2011 International Symposium on Nonlinear Theory and its Applications
NOLTA2011, Kobe, Japan, September 4-7, 2011

Log-Log-Step Method for Detection of Patterns Within Randomness Quantitatively Assessed

Florian Gomez and Ruedi Stoop

Institute of Neuroinformatics, University of Zurich and ETH Zurich
Winterthurerstr. 190, 8057 Zürich
Email: fgomez@ini.phys.ethz.ch, ruedi@ini.phys.ethz.ch

Abstract—It has been demonstrated earlier by us that staircase-like structures in the log-log correlation plot of a time series provide an indicator of underlying patterns, even in conditions of strong noisy or/and jittered data. In this article, we analyze the method under different jitter-noise-configurations and define quantitative measures for the method's applicability. A phase diagram shows the remarkable potential of this method even under very unfavorable conditions of noise and jitter. Moreover, we provide a novel and much more compact analytical derivation of the upper and lower bounds on the number of steps observable in the ideal noiseless case, as a function of pattern length and embedding dimension.

1. Introduction

The detection of patterns against a noisy signal background is an important engineering and neuroscience task. Under these conditions the traditional approaches like Fourier analysis or template matching either quickly break down or are too ambiguous to be helpful from first principles. We provide here an auxiliary tool based on a global quantitative characterization of the data that can guide the search for patterns. The algorithm provides essential information on the structure of putative patterns enclosed in time series data in terms of pattern length and the distances among the events involved in the pattern. Although earlier examples demonstrating the power of the approach have been provided and discussed [2, 3], so far no quantitative overview on the efficacy of the tool has been given. In the present contribution, we provide such a quantification.

We start our presentation by repeating the fundaments of our pattern-detection algorithm. Given a time series $\{a_1, a_2, ...\}$ embedded in m-dimensional space using the standard *coordinate-delay construction* $x_k^{(m)} = (a_k, a_{k+1}, ..., a_{k+n-1})$, in the log-log plot of the correlation integral $C_N^{(m)}(\epsilon)$

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

instead of a straight line needed for the evaluation of the fractal dimension and correlation, steps may emerge These steps emerge if the embedded points follow a simple generating pattern. Simple generating patterns lead to clusters of points in the embedding space that lead to a sudden increase in the log-log plot of the point densities. This can be seen from choosing a random reference data point, around which we enlarge the neighborhood radius ϵ and count the points falling into this neighborhood. After reaching a cluster of points, the count $C(\epsilon)$ quickly increases with ϵ, leading to a step-like structure in the plot of $C(\epsilon)$.

Given a time series generated from a noise-free pattern of length n and using the maximum norm, these steps are sharp, and the number of steps visible decreases with m. We derive upper and lower bounds for the maximal/minimal number of steps appearing under ideal conditions. From the investigation of the way how these steps propagate through the different embedding dimensions, we are able to derive upper and lower bounds to the observable number of steps as follows:
For n odd, the lower bound t / the maximal number s of steps have the expression

$$t(n,m) = (n-1)/2 \cdot \lceil \frac{n}{m} \rceil, \qquad (2)$$

$$s(n,m) = \frac{(n-1)}{2} \cdot (n - (m-1)). \qquad (3)$$

For n even, the lower bound t / the maximal number of steps s have the expression

$$\begin{aligned} t(n,m) &= (\frac{n}{2} - 1) \cdot \lceil n/m \rceil + \lceil \frac{n}{2m} \rceil, \text{if } m \leq n/2, \\ &= (\frac{n}{2} - 1) \cdot \lceil n/m \rceil + 1, \text{ if } m > n/2, \end{aligned} \qquad (4)$$

and

$$\begin{aligned} s(n,m) &= (\tfrac{n}{2} - 1) \cdot (n - (m-1)) + \tfrac{n}{2} - (m-1), \text{if } m \leq \tfrac{n}{2}, \\ &= (\tfrac{n}{2} - 1) \cdot (n - (m-1)) + 1, \text{ if } m > n/2. \end{aligned} \qquad (5)$$

These results extend the original results provided in Ref. [3]. The new results are based on much simpler graphical combinatorial arguments than those given in the original proof. From the steps, we can not only detect situations that are likely to contain pattern regularities within the data, with the help of the table (see Fig. 1), we can can also infer the length of these putative patterns. The method, of course, depends crucially on the ability to extract the correct number of steps from the log-log steps.

t(n,m) / s(n,m)

m/n	1	2	3	4	5	6	7	8	9	10
1	0 / 0	1 / 1	3 / 3	6 / 6	10 / 10	15 / 15	21 / 21	28 / 28	36 / 36	45 / 45
2	0 / 0	1 / 1	2 / 2	3 / 4	6 / 8	8 / 12	12 / 18	14 / 24	20 / 32	23 / 40
3	0 / 0	1 / 1	1 / 1	3 / 3	4 / 6	5 / 9	9 / 15	11 / 20	12 / 28	18 / 35
4	0 / 0	1 / 1	1 / 1	2 / 2	4 / 4	5 / 7	6 / 12	7 / 16	12 / 24	14 / 30
5	0 / 0	1 / 1	1 / 1	2 / 2	2 / 2	5 / 5	6 / 9	7 / 13	8 / 20	9 / 25
6	0 / 0	1 / 1	1 / 1	2 / 2	2 / 2	3 / 3	6 / 6	7 / 10	8 / 16	9 / 21
7	0 / 0	1 / 1	1 / 1	2 / 2	2 / 2	3 / 3	3 / 3	7 / 7	8 / 12	9 / 17
8	0 / 0	1 / 1	1 / 1	2 / 2	2 / 2	3 / 3	3 / 3	4 / 4	8 / 8	9 / 13
9	0 / 0	1 / 1	1 / 1	2 / 2	2 / 2	3 / 3	3 / 3	4 / 4	4 / 4	9 / 9
10	0 / 0	1 / 1	1 / 1	2 / 2	2 / 2	3 / 3	3 / 3	4 / 4	4 / 4	5 / 5

Figure 1: Lower bound t / maximally observable steps s, as a function of pattern length n and embedding dimension m.

2. Method validation

In realistic time series, the regular signal will be contaminated by jitter and noise. Jitter is implemented by addition of signed noise (say, e.g., 10 percent of the smallest interspike interval) to a regular signal. In this way, a period-three signal $\{3200, 7700, 1000\}$ changes into a time series such as $\{3223, 7703, 907, 3203, 7782, 903, 3107, 7603, 1098, \ldots\}$, with some dependence of course on the probability distribution (uniform, Gaussian, e.g.) the noise is drawn from. Noise is implemented by choosing a given percentage of the ISIs according to some random probability distribution. This can be achieved in two different manners that reflect different ways of how the regularity-generating network is linked to the noise-generating part of the network: a) We can choose the next signal event with probability p from the regular pattern and with probability $(1-p)$ from the random distribution. b) Alternatively, with probability p' the whole regular pattern of length n provides the n next signals, whereas with probability $1-p'$ the signal event is drawn from the random distribution to the time series (for a fair comparison among the different paradigms, the probabilities of course must be rescaled as $p' = p/(n-(n-1)p)$). In the presence of jitter and noise, the steps may smear out and finally may no longer be visible. An example of a log-log plot displaying a step-like behavior is shown in Fig. 2. The following approach has also been performed for patterns of length 5 and partially for length 7, yielding compatible results.

In the log-log plot, jitter smoothens out the steps, whereas noise decreases the height of the steps as well as the slope of the stairs. The ability of our method to indicate regular patterns of length n within data contaminated by jitter and noise can be assessed with the help of three criteria: a) How well exactly n steps in embedding dimension $m = 1$ can be detected; b) how well visible a flat plateau at embedding dimension $m = n$ is if compared to that observed at $m = n + 1$; c) how well the predicted decrease of the number of steps with the embedding dimension m can be evidenced.

From these criteria, we derive an overall goodness-of-method measure by adding the measures obtained from the different criteria. To assess the first criterion, we value three height levels that in the correct spaces the derivative of the log-log-step has to overcome. To assess the second criterion, the plateaus at m and $m+1$ were evaluated. A plateau was counted, if the difference quotient of slope was below values $\{0.2, 0.4, 0.7, 1.1\}$. The number of counts were averaged for the four thresholds with weights $\{1/8, 1/4, 1/4, 1/8\}$; the average counts obtained for $m+1$ were then subtracted from the average counts obtained for m. To assess the third criterion, we verified whether the characteristic decrease of the number of steps as a function of m was observed or not. In order to do so, we tested whether a single step was visible when $m = n$ using a difference-quotient-method as for the first criterion. Yet it is observed that inserting noise quickly leads to the reappearance of the theoretically vanishing steps in dimensions $m \geq 1$, rendering such a simple test unsatisfactory. To account for this problem, we also incorporated the detection of two or three steps when $n = 3$ and $m = n$. In such a case, the visibility for the appearing peaks was compared to the visibility of the peaks that result from the evaluation of the same, but randomly permuted timeseries. A randomly permuted timeseries always leads to the full amount of visible steps, independent of the embedding dimension m. Hence, when there is more than one step at $m = n$, the visibility measure of the normal case subtracted from the measure of the permuted case serves as indicator for the characteristic decrease with m.

All three measures were normalized to 1 and a contour-plot with suitable contours was drawn. The resulting Fig. 3 shows the results gained. We defined two or three regions of various visibility for each of the criteria. Not surprisingly, the visibility of exactly n peaks for $m = 1$ is best in

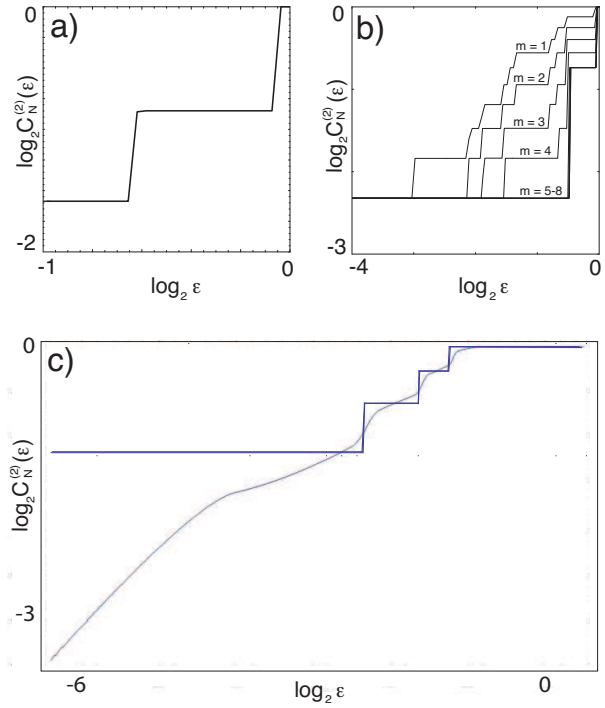

Figure 2: Log-log plots from a pattern of length 3: a) $m = 1$, b) for increasing embedding dimension m, c) modification introduced by the presence of 20 percent jitter and 30 percent noise (pattern $\{3200, 1700, 100\}$.)

the case of little noise and little jitter. Nevertheless, the visibility is considerably good for noise fractions up to 50 or 60 percent. It is natural, however, that results would be worse in the case of longer patterns or steps being more closely located. Clearly, the seven peaks of a length-7 pattern are somewhat less easily identified since with increasing jitter the peaks may overlap. Criteria (b) and (c) are what we consider to be the strongest indicators for the occurrence of patterns. The emergence of the "natural" situation $m = n$ - where patterns are completely inserted but no additional terms spoil the characteristic behavior - is most helpful in the case of little jitter but high noise values. In regions of up to 90 percent of noise, when all other methods normally fail, the plateau occurring at $m = n$ compared to $m = n + 1$ reliably indicates a pattern of length n. We tested criterion (b) for a generic pattern of length 5 comparing the dimension $m = 5$ and $m = 6$ using exactly the same algorithm. Even though there are theoretically two visible steps in this case, the two plateaus quickly merge into a single one. The resulting plot looks very similar with even a slightly extended range of visibility. We thus suppose criterion (b) to be fairly independent of the underlying pattern length. In regions where the criterion (b) fails, i.e. for little noise and high jitter, criterion (c) may serve as indicator of the pattern length. The visibility of one single step in dimension $m = n$ alone yet does not prove a pattern length n, since patterns of length $\leq n$ may also produce such a single step. Comparing to the embeddings $m \leq n$ where more steps should occur helps to exclude these cases. Moreover, high jitter values may merge two steps, if these steps are close together. The possible overlap of neighboring steps sets the natural limit to the method. Yet this happens only in the case of highly jittered signals or specific patterns having two distinct distances very close together. In the latter case, nonetheless still a pattern will be indicated, albeit of the wrong length. To summarize, we emphasize the remarkable performance of the method under very noisy conditions. As a general advice (generally true for time series analysis!) we propose not to rely on one single criterion, but to combine all aspects to obtain a coherent picture. This is best done by embedding the time series in dimensions $m = 1, 2, ..$ up to $m = 10$ for example and plotting the resulting log-log graphs in one single window. While step-like structures already indicate a possible pattern, the very robust criterion (b) might help to determine the pattern length. The slope of the lines in the step-free regions may additionally give interesting insights into the fractal dimension of a possible attractor.

3. Proof of the analytical formula for $s(n, m)$ and $t(n, m)$

For a proof of (2)-(5), we decompose the graph into subgraphs connecting nearest-, next-nearest-, etc., neighbors. The idea underlying the optimized proof with sharper bounds is, as in the old proof, that the choice the maximum norm makes, is restricted to consecutive d_{ij}'s on one distinct subgraph. For $m = 1$, every 'comparison' yields a winner, hence we have $n(n-1)/2$ steps. For larger m, the ordering of d_{ij} on the subgraphs is crucial. In $m = 2$, for $n = 6$, a monotonous ordering $d_{61} > d_{12} > d_{23} > d_{34} > d_{45} > d_{56}$ yields $n - 1$ steps; in $m = 3$, $n - 2$ steps, and so on. If we have a 'regular' distribution of biggest three distances $d_{ij} : d_{34} > d_{12} > d_{56} >$ rest, only 3 steps are contributed. For odd n, each subgraph follows the rules for the monotonous ordering of a maximal number, from where we get $n-(m-1)$ steps, and for a regular ordered set $\lceil n/m \rceil$ steps. From this, we arrive at $t(n, m) = (n-1)/2 \cdot \lceil n/m \rceil$ and $s(n, m) = (n-1)/2 \cdot (n-(m-1))$. For even n, $n/2-1$ subgraphs follow the same rules as above, except for the one with $n/2$ line, which only contributes one step if $m > n/2$.

References

[1] R. Stoop, D.A. Blank, J.-J. van der Vyver, M. Christen, A. Kern, Journal of Electrical & Electronics Engineering, Istanbul University, 3 (1), 693-698 (2003).

[2] M. Christen, A. Kern, A. Nikitchenko, W.-H. Steeb, and R. Stoop, Phys. Rev. E 70, 011901 (2004).

[3] R. Stoop and M. Christen: Detecting Patterns Within Randomness in *Nonlinear Dynamics and Chaos*, Springer Berlin Heidelberg (2010).

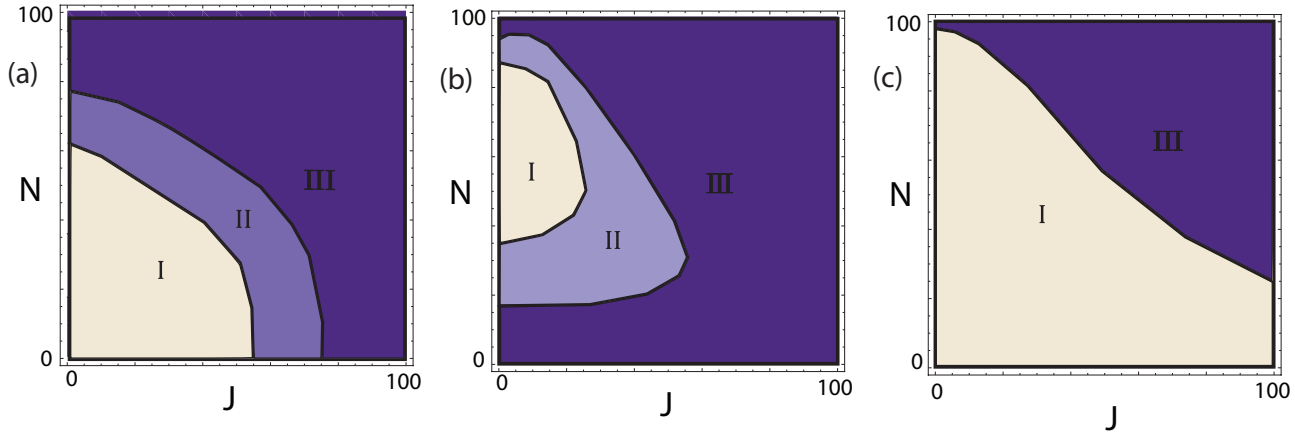

Figure 3: Approximate phase boundaries, for noise N and jitter J in units of percents of events in the data and in percents of the smallest interval in the pattern. a) Fulfillment of the criteria expressed by three degrees: Region I: excellent, region II: fair, region III: ambiguous. a) $m = 1$-criterion; b) difference in plateau visibility for $m = 3$ compared to $m = 4$. Regions I, II and III as in a). c) Measure for the decrease in steps with m (only two regions: I and III).

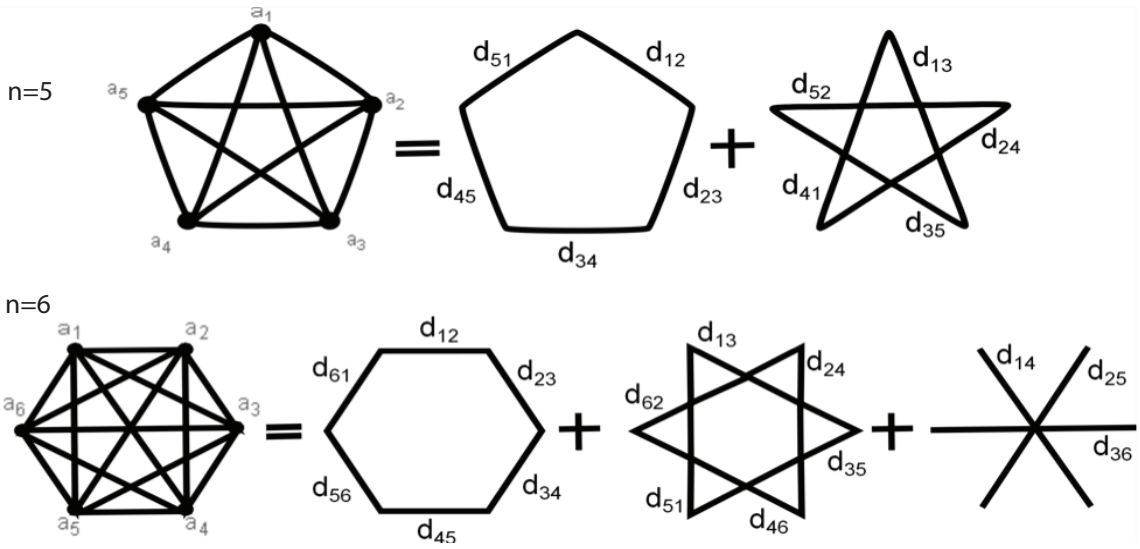

Figure 4: Decomposition of potential distances in the maximum norm for odd and for even pattern lengths n into nearest-, next-nearest-, etc., neighbor subgraphs. Each subgraph can be treated separately.