Characterizing spike trains with Lempel-Ziv complexity

J. Szczepański[1,2], J.M. Amigó[3], E. Wajnryb[1], M.V. Sanchez-Vives[4]

[1]Institute of Fundamental Technological Research, PAS, Warsaw (Poland)

[2]Centre of Trust and Certification "Centrast" Co. (Poland)

[3]Operations Research Centre, Miguel Hernández University, Elche (Spain)

[4]Institute of Neurosciences, Miguel Hernández University-CSIC, Alicante (Spain)

Keywords: Lempel-Ziv complexity, entropy, spike trains, neuronal sources.

Abstract. We review several applications of Lempel-Ziv complexity to the characterization of neural responses. In particular, Lempel-Ziv complexity allows to estimate the entropy of binned spike trains in an alternative way to the usual method based on the relative frequencies of words, with the definitive advantage of no requiring very long registers. We also use complexity to discriminate neural responses to different kinds of stimuli and to evaluate the number of states of neuronal sources.

1. Introduction

Lempel-Ziv complexity measures the generation rate of new patterns along a digital sequence and, furthermore, is closely related to such important information-theoretic properties as entropy, compression ratio and redundancy. Our group has been studying for some time [1,2,3] the characterization of spike trains by means of Lempel-Ziv complexity and other related properties. One of the main interests of this method is that complexity is a fast convergent estimator of the entropy of digital signals (as can be shown by numerical simulation), what speaks for its use especially in those cases in which the registers are too short or the insetting of non-stationary effects in a short time precludes long sequences from being considered. Other use of complexity is the discrimination of neural responses to different kind of stimuli.

We will review in this paper three applications of Lempel-Ziv complexity (not to be mistaken for the complexity measure of the same name used for lossless data compression, which was proposed later) to the analysis of spike trains: (*i*) estimation of the entropy, (*ii*) discrimination of neural responses via complexity curves and (*iii*) discrimination of neural responses via the number of states of the corresponding neuronal sources. The spike trains were intracellularly recorded in the *in vivo* and *in vitro* visual cortex as a response to different stimuli: sinusoidal current injection, visual stimulation and random current injection (only *in vitro*).

2. Lempel-Ziv complexity and entropy

Let S be a source that generates words $x_1^n := x_1 x_2 ... x_n$ of length n whose letters x_i $(1 \leq i \leq n)$ belong to a set $\mathcal{A} = \{a_1, ..., a_\alpha\}$ of size $|\mathcal{A}| = \alpha < \infty$, called the source *alphabet*. Given the word x_1^n, a *block* of length l $(1 \leq l \leq n)$ is just a segment of x_1^n of length l. The *complexity* of the word x_1^n, $C_\alpha(x_1^n)$, counts the number of different blocks (or patterns) it contains according to the following procedure. The first symbol on the left of the word defines the first block or pattern. From there one moves rightward letter by letter, until the string of symbols beginning just after the previous block and ending at the current position happens not to have appeared before. At this point, a new block is defined.

The generation rate of new patterns along x_1^n, a word of length n with letters from an alphabet of size α, is measured by the *normalized complexity* $c_\alpha(x_1^n)$, which is defined by

$$c_\alpha(x_1^n) = \frac{C_\alpha(x_1^n)}{n/\log_\alpha n} = \frac{C_\alpha(x_1^n)}{n} \log_\alpha n.$$

Let $H_b(S)$ denote the entropy of the source S,

$$H_b(S) = -\frac{1}{n} \sum_{x_1^n} p(x_1^n) \log_b p(x_1^n),$$

where $p(x_1^n)$ denotes the probability for the word x_1^n to happen and the sum is over all words of length n (α^n in total, though some of them could have zero probability). If $b = 2$, the entropy is measured in bits per second. If words can be arbitrarily long, one has to let n go to infinity, provided the limit exists. If the source S is stationary and ergodic, then [5]

$$\limsup_{n\to\infty} c_\alpha(x_1^n) = H_\alpha(S) \quad \text{almost surely.}$$

This equation provides a way to estimate the entropy of a (stationary and ergodic) neuronal source via the normalized Lempel-Ziv complexity of a typical (*i.e.* randomly chosen) spike train produced by it.

3. Experimental work

We have studied the complexity of real spike trains. The experimental data were obtained from primary cortex recordings both *in vivo* and in brain slice preparations (*in vitro*). Intracellular recordings *in vivo* were obtained from anesthetized adult cats. For the preparation of slices, 2-4 month old ferrets of either sex were used. Action potentials were detected with a window discriminator and the time of their occurrence was collected with a 10 μsec resolution. The resulting time series were used to analyze the neuronal spiking. Concerning the stimuli, they were of three kinds.

a) *Intracellular periodic current injection.* Intracellular sinusoidal currents were injected *in vivo* and *in vitro*. The frequency of the waveform was 2 Hz and the intensity ranged between 0.2 and 1.5 nA. The cell recorded ensemble comprised of 8 samples.

b) *Visual stimulation with sinusoidal drifting gratings.* The visual stimulus consisted of a 2 Hz sinusoidal drifting grating presented in a circular patch of 3-5 degrees diameter, centered on the receptive field center (*in vivo*). Only simple cells were included in this study. Also 8 samples were analyzed.

c) *Intracellular random current injection.* Random currents with different degrees of correlations were injected during the intracellular recordings from cortical brain slices (*in vitro*). The ensemble consisted of 20 samples. We have further split these responses into two subsets ("slow decay" and "fast decay"), according to whether the autocorrelation function of the stimuli decayed slowly or fast. Only spike trains for which this distinction were clear, were considered for analysis.

4. Codings

Spike trains can be encoded in many ways. We consider henceforth only the following two methods.

a) *Interspike time coding.* Let $\tau_{\min}$ and $\tau_{\max}$ be the minimal and maximal *interspike* times, respectively, in the signal. Divide the interval $[\tau_{\min}, \tau_{\max}]$ into α slots $\Delta\tau_i$ $(1 \leq i \leq \alpha)$ of the same length. If τ_j is the interspike time following spike s_j and τ_j belongs to, say, the k-th slot $\Delta\tau_k$, then assign to the spike s_j the k-symbol a_k from a set $A = \{a_1, ..., a_\alpha\}$ of α symbols. In this way, we get an α-nary message whose length equals the number of spikes.

b) *Time bin* (or *temporal*) *coding.* Let the first spike of a train occur at time 0 and the last one T time units later. The time interval $[0, T]$ is then split in n bins Δt_i $(1 \leq i \leq n)$ of the same length. If there are N_k spikes in the bin Δt_k, then assign the number N_k to Δt_k. The result is a message of length n with no more than n different letters. If, instead, each bin Δt_i is coded by 0 or 1 according to whether it contains no spike (0) or at least one spike (1), the message will be binary.

5. Applications of Lempel-Ziv complexity

5.1. Estimation of the entropy

The spike trains recorded in the lab are digitalized with *binary time bin coding*. Let $\tilde{p}_i$ be the normalized count of the ith word in the ensemble of words of length l ($= L/\Delta\tau$, $L \leq T$) in a set of observations. Then, the estimation of the entropy rate (in bits/sec)

$$H(\Delta\tau) := \lim_{l\to\infty} H(l,\Delta\tau) = -\lim_{l\to\infty} \frac{1}{l\Delta\tau} \sum \tilde{p}_i \log_2 \tilde{p}_i$$

and, hence, of the source entropy $H(S) := \lim_{\Delta\tau\to 0} H(\Delta\tau)$, requires words of increasing length l whereas real spike trains are necessarily finite. In order to avoid undersampling one needs then extrapolation techniques (such as the *standard method*, consisting of extrapolating the linear trend of the graph $H(l,\Delta\tau)$ vs $1/l$ to infinitely long words [4]) or, alternatively, fast convergent entropy rate estimators. Numerical testing with two-state Markov processes shows that the normalized Lempel-Ziv complexity rate $c(x_1^n,\Delta\tau)/\Delta\tau$ counts to the fast estimators of $H(\Delta\tau)$, hence qualifying as a useful and efficient entropy estimator in computational neuroscience. In fact, for only 200 bit long sequences generated by Markov processes (the exact entropy rate being thus known), Lempel-Ziv complexity can outperform the standard method by as much as a 15% reduction in the relative error [3].

As for the experimental validation of the complexity approach against the standard technique, the results are quite satisfactory [3]. By way of illustration, Table I summarizes the results of a typical neuronal response to random current injection (slow decay). We have chosen this experimental case because it turns out to be "intermediate" as regards the agreement between the results by both methods, the agreement being best for periodic visual stimulation and worse for random current injection with a fast decaying autocorrelation function (where stationarity can be an issue). It follows that the relative deviation of the

entropy rate estimations (in bits/sec) goes from 2% for $\Delta\tau = 0.010$ and 0.005 sec to roughly 5% for $\Delta\tau = 0.0033$ sec (corresponding to the *coding frequencies* $1/\Delta\tau = 100$, 200 and 300 Hz, respectively).

Time precision	**Standard**	**Complexity**
$\Delta\tau = 0.010$	52.38	53.38
$\Delta\tau = 0.005$	68.69	67.23
$\Delta\tau = 0.0033$	78.00	74.70

Table I: $H(\Delta\tau)$ for random current injection (slow decay)

5.2. Complexity curves

Let x_1^n be the result of encoding a spike train recorded in any of the four experimental settings considered above. One can then represent graphically $c_\alpha(x_1^n)$ as a function of the number of letters α (for interspike time coding) and also as a function of the length n of the encoded sequence (for binary or α-nary time bin coding). Remember that n is fixed (and equal to the number of spikes) for the interspike time coding and increases with increasing coding frequency $1/\Delta\tau$ for the time bin coding. Sample-averaged complexity curves are smoother.

Typical complexity curves for spike trains encoded with the interspike time coding saturates at a constant level after a fast increasing initial phase. If, instead, the time bin coding is used, the initial growing phase is followed by a convex decay phase (see [2] for details). The important message is that neural responses to different types of stimuli can be distinguished by means of complexity curves. Sometimes, as it happens with the responses to both *in vivo* stimuli (periodic current injection and visual stimulation) and interspike time coding, the complexity curves overlap, so that one has to resort to another coding (*e.g.* binary or multisymbol time bin coding) to tell one response from the other. On the other hand, this

overlapping suggests that the period of the stimuli is a kind of invariant for the interspike time coding.

It is also worth mentioning that our calculations with windows sliding along the responses to periodic stimuli show that the normalized complexity is stable. This implies that the repetition of the same stimulus produces outputs of comparable complexity.

5.3. Neuronal sources and number of states

Once a spike train has been codified into a message, this can be viewed as emitted by an information source, the source comprising everything preceding the message, namely, the stimulus, the neuron or neuronal network and the encoding technique. Consider now an ergodic source with transition probability $p(x_i|x_{-\infty}^{i-1})$, where $x_{-\infty}^{n} := ... x_{n-1}x_n$. We say that the source is Markovian of *finite order* if $p(x_i|x_{-\infty}^{i-1}) = p(x_i|x_{i-k}^{i-1})$ for some integer $k \geq 1$ called the *order* of the source. This means that the probability for the letter x_i at instant i depends directly only on the previous k outcomes: x_{i-1}, ..., x_{i-k}. If all substrings of length k are feasible, the number of states N of such a source is $N < \alpha^k$; otherwise, $N = \alpha^k$. For this reason, we refer also to k (somewhat imprecisely) as the "number of states".

Next let $H(q_{\mathbf{x}}^k)$ be the *k-th order empirical entropy* of an ergodic Markov process as measured at the word $\mathbf{x} = x_1^n$ [6,1]. An estimator k^* for k is then given by

$$k^* = \min\left\{k : H(q_{\mathbf{x}}^k) - \, c_\alpha(\mathbf{x}) \log_2 \alpha \leq \lambda\right\}$$

where $c_\alpha(\mathbf{x})$ is the normalized complexity of $\mathbf{x}$. In the calculations we set $\lambda = 0.02$. Important for us is that k^* is a numerical invariant for neuronal sources. In particular, for a given neuron preparation and coding, k^* depends only on the kind of stimulus (*i.e.* on the experimental subcase considered out of the four experimental cases given in Sect. 3), but not on individual stimuli. This comes as no surprise since the same is true for the normalized complexity.

We refer to [1] for the numerical results and how, in fact, different stimuli and codings lead, in general, to different values (or, rather, estimation intervals) of k^*, hence allowing to discriminate the corresponding neuronal sources.

6. Conclusions

1. Numerical simulations show that Lempel Ziv complexity is a fast convergent estimator of the entropy of sequences and that can provide more accurate estimations than the standard approach for short (say, 200 bit long) sequences.

2. Apart from the estimation of entropy, Lempel-Ziv complexity allows to introduce some useful tools for analyzing spike trains, *e.g.* complexity curves (with several encodings) of a given spike train as well as the number of states of the neuronal source which has produced that spike train.

3. These analytical tools (complexity curves and number of states) have proved useful to separate spike trains that were obtained under different conditions. Thus, we have found significant differences between the spike trains obtained in the cortical neurons *in vivo* versus *in vitro*, such that those *in vivo* had higher information content than those *in vitro*, even when the stimulus was the same (sinusoidal current injection).

References

[1] J.M. Amigó, J. Szczepański, E. Wajnryb, M.V. Sanchez-Vives, On the number of states of the neuronal sources, BioSystems 68 (2003) 57-66.

[2] J. Szczepański, J.M. Amigó, E. Wajnryb, M.V. Sanchez-Vives, Application of Lempel-Ziv complexity to the analysis of neural discharges, Network: Comput. Neural Syst. 14 (2003) 335-350.

[3] J.M. Amigó, J. Szczepański, E. Wajnryb, M.V. Sanchez-Vives, Estimating the entropy

rate of binned spike trains via Lempel-Ziv complexity, Neural Computation 16 (2004) (to appear in issue 3).

[4] S.P. Strong, R. Koberle, R.R. de Ruyter van Steveninck, W. Bialek, Entropy and information in neural spike trains, Phys. Rev. Lett. 80 (1998) 197-200.

[5] J. Ziv, Coding theorems for individual sequences, IEEE Trans. Inform. Theory 24 (1978) 405-12.

[6] J. Ziv, Compression, tests for randomness and estimating the statistical model of an individual sequence, in: R.M. Capocelli, ed., Sequences (Springer Verlag, New York, 1990) 366-373.