Local Correlation's Potential for Noise Reduction and Symbolic Partitions

A. Kern, W.-H. Steeb[a], and R. Stoop

Institut für Neuroninformatik, University of Zürich / ETH Zürich, CH-8057 Zürich
[a] Institute for Applied Mathematics, Rand Afrikaans University, RSA-2006 Johannesburg

Reprint requests to PD Dr. R. S.; e-mail: Ruedi@ini.phys.ethz.ch

Z. Naturforsch. **54 a,** 404–410 (1999); received April 19, 1999

We investigate local correlation dimension-based noise-cleaning of time series, where points having anomalously large dimensions are iteratively removed from the reconstructed attractor. We find an optimal range for the number of iterations in which the algorithm yields good results. Choosing non-local ranges for the linear regression yields a new method for finding nonhyperbolic tangency points. The method is also applicable for noisy systems with unknown dynamics; in this case, noise facilitates the detection of the points.

Key words: Noise Cleaning; Fractal Dimensions; Nonhyperbolic Tangency Points.

1. Introduction

The usual experimental approach to dynamical systems is that a scalar time series of measurements of a variable, $x(t)$, is given, that originates from some system whose underlying dynamics $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ is not known. Often a small number of experimental parameters have a strong and immediate impact on the time series, whereas a possibly larger number of less important parameters is responsible for small-size fluctuations only (referred to as primary and secondary parameters, respectively). The observer often interprets these fluctuations as noise. Usually, for experiments, the relevant system parameters are not directly accessible. Neither the number of parameters nor the dimension of the phase-space is known. However, it is possible to reconstruct the high-dimensional dynamics of the system from the time series $x(t)$ by means of the coordinate delay method. It is a consequence of abstract differential geometry (essentially, Whitney's embedding theorem [1]), that if the time series $x(t)$ are embedded in m dimensions,

$$\boldsymbol{y}(t) = (x(t), x(t-T), \ldots, x(t-(m-1)T)), \quad (1)$$

where m must be chosen large enough and T is a suitably chosen delay time, then, generically, a diffeomorphism between the original attractor, and the attractor in reconstructed phase-space, is established. The exact statement is that the relation $m \geq 2d + 1$ should be satisfied, where d is the box-counting dimension. The delay-reconstructed attractor displays the same geometric and dynamic properties as the attractor in the original, inaccessible, phase-space. For example, the Lyapunov exponents [2] or the fractal dimensions [3 - 5] are conserved under the embedding. Therefore, for most purposes it is sufficient to investigate the behavior of the system in the reconstructed phase-space. Unfortunately, the determination of a closed form of the dynamical law in the embedding space, $\dot{\boldsymbol{x}}_{\text{emb}} = \boldsymbol{f}_{\text{emb}}(\boldsymbol{x}_{\text{emb}})$, is generally impossible, due to the usually complicated nature of the connecting diffeomorphism. In a few examples, successful matching of the experimental data to a pre-chosen form of the map $\boldsymbol{f}_{\text{emb}}$ has been reported (e. g., by using as an ansatz a polynomial of a pre-determined order). However, especially when the measured system shows traces of noise, it is unknown what ansatz should be used and, therefore, one is usually forced to work with the embedded time series.

An important task in this context is the a posteriory elimination of "noise" that may either have originated from secondary parameters or from measurement errors. The goal is to reveal the structure of the principal dynamics; this should lead to a simplified description of the system. In the present paper, we are interested in the extent to which this structure can be recovered from the noisy situation, without allowing for a modification of the system itself. The testing paradigm for our investigations will be the Hénon map [6 - 7]

0932–0784 / 99 / 0600–0404 $ 06.00 © Verlag der Zeitschrift für Naturforschung, Tübingen · www.znaturforsch.com

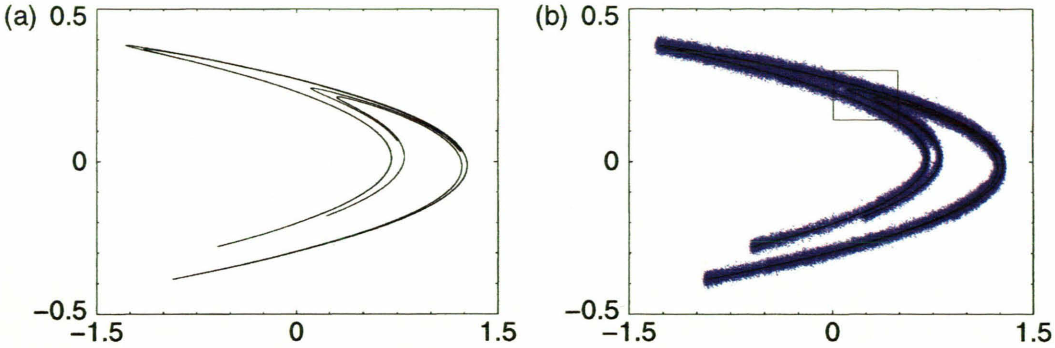

Fig. 1. (a) Original and (b) noisy Hénon attractor, noise level 0.01.

$$x_{n+1} = y_n + 1 - ax_n^2, \tag{2}$$

$$y_{n+1} = bx_n, \tag{3}$$

at the classical parameter values $a = 1.4, b = 0.3$. At these values, a strange attractor is obtained that has an information dimension of about 1.26 (correlation dimension about 1.22). Time series generated by this system were contaminated by additive or dynamical Gaussian white noise. Figure 1 shows the Hénon attractor, in its original form and with additive noise, respectively. Our noise-cleaning method will be based on local correlation (or: pointwise) dimensions. These dimensions are evaluated by extrapolations of the logarithmic population growth in balls $B(\epsilon, \boldsymbol{x}_0)$ for the local limit $\epsilon \to 0$ (where $\boldsymbol{x}_0$ denotes a point of the embedded time series and ϵ denotes the radius). We will also apply similar calculations without going to the local limit. In principle, this amounts to an abuse of the original method. However, it will allow us to detect the nonhyperbolic tangency points [7 - 8] of the attractor, where noise facilitates the finding of these points. Nonhyperbolic tangency points are of interest since they are crucial for the introduction of symbolic dynamics of the system.

2. Correlation Integrals

The correlation integral $C(\epsilon)$ is based on the probability that two randomly chosen points are separated by a distance less than ϵ, which has the form

$$C^{(m)}(N, \epsilon) = \frac{2}{N(N-1)} \sum_{j=1}^{N} \sum_{i=j+1}^{N} \Theta(\epsilon - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|), \tag{4}$$

where $\Theta(\cdot)$ is the Heavyside step function and $\boldsymbol{x}_i = \boldsymbol{x}(t_i), i = 1,..,N$ are the sampled points in m-dimensional embedding space. It is assumed that the limit $N \to \infty$: $C^{(m)}(N, \epsilon) \to C^{(m)}(\epsilon)$, exists, for all ϵ. The correlation dimension of the reconstructed attractor is defined by

$$D_2^{(m)} = \lim_{\epsilon \to 0} \frac{\log C^{(m)}(\epsilon)}{\log \epsilon}. \tag{5}$$

This value is approximated by the slope of the logarithmic plot of the correlation integral $C^{(m)}(N, \epsilon)$. The value of $D_2^{(m)}$ can be shown to increase with m. For m larger than a saturation dimension $\tilde{m}$, the embedding condition is met, a diffeomorphism between the original and the reconstructed attractor is obtained, and the value of $D_2^{(m)}$ should therefore be constant[1]. In practical applications, at increased embedding dimensions the numerical approximations to $D_2^{(m)}$ change again. In order to provide a faithful representation of the submanifold to which the dynamics are restricted, an increasing number of points is required. This quickly exhausts the data set [11]. Thus, the plateau value of numerical approximations to $D_2^{(m)}$, if it exists, can be assumed to give a reliable estimator for the correlation dimension.

These observations apply for deterministic systems; for a white noise system no saturation occurs and a plateau onset for the correlation dimension is not observed. This fact may be used to decide whether a given time series is generated by a deterministic law

[1] It has been shown in [9] that for infinitely long noise-free data, $D_2^{(m)}$ saturates already at $\tilde{m}$ = Ceil(D_2) (where Ceil(D_2) stands for the smallest integer greater than or equal to D_2), and that for short data sets no saturation may occur.

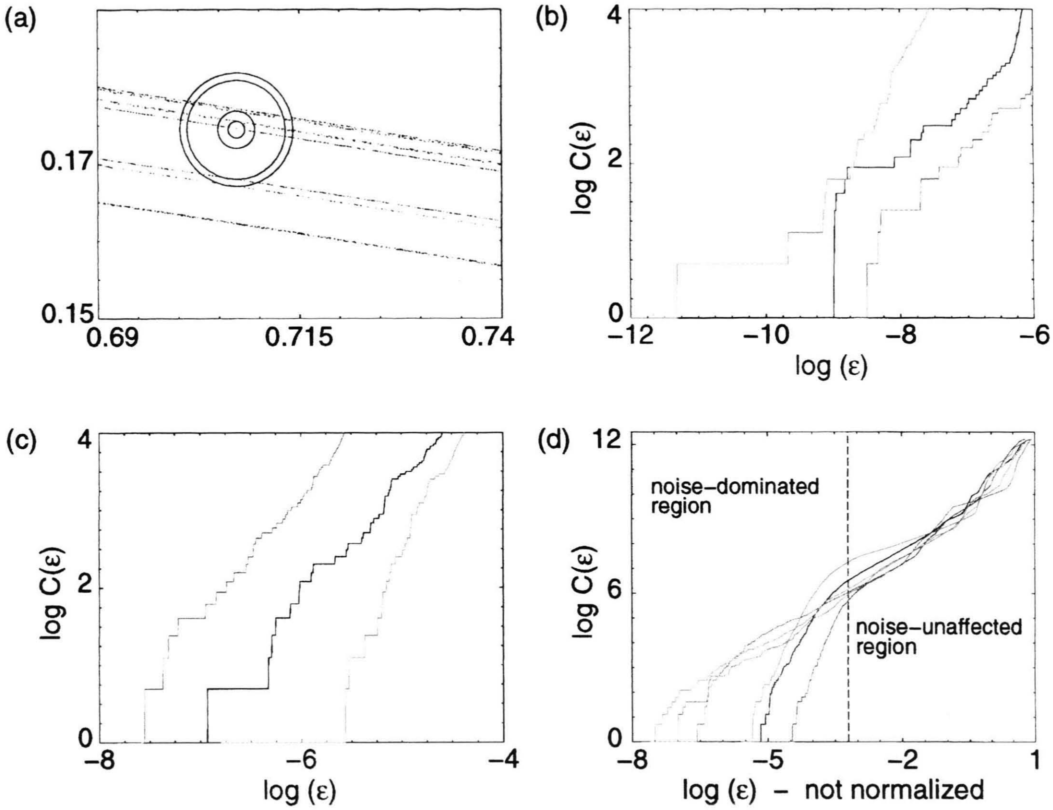

Fig. 2. Enlargement of an ϵ-neighborhood. For certain values of ϵ, new pieces of expanding manifolds are included. (b) Growth onset of three curves for the noise free and (c) for the noisy case. (d) Logarithmic plots of local correlation integral curves. The three rightmost curves are from the noisy Hénon attractor.

or by a stochastic process. As will be discussed in detail below, in noisy deterministic systems the growth onset of the correlation integral curve is moved towards larger values of ϵ. This leads to an increase of the slope of the logarithmic correlation integral curve on scales ϵ below the standard deviation s_n of the noise. As a consequence, the logarithmic plot of the noisy system reveals two regimes: A noise-affected regime $\epsilon \in [0, s_\mathrm{n}]$, and an essentially noise-unaffected regime $\epsilon \in [s_\mathrm{n}, s_\mathrm{v}]$, where s_v is the saturation value introduced by the finiteness of the system. Numerical estimators $D_2^{(m)}$ therefore yield values that depend on the regime of the logarithmic plot in which the slope is measured. In the first part of our approach, numerical dimension estimators from the noise-affected regime are used for the recovery of local attractor fine-structures. In the second part, dimension estimators from the noise-unaffected regime will be used to uncover the nonhyperbolic tangency points of the system.

For an understanding of these approaches, it is necessary to shift from the global average provided by $D_2^{(m)}$ to a more localized point of view. Equation (4) can be interpreted as an average over local ϵ-balls if we hold one of the two points that are summed over, fixed [4 - 5]:

$$C^{(m)}(N, \epsilon, \boldsymbol{x}_n) = \frac{1}{N-1} \sum_{i=1, i \neq n}^{N} \Theta(\epsilon - \|\boldsymbol{x}_i - \boldsymbol{x}_n\|), \quad (6)$$

$$D_2^{(m)}(N, \boldsymbol{x}_n) = \lim_{\epsilon \to 0} \frac{\log C^{(m)}(N, \epsilon, \boldsymbol{x}_n)}{\log \epsilon} \quad (7)$$

These *pointwise* [7], or *local correlation* dimensions reveal the reasons for the technical comments made for global correlation dimensions. For this, the observation of how points are recruited for the correlation integral is crucial. For the local correlation integral, the number of points around a pre-chosen central point needs to be counted, as a function of the radius (c. f. (6) and (7)). For some radii, a whole new section of a neighboring expanding manifold is collected, which leads to a sudden jump in the number of points included in the ball. Figure 2 (a) gives an illustration of this principle for the noise-free Hénon attractor. Since in the local approach the averaging process over many balls is not performed, the local logarithmic correlation integral plots do not show straight-line segments; instead, the traces of the fractal structure of the chaotic attractor become apparent (see Figure 2 (b)). The increase of the local correlation integral appears as the first neighboring points are recruited. For the black and the blue curves, this occurs at rather large ϵ-scales. The black curve, however, starts to recruit points already on small scales, suggesting that neighboring pieces of the expanding manifold closely pass by and are included early in this stage. It is easily verified that the jumps in the correlation integral occur exactly at the typical radii plotted in Figure 2 (a). As more and more points are included in the ball, the effect is smoothed.

Smearing by noise leads to a qualitatively different situation (c. f. Figure 2 (c)). The increase of the correlation integral is now much smoother from the start. A comparison of the behavior of local correlation integrals for noisy and noise-free cases is shown in Fig. 2 (d), where the three leftmost curves originate from the noise-free, the others from the noisy case. These plots reveal the local origin of the two observed regimes of the correlation integral and their relation to s_n. Only for $\epsilon < s_n$, we observe a strong influence of noise, which results in an increase of the local dimensions. Removing points with this property from the data set is a reasonable approach for noise cleaning.

3. Noise Detection

When our noisy scalar time series were embedded in increased embedding dimensions, attractor structures that in lower dimensions were hidden by the noise, became visible. Due to the limited number of points available, this process, however, must come to an end. In practice, the best choice of the embedding dimension is indeed $\tilde{m}$, provided that the time series is large enough to let the saturation occur. Below, we focus on results obtained from a time series of the Hénon map of length 200000, to which Gaussian white noise of standard deviation 0.01 was added. At this size, the length of the time series is uncritical; a time series of half the length gives identical results (it was for reasons of graphical presentation that we chose the largest investigated file). Dynamical noise was also examined, with similar results. However, in this case the standard deviation had to be restricted in size in order to prevent the dynamical system from escaping to infinity and the transition point between the two regimes is not as clear as in the case of additive noise. We obtained the best results, when in the first round of the noise-cleaning, 10% of the highest-dimensional points of the embedded attractor were removed. Other iterations are needed because a removal of points clearly affects the dimensions of neighboring points. For the same reason, the range of regression needs to be adjusted in each iteration, since the shapes of the correlation integral curves change after the removal of a fraction of attractor points. To optimize the procedure, we ran the computation using $n = 25$ iterations with different choices of the linear regression range. Fortunately for the approach, the individual variations in the shapes of the correlation integral curves disappear after a few iterations. Optimal embedding dimensions for the data sets were from the range $m \in \{4, 5, 6\}$. For $m = 5$, the choice of the regression range was optimized in the following way:

- For the first few iterations, we chose a region near to the end of the noise-dominated region (from –4.5 to –3.0) for the regression. This ensures that the most severely noise-affected points are immediately eliminated.
- The regression was restricted to the support of the correlation integral (i. e. to values larger than -4.5). This avoids missing noise-affected points with late growth onsets.
- A lower boundary was introduced for the local correlation dimensions. For points with unreasonably low dimensions, the growth onset occurs past the right boundary of the regression range. Such points are also heavily noise affected.
- After some iterations of noise-cleaning, the border between the noise-dominated and the noise unaffected regime moves to the left. As a consequence, the range of regression has to be adapted.

This iterative noise-cleaning approach was tested in embedding dimensions 2, 4, 5, and 6. In dimension two, we were able to recover little of the fractal structure. The retained points tend to form huge clouds that obstruct further recovery of the one-dimensional

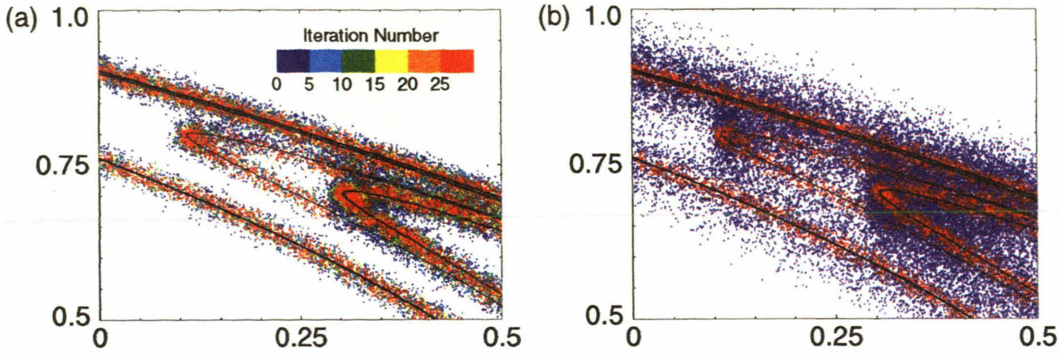

Fig. 3. Main results of noise-reduction. (a) Color-coding shows the iteration number at which the points are removed (Hénon system, embedded in 5 dimensions, projection on dimension 2). (b) Final effect of noise-cleaning on the original noisy data set. Blue: points removed by the noise-cleaning. Red: remaining points, revealing the geometrical form of the underlying attractor.

expanding manifolds. This effect is strongest when the range of regression remains the same for all iterations. For embedding dimension $m = 5$, results from our optimized noise-cleaning approach are presented in Figure 3. Figure 3 (a) shows a projection from embedding dimension $m = 5$ onto dimension 2. To follow our iterative approach, colors indicate the number of the iteration in which the point was removed from the data set (color changes in steps of five iterations). The strongest noise-affected points (in dark blue) are removed in the first few iterations. Light blue and green points form an onion-skin structure. They are removed in the next 10 iterations. The red points indicate survival of the whole noise reduction process, leaving about 50000 points out of 200000 at the beginning. The procedure achieves its strongest effect in the first 10 - 15 iterations (dark blue, light blue, and green points). Iterations 15 - 25 (yellow and orange points) do not further improve the results. Points that lie quite reliably on a piece of the expanding manifold would be rejected in this phase, where points from regions of low population are most severely affected. Whereas for pieces of rather isolated expanding manifolds with medium density the procedure worked perfectly, approaching pieces of the manifolds tended to be merged into one single structure. In such cases, the branch with higher point density wins the competition, as can be observed in the lower right part of Figure 3 (a). The final results of our iterative noise-cleaning approach are shown in Figure 3 (b). In this figure, we marked those points of the original (2-dimensional) non-embedded data set that survived the noise-cleaning procedure and therefore can be assumed to describe the underlying attractor most precisely. The substantial noise-cleaning effect (red points against blue points) is evident. Essential parts of the fractal structure that were originally hidden by the noise, could be recovered. Only around turning points of the manifold, where the point density is especially large, clustering is still a major problem. This effect is due to the presence of nonhyperbolic tangency points [7 - 8, 10], which cannot be removed and will be discussed in more detail in the next section. The next important factor obstructing even better results is in the removal of the noisy points itself which not only leads to a significant loss of data points, but also is responsible for smaller effects of clouding that cannot be attributed to nonhyperbolic tangency points. We expect that a projection of noisy points onto estimated expanding manifolds, instead of removal, would drastically improve the results. The whole approach is simple to implement and its efficacy is comparable to other approaches of noise-cleaning.

4. Nonhyperbolic Tangency Points

In this section we investigate what information is contained in the noise-unaffected regime of the correlation integral curves. That is, we apply the correlation-dimension algorithm by abuse to non-local scales. For our numerical tests, ranges of regression between –3.8 and –1.3 were chosen. We found it convenient to add Gaussian white noise having a doubled standard deviation 0.02 to the Hénon time series. The increased level of noise considerably widens the

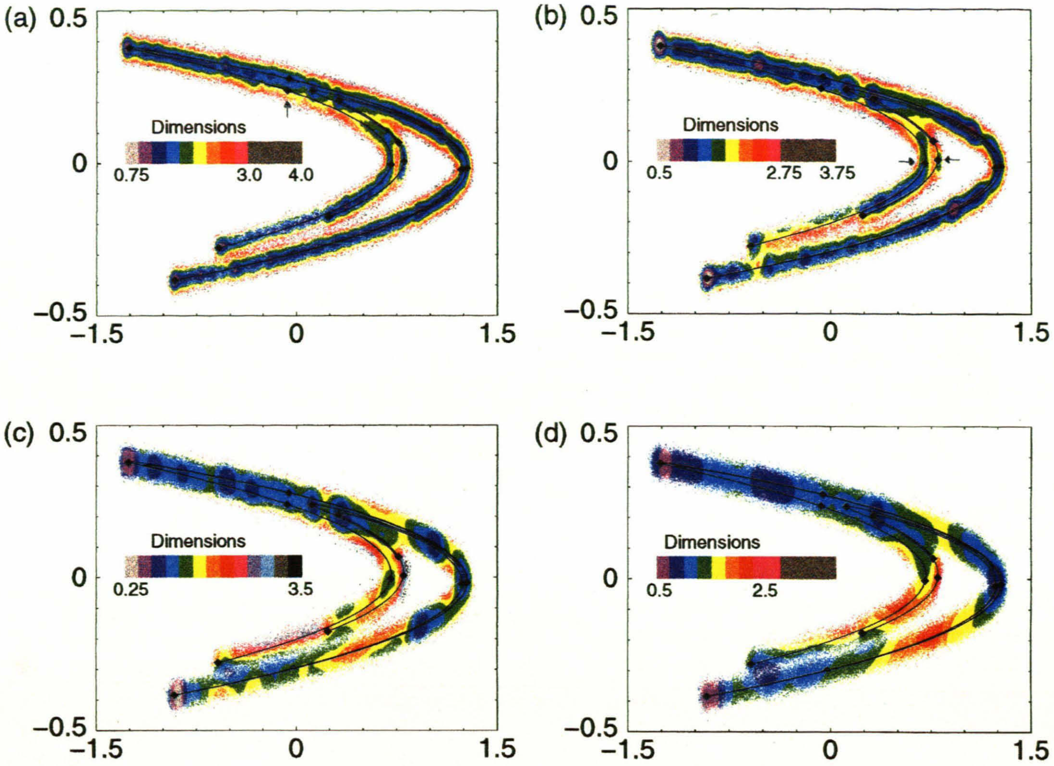

Fig. 4. Location of nonhyperbolic tangency points, within areas of minimal local correlation dimensions. For different choices of the regression interval, local correlation dimensions were calculated and color-coded: (a) Interval $[-3.8, -2.8]$, (b) $[-3.3, -2.3]$, (c) $[-2.8, -1.8]$, (d) $[-2.3, -1.3]$.

attractor structures which enhances the visibility of interesting features. Due to the logarithmic scale that is used, the boundary between the noise-affected and noise-unaffected parts of the curves (see Fig. 2 (d)) is relatively stable towards this increase.

Figure 4 presents the distribution of local correlation dimensions for nonlocal regression intervals, where, starting at $[-3.8, -2.8]$, the regression interval is shifted to the right in steps of 0.5. For the different choices, the values of the dimensions are color-coded according to the panels shown in the figure. On top of these results, a plot of the noise-free Hénon attractor is superimposed, in which nonhyperbolic tangency points are represented as diamonds. In each of the figures, several violet and dark blue spots mark local minima of the correlation dimensions. It is known that in the vicinity of nonhyperbolic tangency points, the phase-space density is significantly increased (for a theoretical explanation see [7 - 8]). Nonhyperbolic tangency points can therefore be expected to be situated at local minima of the local correlation dimension. Indeed, we find that the minima of the non-local correlation dimension often coincide with nonhyperbolic tangency points, but that this is not always the case. Figure 4 (a) is dominated by local features (including noise), which is reflected by the pronounced onion-skin distribution of dimensions. Most nonhyperbolic tangency points are reported as local minima of the dimensions (violet spots). Unfortunately, some tangencies are missed that later on will be uncovered by shifted regression intervals (for an example, the point marked by an arrow). As the regression interval is shifted, the regions that are generated by the dominant nonhyperbolic tangency points increase in size. Optimal results are obtained for ranges that are adjacent to the boundary dividing the noise-affected and the noise-unaffected regions (i. e., near s_{n}). Here, most of the prominent nonhyperbolic tangency points are obtained from the major regions of minimal correlation dimension. Figure 4 (b) is based on an interval that is slightly above the optimal range. As a consequence, some tangencies that before were detected, are now lost (see the two arrows in (b)). This is due to the special location of these points in the center of the attractor: The distance from these points to the outer lobe is about 0.1, the logarithm thereof about −2.3. The regression interval of Fig. 4 (b) includes this distance, so that the measure of points lying on the outer lobe outbalances the measure of points lying near these two tangencies. In this way, the tangency point becomes undetectable. For Fig. 4 (d), the regression interval is [−2.3, −1.3]. Here, the correlation dimensions are entirely determined by the large-scale structure of the Hénon attractor and reflects essentially its global geometrical structure.

5. Conclusions

Our investigations show that the noise-cleaning potential by local dimensions is considerable. Based on extended numerical simulations, we believe we have obtained optimal results for the correlation-based method, in the sense that no further potential for optimization is left, if the constraint of an unmodified local attractor structure is to be respected. For even "better" results of the approach, this constraint must be given up [12]. This, however, is only at the price of an estimation of the šidealÏ local geometrical form of the attractor, which needs a careful consideration of the ambiguities introduced in this way.

When the local correlation dimension algorithm is applied to the non-local regime, we find that the determination of primary [10] nonhyperbolic tangency points is as simple, or even simpler than in the noise-free case. This shows that the influence of nonhyperbolic tangency points is preserved under noisy conditions. From the positions of the most prominent nonhyperbolic tangency points, an efficient symbolic partition of the system could be introduced [10, 7]. As a consequence, the description of noisy systems by symbolic dynamics may be less difficult than is generally suspected. This may help to gain further insight into the statistical behavior of both noise-affected and noise-unaffected dynamical systems.

Acknowledgements

We acknowledge support from the Swiss National Science Foundation (KTI, Phonak AG).

[1] H. Whitney, Ann. Math. **37**, 645 (1936).
[2] V. I. Osceledec, Moscow Math. Soc. **19**, 197 (1968); G. Benettin, L. Galgani, and J. M. Strelcyn, Phys. Rev. A **14**, 2338 (1976); R. Stoop and P. F. Meier, J. Opt. Soc. Amer. B **5**, 1037 (1988).
[3] B. B. Mandelbrot, The Fractal Geometry of Nature, Freeman, New York 1982.
[4] J. D. Farmer, E. Ott, and J. A. Yorke, Physica D **7**, 153 (1983).
[5] P. Grassberger and I. Procaccia, Physica D **13**, 34 (1984).
[6] M. Hénon, Comm. Math. Phys. **50**, 69 (1976).
[7] J. Peinke, J. Parisi, O. E. Roessler, and R. Stoop, Encounter with Chaos, Springer, Berlin 1992.
[8] C. Grebogi, E. Ott, and J. A. Yorke, Phys. Rev. A **37**, 1711 (1988).
[9] M. Ding, C. Grebogi, E. Ott, T. Sauer, and J.A. Yorke, Physica D **69**, 404 (1993).
[10] P. Grassberger and H. Kantz, Phys. Lett. A **113**, 235 (1985).
[11] J.-P. Eckmann and D. Ruelle, Physica D **56**, 185 (1992).
[12] A. Kern and R. Stoop, in preparation.