OPEN ACCESS Freely available online

PLoS one

Quantitative Assessment of the Log-Log-Step Method for Pattern Detection in Noise-Prone Environments

Florian Gomez, Ruedi Stoop*

Institute of Neuroinformatics, ETH Zurich and University of Zurich, Zurich, Switzerland

Abstract

Staircase-like structures in the log-log correlation plot of a time series indicate patterns against a noisy background, even under condition of strong jitter. We analyze the method for different jitter-noise-combinations, using quantitative criteria to measure the achievement by the method. 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 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. The quantitative measure developed combined with the ideal bounds can serve as guiding lines for determining potential periodicity in noisy data.

Citation: Gomez F, Stoop R (2011) Quantitative Assessment of the Log-Log-Step Method for Pattern Detection in Noise-Prone Environments. PLoS ONE 6(12): e28107. doi:10.1371/journal.pone.0028107

Editor: Stefano Boccaletti, Technical University of Madrid, Italy

Received October 12, 2011; **Accepted** November 1, 2011; **Published** December 12, 2011

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ruedi@ini.phys.ethz.ch

Introduction

The detection of patterns against a noisy signal background is a particularly important task for engineering and neuroscience [1–7]. Traditional approaches like Fourier analysis quickly break down under these conditions, or are far too ambiguous to be helpful from first principles (template-matching methods). Here, we assess the usefulness of an auxiliary tool. By providing information on the length and on metric aspects of putative patterns enclosed in a time series, the tool can guide the search for patterns. Although earlier [8,9] the power of this method has been exemplified, so far no quantitative overview on its efficacy could be provided. In the present contribution, we introduce such a quantification.

Given a time series $\{a_1,a_2,...\}$ embedded in m-dimensional space using the standard *coordinate-delay construction* $x_k^{(m)}=(a_k,a_{k+1},...,a_{k+m-1})$ [10–14], in the log-log plot of the correlation integral $C_N^{(m)}(\epsilon)$

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

instead of a straight line needed for the evaluation of the fractal dimension and correlation [15–18], 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, in turn, lead to a sudden increase in the log-log plot of the point densities. This can be seen by choosing a random reference data point. Around this point, we enlarge the neighborhood radius ϵ, counting the points that fall into this neighborhood. After reaching a cluster of points, the count $C(\epsilon)$ quickly increases with ϵ, which leads 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 visibly decreases with m. From 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 appearing under ideal conditions as follows:

For n odd, the lower bound t and the maximal number s of steps have the expression

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

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

For n even, the lower bound t and the maximal number of steps s have the form

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

and

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

These results extend and detail insights from previous approaches [9].

By searching for steps, we can not only pin down data that are likely to contain patterns. With the help of the table presented in Fig. 1, we can also infer the length of putative patterns.

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. Ideal case. Lower bound t/maximally observable steps s, as a function of pattern length n and embedding dimension m.
doi:10.1371/journal.pone.0028107.g001

Results

Method validation

To what extent is the method reliable? In realistic time series, especially in neuroscience, a regular signal will be contaminated by jitter and noise. Jitter is commonly defined as the addition of an amount of signed (or unsigned) noise to the signal. Under its influence, a period-three signal of interspike intervals (ISIs) {3200,7700,1000} may change into a time series such as {3223,7703,907,3203,7782,903,3107,7603,1098,...}. For this example, we added a jitter of 10 percent of the smallest ISI to the data, drawn from a uniform probability distribution. Alternatively, Gaussian or long-tailed distributions can be considered, which leads, in the range of interest, only to negligible differences. Noise is implemented by choosing a given percentage of the ISIs according to some random probability distribution. This can be achieved in two 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 a probability p from the regular pattern and with a 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 (for a fair comparison among the different paradigms, the probabilities must be rescaled as $p'=p/(n-(n-1)p)$). Motivated by neuroscience applications, here we focus for our results on the second paradigm.

Upon the addition of jitter and noise, the steps gradually 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 analysis focuses on a pattern of length $n=3$. The analysis has, however, also been performed for patterns of length 5 and partially for length 7, with comparable results. Longer patterns have obtained little attention in experimental time series [19], [20].

In the log-log plot, jitter predominantly smoothens out the steps, whereas noise decreases the heights as well as the slopes of the stairs. We assess the ability of our method to highlight regular patterns of length n in jitter and noise contaminated data with the help of three criteria: a) How well can the predicted decrease of the number of steps with the embedding dimension m be evidenced? b) How well can exactly n steps in the embedding dimension $m=1$ be detected? c) How well is a plateau, the flat part of the graph prior to the step, at embedding dimension $m=n$ expressed if compared to that observed at $m=n+1$ [9]?

For the first criterion, we verified whether the predicted decrease of the number of steps as a function of m was observed or not. To this end, we tested whether a single vertical step was visible at $m=n$. For this we preset three height levels h with corresponding quality weights w (denoted $\{h,w\}$) $\{\{1.0,2\}, \{1.2,3\},\{1.5,3\},\{2.0,2\},\{3.0,1\}\}$ that in the ideal case the derivative of the log-log plot would all exceed. Given a particular preset height level, we rewarded the detection of exactly one peak in the derivative (corresponding to a sharp step-like increase in the original plot) with the level's corresponding weight and used the resulting sum over the height levels as 'quality' measure. Added noise, however, may trigger a reappearance of the theoretically vanishing steps at and beyond the embedding dimension at which only one step should emerge. To eliminate this problem, if two or three steps emerged in the data, we compared the time series vs. surrogate (i.e. randomly permuted) series, in which the repeated steps emerge most pronouncedly. To characterize the distance from the surrogate case, the quality of the time series data was subtracted from the surrogate quality. The final 'quality' measure was thus composed as sum of a first measure for the visibility of exactly one step and a second term which, being nonzero only in the case of two or three observed steps, reflects the distance from the surrogate case.

For the second criterion, in order to quantify the visibility of exactly three derivative peaks at $m=1$ we proceeded with a peak-detection algorithm as in criterion (a) yet with a slightly different attached level-weight-vector $\{\{0.75,2\},\{1.0,3\},\{1.5,3\},\{2.0,2\}, \{3.0,1\}\}$. For the third criterion, the plateau flatness at $m=n$ was compared to $m=n+1$. A plateau was counted, if the derivative of the plot was below values $\{\{0.2,1\},\{0.4,2\}, \{0.7,2\},\{1.1,1\}\}$ (again with corresponding weights w, $\{h,w\}$). The weighted average counts obtained for $m+1$ were then subtracted from the weighted average counts obtained for m.

For all criteria assessments (a), (b) and (c), we approximated the derivative values as difference quotients between two consecutive

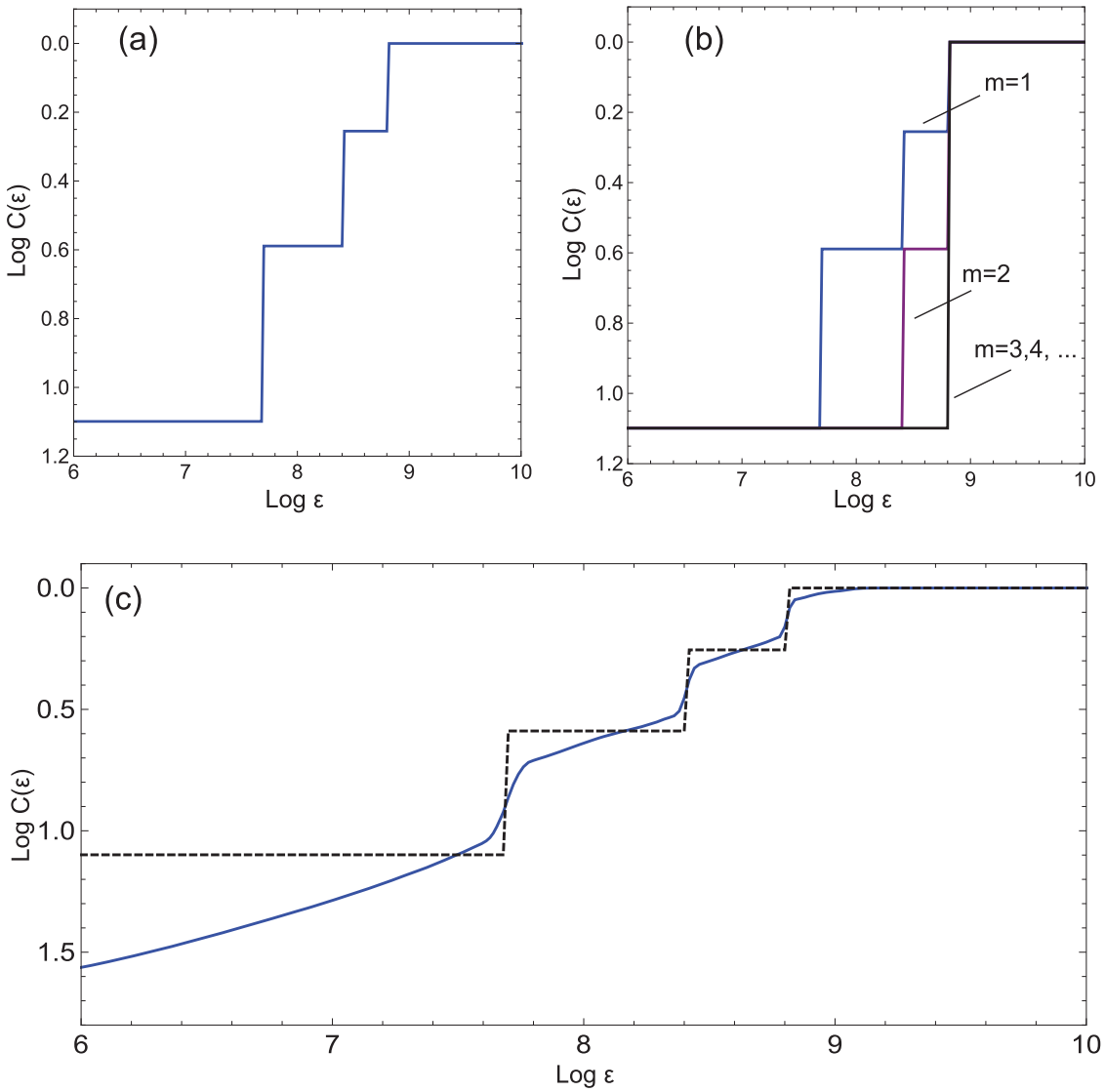

Figure 2. Log-log plots from a pattern of length 3. a) $m=1$, b) for increasing embedding dimension m, c) $m=1$, modification introduced by the presence of 20 percent jitter and 30 percent noise (pattern $\{3200,1700,1000\}$.)
doi:10.1371/journal.pone.0028107.g002

data points, for which $\log\varepsilon$ was increased in steps of 0.02. Certainly the above described algorithms are not the unique possibility to reasonably quantify the proposed criteria. We however argue that the algorithms, together with the carefully selected weight vectors, do provide a measure which is in accordance with the human eye's perception of peaks and plateaus.

By dividing through the observed maximal measure, the three measures were normalized and a contour-plot with suitable contours was drawn. Fig. 3 shows the results obtained. We defined two or three regions of various visibility for each of the criteria. Not surprisingly, the visibility of exactly n peaks for $m=1$ (Criterion (b)) is best in 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 steps of a length-7 pattern are more difficult to distinguish since with increasing jitter the peaks in the derivative may overlap. Criteria (a) 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 (c) for a generic pattern of length 5 comparing the dimensions $m=5$ 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 (c) to be fairly independent of the underlying pattern length. In regions where the criterion (c) fails, i.e. for little noise and high jitter, criterion (a) 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 lead to such a single step. Comparing to the embeddings $m\leq n$ where more steps should

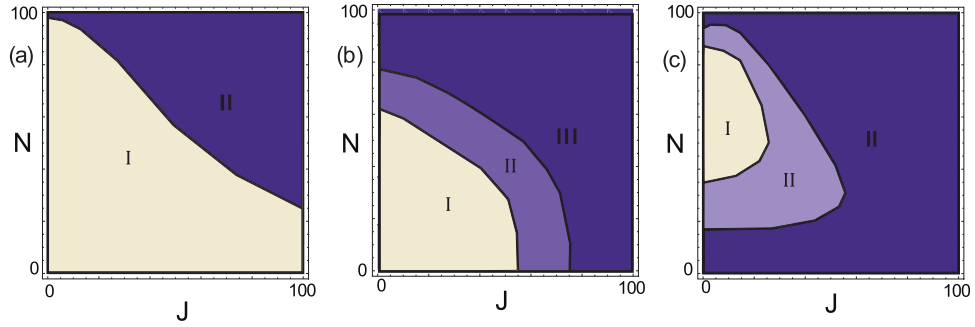

Figure 3. Approximate phase boundaries, for noise N and jitter J in units of percents of events in the data and in percents of the smallest interval in the pattern. Fulfillment of the criteria is expressed by three degrees: Region I: excellent, region II: fair, region III: ambiguous. a) Measure for the decrease in steps with m (only two regions: I and III). b) $m=1$-criterion; c) difference in plateau visibility for $m=3$ compared to $m=4$. Regions I, II and III as in a).
doi:10.1371/journal.pone.0028107.g003

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 thus 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.

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

For a proof of (2)–(5), we decompose the graph of componentwise distances $d_{ij} := |a_i - a_j|$ into subgraphs connecting nearest-, next-nearest-, etc. neighbors, see Fig. 4. The idea underlying the optimized proof with sharper bounds is, as in the old proof in [9], the following: 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. When $m=2$ and $n=6$, 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. Contrarily, if we have a 'regular' distribution of the biggest three distances $d_{ij}: d_{12}>d_{34}>d_{56}>\text{remaining}\, d_{ij}$, only 3 steps are contributed when $m=2$. 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 $n/2$ lines, which only contributes one step if $m>n/2$.

Discussion

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. The reader may thus derive an overall goodness-of-method measure by adding the measures obtained from the different criteria. This might help to *a priori* evaluate the applicability of the method to a user's problem. As guideline for the practical use of the method, we suggest to embed a given time series in spaces of multiple dimensions m up to $m \geq 10$ in order to capture possible pattern lengths of such order. Regarding the computation of the correlation integral, it is important to sample densely enough among randomly selected reference points (e.g., for 10.000 data points, we recommend something above 200 reference points). Equipped with the log-log curves for multiple m, a significant plateau flatness difference between to consecutive m (criterion (c)) can serve as first indicator of the pattern length [9]. Criteria (a) and (b) may be helpful to confirm such a suspicion and to gain additional, metric information about the pattern.

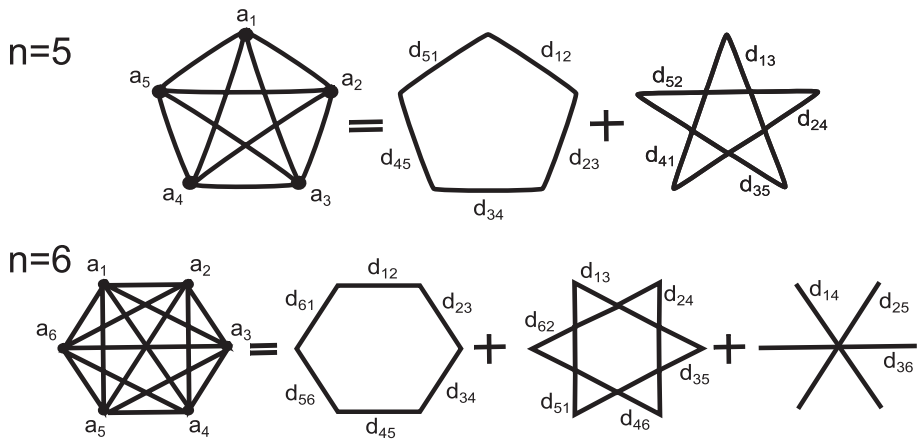

Figure 4. Graphs of componentwise distances. Decomposition of potential distances in the maximum norm for odd and for even pattern lengths n into nearest-, next-nearest-, etc., neighbor subgraphs. Each subgraph can be treated separately.
doi:10.1371/journal.pone.0028107.g004

Moreover, the slope of the lines in the step-free regions may give interesting insights into the fractal dimension of a possible attractor.

Materials and Methods

All computations were performed in a C++ and *Mathematica* environment on a custom laptop. The method validation was based on the length-3 pattern {3200,7700,1000}. The correlation integral was evaluated for a total of 9900 embedded points, where 1000 points were used as reference points. For a total of 11×11 jitter-noise-configurations (from 0% to 100% in steps of 10%), we evaluated the three described criteria. A set of levels appropriate for the classification into the 'excellent', 'fair' and 'ambiguous' evaluation regimes resulted in the contour-plots shown in Fig. 3.

Author Contributions

Conceived and designed the experiments: RS FG. Performed the experiments: RS FG. Analyzed the data: FG. Contributed reagents/materials/analysis tools: RS. Wrote the paper: RS FG.

References

1. Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W (1999) Spikes. Exploring the Neural Code. Cambridge, MA: MIT Press.
2. Stoop R, Blank DA, Kern A, van der Vyver JJ, Christen M, et al. (2002) Collective bursting in layer iv: Synchronization by small thalamic inputs and recurrent connections. Cog Brain Res 13: 293–304.
3. Stoop R, Blank DA, van der Vyver JJ, Christen M, Kern A (2003) Synchronization, chaos and spike patterns in neocortical computation. Journal of Electrical and Electronics Engineering, Istanbul University 3: 693–698.
4. Stoop R, Bunimovich L, Steeb WH (2000) Generic origins of irregular spiking in neocortical net- works. Biol Cybern 83: 481–489.
5. Dayhoff JE, Gerstein GL (1983) Favored patterns in spike trains. J Neurophysiol 49: 1334–1348.
6. Abeles M, Gerstein GL (1988) Detecting spatiotemporal firing patterns among simultaneously recorded single neurons. J Neurophysiol 60: 909–924.
7. Lestienne R, Tuckwell HC (1997) The significance of precisely replicating patterns in mammalian cns spike trains. Neuroscience 82: 315–336.
8. Christen M, Kern A, Nikitchenko A, Steeb WH, Stoop R (2004) Fast spike pattern detection using the correlation integral. Phys Rev E 70: 011901.
9. Stoop R, Christen M (2010) Detection of patterns within randomness. In: Nonlinear Dynamics and Chaos: Advances and Perspectives, Springer Berlin/Heidelberg.
10. Eckmann JP, Ruelle D (1985) Ergodic theory of chaos and strange attractors. Rev Mod Phys 57: 617–656.
11. Kantz H, Schreiber T (2003) Nonlinear Time Series Analysis. Cambridge, UK: Cambridge Univ. Press.
12. Peinke J, Parisi J, Roessler OE, Stoop R (1992) Encounter with Chaos Springer Berlin/Heidel- berg.
13. Sauer T (1994) Reconstruction of dynamical systems from interspike intervals. Phys Rev Lett 72: 3811–3814.
14. Pecora L, Moniz L, Nichols J, Carroll T (2009) A unified approach to attractor reconstruction. In: Dana S, Roy P, Kurths J, eds. Complex Dynamics in Physiological Systems: From Heart to Brain, Springer Berlin/Heidelberg, volume 41 of *Understanding Complex Systems*. pp 3–19.
15. Grassberger P, Procaccia I (1983) Measuring the strangeness of strange attractors. Physica D 9: 189–208.
16. Ding M, Grebogi C, Ott E, Sauer T, Yorke JA (1993) Plateau onset for correlation dimension: When does it occur? Phys Rev Lett 70: 3872–3875.
17. Castro R, Sauer T (1997) Correlation dimension of attractors through interspike intervals. Phys Rev E 55: 287–290.
18. Kern A, Steeb WH, Stoop R (1999) Local correlations potential for noise reduction and symbolic partitions. Z Naturforschung A. pp 404–410.
19. Lestienne R, Strehler BL (1987) Time structure and stimulus dependence of precisely replicating patterns present in monkey cortical neuronal spike trains. Brain Res 437: 214–238.
20. Prut Y, Vaadia E, Bergman H, Haalman I, Slovin H, et al. (1998) Spatiotemporal structure of cortical activity: Properties and behavioral relevance. J Neurophysiol 79: 2857–2874.