PATTERN DETECTION IN NOISY SIGNALS

M. Christen, A. Kern, J.-J. van der Vyver and R. Stoop

Institute of Neuroinformatics, University / ETH Zürich
Winterthurerstrasse 190, 8057 Zürich, Switzerland

ABSTRACT

Methods for detecting patterns in noisy signals are often template based. As a consequence, *a priori* selections of potential pattern structures have to be made. To avoid this shortcoming, we propose a novel statistical approach based on the correlation integral. The method significantly reduces the set of appropriate templates, and also works under noisy conditions.

1. INTRODUCTION

Biological neural systems can be viewed as an alternative information processing paradigm, that often proves far more efficient than conventional approaches. Although the underlying structures (neurons and their connectivity) can be accurately modelled by electronic circuits [1], the principles according to which they process information are not well understood. However, growing evidence suggests that neuronal circuits work according to distributed parallel processing principles, and that information encoding differs from than of traditional signal processing [2].

In neural information processing systems, activity is manifested as spikes. Temporal recordings of firing events provide interspike interval (ISI) series as the empirical material to work with. It is expected that the information processed in the network is encoded as some structure in the ISI series. This implies that pattern detection should constitute the first step in the investigation of neural information processing. The simplest starting point are single ISI series. Patterns can be defined as parts of the series that repeat significantly more often than they would in random distributions [3]. For the identification of patterns, pattern templates are usually predefined and their frequency of occurrence is counted (*template-based methods* [4]).

Template-based methods suffer from two fundamental problems. First, the detection relies on the set of pre-chosen templates. As the patterns are *a priori* unknown, the inclusion of appropriate templates could be considered a matter of luck. Unbiased guessing will, therefore, require exceedingly large template sets. The second problem is the omnipresence of noise, implying that patterns cannot be expected to repeat perfectly. As a consequence, the question arises of how to choose the accuracy required for template matching. To avoid these problems, we propose an unbiased, purely statistical approach to pattern detection.

This work was partially supported by the Swiss National Science Foundation grant 247076 to R. Stoop and by a KTI contract with Phonak AG Hearing Systems.

2. CORRELATION INTEGRAL METHOD

The correlation integral was originally designed for the determination of the correlation dimension [5]. The purpose of this paper is to explore its potential for the detection of patterns in ISI-series. First, we briefly introduce the correlation integral, and its ability to detect point clusters in the embedding space. Second, we discuss the utilizability of the method for ISI patterns. Third, we evaluate, to what extent the method is able to detect the length of ISI patterns.

Consider an arbitrary scalar time series of measurements $\{x_i\}$, $i = 1, \ldots, L$, where L denotes the length of the time series. From these data, embedded points $\xi_k^{(m)}$ are constructed: $\xi_k^{(m)} = \{x_k, x_{k+1}, \ldots, x_{k+(m-1)}\}$, where m is called the embedding dimension. This *coordinate-delay construction* is standard in nonlinear dynamics. Its purpose is to reconstruct the complete underlying dynamics from partial, mostly scalar, measurements [6, 7]. From the embedded data, the *correlation integral* is calculated as

$$C_N^{(m)}(\epsilon) = \frac{1}{N(N-1)} \sum_{i \neq j} \theta(\epsilon - \|\xi_i^{(m)} - \xi_j^{(m)}\|),$$

where $\theta(x)$ is the Heavyside function ($\theta(x) = 0$ for $x \leq 0$ and $\theta(x) = 1$ for $x > 0$) and N is the number of embedded points ($N \leq L - m + 1$). For the actual computation of $C_N^{(m)}(\epsilon)$, different norms can be used. In most cases, the maximum norm is of advantage, as this choice speeds up the computation, and allows an easy comparison of results obtained for different embedding dimensions. Degeneracies introduced by this choice are removed by adding a small amount of noise (in general uniformly distributed within 1% of the interspike intervals).

The correlation integral $C_N^{(m)}(\epsilon)$ allows the detection of clusters, which are formed by the embedded points: For the calculation of $C_N^{(m)}(\epsilon)$, an embedded point $\xi_0^{(m)}$ is randomly chosen. Then, the number of points in its ϵ-neighborhood is evaluated, as ϵ is enlarged. If the point belongs to a cluster, many points will join the ϵ-neighborhood. Once the cluster size is reached, less points are recruited, leading to a slower increase of $C_N^{(m)}(\epsilon)$. When an average over different points of the cluster is performed, as required by the correlation integral, pieces of fast increase of $C_N^{(m)}(\epsilon)$ interchange with pieces of slow increase. The denser the clustered region, the more prominent this step-wise structure. When $C_N^{(m)}(\epsilon)$ is displayed in a log-log plot ($\log C_N^{(m)}(\epsilon)$ vs. $\log \epsilon$), the step-like structures are preserved and extends over a larger region, as compared to $C_N^{(m)}(\epsilon)$ vs. ϵ. To demonstrate our predictions, we constructed a series as a repetition of the sequence {1,2,4}, where the sequence numbers can be interpreted as ISI durations. The embedding of this series for $m = 2$ leads to three clusters, represented by the points $P_1 = \{1, 2\}$, $P_2 = \{2, 4\}$ and $P_3 = \{4, 1\}$. Calculating the correlation integral and plotting $\log C_N^{(m)}(\epsilon)$ against $\log \epsilon$ leads to a clean-cut staircase structure in the log-log plot (Fig. 1a).

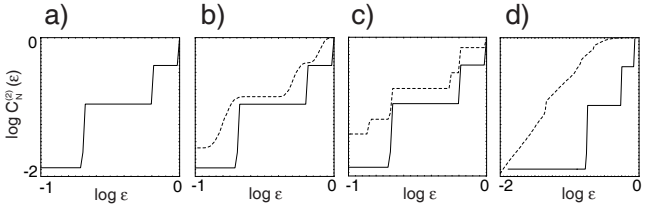

Fig. 1. Log-log steps for different classes of data ($m = 2$). a) Virtually noise free (noise $\pm 1\%$, solid line in all four plots), b) noisy (noise $\pm 10\%$, dashed line), c) unstable periodic (dashed line), d) random background (dashed line).

In practical applications of the method, the log-log steps generally become less salient due to influences that will be discussed below. In this case, the difference quotient $\Delta \log C_N^{(m)}(\epsilon_i) := \log C_N^{(m)}(\epsilon_{i+1}) - \log C_N^{(m)}(\epsilon_i)$, which approximates the derivative of the correlation integral, may provide an improved indication for the occurrence of clusters. For small ϵ-neighborhoods, the log-log plot is affected by strong statistical fluctuations. These regions, however, are easily identified and excluded from the analysis.

3. BLURRED LOG-LOG STEPS

The ISI series used for Fig. 1a is the simplest and most ideal case of a ISI pattern – the repetition of a single sequence. However, in natural systems, three influences contribute towards a blurring of the steps, if only one pattern is present.

First, the ISI series could be affected by noise. This can be modelled by adding uniform noise to our ISI series {1,2,4,1,2,4,...}. As the point clusters are less dense in the embedding space, we expect noise to first affect the boundaries of the log-log steps, and then penetrate towards the centers, as the amount of the noise increases. When noise is added to our ISI series, the obtained results in the log-log plot (dashed line) clearly corroborate our predictions: the horizontal parts of the steps are less broad, and the vertical parts are less steep (Fig. 1b).

Second, the system that generates the ISI series could be chaotic in nature. In this case, a distance from a given unstable periodic orbit is expected to grow as $e^{t\lambda}$, where t denotes the time and λ is the (positive) Lyapunov exponent of the orbit. This implies that a continued perfect repetition of a sequence is unlikely. Moreover, because the decay from the unstable orbit occurs in a deterministic manner, additional (pseudo-) orbits will emerge and lead to an increased number of steps. We illustrate this with a simplified model: With probability $p_1 = 0.5$ we take the whole sequence {1,2,4}, with probability $p_2 = 0.31$ we select the subsequence {1,2}, and with probability $p_3 = 0.19$ the subsequence {1} (this choice leads to $\frac{p_1}{p_2} \simeq \frac{p_2}{p_3}$). The obtained results in the log-log plot (dashed line) corroborate our assertions: instead of three steps, five steps appear (Fig. 1c).

Third, patterns could occur within a noisy background. In this case, the characteristic sequence of ISIs is only produced occasionally, otherwise, the length of the ISIs are random. Therefore, the fraction of points belonging to clusters in the embedding space will be diminished, which leads us to expect that the steps in the log-log plot become less prominent. To simulate this situation, we took with probability $p = 0.5$ the sequence {1,2,4}, otherwise three interspike intervals were randomly drawn from the interval $(0, 4]$. The obtained results in the log-log plot (dashed line) corroborate our assertions: the number of steps is unaffected, but the steps themselves are much less pronounced (Fig. 1d).

ISI data of natural systems can also be expected to contain more than one pattern. To analyze this situation, we extended our investigations to data composed of 3 sequences ({2,6,10}, {8,2,1}, {2,7,5}). To contrast multiple patterns against random firing, two series were constructed: the first by randomly selecting among the three sequences, the second by randomly selecting intervals from the concatenated set {2,6,10,8,2,1,2,7,5}. Thus the first series was composed of patterns, whereas the second series was purely random. Both series, however, were based on identical probability distributions (Fig. 2a,b, respectively). Our analysis shows that steps (peaks for the difference quotient plot) emerge only if patterns are present. Thus our method is able to distinguish series with patterns from series without.

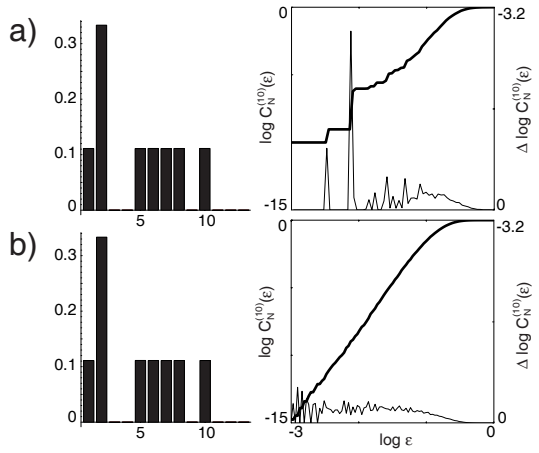

Fig. 2. Comparison between a series composed of patterns (a) and a series composed from random selection of intervals (b), based on identical ISI distributions. The comparison shows that log-log plot steps (y-axis: $\log C^{(10)}(\epsilon)$, thick line) emerge only in the presence of patterns. A more sensitive indicator are the peaks of the difference quotient plot (y-axis: $\Delta \log C^{(10)}(\epsilon)$, thin line), respectively ($m = 10$, Euclidean norm).

4. PATTERN LENGTH ESTIMATION

Once the presence of patterns is established, an estimate of the pattern length, defined as the number of ISIs involved, is desirable. That this is achievable is motivated by the following argument. For the calculation of the distance between two points, the differences between corresponding coordinates need to be calculated. Using the maximum norm, the distance between the points is defined as the largest difference between corresponding coordinates. As an increase of the embedding dimension yields ever more pairs of coordinates, the presence of a particularly large difference (which has an increased probability for being present in higher dimensions) will dominate. As a consequence, the number of steps calculated for data with pattern length n will decrease with increased embedding dimension m.

For ideal toy systems, the maximal number of occurring steps $s(m, n)$ can be computed numerically: We start from a series generated by a repetition of a sequence of length n. Additionally, we require that the elements $\{x_1, \ldots, x_n\}$ yield distinct coordinate differences $|x_i - x_j|$. After choosing an embedding dimension m, n distinct embedded points are generated. On this set of points, the maximum norm induces classes of equal point-point distances. The number of these classes equals $s(m, n)$. The actual calculation of $s(n, m)$ proceeds via a computer program (that exhausts the capabilities of an ordinary computer very quickly) or an analytical calculation, which however is unexpectedly involved. The obtained closed expression for $s(m, n)$ is beyond the scope of this paper, even in the simple case considered [8].

For the series generated from the sequence $\{5, 24, 37, 44, 59\}$, our correlation integral approach was able to reproduce the predicted decrease of $s(n, m)$ (Fig. 3a): In embedding dimension $m = 1$, all ten possible nonzero differences are reflected. As m increases towards 5, the number of steps decreases in accordance with the analytical result [8], remaining constant for $m > 5$.

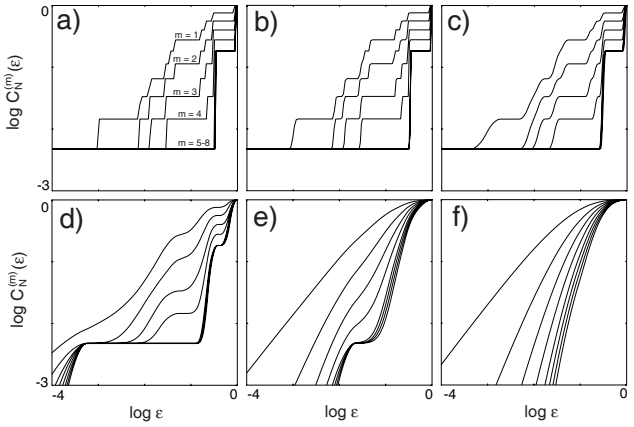

Fig. 3. a): Decrease of the number of steps as a function of the embedding dimension ($m = 1, \ldots, 8$, sequence length: 5). b-f): Influence of additive noise to the data (noise level: 8%, 32%, 128%, 512% and 1024% of the smallest interval of the sequence, uniform distributed). The number of steps for $m = 1, \ldots, 4$ decreases for increased noise, whereas the clearest step always emerges for $n = 5$, even for a noise level of 512%, indicating a sequence of length 5.

This behavior, however, only holds if the series are created by repeating a single sequence with distinct inter-coordinate differences. The exact determination of the pattern length in the more general cases is hampered by an obvious basic difficulty: If, for the simplest example, one step emerges in the log-log plot, this can either be the consequence of one pattern consisting of two consecutive ISIs, or two "patterns" of one ISI each. A greater number of steps, as is obtained in the presence of a multitude of patterns, further complicates this problem. Moreover, in the presence of noise, the identification of steps becomes less reliable.

Fortunately, a helpful indicator for the pattern length can be obtained from the following reasoning. A pattern will emerge in the embedded ISI series in its most genuine form (neither cut into pieces, nor spoilt by points that do not belong to the pattern), if the pattern length equals the chosen embedding dimension ($m = n$). In fact, in Fig. 3a, the most pronounced step appears at $m = 5$, correctly indicating a pattern length of $n = 5$.

To investigate the influence of noise, we applied additive noise of increased magnitude (Fig. 3b-f) to the series

generated from the sequence {5, 24, 37, 44, 59} (the magnitude of noise is measured as a percentage of the smallest interval). The results demonstrate that the estimation of the pattern length is possible for a noise level of 512% (Fig. 3e), where the most pronounced step still appears at $m = 5$. The number of steps for $m < 5$, however, is much more sensitive to noise: For $m = 1$, for example, 9 steps are present for 8% noise (Fig. 3b), 7 steps for 32% noise (Fig. 3c) and 3 steps for 128% noise (Fig. 3d). As expected, the step-structure disappears, when the noise level reaches the same magnitude as the longest ISI of the sequence (Fig. 3f). Consequently, the observation that a pattern emerges in the embedded ISI series in its most genuine form for $m = n$, provides a criterion for estimating the pattern length.

To investigate its potential in less idealistic settings, we performed a number of experiments. First, we include the sequences $\{5, 25, 10, 2\}$ and $\{5, 25, 10, 2, 17, 33\}$, respectively, with probability $p = 0.06$ into a noisy background. The background was provided by a homogenous Poisson spike generator with refractory period. Additionally, the Poisson distribution is chosen so as to produce a mean interspike interval identical with the one generated by the patterns alone. Consistent with our expectations, the clearest steps emerge at the embedding dimensions 4 and 6 (Fig. 4a,b), which implies that the pattern length can be estimated even in this case.

We refined this investigation by varying the individual pattern inclusion probabilities (using the sequences $\{4, 17, 12\}$ and $\{5, 25, 10, 2\}$). For the generation of the first series, the first pattern was chosen with $p = 0.12$ and the second with $p = 0.04$. For the generation of the second series, the probabilities were exchanged. The obtained results imply that even in this setting, the clearest steps emerge for m being equal to the pattern length n, but the influence of a particular pattern is weighted by its probability of occurrence (Fig. 4c). If the two probabilities are close, the extraction of the pattern lengths is still possible, but may be hampered by effects of interference among the patterns. A possible means of quantifying the "degree of clearness" of a step is by calculating the ratio between the slopes of the flat and of the steep part of the steps in each embedding dimension. The embedding dimension for which this ratio reaches a minimum indicates the pattern length.

5. DISCUSSION

As shown, our method allows unbiased testing for pattern occurrence. Furthermore, the presence of patterns can be detected against a random environment. Although the method does not deliver the patterns themselves, robust indicators for the lengths of patterns are provided. As soon as the existence of patterns is established, pattern size estimation combined with the locations of the steps can be used to substantially minimize the set of possible trial templates when working with template-based methods. This pattern detection method gains importance in robotics and signal processing, where biological signals are taken as basis for reverse engineering approaches.

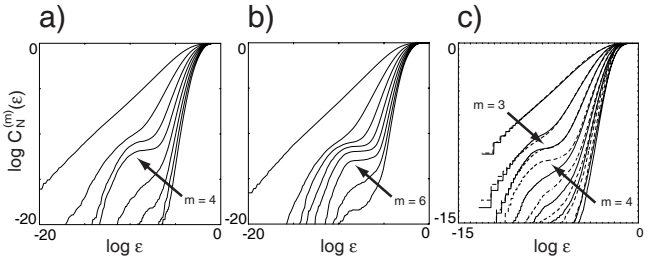

Fig. 4. Log-log plots for varying embedding dimensions ($m = 1, \ldots, 8$) indicate the pattern size. a) Sequence of length 4 immersed in a homogenous Poisson spike train: Most pronounced step for $m = 4$. b) Sequence of length 6: Most pronounced step for $m = 6$. c) Sequences of length 3 and 4 immersed at different ratios (solid line: ratio 3:1, dashed line: ratio 1:3): Most pronounced step is where m equals the length of the dominating sequence.

6. REFERENCES

[1] C. Koch, *Biophysics of Computation*, Oxford University Press, Oxford, 1999.

[2] M. A. Arbib, Ed., *The Handbook of Brain Theory and Neural Networks*, MIT Press, Cambridge, 1998.

[3] F. Rieke, D. Warland, de Ruyter van Steveninck, and W. Bialek, *Spikes. Exploring the Neural Code*, MIT Press, Cambridge, 1999.

[4] I.V. Tetko and A.E.P. Villa, "A pattern-grouping algorithm for analysis of spatiotemporal patterns in neuronal spike trains. 1. Detection of repeated patterns," *J. Neurosci. Methods*, vol. 105, pp. 1–14, 2001.

[5] P. Grassberger and I. Procaccia, "Dimensions and entropies of strange attractors from a fluctuating dynamics approach," *Physica D*, vol. 13, pp. 34–54, 1984.

[6] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge University Press, Cambridge, 2000.

[7] J. Peinke, J. Parisi, O.E. Rössler, and R. Stoop, *Encounter with Chaos*, Springer, Berlin, 1992.

[8] A. Nikitchenko, M. Christen, and R. Stoop, *In preparation*.