2008 International Symposium on Nonlinear Theory and its Applications
NOLTA'08, Budapest, Hungary, September 7-10, 2008

Detection of regular patterns within randomness

Ruedi Stoop and Markus Christen†

†Institute of Neuroinfomatics, University / ETH Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
email: {ruedi}@ini.phys.ethz.ch

Abstract—The identification of slightly jittered regular signals (="patterns") embedded in strongly noisy background is a nontrivial important task, particularly in the neurosciences. Whereas traditional methods generally fail to capture such signals, staircase-like structures in the log-log correlation plot are reliable indicators. We provide the analytic relationship between the length of the pattern n and the maximal number of steps $s(n, m)$ that are observable at a chosen embedding dimension m. For integer linearly independent patterns, the length of the embedded pattern can be calculated from the number of steps. We, moreover, discuss several applications of this concept that demonstrate the power of this concept.

1. Introduction

It has been speculated that spike trains from nervous systems, in particular the human brain, fall into this class of mixed regular-noisy signals [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. As a main generation principle, they are believed to originate from neurons that are mostly driven by complex processes, but occasionally get recruited by more locally defined, simpler circuits of regular firing. Similar signal characteristics may result from neuronal multi-electrode recordings, where signals from different (randomly, regularly, or in a mixed mode firing) sources arrive at an electrode. After spike sorting by which the events of relevance on the time axis are defined, we may be left with regular signal components embedded in a noisy background.

Both parts of such signals could play equally important roles in cortical signal processing and computation [11]. Here, however, we will be primarily interested in the regular components and how they can reliably be extracted from the data. Past approaches dealing with this task contained several critical tuning parameters, expressing expectations of how the pattern to be searched for should look like, rendering it difficult to assess the validity of the obtained results. This may be one of the reasons why mostly patterns of relatively short length (1-5) have been identified [4, 7], although it may rightfully be argued that information-bearing signals in neuroscience cannot be too long, as actions on relatively short time scales are generally required. Here, we discuss application and proof of a method for a fast and unbiased detection of patterns in noisy contexts that does not share these shortcomings. The method is based upon the observation that in the presence of patterns, in the correlation integral plots used for the evaluation of fractal dimensions [12, 13, 14, 15], step-like structures emerge. The method works with very modest data size of the kind obtainable in most experimental contexts in neuroscience, and has, in principle, no pattern length limitation. In the presence of a noisy signal component or jitter, the traditional Fourier method, e.g., quickly fails, whereas the characteristic decrease of the number of steps with the embedding dimension is conserved, even for strong noise components. As the only variance with the noise-free case, the most prominent step reappears in a weakened form at multiples of the pattern length, a phenomenon that is simple to understand along the proof of the main theorem given below. The comparison between our method and the traditional Fourier approach provided in Fig. 2 illustrates these facts.

Consider an arbitrary scalar time series of measurements $\{x_i\}$, $i = 1..L$. From this data, embedded points $\xi_k^{(m)}$ are constructed as

$$\xi_k^{(m)} = \{x_k, x_{k+1}, \ldots, x_{k+(m-1)}\}, \qquad (1)$$

where m is called the embedding dimension [16]. This *coordinate-delay construction* is standard in nonlinear dynamics [13, 14]. Its purpose is to reconstruct the complete underlying (in general: high-dimensional) dynamics from partial, usually scalar, measurements. The reconstruction of the phase space is generically successful if a sufficiently large data set a sufficiently high embedding dimension is chosen [13, 14]. Using the embedded points, the *correlation integral* [12, 13, 14, 15] is calculated as

$$C_N^{(m)}(\varepsilon) = \frac{1}{N(N-1)} \sum_{i \neq j} \theta(\varepsilon - \|\xi_i^{(m)} - \xi_j^{(m)}\|), \qquad (2)$$

where $\theta(x)$ is the Heaviside function ($\theta(x) = 0$ for $x \leq 0$ and $\theta(x) = 1$ for $x > 0$) and N is the number of embedded points ($N \leq L - m + 1$). The correlation integral $C_N^{(m)}(\varepsilon)$ averages the probability of measuring a distance smaller than ε between two randomly chosen points $\xi_i^{(m)}$ and $\xi_j^{(m)}$. In practical applications, $\log C_N^{(m)}(\varepsilon)$ is plotted against $\log \varepsilon$ (the so-called *log-log plot*). The correlation dimension $d_C^{(m)}$ is defined as the limit $d_C^{(m)} = \lim_{\varepsilon \to 0} \frac{\log C_N^{(m)}(\varepsilon)}{\log \varepsilon}$ [12, 13, 14]. If an embedding dimension $m > 2d_C^{(m)}$ is chosen, the slope of $\log C_N^{(m)}(\varepsilon)$ versus $\log \varepsilon$ for small ε provides a good estimate of the correlation dimension. For the evaluation of the distances, any norm could be used. Instead of the 'natural' Euclidean norm, often the maximum norm is used,

in order to simplify numerical computations and theoretical arguments. Degeneracies introduced by this choice are removed upon the addition of a small amount of jitter.

We first demonstrate how the presence of patterns leads to a step-like behavior of the log-log correlation dimension plots. Patterns manifest themselves as a clustering of the embedded data. For the calculation of $C_N^{(m)}(\varepsilon)$, an embedded point $\xi_0^{(m)}$ is chosen at random. As the radius ε of its neighborhood $U(\xi_0^{(m)}, \varepsilon)$ is enlarged, we keep track of the number of points that fall into this neighborhood. If a point newly entering the neighborhood belongs to a cluster, upon a small enlargement of ε, many points will join. I.e., the number of points $C_N^{(m)}(\varepsilon)$ quickly increases with ε. Once the cluster size is reached, fewer points are recruited, and $C_N^{(m)}(\varepsilon)$ increases but slowly. In this way, step-like structures emerge. The denser the clustering regions, the more prominent the step-like structure. To demonstrate this effect, artificial (noise-free) time series were constructed from a repetition of a sequence of length n. The series were then embedded (embedding dimension m) and the correlation integrals were evaluated. The results shown in Fig. 1 demonstrate a clean emergence of stairs, the number of which increases with the length of the embedded pattern n, and decreases with the embedding dimension m. In the presence of patterns, the step-like behavior emerges stably even for a few hundred scalar measurements. If in an experiment single trials generate less data than needed (say, in neuroscience, because of adaptation), data from several trials under identical conditions can be concatenated. Although in this case the embedded data will contain some points that violate the continuous dependence on time, this has normally no statistical influence.

2. Analytical approach

After having shown that the presence of patterns is reflected in the emergence of log-log correlation integral steps, our next goal is an estimate of the pattern length from the number of steps. That this might be achievable is motivated by the following argument. Using the maximum norm, the distance between two points in the embedding space is defined as the maximum of the component differences. As the dimensionality of the embedding space is increased, ever more of the possible differences will be present. A few large differences will, however, prevent the smaller ones from winning the competition for the maximum. As a consequence, the number of steps $s(n,m)$ obtained for a pattern of length n can be expected to decrease with increased embedding dimension m. That this indeed is the case is demonstrated in Fig. 1b.

The precise way how this decay proceeds depends on the pattern length n. For toy systems, the maximal number of occurring steps $s(n,m)$ can be computed numerically as follows. A time series generated by repeating a sequence of length n composed of elements $\{x_1, \ldots, x_n\}$, generates distinct coordinate differences $d_{ij} := |x_i - x_j|$. By shifting a window of length m along the time series, we repeatedly generate embedded points of embedding dimension m. On the set of the generated points, the maximum norm induces classes of equal distances, the number of which equals $s(n,m)$. Unfortunately, this numerical calculation quickly exhausts computing time, calling for an analytical way to compute $s(n,m)$. The values of $s(n,m)$ that can be corroborated with the help of a desktop computer are shown in Table I. In Fig. 1 we demonstrate how for the toy system generated from the sequence $\{5, 24, 37, 44, 59\}$, the correlation integral method is able to reproduce the decrease of $s(n,m)$ predicted by Table I: In embedding dimension $m = 1$, all ten possible differences are detected. As m increases towards 5, the number of steps decreases in accordance with Table I, before remaining constant for $m > 5$. The basis of the relationship between number of steps s, pattern length n and embedding dimensions m is provided by the following Proposition.

Proposition The number of correlation-dimension log-log steps $s(n,m)$ generated from an embedded repeated pattern, equals the number of distinct distances among the embedded points using the maximum norm.
Proof: For the correlation integral, all $\frac{n(n-1)}{2}$ distances between points are calculated, where classes of equal distances $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_\kappa\}$ are generated. Around a point $\xi_0^{(m)}$ in the embedding space, $C_N^{(m)}(\varepsilon, x_k)$ changes whenever $\varepsilon \in \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_\kappa\}$. As this is true for any point, also for the averaged correlation integral $C_N^{(m)}(\varepsilon, \xi_0^{(m)})$ the number of steps is $s(n,m) = \kappa$. This proves the proposition. □

For the analytical derivation of $s(n,m)$ we start from a time series $\{x_i\}_{i=1\ldots N}$ generated by the repetition of a pattern of length n. The pattern is supposed to be general in the sense that the elements $\{x_1, \ldots, x_n\}$ yield $\frac{n}{n-1}$ distinct coordinate differences $d_{ij} = |x_i - x_j|$ (="integer linear independence")). Embedded points are generated by the shift of a window of length m along the data series $\{x_i\}_{i=1..N}$. As was previously pointed out, on the set of the embedded points the maximum norm induces classes of equal distances, the number of which equals $s(n,m)$. The goal of the rest of this paper is to analytically compute $s(n,m)$.

We start by calculating the number of different distance vectors. Since different distance vectors will not necessarily imply different distances between points, this is but a preliminary task that can be achieved without specifying the metric. Using this information, we will calculate the number of different distances, specifying the maximum norm as the relevant one.

We will focus on the case $m \leq n$ (case $m > n$ is trivial). As an example of $n = 4, m = 3$, we start with the repeated sequence $\{x_1, x_2, x_3, x_4\}$, which generates the time series

$$\{x_1, x_2, x_3, x_4, x_1, x_2, x_3, x_4, x_1, x_2, x_3, x_4, x_1, x_2, ..\}.$$

By the embedding process in $m = 3$, we obtain the set of

embedded points

$$\{\{x_1, x_2, x_3\}, \{x_2, x_3, x_4\}, \{x_3, x_4, x_1\}, \{x_4, x_1, x_2\}\}.$$

The distance vector components d_{ij} between the embedded points then become

$$\begin{aligned} |\{x_1, x_2, x_3\} - \{x_2, x_3, x_4\}| &= |\{x_1 - x_2, x_2 - x_3, x_3 - x_4\}|, \quad (3)\\ &=: \{d_{12}, d_{23}, d_{34}\},\\ |\{x_2, x_3, x_4\} - \{x_3, x_4, x_1\}| &= |\{x_2 - x_3, x_3 - x_4, x_4 - x_1\}|,\\ &=: \{d_{23}, d_{34}, d_{41}\},\\ |\{x_3, x_4, x_1\} - \{x_4, x_1, x_2\}| &= |\{x_3 - x_4, x_4 - x_2, x_1 - x_2\}|,\\ &=: \{d_{34}, d_{41}, d_{12}\}, \end{aligned}$$

where $|\,.\,|$ indicates the componentwise 'absolute value' operation. The emerging distance vectors can be collected in the form of a 2-torus:

$$D_{(n)} = \left|\begin{array}{cccccc} d_{12} & d_{13} & \cdots & . & \cdots & d_{1n} \\ d_{23} & d_{24} & \cdots & . & d_{2n} & d_{21} \\ \vdots & \cdots & \cdots & . & \cdots & \vdots \\ d_{(n-1)n} & d_{(n-1)1} & \cdots & . & \cdots & d_{(n-1)(n-2)} \\ d_{n1} & d_{n2} & \cdots & . & d_{n(n-2)} & d_{n(n-1)} \\ \hline d_{12} & d_{13} & \cdots & . & \cdots & d_{1n} \\ d_{23} & d_{24} & \cdots & . & d_{2n} & d_{21} \\ \vdots & \cdots & \cdots & . & \cdots & \vdots \\ d_{(n-1)n} & d_{(n-1)1} & d_{(n-1)2} & . & d_{(n-1)(n-3)} & d_{(n-1)(n-2)} \end{array}\right|.$$

In this distance matrix D, distance vectors of the m-dimensional embedding space are represented by subcolumns of dimension m. Because of the two-torus nature of D, by starting the vectors at arbitrary positions, we observe multiple repeats of the vectors (even though we have requested that $d_{ij} \neq d_{kl}$, unless $_{ij} =_{kl}$ or $_{kl} =_{ji}$). By carefully considering the different origins of such repetitions and by describing the number of remaining distances in terms of combinatorics, we arrive at the following result: The maximal number of steps $s(n, m)$ emerging in a log-log plot of a time series from a repeated pattern of length n in embedding space dimension m has the expression

$$n \quad \text{even:} \quad s(n,m) = \begin{cases} \frac{n(n-m)}{2} & : \quad 1 \le m \le \frac{n}{2} \\ \frac{n(n-m)+2m-n}{2} & : \quad \frac{n}{2} < m \le n \\ \frac{n}{2} & : \quad m > n, \end{cases} \quad (4)$$

$$n \quad \text{odd:} \quad s(n,m) = \begin{cases} \frac{n(n-m)+m-1}{2} & : \quad 1 \le m \le n \\ \frac{n-1}{2} & : \quad m > n \end{cases} \quad (5)$$

Using this analytic expression, we obtain the following table of the maximal number of of observable steps $s(n, m)$:

m\n	1	2	3	4	5	6	7	8	9	10
1	0	1	3	6	10	15	21	28	36	45
2	0	1	2	4	8	12	18	24	32	40
3	0	1	1	3	6	9	15	20	28	35
4	0	1	1	2	4	7	12	16	24	30
5	0	1	1	2	2	5	9	13	20	25
6	0	1	1	2	2	3	6	10	16	21
7	0	1	1	2	2	3	3	7	12	17
8	0	1	1	2	2	3	3	4	8	13
9	0	1	1	2	2	3	3	4	4	9
10	0	1	1	2	2	3	3	4	4	5

TABLE I: $s(n, m)$ for $n, m = 1, \ldots, 10$. In experimental applications, one might expect more than one pattern to be present in a time series. This leads to complications in the application of $s(n, m)$. If, for the simplest case, one single step emerges in the log-log plot, this could either be due to one pattern composed of length two, or to two "patterns" of length one each. A greater number of steps, as obtained from a multitude of patterns, will further complicate this problem. Once the presence of patterns is indicated and precise alternatives for the patterns are posed by the method, the existence/non-existence of a particular alternative, can be corroborated by direct methods that under such conditions are justified. The number of steps predicted by $s(n, m)$ is reliable up to very strong noise components (this will be demonstrated in Fig. 2), or up to the point where the jitter of the pattern conflicts with its nature. Even in very difficult conditions, the method is able to indicate the presence of patterns, where $s(n, m)$ can still serve as a guideline for further processing (see the final discussion).

3. Applications

For applications, the stability of the method with respect to jitter on the regular pattern is of importance. Jitter on the repeated patterns modifies the density of the point clusters in the embedding space and, therefore, the distribution of the distances. In the log-log plot this primarily leads to a smearing of the steps. A pattern, however, will always emerge in the embedded time series in its most genuine form (it is neither cut into pieces, nor spoiled by foreign points) if the embedding dimension equals the pattern length ($n = m$). In the absence of a noise component, the decrease of steps fully stops at $n = m$. In the presence of a noise component, the steps repeat at multiples of n, in a softened fashion. As a consequence, the most prominent step provides a reliable indicator of the pattern length.

That this is indeed the case is demonstrated by two characteristic examples. In the first one, to the series generated from the sequence $\{5, 24, 37, 44, 59\}$ (c.f. Table I), jitter was added, where the jitter strength is defined as the ratio of the interval size from which we uniformly sample the jitter, over the shortest pattern interval. The results (Fig. 1b-f) demonstrate that the pattern length can be reliably esti-

mated up to a jitter of 512% (Fig. 1e), where the most pronounced step still appears at $m = 5$. The number of steps for $m < 5$, however, are affected by the jitter: For $m = 1$, for example, 9 steps are identifiable at 8% jitter (Fig. 1b), 7 steps at 32% (Fig. 1c) and 3 steps at 128% (Fig. 1d). The step-like structure disappears if the jitter reaches the size of the largest element of the pattern (Fig. 1f). Thus, the criterion that the most pronounced step appears at $m = n$, still yields a valuable indicator for the pattern length at very strong jitter. The method is able to detect the presence of

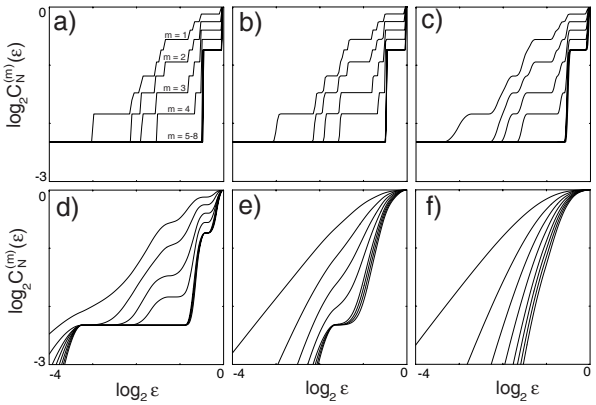

Figure 1: Decrease of the number of steps with increasing embedding dimension, with jitter on the regular signal component ($n = 5$, $m = 1, \ldots, 8$). For $m = 1$: 10 steps, in agreement with Table I. Panels b)-f): Jitter levels 8%, 32%, 128%, 512% and 1024%. Overall, for increased jitter, the number of identifiable steps decreases. The clearest step, however, always emerges for $n = 5$, indicating a sequence of length 5.

patterns with ease in cases where the Fourier / Power spectrum method fails, see Fig. 2.

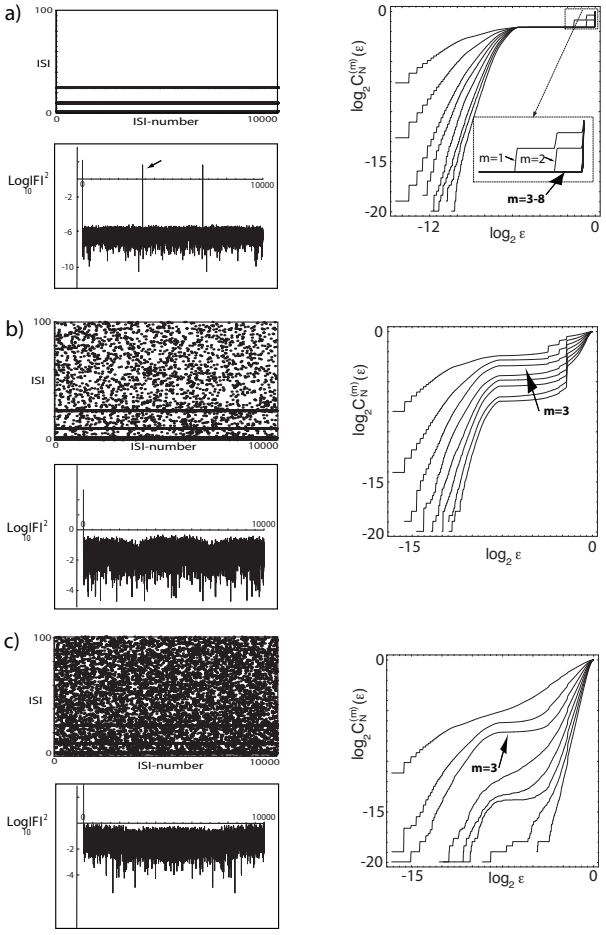

Figure 2: Comparison with power spectrum: Pattern $\{2, 25, 10\}$, jitter ±1 % of the smallest interval. a) Pattern only: Both methods predict periodicity; our method correctly predicts pattern length 3. b)-c) Time series composed of whole patterns and random intervals uniformly distributed in $[0, 100]$, so that 25% (75%, respectively) of all intervals are random: Power spectrum fails, whereas the log-log plot shows the characteristic decrease of steps with on large step remaining at dimension 3 (arrows).

References

[1] B. L. Strehler, Perspect. Biol. Med. **12**, 584 (1969).

[2] J. E. Dayhoff, G. L. Gerstein, *J. Neurophysiol.* **49**, 1334 (1983).

[3] J. E. Dayhoff and G. L. Gerstein, *J. Neurophysiol.* **49**, 1349 (1983).

[4] R. Lestienne, B. L. Strehler, *Brain Res.* **437**, 214 (1987).

[5] M. Abeles and G. L. Gerstein, J. Neurophysiol. **60**, 909 (1988).

[6] R. Lestienne and H. C. Tuckwell, Neuroscience **82**, 315 (1998).

[7] Y. Prut, E. Vaadia, H. Bergman, I. Haalman, H. Slovin, and M. Abeles, J. Neurophysiol. **79**, 2857 (1998).

[8] R. Stoop, K. Schindler, and L. A. Bunimovich, Acta Biotheor. **48**, 149 (2000).

[9] I. V. Tetko and A. E. P. Villa, J. Neurosci. Methods **105**, 1 and 15 (2001).

[10] R. Stoop, D. A. Blank, A. Kern, J.-J. van der Vyver, M. Christen, S. Lecchini, and C. Wagner, Cog. Brain Res. **13**, 293 (2002).

[11] R. Stoop, K. Schindler, and L. Bunimovich, Biol. Cybern. **83**, 481 (2000).

[12] P. Grassberger and I. Procaccia, Physica D **13**, 34 (1984).

[13] J. Peinke, J. Parisi, O. E. Roessler and R. Stoop, *Encounter with Chaos* (Springer, Berlin, 1992).

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

[15] A. Kern, W.-H. Steeb, and R. Stoop, Z. Naturforsch. **54**a, 404 (1999).

[16] F. Takens, in: *Dynamical Systems and Turbulence*, Lecture Notes in Mathematics 898, eds. D. A. Rand and L. S. Young (Springer, Berlin, 1981), pp. 366-381.