2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Spike train clustering using a Lempel-Ziv distance measure

M. Christen†, T. Ott†, and R. Stoop†

†Institute of Neuroinformatics, University / ETH Zürich,
Winterthurerstrasse 190, 8057 Zürich, Switzerland
Email: {markus,tott,ruedi}@ini.phys.ethz.ch

Abstract—Multi-electrode array recordings reveal complex structures in the firing of spatially distributed neurons. The analysis of this neuronal network activity demands a classification of neurons according to similarities in their firing behavior. If similar spike patterns do not occur synchronously, but have unknown delays within spike trains, this processing step is difficult. To solve this problem, we introduce a Lempel-Ziv complexity-based distance measure. Using our distance measure as the input for a superparamagnetic clustering algorithm, we achieve an efficient classification of spike trains.

1. Introduction

In recent years, multi-electrode arrays have become a standard tool in neuroscience [1]. They open up new horizons for the investigation of the input-output relationship of neuronal networks, as the spiking behavior of dozens up to hundreds of cortical neurons can be measured simultaneously. This development has led to an increased interest in spatio-temporal spike patterns [2]. Up to now, mostly patterns of synchronized spikes have been investigated [3]. However, due to the complex neuronal connectivity, the same pattern may occur at different times in different neurons. Such neurons may be assumed to be receiving similar input or/and performing similar computations. A method to classify spike trains should thus be able to group together neurons with similar temporal structure, irrespective of the delays between the patterns (Fig. 1).

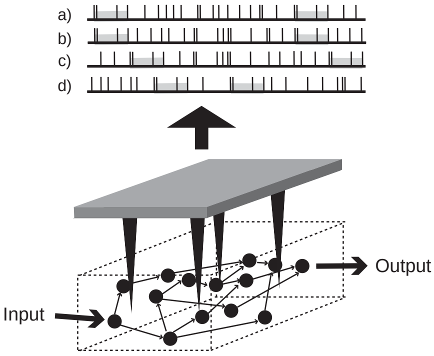

Figure 1: Outline of the spike train clustering problem: Multi-electrode array recordings probe the spatio-temporal activity within a neural net. Correlation-based distance measures allow to group together spike trains with synchronous spike patterns (a,b; pattern marked with grey bar), but fail to group trains with delayed patterns (c,d).

A thorough classification of spike trains thus requires two ingredients: The choice of an appropriate distance measure, and an efficient clustering algorithm [4]. In our contribution, we focus on the first step. We introduce a distance measure based on the Lempel-Ziv complexity (LZ-distance) [5]. This measure allows the grouping of neurons with distributed patterns, in contrast to customary correlation-based distance measures. We will proceed by defining the LZ-distance and introducing two alternative bitstrings parsing schemes. Thereafter, using data of artificial models and of *in vivo* neuronal data, we will assess the suitability of the LZ-distance. For this, we will use the LZ-distances among spike trains as the input to our sequential superparamagnetic clustering algorithm. The latter will not be explained in this contribution, we refer to Ref. [6] for a detailed description. We will finally apply the classification method to multi-electrode recordings obtained from the olfactory system of the rat.

2. The LZ-distance

The starting point of our investigation are sequences of neuronal spike-times $\{t_1, \ldots, t_n\}$. These trains are translated into bitstrings. For this translation, the interval $[0, T]$ covering the whole measurement time is partitioned into n bins of width $\Delta\tau$ ($n\Delta\tau = T$). If at least one spike falls into the i-th bin, the letter "1" (and otherwise the letter "0") is written to the i-th position of the string. Usually, $\Delta\tau$ is chosen so that maximally one spike falls into one bin. This can be achieved by setting $\Delta\tau = 1$ ms, as the refractory period of neurons is of the same magnitude. The resulting bit-string can be viewed as being generated by a more general information source. For this source, we want to find the optimal coding [7]. This coding is based on *parsing*, a procedure to partition the string into non-overlapping substrings, according to some procedure. Based on the concept of LZ-complexity, two distinct parsing procedures have been introduced [5, 8]. Both of them follow the same basic idea: strings are sequentially parsed into sequences that have not occurred.

To explain the differences among the two procedures, let $X_n = x_1 \ldots x_n$ be a bit-string of length n ($x_i \in \{0, 1\}$). Let

$X_n(i, j)$ be a substring starting at position i and terminating at position j, called a *phrase*, if it is the result of a parsing procedure. Let P_{X_n} denote the set of phrases generated by parsing X_n, and $c(X_n)$ is the number of phrases detected ($c(X_n) = |P_{X_n}|$). A *vocabulary* V_{X_n} of a string X_n is the set of all possible substrings of X_n. We assume that X_n has been parsed up to position i, such that $P_{X_n(1,i)}$ is the set of phrases generated so far and $V_{X_n(1,i)}$ is the vocabulary of the parsed substring $X_n(1, i)$. The question is: which will be the next phrase $X_n(i + 1, j)$? According to the originally proposed parsing procedure [5], it will be the first substring which is not yet an element of $V_{X_n(1,i)}$ (LZ-76). According to the second, later, proposed parsing procedure [8], it will be the first substring which is not an element of $P_{X_n(1,i)}$ (LZ-78). As an illustration, take the string 0011001010100111. Using the LZ-76 procedure, it will be parsed as 0|01|10|010|101|00111, whereas it will be parsed as 0|01|1|00|10|101|001|11 using the LZ-78 procedure.

Both procedures belong to the class of distinct parsings (all elements in P_{X_n} are distinct, respectively). Distinct parsings of strings X_n which are the result of stationary ergodic processes with entropy rate H, have the property of asymptotic optimality [7]

$$K(X_n) = \limsup_{n\to\infty} \frac{c(X_n) \log c(X_n)}{n} \leq H \qquad (1)$$

with probability 1. Equation (1) defines the LZ-complexity $K(X_n)$ of the string X_n, which can be calculated using either parsing procedures. The application of (1) for estimating the entropy of spike trains is based on the assumption of stationarity – which is hardly fulfilled in biological processes. We nevertheless base our distance measure on the application of (1), showing that it will not be affected by this constraint.

Consider two strings X, Y of equal length n. From the perspective of LZ-complexity, the amount of information Y knows about X is given as $K(X) - K(X|Y)$, where, for the calculation of $K(X|Y)$, $c(X|Y)$ is the size of the difference set $P_X \setminus P_Y$. If Y provides no information about X, then the sets P_X and P_Y are disjoint, and $K(X) - K(X|Y) = 0$. If Y provides complete information about X, then $P_X \setminus P_Y = \emptyset$ and $K(X) - K(X|Y) = K(X)$. This leads to the following definition of the *LZ-distance*:

$$d(X, Y) = 1 - \min\left\{\frac{K(X) - K(X|Y)}{K(X)}, \frac{K(Y) - K(Y|X)}{K(Y)}\right\} \quad (2)$$

We use *min* to ensure $d(X_n, X_m) > 0$ for $n \neq m$ (it can be shown, that the definition satisfies the axioms of a metric). The LZ-distance thus compares the set of phrases generated by a LZ parsing procedure of two bitstrings originating from corresponding spike trains. A large number of patterns appearing in both spike trains should lead to a large overlap of the sets of phrases. This leads to the prediction that distances between spike trains with similar patterns are small, whereas distances between trains with different patterns are large. Thus, the LZ-distance should allow a classification of spike trains according to temporal similarities (unrestricted with regards to temporal synchrony).

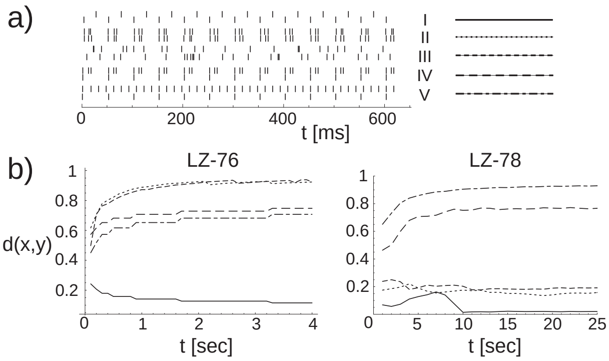

Figure 2: Distances of pairs of spike trains. a) Raster plot of pairs I-V. b) Distances of pairs I-V in dependence from the length of the spike train for LZ-76 and LZ-78 parsing.

3. Assessment of the LZ-distance

To verify our predictions, we shall evaluate a number of test cases. The appropriateness of our distance is assessed in three steps: We first compare the parsing procedures LZ-76 and LZ-78 from the practical aspects point of view of. We then analyze whether the LZ-distance classifies spike trains into physiologically meaningful classes. We finally compare the LZ-distance with correlation-based distance measures, showing its superiority in the presence of delayed patterns.

1) Choosing the parsing procedure: We calculated the LZ-distance for five pairs of model spike trains (see Fig. 2.a) using both parsing procedures. I: two period-1 spike trains (interspike interval, ISI, = 50 ms) with equal period length and phase shift (25 ms). II: two period-3 spike trains (ISI-pattern (10,5,35)) with no phase shift but spike jitter (±1 ms). III: two spike trains obtained from an uniform random process on the interval $[1, 50]$. IV: a period-1 (ISI = 50 ms) and a period-3 (ISI-pattern (10,5,35)) spike train with coincident spikes. V: two period-1 spike trains with different period lengths (50 ms,15 ms). In order to analyze the convergence behavior, for LZ-76 parsing, the length of the trains have been increased from 100 ms up to 4000 ms in steps of 100 ms. For LZ-78 parsing, the lengths of the trains have been increased from 1 sec up to 25 sec in steps of 1 sec.

The results (Fig. 2.b) show that the distance based on LZ-76 parsing converges faster than LZ-78 parsing. Furthermore, same pairs lead to different distances: Using LZ-76 parsing, the trains I are close, the trains IV and V are rather distant and the trains II and III are most distant. Using LZ-78 parsing, the trains I are close as well, but the trains II and III are less close, and the trains IV and V are most distant. This demonstrates, that the LZ-distance based on LZ-78 parsing is more noise-robust than LZ-76 parsing. The latter considers similar, but noisy trains (II) as most distant, whereas the first measure considers spike trains with distinct firing behaviors (IV, V) as most distant. The evaluation of the parsing procedure explains this difference: In LZ-76 parsing, the lengths of the phrases increase much faster during the procedure than in the LZ-78 parsing, because $|V_{X_n(1,i)}| \gg |P_{X_n(1,i)}|$. Therefore, the probability that two similar, but noisy, strings contain many different phrases is higher for the LZ-76 if compared to the LZ-78 parsing. As noise robustness is important when dealing with neuronal data, LZ-78 is better suited for practical purposes. It is furthermore computationally cheaper and faster than LZ-76 parsing.

2) Spike train classification: To test whether the LZ-distance sorts spike trains in physiologically meaningful categories, we use different model (A,B,C) and *in vivo* (macaque monkey visual cortex data (D,E): A) Poisson spike trains with refractory period. B) Poisson spike trains with refractory period driven by a step function of 12.5 Hz. C) Noisy burst-pattern spike trains. D) Spike trains of a complex cell driven by drifting gratings of 6.25 Hz. E) Spike trains of a simple cell driven by drifting gratings of 12.5 Hz. Each class consists of nine different trains of 2400 ms length each, and comparable firing rates (80-90 spikes/second). The order of the trains was randomized (Fig. 3.a).

After calculating the LZ-distance between all trains, clustering led to the following result: The classes B, C and E have been separated in the first run. Sequential clustering of the remaining cluster led to an incomplete separation between spike trains of the classes A and D (Fig. 3.b). Two conclusions can be drawn: First, a classification of spike trains based on differing temporal structures, but with similar firing rates, is possible. Second, the method also allows the evaluation of the quality of neuron models: Whereas simple cell firing is not captured by a Poisson model driven by a periodic function of equal frequency as the visual stimulus, the firing behavior of the complex cell appears to be (in a first approximation) properly modeled by a Poisson model.

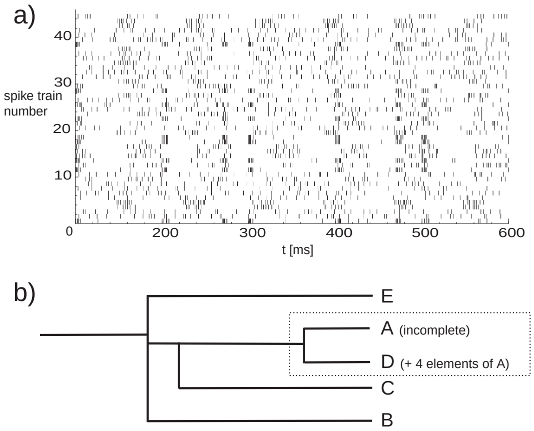

Figure 3: Clustering of simulated multi-train data: a) Raster plot of initial spike set. b) Dendrogram outlining the result of clustering. Dashed box: result of sequential clustering.

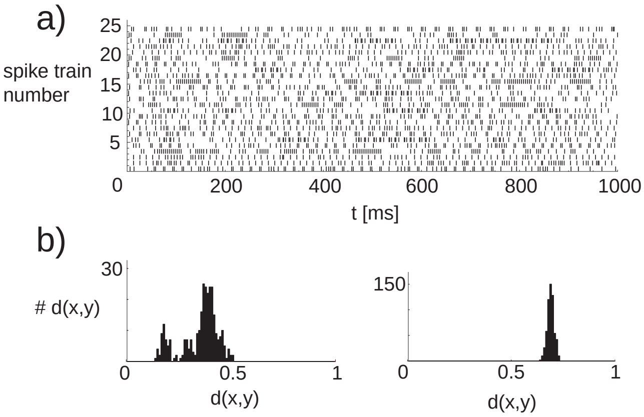

Figure 4: Comparison of distance measures. a) Test trains b) Dynamic range of LZ-distances (left) and correlation-based distances.

3) Comparison with correlation-distance: A variety of distance measures for spike trains have been proposed in the past (e.g. in [4]). In the test case of Fig. 2, such measures lead to results similar to our LZ-distance. We now investigate the more general case, where repeated ISI sequences have been *arbitrarily* placed in a random background. We generated five classes of spike trains, characterized by the ISI-patterns A: (4,4), B: (13,13,13), C: (5,20,3), D: (3,16,3,16), and E: (1,4,7,2,6,11). The spike trains (five per class) were generated such that 50% of the ISI's originate from the sequence and 50% from a homogeneous Poisson (background) process. The latter has been tuned so as to generate comparable mean firing rates for all spike trains (92-94 spikes per second). The order of the spike trains was again randomized (Fig. 4.a).

To these spike trains, sequential superparamagnetic clustering has been applied, using the LZ- and a correlation-based distance measure [4]. Whereas the application of our distance measure allowed a clear-cut separation of all five classes, the use of the correlation-based measure did not allow any classification at all. This noticeable difference in performance becomes transparent, if the dynamic range of the distances is compared (Fig. 4.b). For the LZ-distance, the range is ~ 0.4 with a multimodal distribution (indicating the structure within the dataset), whereas for the correlation-based distance the range is ~ 0.1 with a unimodal distribtion. The latter observation implies that a rescaling is useless for a performance increase. Thus, in

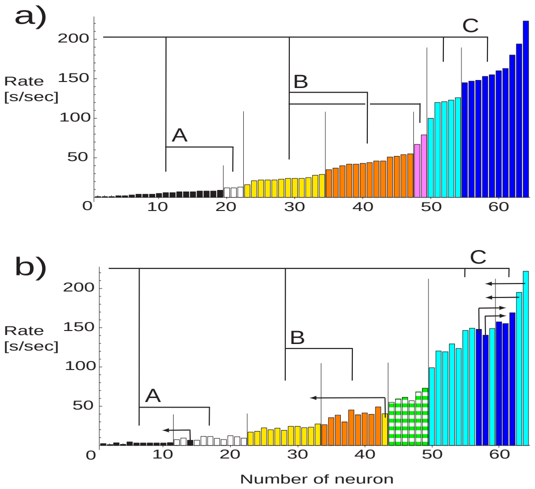

Figure 5: Classification of multi-electrode recordings. a) Neurons are ordered according to the firing rate before stimulus presentation. Dendrograms indicate the result of clustering. Clusters are color-coded, borders are indicated by lines. b) Classification after stimulus presentation. Arrows indicate neurons belonging to different clusters. Striped bars indicate neurons belonging to no cluster.

contrary to the correlation-based distance, the LZ-distance allows a classification of spike trains with delayed patterns.

4. Application to multi-electrode recordings

Finally, we apply our method to a multi-electrode recording in the rat olfactory bulb. We analyzed the spike trains of 64 neurons obtained from a 6 × 7 electrode array in a pre-stimulus condition (= no odor present) and during presentation of an odor. Each measurement lasted 5 seconds. The neurons were labeled according to the firing rate in the pre-stimulus condition (neuron #1: lowest firing rate). The results (Fig. 5) show that the application of our method in the pre-stimulus condition leads to a classification according to the firing rate (Fig. 5.a): First, the classes A (low firing rate), B (medium firing rate) and C (high firing rate) appear, splitting into even more classes, when sequential clustering is applied. In the stimulus-condition a more complex picture emerges (Fig. 5.b) : Some neurons (dashed bars) fall in no clusters at all, and the sequential clustering of the classes A, B and C leads to a fine classification that does no more fit into the previous picture (classification according to the firing rate). LZ-distance based clustering thus takes account of the changes in the temporal structure of the firing of neurons before and during stimulus-presentation.

5. Conclusion

We have combined the LZ-distance with sequential superparamagnetic clustering. This method is able to group spike trains with similar, but not necessarily synchronous, patterns. It therefore has a broader range of applications than other distance measures. Moreover, it does not rely on prior information, is easy to implement, and computationally cheap. Our method is helpful in the context of a number of multi-spike train problems: 1) It can identify neurons with similar firing behaviors in a fast and unbiased way. Instead of the analysis of all neurons, a refined analysis can be restricted to one representative of each class. 2) It is able to assess the degree of precision work to which a neuron model reproduces the temporal structure of a biological neuron. 3) By comparing different spike trains from one neuron, its neuronal firing reliability can be measured (and compared with other neurons).

Acknowledgments: The authors thank Adam Kohn (Center of Neural Science, New York University) for V1 and Alister Nicol (Cognitive & Behavioural Neuroscience Lab, The Babraham Institute, Cambridge) for rat olfactory bulb data used in this analysis. This work was supported by the SNF grant 247076 and by the KTI contract 5007.2 with Novartis.

References

[1] G. Buzsáki, “Large-scale recordings of neuronal ensembles”, *Nat. Neurosci.*, 7(5), pp.446–451, 2004.

[2] Y. Ikegaya, G. Aaron, R. Cossart, D. Aronov, I. Lampl, D. Ferster, R. Yuste, “Synfire chains and cortical songs: temporal modules of cortical activity”, *Science*, 304, pp.559–564, 2004.

[3] S. Grün, M. Diesmann, A. Aertsen, “Unitary events in multiple single-neuron spiking activity”, *Neural Computation*, 14, pp. 43–119, 2002.

[4] J.-M. Fellous, P. H. E. Tiesinga, P. J. Thomas, T. J. Sejnowski, “Discovering spike patterns in neuronal responses”, *J. Neurosci.*, 24(12), pp.2989–3001, 2004.

[5] A. Lempel, J. Ziv, “On the complexity of finite sequences”, *IEEE Trans. Inform. Theory*, IT-22, pp.75–81, 1976.

[6] T. Ott, A. Kern, A. Schuffenhauer, M. Popov, P. Acklin, E. Jacoby and R. Stoop, “Sequential Superparamagnetic Clustering for Unbiased Classification of High-dimensional Chemical Data”, *J. Chem. Inf. Comput. Sci.*, 44(4), pp.1358-1364, 2004.

[7] T. M. Cover, J. A. Thomas, “Elements of Information Theory”, John Wiley & Sons Inc., New York, 1991.

[8] J. Ziv, A. Lempel, “Compression of individual sequences by variable rate coding”, *IEEE Trans. Inform. Theory*, IT-24, pp.530–536, 1978.