ROYAL SOCIETY OPEN SCIENCE

Research

Author for correspondence:
Conor Houghton
e-mail: conor.houghton@bristol.ac.uk

Calculating mutual information for spike trains and other data with distances but no coordinates

Conor Houghton

Department of Computer Science, University of Bristol, Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, UK

CH, 0000-0001-5017-9473

1. Summary

Many important data types, such as the spike trains recorded from neurons in typical electrophysiological experiments, have a natural notion of distance or similarity between data points, even though there is no obvious coordinate system. Here, a simple Kozachenko–Leonenko estimator is derived for calculating the mutual information between datasets of this type.

2. Introduction

This article describes a simple formula for calculating the mutual information between random variables where one or both of the variables take values in a metric space. This is relevant to neuroscience because electrophysiological data, whether spike trains from single neurons or collections of spike trains from a population of neurons, can be naturally considered to take values in a metric space [1,2]. It is, in turn, useful to be able to calculate information theory quantities for these data as part of an investigation into effective coding theories of neurodynamics or as a tool for quantifying the relationship between the activity of different neurons or different neuronal populations. However, the relevance is not limited to spike trains, it extends to other electrophysiological data types, such as calcium or electroencephalogram traces. Indeed, non-coordinate metrics or similarity measures are also used, for example, in genetics and biochemistry [3,4], in image analysis [5] and in information retrieval [6].

The aim here is to address two difficulties associated with estimating mutual information on metric spaces. The first difficulty is that information theory is most typically applied to problems where the data are either discrete or take values

THE ROYAL SOCIETY PUBLISHING

in an integrable manifold. The second difficulty is that many approaches to estimating information theory quantities can demand an unrealistically large amount of data. The second of these difficulties has been addressed in the past by the Kozachenko–Leonenko estimator [7–10], but, in line with the first difficulty, this estimator is derived specifically for an integrable manifold or using a local effective dimension. The aim here is to provide a simple approach to a Kozachenko–Leonenko estimator which applies to metric spaces. With this extension, mutual information can be calculated for the broad class of important data where there is a similarity measure, but no coordinates.

The particular application which motivates this article is the problem of calculating information theory quantities for spike trains. Three different types of neuroscience experiment could be considered typical. In the first, the activity of a neuron or group of neurons is recorded from the brain of an anesthetized or restrained animal while that animal is being presented with a series of stimuli. The challenge is to estimate the mutual information between the stimulus and the neuronal activity during the presentation. In the second, neuronal activity is recorded while an animal is moving freely in an arena and the mutual information is to be estimated between the position of the animal at a given time and the neuronal activity in a temporal window centred on that time. In the third example, the mutual information is to be estimated between temporal slices of the spike trains produced by different neurons so that this can be used to measure the relationship between those neurons, for example, at different times during development.

Estimating mutual information is not straightforward in any of these examples because there is no obvious coordinate system for describing neuronal activity. One approach to solving this difficulty is to discretize the spike trains, turning individual fragments of spike train into sequences of ones and zeros with each bit accounting for the presence or absence of a spike in a corresponding time slot [11]. However, the amount of possible words is huge and so this approach is bedevilled by the large amount of data it requires. One response to this problem is to exploit what is sometimes called the birthday problem and to look at coincidences [12,13]. However, the proximity structure of the space of spike trains is poorly approximated by examining the coincidence whereby two spike trains are discretized to the same sequence of ones and zeros. Here, it is proposed that, instead, one of the many metrics or similarity measures on the space of spike trains be used to define proximity [14–20].

3. Method

Two simple formulae for mutual information are presented in this article, one for the mutual information between a discrete space and a metric space and one for the mutual information between two metric spaces. These two formulae are intended to cover, respectively, the first and the second and third neuroscientific examples described above. There are two main steps to deriving these formulae. Firstly, probabilities are estimated using a simplified version of the Kozachenko–Leonenko approach [7–10]; this estimate of probability involves terms that depend on the volumes. In the second step, these volumes are estimated using the probability distribution as a measure.

3.1. A formula for the entropy

Consider estimating entropy for a random variable X which takes values in a space $\mathcal{X}$ with probability mass density $p_X(x)$. Given a set of N outcomes, $\{x_1, x_2, \ldots, x_N\}$, the entropy is estimated by

$$H(X) \approx -\frac{1}{N} \sum_{i=1}^{N} \log_2 p_X(x_i).$$

(3.1)

The problem here is how to calculate this quantity when $p_X(x)$ is not known. To do this, the approach given in [7,9] is followed in spirit but modified to avoid any quantities that rely on coordinates; the aim is to derive a formula for metric spaces. This will require that a region $B(x_i, V)$ with volume V is chosen around each data point; this region will ultimately be specified using the metric, but for now it is supposed only that there is a such a region for each data point.

Now, consider the probability $P_k(x_i)$ that the region $B(x_i, V)$ contains precisely k points; it is

$$P_k(r_i) = \binom{N}{k} F_i^k (1 - F_i)^{N-k},$$

(3.2)

where F_i is the probability mass contained in $B(x_i, V)$. This means that

$$\langle k \rangle = NF_i. \tag{3.3}$$

This quantity can be estimated from the data

$$\langle k \rangle \approx \#[B(x_i, V)], \tag{3.4}$$

where $\#[B(x_i, V)]$ denotes the number of data points in $B(x_i, V)$; that is for any $B \subseteq \mathcal{X}$

$$\#[B] = |B \cap \{x_1, x_2, \ldots, x_N\}|. \tag{3.5}$$

This means that

$$NF_i \approx \#[B(x_i, V)]. \tag{3.6}$$

Now, the probability mass function is approximated by assuming that it is constant in the ball

$$F_i \approx Vp_X(x_i). \tag{3.7}$$

This means

$$NVp_X(x_i) \approx \#[B(x_i, V)] \tag{3.8}$$

so

$$\log_2 p_X(x_i) \approx \log_2 \#[B(x_i, V)] - \log_2 N - \log_2 V \tag{3.9}$$

or

$$H(X) \approx \log_2 N + \log_2 V - \frac{1}{N} \sum_{i=1}^{N} \log_2 \#[B(x_i, V)]. \tag{3.10}$$

This assumption is the same as the one used in [7,9]; in [9], some care is given to justifying this as an approximation; here, though, it is introduced only with the general justification that the variation in $p_X(x)$ should be modest if the region spanned by $B(x_i, V)$ is small.

The formula for the entropy, equation (3.10), is similar in spirit to the one given in [7,9]. However, it is not identical, and is, in fact, simpler, because here the probability is estimated using the expected number of points in a ball rather than the size of the ball that contains a given number of points; the latter requires the trinomial, as opposed to the binomial, expansion. Also, of course, the approach here is chosen because it makes it possible to avoid quantities that are only defined on integrable manifolds.

3.2. Estimating the volume

The problem with this is that there may not be an obvious measure. Certainly in the case of spike trains there are no good coordinates and so there is no way to calculate the volume of a region based on the usual sort of coordinate-based measure. However, a probability distribution always defines a measure: the volume of a region can be defined as being equal to the probability mass it contains,

$$\text{vol } B = P(x \in B). \tag{3.11}$$

This volume can be estimated from the data

$$\text{vol } B \approx \frac{\#[B]}{N}. \tag{3.12}$$

This gives a trivial estimate of the entropy: differential entropy is zero if the probability distribution is used as the measure and the approximation used here is exact in this case: if $V = h/N$ for some integer $h \leq N$

$$H(X) \approx \log_2 N + \log_2 \frac{h}{N} - \frac{1}{N} \sum_{i=1}^{N} \log_2 h = 0 \tag{3.13}$$

as $\#[B(x_i, h/N)] = h$ by definition. Thus, using the probability as a measure for estimating information theory quantities would be useless if the aim was to estimate entropy. In fact, the entropy is a measure-dependent quantity whose value changes if the measure is changed. This is perhaps less important when a particular relevant coordinate system distinguishes a measure, but generally the significance of the entropy is not clear. The mutual information, however, does not suffer from this problem, its value is independent of the measure used. Furthermore, from the perspective of the approach taken in this article, it involves more than one distribution which means that one distribution can be used to give a measure while calculating estimates for the other distributions.

4

Here, two cases will be considered: in the first case one random variable is discrete and the other takes its values in a metric space; in the second case both random variables take values in metric spaces. In addition, an estimate is derived for the Kullback–Leibler (KL) divergence between two distributions over the same metric space.

3.3. The mutual information where one random variable is discrete

In many electrophysiological experiments, stimuli from a discrete corpus are presented repeatedly while spike trains are recorded. In this situation, the stimuli are represented by a discrete random variable and the response by a random variable taking values in a metric space. This situation is considered here; the situation is more general than the neuroscientific application, but, for convenience, the terminology is based on that application.

Let $\mathcal{S}$ be a discrete set representing the stimuli and let $\mathcal{R}$ be a set of responses, which may be spike trains or sets of spike trains recorded from multiple neurons. For simplicity, consider the situation where each element of $\mathcal{S}$ is presented an equal number of times, n_t. Let $n_s = |\mathcal{S}|$ be the number of stimuli, so the total number of data points is $N = n_t n_s$.

The metric on $\mathcal{R}$ is used to define the regions that are required around the data points. For a point r in $\mathcal{R}$ define an open ball

$$B_\epsilon(r) = \{t \in \mathcal{R} : d(r, t) < \epsilon\}. \tag{3.14}$$

Next, let $B(r, V)$ be the open ball $B_\epsilon(r)$ with ϵ chosen so that $B_\epsilon(r)$ has volume V. The total probability $p_R(r)$ will be used as a measure and the volume V fixed at $V = h/N$ for some $h \leq N$. This means that $B(r, h/N)$ is the open ball around r which contains h points.

With this measure $H(R) = 0$. The same measure can be used for the conditioned probabilities; that is, for calculating $H(R|S = s)$ using the conditioned probability $p_{R|S=s}(r)$. Hence,

$$H(R|S = s) \approx \log_2 n_t + \log_2 \frac{h}{N} - \frac{1}{n_t} \sum_{i=1}^{N} \log_2 \#\left[B\left(r_i, \frac{h}{N}\right)\right]$$

$$\approx -\log_2 n_s + \log_2 h - \frac{1}{n_t} \sum_{i=1}^{N} \log_2 \#\left[B\left(r_i, \frac{h}{N}\right)\right]. \tag{3.15}$$

Thus, averaging over $s \in \mathcal{S}$

$$I(R; S) \approx \log_2 n_s - \log_2 h + \frac{1}{N} \sum_{i=1}^{N} \log_2 \#\left[B\left(r_i, \frac{h}{N}\right)\right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \log_2 \frac{n_s \#[B(r_i, h/N)]}{h}. \tag{3.16}$$

This formula is the same as the one proposed in [21]. However, it is derived there in a convoluted way which leaned heavily on intuition, whereas here the derivation is straightforward and can be easily extended to the case where both $\mathcal{S}$ and $\mathcal{R}$ are metric spaces. It is pointed out there that although the estimate given in [9] is derived using coordinate-based quantities, it can be used to give a formula that applies in this case. Numerical experiments in [21] comparing the two formulae gave very similar results.

3.4. The mutual information where both random variables take values in metric spaces

If $\mathcal{S}$ and $\mathcal{R}$ are both metric spaces, the marginal probability mass functions $p_R(r)$ and $p_S(s)$ give volume measures and with these measures $H(R) = H(S) = 0$. These same measures also induce a measure on $\mathcal{R} \times \mathcal{S}$, the space where (R, S) takes its values. In other words, $p_{R,S}(r_i, s_i)$ is estimated by considering regions around (r_i, s_i) whose volumes are calculated using the measure $p_R(r)p_S(s)$ induced from the marginal spaces $\mathcal{R}$ and $\mathcal{S}$. Thus, a square is used to define the regions:

$$S\left(r_i, s_i, \frac{h_1}{N}, \frac{h_2}{N}\right) = \left\{(r, s) \in \mathcal{R} \times \mathcal{S} : r \in B_R\left(r_i, \frac{h_1}{N}\right), \; s \in B_S\left(s_i, \frac{h_2}{N}\right)\right\}, \tag{3.17}$$

where h_1/N and h_2/N are the volumes chosen for $\mathcal{R}$ and $\mathcal{S}$. Now, under the induced measure

$$\text{vol } S\left(r_i, s_i, \frac{h_1}{N}, \frac{h_2}{N}\right) = \text{vol } B_R\left(r_i, \frac{h_1}{N}\right) \text{vol } B_S\left(s_i, \frac{h_2}{N}\right) \approx \frac{h_1 h_2}{N^2} \tag{3.18}$$

so

$$I(R; S) \approx \frac{1}{N} \sum_{i=1}^{N} \log_2 \frac{N \#[S(s_i, r_i, h_1/N, h_2/N)]}{h_1 h_2}. \tag{3.19}$$

Thus the mutual entropy depends on $\#[S(s_i, r_i, h_1/N, h_2/N)]$ which counts the number of stimulus–response pairs (s, r), where s is one of the h_1 points closest to s_i and r is one of the h_2 points closest to r_i.

It is instructive to consider what happens when the two variables are independent. By definition, $B_R(r_i, h_1/N)$ contains h_1 points out of N; as R and S are independent this means the average number of points in $B_S(s_i, h_2/N)$ which are also in $B_R(r_i, h_1/N)$ is $h_1 h_2/N$ so

$$\frac{N \#[S(s_i, r_i, h_1/N, h_2/N)]}{h_1 h_2} \approx 1. \tag{3.20}$$

The formula includes two integer parameters, h_1 and h_2, which need to be chosen; large values for these parameters reduce the accuracy of the approximation in equation (3.7) where the probability is taken as constant throughout the region, whereas taking a smaller value reduces the accuracy of the approximation in equations (3.6) and (3.12) where the mean or volume is estimated by counting.

3.5. The Kullback–Leibler divergence on a metric space

The approach described in this article also gives an estimate for the KL divergence. Consider two random variables R and S on a metric space $\mathcal{X}$ with probability mass functions $p_R(x)$ and $p_S(x)$. If $\{r_1, r_2, \ldots, r_M\}$ and $\{s_1, s_2, \ldots, s_N\}$ are sampled from R and S then the KL divergence is estimated by

$$d(R|S) \approx \frac{1}{M} \sum_{i=1}^{M} \log_2 \frac{p_R(r_i)}{p_S(r_i)}. \tag{3.21}$$

Now, as before

$$MVp_R(r_i) \approx \#[B(r_i, V)]. \tag{3.22}$$

However, in this case, the other distribution on the same space is used to measure the volume. If the volume is chosen as h/N, then $B(r_i, h/N)$ is the ball around r_i chosen to be large enough to include h points from $\{s_1, s_2, \ldots, s_N\}$ and $\#[B(r_i, h/N)]$ is the number of points from $\{r_1, r_2, \ldots, r_M\}$ in the ball. The usual formula then gives

$$d(R|S) \approx \frac{1}{M} \sum_{i=1}^{M} \log_2 \frac{N \#[B(r_i, h/N)]}{Mh}. \tag{3.23}$$

It is easy to check that this formula gives an alternative derivation of the formula above for the mutual information between two random variables on metric spaces.

4. Conclusion

The formula derived here is very simple and is derived without any reference to coordinates on the sample space. It is intended that this demonstrates that the Kozachenko–Leonenko approach applies to metric spaces. The Kozachenko–Leonenko formula presented in [9] relies on a manifold structure in its derivation, but in its final form it is also applicable to metric spaces. It seems unlikely that the performance of the estimates here will be different from the performance of that formula.

In the estimates here, the distance or similarity measure is only required to order the points. Although electrophysiological data are used as a paradigmatic example, the formula can be applied to any pair of random variables taking values in metric spaces or indeed any space with a similarity or distance function suitable for defining regions surrounding each data point. As such it could be used in a straightforward way to calculate, for example, the mutual information between local field potentials and spike trains, or between the position of an animal in a maze and a neuronal population response, or, indeed, between two collections of images, or between images and text.

The estimate provided here relies on an approximation in which probabilities are replaced by counting how many data points fall within a region in space. This makes sound intuitive sense, but it has not been proved here that the estimate is a good one. Indeed, no principle has been described that would allow the volume of the regions to be chosen to sensibly balance the two competing requirements: small regions to reduce the error in assuming the probability is constant throughout the region and large regions so that counting-based estimates are robust. There are two ways in which these difficulties will need to be addressed: theoretically, in demonstrating that the estimate converges under sensible conditions, and practically, in demonstrating that the estimate is accurate and not overly sensitive to the choice of volumes for the sorts of data that are likely to be of interest. Hopefully, the simplicity of the approach described here will aid further development.

References

1. Victor JD. 2005 Spike train metrics. *Curr. Opin. Neurobiol.* **15**, 585–592. (doi:10.1016/j.conb.2005.08.002)

2. Houghton C, Victor J. 2012 Spike rates and spike metrics. In *Visual population codes: toward a common multivariate framework for cell recording and functional imaging* (eds N Kriegeskorte, G Kreiman), pp. 391–416. Cambridge, MA: MIT Press.

3. Sellers PH. 1974 On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* **26**, 787–793. (doi:10.1137/0126070)

4. Sankoff D, Kruskal JB. 1983 *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Reading, MA: Addison Wesley.

5. Rubner Y, Tomasi C, Guibas LJ. 2000 The Earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **40**, 99–121. (doi:10.1023/A:1026543900054)

6. McGill G, Salton MJ. 1983 *Introduction to modern information retrieval*. New York, NY: McGraw-Hill.

7. Kozachenko LF, Leonenko NN. 1987 Sample estimate of the entropy of a random vector. *Probl. Pereda. Inf.* **23**, 9–16.

8. Victor JD. 2002 Binless strategies for estimation of information from neural data. *Phys. Rev. E* **66**, 051903. (doi:10.1103/PhysRevE.66.051903)

9. Kraskov A, Stögbauer H, Grassberger P. 2004 Estimating mutual information. *Phys. Rev. E* **69**, 066138. (doi:10.1103/PhysRevE.69.066138)

10. Shapira AH, Nelken I. 2013 Binless estimation of mutual information in metric spaces. In *Spike timing: mechanisms and function* (eds P DiLorenzo, J Victor), pp. 121–136. Boca Raton, FL: CRC Press.

11. Bialek W, Rieke F, de Ruyter van Steveninck R, Warland D. 1991 Reading a neural code. *Science* **252**, 1854–1857. (doi:10.1126/science.2063199)

12. Nemenman I, Shafee F, Bialek W. 2001 Entropy and inference, revisited. In *Advances in neural information processing systems 14* (eds TG Dietterich, S Becker, Z Ghahramani), pp. 471–478. Cambridge, MA: MIT Press.

13. 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**, 056111. (doi:10.1103/PhysRevE.69.056111)

14. Victor JD, Purpura KP. 1996 Nature and precision of temporal coding in visual cortex: a metric-space analysis. *J. Neurophysiol.* **76**, 1310–1326.

15. van Rossum M. 2001 A novel spike distance. *Neural Comput.* **13**, 751–763. (doi:10.1162/089976601300014321)

16. Aronov D, Reich DS, Mechler F, Victor JD. 2003 Neural coding of spatial phase in v1 of the macaque monkey. *J. Neurophysiol.* **89**, 3304–3327. (doi:10.1152/jn.00826.2002)

17. Schreiber S, Fellous JM, Whitmer D, Tiesinga P, Sejnowski TJ. 2003 A new correlation-based measure of spike timing reliability. *Neurocomputing* **52–54**, 925–931. (doi:10.1016/S0925-2312(02)00838-X)

18. Kreuz T, Haas JS, Morelli A, Abarbanel HDI, Politi A. 2007 Measuring spike train synchrony. *J. Neurosci. Methods* **165**, 151–161. (doi:10.1016/j.jneumeth.2007.05.031)

19. Houghton C, Sen K. 2008 A new multi-neuron spike-train metric. *Neural Comput.* **20**, 1495–1511. (doi:10.1162/neco.2007.10-06-350)

20. Kreuz T, Chicharro D, Houghton C, Andrzeja RG, Mormann F. 2013 Monitoring spike train synchrony. *J. Neurophysiol.* **109**, 1457–1472. (doi:10.1152/jn.00873.2012)

21. Tobin RJ, Houghton CJ. 2013 A kernel-based calculation of spike train information. *Entropy* **15**, 4540–4552. (doi:10.3390/e15104540)