varicond}}
$$

and the joint probabilities of previous and next reorientations

$$
\begin{aligned}
P(+,+) &= P(+|+)P(+) = \bar{p}^2 + \sigma_p^2, \\
P(+,-) &= P(-|+)P(+) = \bar{p}(1-\bar{p}) - \sigma_p^2, \\
P(-,+) &= P(+|-)P(-) = \bar{p}(1-\bar{p}) - \sigma_p^2, \\
P(-,-) &= P(-|-)P(-) = (1-\bar{p})^2 + \sigma_p^2.
\end{aligned}
\tag{21}
$$

The entropy of single reorientations is then

$$
H_1 = -\sum_{a \in \{+,-\}} P(a) \log P(a) = -\bar{p} \log \bar{p} - (1-\bar{p}) \log(1-\bar{p}),
\tag{22}
$$

and the mutual information is

$$
\begin{aligned}
I_2 &= \sum_{a,b \in \{+,-\}} P(a,b) \log \frac{P(a,b)}{P(a)P(b)} \\
&= \left(\bar{p}^2 + \sigma_p^2\right) \log\left(1 + \frac{\sigma_p^2}{\bar{p}^2}\right) + 2\left(\bar{p}(1-\bar{p}) - \sigma_p^2\right) \log\left(1 - \frac{\sigma_p^2}{\bar{p}(1-\bar{p})}\right) \\
&\quad + \left((1-\bar{p})^2 + \sigma_p^2\right) \log\left(1 + \frac{\sigma_p^2}{(1-\bar{p})^2}\right),
\end{aligned}
\tag{23}
$$

which vanishes when $\sigma_p = 0$.

3 Practical issues

3.1 Correcting systematic errors

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 marginal distribution, we will estimate $P_1(X)$ from the pooled data, rather that estimating the $P_1(X_i)$ separately.

The lower bounds on mutual information, (7) and (8), 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 the reciprocal of the sample size and estimating the leading order correction, see [3]. 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

8

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.3, 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.4. This can be used as a guide for when N is large enough to trust the results.

3.2 Estimating animal-to-animal variability

Estimating σ_p in (16) is tricky, as we can't keep track of animal identity during collisions. What we will do is to restrict attention to those tracks that start during the first $20\,\mathrm{min}$ and are at least $20\,\mathrm{min}$ long. This ensures that each animal is sampled at most once and that we have a decent amount of data for each animal. We will then assume that the estimate for σ_p applies to the whole experiment. Whilst the turning rates do change during the experiment, we only care about the ratio of turning rates between different animals, which we will assume remains constant.

For reorientation type, we count the number of omegas in each track, Ω_i, and the number of reversal-omegas, R_i. If the track length is t_i,

$$r_i = \frac{\Omega_i + R_i}{t_i}, \qquad q_i = \frac{r_i}{\sum_j r_j}, \qquad p_i = \frac{\Omega_i}{\Omega_i + R_i}. \tag{24}$$

For reorientation direction, we count the number of right turns in each track, R_i, and the number of left turns, L_i. If the track length is t_i,

$$r_i = \frac{R_i + L_i}{t_i}, \qquad q_i = \frac{r_i}{\sum_j r_j}, \qquad p_i = \frac{R_i}{R_i + L_i}. \tag{25}$$

In both cases, $\bar{p}$ and σ_p can be calculated by substituting (24) or (25) into (16).

The estimators for σ_p may be biased, but they should be consistent. As the sample sizes are large, the bias should be negligible. We will not attempt to calculate standard errors for these estimators.
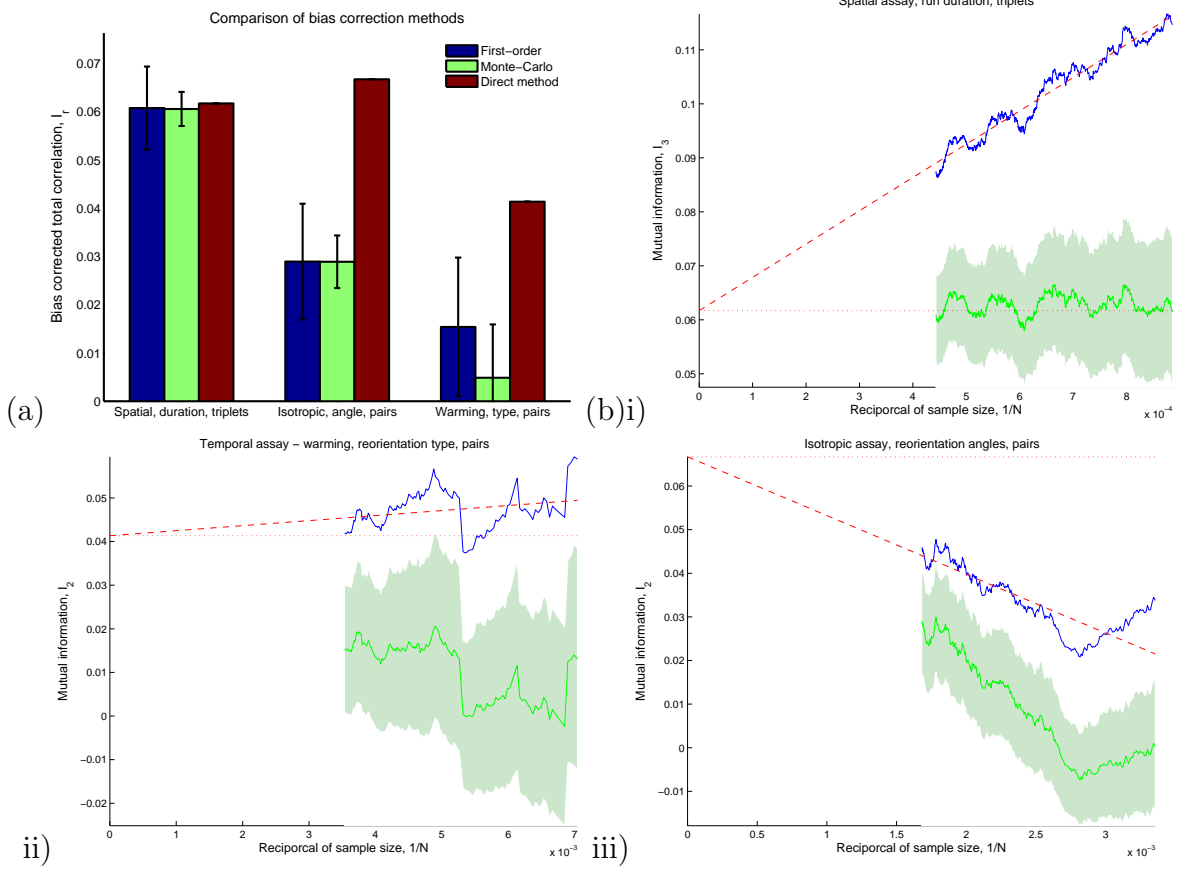
Figure 3: Comparison of different methods for removing bias. (a) Three examples of the three unbiased estimates of mutual information/total correlation. First-order refers to the methods of appendix A. Monte-Carlo refers to subtracting the mean of 1000 nonparametric bootstrap simulations, the error-bars 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).
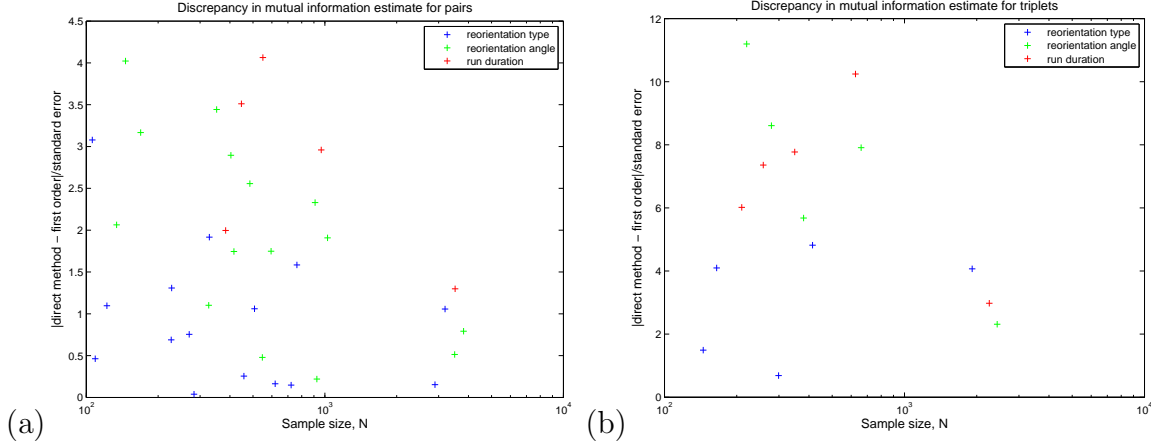
fig:biascomp

Figure 4: 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).

The results of this analysis for worms without any stimulus is shown in fig.5.

4   Results

sec:results

4.1   Binary cases

naryresults

Now we will restrict attention to binary cases, i.e. when looking at reorientation types, we will restrict attention to the 2 most common types - *omega* and *reversal-omega* (the probabilities of the different types are shown in fig.6). For reorientation angle we only pay attention to the sign, i.e. whether it is to the right or to the left.

The analysis was performed for isotropic assays (where the temperature was held uniform and constant at 20°C) and spatial assays (where the temperature was constant in time but varied linearly in space from $18 - 23$°C over 22 cm). These were done with worms cultivated at 20°C (neutral), 15°C (cryophilic) ant 25°C (thermophilic). It was also done for worms grown on 50 mM salt concentration on a spatial salt gradient from $0 - 50$ mM over 22 cm (saltphilic) and on a gradient from $50 - 100$ mM also over 22 cm (saltphobic).

We only looked at the cases $r = 2$ (pairs of consecutive reorientations). The results are in fig.7.

4.2   Continuous cases

ctsresults

We computed the mutual information in successive reorientation angles (the total change in heading direction during the reorientation) and run durations (the time between this reorientations and the previous one). As these are continuous variables, the data has to
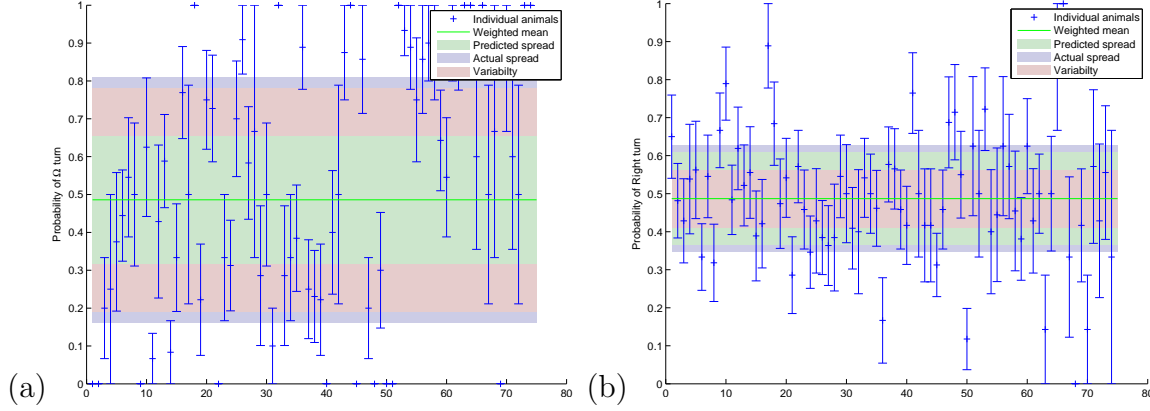
11

Figure 5: Estimating animal-to animal variability. (a) probability of Ω turns for individual animals and the weighted mean. (b) probability of Right turns for individual animals and the weighted mean. The weighted mean is $\bar{p}$ and its errorbar is σ_p (16).

`fig:variabil`

be binned. We will follow the approach of [5] 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 results of this analysis are presented in fig.8. The analysis was performed for isotropic assays (where the temperature was held uniform and constant at 20°C) and spatial assays (where the temperature was constant in time but varied linearly in space from $18 - 23$°C over 22 cm). These were done with worms cultivated at 20°C (neutral), 15°C (cryophilic) ant 25°C (thermophilic).

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 30 s as well as pairs separated by more than 30 s.

Note that when we impose a restriction on the separation of reorientations, it is no longer true that the marginal distributions of the first and second run durations in a pair are identical. If run duration can affect reorientation type or angle, this same problem could exist in those cases. Therefore we performed the analysis using (6) instead of (8) for these cases only, in which case the uncertainty and bias can be calculated using the methods of [3] without modification.

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 bounds, such as (7).
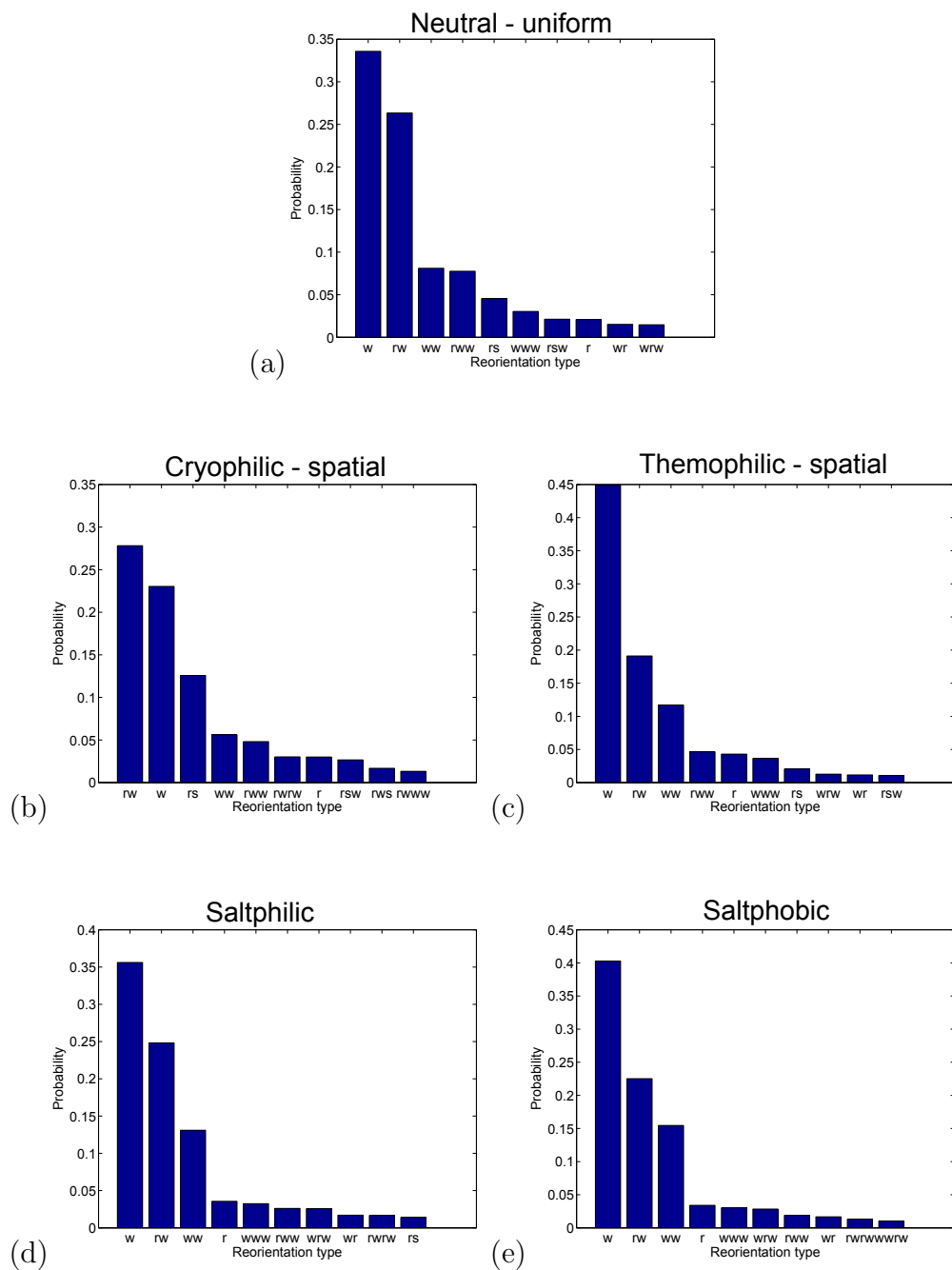
Figure 6: Probabilities of different reorientation types for (a) worms grown at 20°C on no gradient at 20°C, (b) worms grown at 15°C on spatial gradient from $18 - 23$°C, (c) worms grown at 25°C on spatial gradient from $18 - 23$°C, (d) worms grown on 50 mM salt on spatial gradient from $0 - 50$ mM, (e) worms grown on 50 mM salt on spatial gradient from $50 - 100$ mM. Reorientation labels: w - Omega turn, r - reversal, s - unreversal.
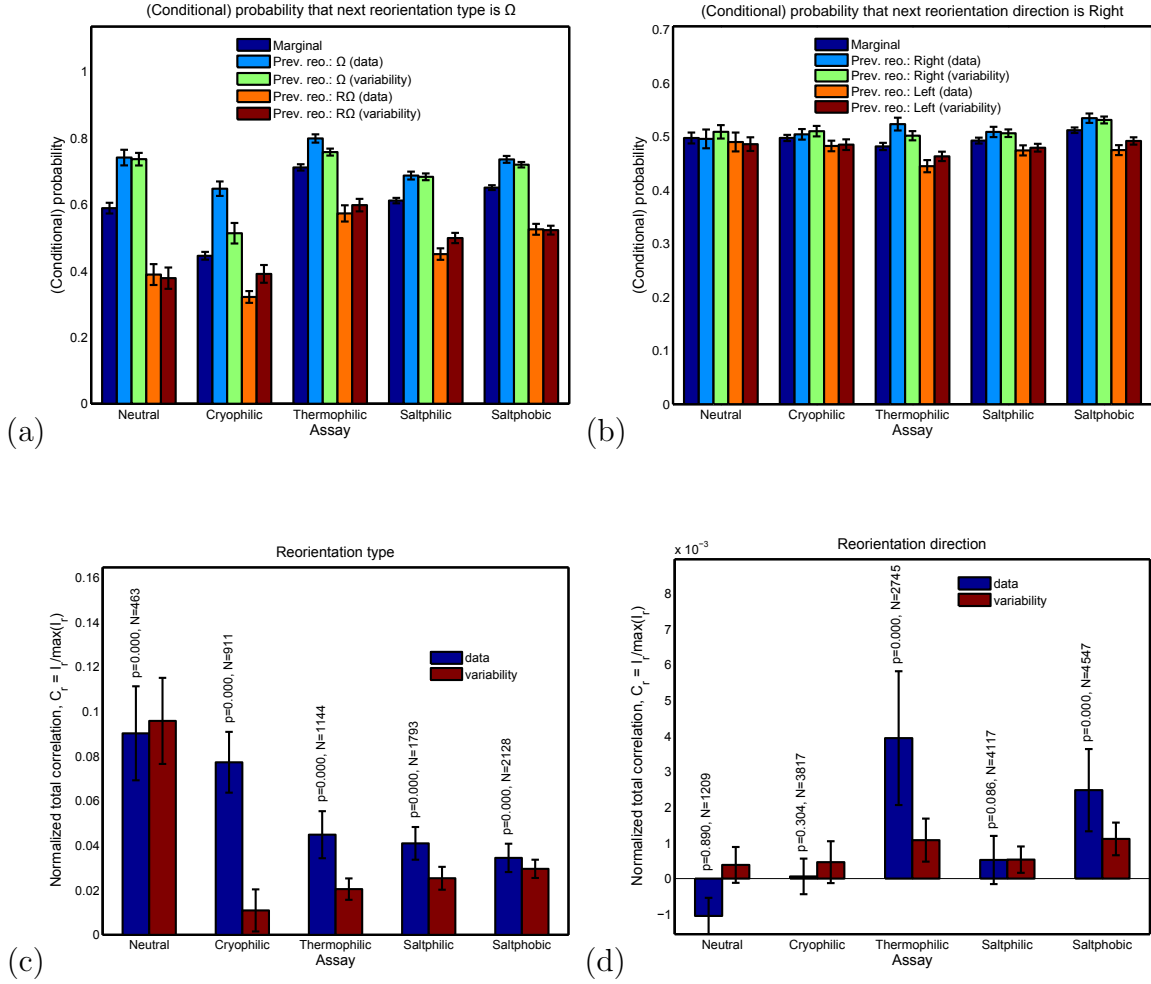
`fig:typeprob`

Figure 7: Conditional probabilities and mutual information for worms grown at 20°C on no gradient at 20°C, worms grown at 15°C on spatial gradient from $18 - 23$°C over 22 cm, worms grown at 25°C on spatial gradient from $18 - 23$°C over 22 cm, worms grown on 50 mM salt on spatial gradient from $0 - 50$ mM over 22 cm, worms grown on 50 mM salt on spatial gradient from $50 - 100$ mM over 22 cm. (a-b) Conditional probabilities for (a) reorientation types and (b) reorientation directions. Data refers to the actual measured values, variability refers to the predictions assuming that the only effect is due to animal-to animal variability as discussed in §2.3, with the variability measured as discussed in §3.2. (c-d) Normalised total correlation/mutual information in successive (c) reorientation types and (d) reorientation angles. 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 were 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.
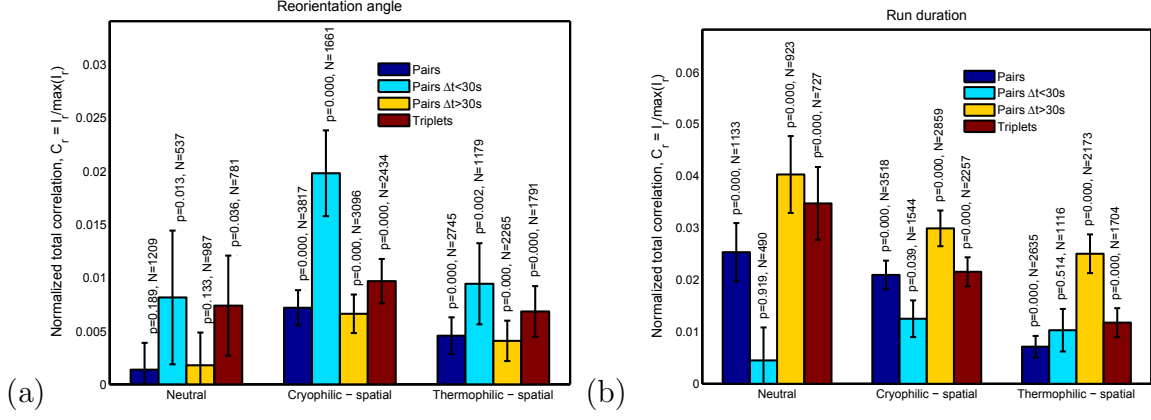
`fig:binaryre`

14

Figure 8: Normalised total correlation/mutual information in successive (a) reorientation angles and (b) run durations for worms grown at 20°C on no gradient at 20°C, worms grown at 15°C on spatial gradient from $18-23$°C over 22 cm and worms grown at 25°C on spatial gradient from $18-23$°C over 22 cm. 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 were 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:ctsresul`

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

$$q_{i_1\ldots i_r} = \frac{n_{i_1\ldots i_r}}{N}, \qquad N = \sum_{j_1\ldots ji_r} n_{j_1\ldots j_r}. \tag{26}$$

`eq:tupprob`

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

$$q_j = \sum_{i_1\ldots i_r} \left(\frac{q_{i_1\ldots i_r}}{r} \sum_{a=1}^{r} \delta_{j,i_a} \right). \tag{27}$$

`eq:singprob`

From now on, we will use A to denote the estimate of $A(p)$ with p replaced by q and $\widehat{A} = A - \text{Bias}(A)$, 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:

$$B_1 = \text{Bias}(H_1) = -\frac{\#b_1}{2rN}, \qquad B_r = \text{Bias}(H_r) = -\frac{\#b_r}{2N}, \tag{28}$$

`eq:biasH`

where $\#b_r$ is the number of non-empty bins (e.g. in fig.2, $\#b_1 = 3$, $\#b_2 = 6$). The bias estimate for I and C follow in the usual way.

15

We can estimate the standard errors with

$$\operatorname{Var}(A) \approx \sum_{i_1 \ldots i_r} \left(\frac{\partial A}{\partial n_{i_1 \ldots i_r}} \right)^2 \operatorname{Var}(n_{i_1 \ldots i_r}), \qquad \operatorname{Var}(n_{i_1 \ldots i_r}) \approx N q_{i_1 \ldots i_r}(1 - q_{i_1 \ldots i_r}). \qquad (29) \quad \boxed{\texttt{eq:stderr}}$$

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

$$
\begin{aligned}
\frac{\partial q_{i_1 \ldots i_r}}{\partial n_{j_1 \ldots j_r}} &= \frac{(\prod_a \delta_{i_a, j_a}) - q_{i_1 \ldots i_r}}{N}, & \frac{\partial B_r}{\partial n_{j_1 \ldots j_r}} &= -\frac{B_r}{N}, \\
\frac{\partial q_i}{\partial n_{j_1 \ldots j_r}} &= \frac{\frac{1}{r}(\sum_a \delta_{i, j_a}) - q_i}{N}, & \frac{\partial B_1}{\partial n_{j_1 \ldots j_r}} &= -\frac{B_1}{N},
\end{aligned}
\qquad (30) \quad \boxed{\texttt{eq:dqbydn}}
$$

which leads to

$$
\begin{aligned}
\frac{\partial H_r}{\partial n_{j_1 \ldots j_r}} &= -\frac{\log q_{j_1 \ldots j_r} + H_r}{N}, & \frac{\partial I_r}{\partial n_{j_1 \ldots j_r}} &= -\frac{(\sum_a \log q_{j_a}) - \log q_{i_1 \ldots i_r} + I_r}{N}, \\
\frac{\partial H_1}{\partial n_{j_1 \ldots j_r}} &= -\frac{\frac{1}{r}(\sum_a \log q_{j_a}) + H_1}{N}, & \frac{\partial C_r}{\partial n_{j_1 \ldots j_r}} &= \frac{\log q_{i_1 \ldots i_r} H_1 - \frac{1}{r}(\sum_a \log q_{j_a}) H_r}{(r-1) N (H_1)^2}.
\end{aligned}
\qquad (31) \quad \boxed{\texttt{eq:dhbydn}}
$$

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

References

$\boxed{\texttt{Cover:2006}}$ [1] T. M. Cover and J. A. Thomas, *Elements of information theory.* John Wiley & Sons, Hoboken, New Jersey, USA, 2nd ed., 2006.

$\boxed{\texttt{Watanabe:1960}}$ [2] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM J. Res. Dev.* **4** (January, 1960) 66–82.

$\boxed{\texttt{Roulston1999}}$ [3] M. S. Roulston, "Estimating the errors on measured entropy and mutual information," *Physica D Nonlinear Phenomena* **125** (Jan., 1999) 285–294.

$\boxed{\texttt{Strong1998}}$ [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.

$\boxed{\texttt{Slonim:2005}}$ [5] N. Slonim, G. S. Atwal, G. Tkacik, and W. Bialek, "Estimating mutual information and multi–information in large networks," `arXiv:cs/0502017`.