Dynamical Correlations on Reconstructed Invariant Densities and their Effect on Correlation Dimension Estimation

Andreas Galka* *and* Gerd Pfister

Institute of Experimental and Applied Physics, University of Kiel, 24098 Kiel, Germany

International Journal of Bifurcation and Chaos **13** (2003), 723 – 732

Abstract

We investigate the structure of dynamical correlations on reconstructed attractors which were obtained by time-delay embedding of periodic, quasi-periodic and chaotic time series. Within the specific sampling of the invariant density by a finite number of vectors which results from embedding, we identify two separate levels of sampling, corresponding to two different types of dynamical correlations, each of which produces characteristic artifacts in correlation dimension estimation: the well-known trajectory bias and a characteristic oscillation due to periodic sampling. For the second artifact we propose random sampling as a new correction method which is shown to provide improved sampling and to reduce dynamical correlations more efficiently than it has been possible by the standard Theiler correction. For accurate numerical analysis of correlation dimension in a bootstrap framework both corrections should be combined. For tori and the Lorenz attractor we also show how to construct time-delay embeddings which are completely free of any dynamical correlations.

Keywords: time series analysis, correlation dimension, dynamical correlations, resampling, invariant density.

1 Introduction

Deterministic dynamical systems can be classified according to the qualitative properties displayed by the dynamics of the systems, after initial transients have died out. These properties may be described by the concept of *attractors*, i.e. geometrical objects in an appropriately chosen state space, and by probability densities, concentrated on these attractors, which are generally known as *invariant densities*. Four basic classes of attractors can be discriminated: Point attractors, limit cycles, N-tori and *chaotic attractors*. Whereas the first three classes may also be identified and characterised by traditional linear quantities such as the power spectrum, for the case of chaotic attractors it is necessary to employ nonlinear tools directly referring to the state space of the system. Some well-known examples of such tools that due to their invariance with respect to embedding are frequently used, are fractal dimensions (such as correlation dimension), Lyapunov exponents and entropies [Kantz & Schreiber, 1997].

* Corresponding author. E-mail: galka@physik.uni-kiel.de

In many cases, especially in the case of chaotic attractors, no analytical description of the invariant density is available, such that the analysis of the underlying dynamics has to be based on sampling specific realisations of the dynamics, i.e. on a finite amount of numerical data which usually is available in the form of *time series* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ where $\mathbf{x}_i$ represents the state of the system at time i, and n denotes the finite length of the time series; $\mathbf{x}_i$ may be univariate or multivariate. While being true also for many systems which are accessible to simulation, this is the standard situation for systems which are investigated through experiments or observations.

From such time series trajectories in a state space can be constructed, either directly or by application of an embedding technique, such as time-delay embedding [Takens, 1981]. We therefore have the situation of a twofold sampling: the invariant density is sampled by trajectories and the trajectories themselves are further sampled by isolated vectors, which are the result of the inability to store more than a finite amount of numerical information in digital hardware.

It is obvious that in most cases this approach to sampling a probability density will not be optimal. The most severe problem results from the fact that the reconstructed vectors on the trajectories will typically display *dynamical correlations*, whence the set of these vectors does not constitute an independent sample from the underlying density. Such correlations are known in particular to disturb the estimation of fractal dimensions, such as correlation dimension [Theiler, 1986]. The concept of correlation dimension was originally introduced in order to quantify purely *spatial* correlations, as opposed to dynamical (i.e. temporal) correlations.

It is the purpose of this paper to contribute to the investigation of these dynamical correlations by applying correlation dimension estimation to simulated time series, and to decompose these correlations into two constituents which correspond to the two levels of sampling mentioned above. For this purpose we will start by applying correlation dimension estimation to limit cycles and N-tori. As a side result this will lead us to a better understanding of the somewhat disturbing fact that correlation dimension estimation does not work well for N-tori [Jedynak *et al.*, 1994].

2 Correlation Dimension Estimation

We shall now very briefly review the estimation of correlation dimension according to the widely used Grassberger-Procaccia algorithm [Grassberger & Procaccia, 1983] (GPA).

We start from a given set of n vectors in a d-dimensional state space

$$\mathbf{x}(it_s) =: \mathbf{x}_i \ , \quad i = 1, \ldots, n \ . \tag{1}$$

These vectors may either directly describe the state of a system, or they may result from a time-delay embedding of a univariate time series x_i, as given by

$$\mathbf{x}_i = (x_i, x_{i-\tau}, x_{i-2\tau}, \ldots, x_{i-(d-1)\tau}) \ , \tag{2}$$

where τ and d denote the time delay and the embedding dimension, respectively. From these vectors the *correlation sum* is formed:

$$C(r) = \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} H(r - \|\mathbf{x}_i - \mathbf{x}_j\|) \ , \tag{3}$$

where $H(x) = 1$ for positive x and $H(x) = 0$ otherwise. $\|.\|$ denotes maximum norm, though other norms could also be employed. $C(r)$ estimates the probability of finding two vectors in the set which are separated by a distance not larger than a radius r.

For sufficiently small r the correlation sum is expected to display a scaling

$$C(r) = a\, r^{d_c} \quad , \quad a \text{ const.}, \tag{4}$$

whence the correlation dimension d_c can be obtained by

$$d_c = \lim_{r\to 0} d_c(r) = \lim_{r\to 0} \frac{\partial \log C(r)}{\partial \log r} \quad . \tag{5}$$

The derivative is carried out numerically and yields a "scale-dependent dimension" $d_c(r)$. The details of our implementation are the same as in [Galka *et al.*, 1998]. In the usual case of only a finite amount of finite-precision data being available the limit cannot be formed, therefore estimates of dimension inevitably will have to reflect scaling behaviour at finite length scales.

One then tries to detect a *scaling region*, i. e. a radius interval of sufficient length at small r where $d_c(r)$ remains approximately constant, and uses this value as an estimate of d_c. However, in this paper we will prefer to present the complete curve $d_c(r)$ as the result of the dimension estimation.

Although this algorithm seems to be straightforward at first sight, it is nevertheless susceptible to various artifacts and systematic errors; reviews of known problems are given in [Theiler, 1990a] and [Galka, 2000]. Statistical errors have been investigated by [Theiler, 1990b] and [Frank *et al.*, 1996]. Clearly the main source of systematic error results from impossibility to perform the limit $r \to 0$, but we shall assume that the data sets we analyse are large enough to make sufficiently small scales accessible.

Soon after the introduction of the GPA it was observed that serious artifacts may result from dynamical correlations on reconstructed attractors [Theiler, 1986]; by the term 'dynamical correlation' we refer to correlations within the set of reconstructed vectors, which result both from using a constant sampling frequency (i.e. periodic sampling) and from the alignment of reconstructed vectors on trajectories. Whereas for specific time series sampled from periodic and quasiperiodic dynamics such correlations will persist over arbitrarily long time spans, they can be expected to decay for time series from chaotic dynamics, as well as from stochastic dynamics. As a tool to detect and quantify such correlations space-time-separation plots have been proposed [Provenzale *et al.*, 1992]; when applied to time series from chaotic dynamics they tend to reveal sizable correlations over surprisingly long time spans [Smith, 1997].

3 Limit Cycles

Assume that we are observing the behaviour of a given dynamical system by periodically sampling its state $\mathbf{x}(t)$, i.e. only those values of $\mathbf{x}(t)$ are stored for which t is an integer multiple of a fixed sampling time t_s, as indicated in Eq. (1). Obviously by periodic sampling a spurious frequency $1/t_s$, the *sampling frequency*, is introduced into the time series, which usually does not correspond to any inherent properties of the dynamical system.

For the case of limit cycles we can model this situation by forming a set of $n = 10,000$ two-dimensional vectors which sample a circle:

$$\mathbf{x}_i = \left(\sin \frac{2\pi i}{n}, \cos \frac{2\pi i}{n} \right) \quad , \quad i = 0, \ldots, (n-1) \; . \tag{6}$$

We could regard this circle as the state space representation of a harmonic oscillator; then this choice of vectors corresponds to periodic sampling.

The correct dimension of this object is obviously $d_c = 1$; if we apply correlation dimension estimation, as described in Sec. 2, to this set of vectors, we obtain the result shown in Fig. 1 (solid line). At large radius we see a peak (known as *overfolding peak*) which can easily be explained [Theiler, 1988a] and does not matter for our discussion since correlation dimension is defined for the limit of small radius; but at small radius we see after a reasonable scaling region at the correct dimension another rise of the estimate, followed by an oscillation of rapidly increasing amplitude and finally even a divergence of the radius-dependent estimate. Oscillations of $d_c(r)$ also occur for many fractals displaying lacunarity [Theiler, 1988b]; but those oscillations usually have constant amplitude in contrast to the increasing amplitude of Fig. 1.

There is a simple way to explain the occurrence of these oscillations and spuriously high estimates for periodic sampling: Since the trajectory locally will look like a straight line, all sampled vectors will have approximately the same distance to their first temporal neighbours (both forward and backward in time), approximately twice that distance to their second temporal neighbours, etc. But then in the set of all distances between pairs of vectors these values occur much more often than intermediate values. Upon reaching such a value of the radius the correlation sum increases rapidly, whence its slope can become very large within a short interval. The effect would be even more pronounced, if Euclidean norm had been used.

It is possible to remove such artifacts by a simple modification of the sampling procedure: we replace periodic sampling by *random sampling*, i.e. the times at which $\mathbf{x}(t)$ is stored, are chosen completely randomly instead of being integer multiples of a fixed sampling time t_s.

For the circle of Eq. (6) we therefore draw $n = 10,000$ random numbers from a uniform distribution between 0 and 1, and insert them for i/n into Eq. (6). The dimension estimation of this set of vectors yields the dashed line in Fig. 1. It can clearly be seen that the scaling region is much longer now and that the pronounced oscillation and the divergence of the estimate are absent.

This result illustrates the benefit of random sampling as compared to traditional periodic sampling. It also agrees well with the well-known fact that the resolution of the correlation sum with respect to small length scales will be worst for the case of sampling by a periodic array of vectors [Grassberger *et al.*, 1991].

It is obvious that vectors resulting from periodically sampling a circle will form a highly correlated set; in fact, the Grassberger-Procaccia algorithm for correlation dimension estimation was not designed for such correlated sets, rather it assumes a set of vectors sampled independently from the invariant density. By random sampling we have removed the correlation and thereby made limit cycles accessible for correlation dimension estimation.

4 N-tori

Limit cycles are somewhat extreme examples of dynamical behaviour since their invariant measure is completely concentrated on one closed trajectory. More interesting attractors can be constructed from the sum of several periodic processes with mutually incommensurate frequencies ω_i:

$$x(t) = \sum_{i=1}^{N} c_i \sin(\omega_i t + \phi_i) \quad . \tag{7}$$

Time series generated by such systems (known as *quasi-periodic* time series) will, after time-delay embedding (see Eq. (2)), sample N-tori, i.e. geometrical objects the local dimensionality of which is precisely N, provided the dimension of the embedding space was chosen sufficiently large. In the case of N-tori an embedding dimension of at least $m = 2N$ is necessary in order to obtain a reconstructed attractor without self-intersections.

Jedynak *et al.* [1994] have reported the failure of correlation dimension estimation applied to 32000 points sampled from a 5-torus. Although there is no specific need to estimate correlation dimensions from quasi-periodic time series, since their structure can be analysed more conveniently by their power spectrum, this failure nevertheless remains disturbing and deserves further attention.

However, since a dimension of $d_c = 5$ represents indeed a fairly high value, requiring a large data set size in order to resolve sufficiently small scales [Galka *et al*, 1998], here we prefer to consider the case of a 3-torus, given by inserting $N = 3$, $c_1 = 12000$, $c_2 = 10000$, $c_3 = 8000$, $\omega_1 = \sqrt{2}$, $\omega_2 = \sqrt{3}$ and $\omega_3 = \sqrt{5}$ into Eq. (7); the choice of phases ϕ_i is irrelevant. The values of the amplitudes c_i were chosen such that the resulting time series would fit well into a 16bit data format.

We create a time series of $n = 10^5$ points length using a fixed sample time of $t_s = 0.15$ (which is a reasonable choice for a sufficient but not too dense sampling of the trajectory) and embed it by time-delay embedding in a 10-dimensional embedding space, using a time delay of $\tau = 6t_s$; this value provides a good unfolding of the torus, as can be checked by the ILD-criterion of Buzug and Pfister [1992]. We then apply correlation dimension estimation to the resulting set of vectors; the result is shown in the left panel of Fig. 2 (solid line).

In the figure we see a similar behaviour as in Fig. 1: at large radius there is an overfolding peak, then the estimate converges towards the correct value

$d_c = 3$, starts to oscillate with rapidly growing amplitude and finally diverges. Again we try to remove the dynamical correlations resulting from periodic sampling by switching to random sampling. Since the geometrical object under investigation, the 3-torus, results from time-delay embedding using a constant time delay, the random sampling obviously has to be applied *after* embedding the periodically sampled time series, therefore we might also speak of *random resampling*. For this purpose 10^5 times are chosen randomly from the same total time interval covered by the original time series, and the corresponding resampled vectors are obtained from the embedded trajectories by appropriate interpolation. Applying correlation dimension estimation to this set of vectors yields the dashed curve shown in the left panel of Fig. 2.

Now we do not see a divergence of the estimate, but one period of a slower oscillation and then a convergence to $d_c = 1$. In order to understand this incorrect result we have to remember that there are two levels of sampling present in the embedded object; by random resampling we have removed the characteristic dynamical correlations resulting from the second level, as quantified by the difference between the two curves in the left panel of Fig. 2. But unlike with the limit cycle of the Sec. 3 this does not suffice for the case of a 3-torus, since this 3-dimensional object is still sampled by trajectories. Since these trajectories locally are indeed one-dimensional, the result $d_c = 1$ is actually correct. The total amount of dynamic correlations has been reduced by random resampling, but still the sampled vectors are confined to certain trajectories and therefore correlated. These trajectories fill the 3-dimensional object in a regular pattern, such that the presence of local periodicity perpendicular to the trajectories has to be expected; we presume that this periodicity is reflected in the slow oscillation of the dashed curve in the left panel of Fig. 2.

The spurious underestimation of correlation dimension due to sampling an invariant density by trajectories has been described already in 1986 by Theiler [1986]; by switching from periodic sampling to random sampling this effect is even enhanced since the resolution of the correlation sum with respect to small length scales will be increased. For periodic sampling there was a minimum time of t_s between temporally consecutive vectors on the trajectory, but for random sampling many pairs of temporally consecutive vectors will be much closer in time. This situation clearly invites a pronounced trajectory bias.

Theiler [1986] has proposed a simple correction method against trajectory bias: Simply omit all those distances $\|\mathbf{x}_i - \mathbf{x}_j\|$ from the sum in Eq. (3) for which $\mathbf{x}_i$ and $\mathbf{x}_j$ are too close together in time, i. e. for which $i - j < W$ with a fixed integer parameter W. The absence of the correction corresponds to $W = 1$.

However, this correction fails for the case of the 3-torus; regardless of how large we choose the parameter W, the convergence to $d_c = 1$ persists. The simple reason for this result is that in a quasi-periodic time series dynamic correlations will be present without any decay over arbitrarily long distances in time, such that no choice of W would be large enough to remove them. Only for chaotic time series we can expect to find a decay of dynamic correlations with time, such that a finite W would suffice.

If we intend to completely remove the dynamic correlations responsible for trajectory bias we have to introduce a random element already in the creation of the 3-torus. This can be done by inserting randomly chosen phases ϕ_i, $i = 1, 2, 3$, into Eq. (7) and creating a time series just long enough to reconstruct one single vector, i.e. of length $n_v = \tau(m-1)+1$. In this way for each vector a different set of random phases is chosen and a new short time series is created. If we create 10^5 vectors by this method, using the same embedding parameters as before, and estimate the correlation dimension of the resulting set, we obtain the result shown by the solid line in the right panel of Fig. 2. As can be seen in the figure, this set of vectors yields the correct result $d_c = 3$.

5 Confidence Profiles for Dimension Estimates

It is desirable to provide estimates for the statistical error of dimension estimates and thereby assess the significance of scaling regions or other patterns, such as oscillations of the estimates. A simple method to quantify statistical errors is given by repeating the estimation for several realisations sampled from the system, such that a distribution of results can be obtained. In spectral estimation a similar idea is employed as *block-averaging*; the well-known statistical technique of *bootstrapping* is also closely related.

The bootstrap approach has been applied to correlation dimension estimation only rarely, notable exceptions are [Judd, 1994] and [Mikosch & Wang, 1995]. In [Galka *et al.*, 1998] a simple bootstrap approach was employed for an analysis of data set size requirements in high-dimensional systems.

In systems involving a random element it is easy to generate a set of different realisations by randomly choosing different values for this element. For chaotic systems this random element is given by the initial conditions, for stochastic systems it could be given by the innovations, and for N-tori we can employ the sets of phases ϕ_i as such random element. Changing the phases would not make sense in the context of time series, but if we are only interested in invariant densities in state spaces, these sets of phases can be regarded as equivalent to initial conditions or innovations.

Since the correlation sum according to Eq. (3) is known to be an unbiased estimator of the correlation integral (i.e. the limit for infinitely long time series) [Denker & Keller, 1986], [Cutler, 1994], we estimate the correlation sum from each realisation, average over the results, and then form the logarithmic derivative according to Eq. (5). In this way the solid curve of Fig. 2 was obtained, using $N_r = 100$ realisations of the 3-torus.

However, the standard deviations $\sigma(r)$ cannot be obtained in the same way, since the pointwise averages (i.e. for fixed radius) of the $C(r)$ do not contain any information about local slopes. So now we have *first* to apply Eq. (5) to the correlation sum of each individual realisation, and *then* to form the pointwise standard deviations of the distribution of the resulting $d_c(r)$; it is also possible to form averages in this step, but they will show a bias, especially at small radius, where due to poor statistics the distribution becomes broad.

The standard deviations $\sigma(r)$ resulting from this approach describe the error of the result of analysing an individual realisation; as it is well known the error of the mean of the corresponding distribution, i.e. the average curve $d_c(r)$, is much smaller and given by $\sigma(r)/\sqrt{N_r}$. By multiplying with 2 we obtain 95% confidence profiles, which we can use for assessing the significance of patterns in $d_c(r)$. If we apply this approach to the dimension analysis of the 3-torus with random phases, we obtain the thin dash-dotted lines shown in the right panel of Fig. 2; they confirm well the significance of the observed scaling region.

Closer analysis reveals that $\sigma(r)$ depends on r by a power law $\sigma \propto r^{\alpha}$ with exponent $\alpha \approx -d_c/2$, as should be expected from [Galka *et al.*, 1998].

Finally it should be admitted that in the case of correlation dimension estimation our approach to estimating statistical errors is somewhat inefficient. Since not vectors in state space themselves, but the distribution of the distances between pairs of vectors is analysed by the Grassberger-Procaccia algorithm, by generating 100 realisations and evaluating only distances between pairs of vectors within each realisation, but not across realisations, we leave much information unused. This does not constitute a problem as long as a sufficient amount of data is available, but may become relevant for analyses of experimental or observational data.

6 Lorenz System

We will now investigate in which way the two levels of sampling discussed above and the corresponding dynamical correlations are also present in attractors reconstructed from chaotic time series.

To this end we consider the well-known Lorenz equations

$$
\begin{aligned}
\dot{x}(t) &= \sigma(y(t) - x(t)) \\
\dot{y}(t) &= rx(t) - y(t) - x(t)z(t) \\
\dot{z}(t) &= x(t)y(t) - bz(t)
\end{aligned}
\tag{8}
$$

(where the parameters are chosen as $\sigma = 10.0$, $r = 28.0$ and $b = 8/3$); they represent a deterministic nonlinear system which produces nonperiodic trajectories evolving according to an invariant density which is believed to have a noninteger correlation dimension of $d_c \approx 2.06$ [Grassberger & Procaccia, 1983].

A standard fourth-order Runge-Kutta integrator with integration step size $t_i = 10^{-3}$ is used; initial transients are discarded. Again we create $N_r = 100$ time series, each of $n = 10^5$ points length, using a fixed sample time of $t_s = 25t_i$ (which again is a reasonable choice for good sampling of the trajectory); we keep only the x-coordinate, while discarding y and z, and embed it by time-delay embedding in a 20-dimensional embedding space, using a time delay of $\tau = 5t_s$. The rather high value of the embedding dimension was chosen in order to make the effects which we intend to demonstrate more prominent. In the same way as in the case of the 3-torus we then apply correlation dimension estimation to the resulting sets of vectors; the average of the resulting $d_c(r)$ is shown in the upper left panel of Fig. 3 (solid line); the corresponding 95% confidence profiles are shown by thin dash-dotted lines.

In the figure we see a large overfolding peak at large radius and a somewhat irregular oscillation towards smaller radius; at radius values smaller than those shown in the figure the distribution of the results of individual realisations becomes very broad, such that the true behaviour of $d_c(r)$ cannot be inferred.

The oscillation takes place mainly in the interval $2.05 < d_c < 2.08$ which could be regarded as still a sufficiently small error interval of an estimate of correlation dimension; however, with shorter time series the oscillation is found to be considerably more pronounced.

If we repeat the same analysis with random resampling of trajectories, we obtain the result shown in the upper left panel of Fig. 3 by a dashed line; here the confidence profiles follow closely the average curve. Again the random resampling removes the oscillation due to periodic sampling, but causes also a spurious convergence to $d_c = 1$ due to enhanced trajectory bias. It is remarkable that this bias can already be removed completely by applying a Theiler correction with $W = 2$ instead of $W = 1$.

It is frequently claimed that by choosing W sufficiently high all problems with dynamical correlations could be removed; Grassberger [1990] recommends to "*be very generous with the Theiler correction parameter*". But if we increase W to much larger values ($W = 100$ and $W = 1000$), we obtain the results shown in the right panels of Fig. 3 by solid lines. It can be seen that the oscillation is reduced (especially at larger radius), but not removed. The reason is probably given by the fact that correlation dimension estimation by GPA is based on evaluating distances between each vector (chosen as a reference vector) and the vectors within a small neighbourhood around that reference vector. The Theiler correction succeeds in removing the dynamical correlations on that part of the trajectory where the reference vector itself is located, but not those on other parts or other trajectories which are distant in time, but close in space. They will still be periodically sampled and hence give rise to some oscillations of the average $d_c(r)$.

If we apply both random resampling and the Theiler correction with large W to the same analysis, we obtain the results shown in the right panels of Fig. 3 by dashed lines. The oscillations are further reduced and the error intervals are much smaller as compared to the case without random resampling. Any remaining oscillations will probably be due to weak periodicity effects perpendicular to the trajectories. At still smaller radius these estimates sometimes tend to deviate towards large dimensions, but in the radius interval shown in the figure they evidently succeed in reducing the dynamical correlations much more than it is possible by pure Theiler correction.

Finally we show the case of the dimension analysis of a set of vectors with absolutely no dynamical correlations. We create sets of $n = 10^5$ vectors in a similar way as done in Sec. 4 for the 3-torus, i.e. choosing new random initial conditions for each vector, integrating the Lorenz system until transients have died out, and then sampling (starting at a randomly chosen time) only $n_v = \tau(m-1) + 1$ values (where $\tau = 5$ and $m = 20$). The result of applying correlation dimension estimation to $N_r = 100$ sets of vectors generated by this approach is shown in the lower left panel of Fig. 3. As can be seen in the figure, the result is an almost perfect scaling region and a very narrow distribution of individual realisations. This curve provides the yardstick for comparison with the other curves shown in Fig. 3.

7 Discussion and Conclusion

In this paper we have proposed the idea that dynamical correlations in invariant densities reconstructed from a finite number of measurements of the state of a deterministic dynamical system can be decomposed into two contributions, as represented by the two levels of sampling which are given by sampling the invariant density by isolated trajectories and furthermore sampling these trajectories *periodically* by isolated vectors; and we have presented results of several numerical experiments supporting this view.

We have shown that each of these two contributions may lead to characteristic artifacts in the results of correlation dimension estimation by the Grassberger-Procaccia algorithm; therefore dimension analysis provides a tool to investigate these dynamical correlations and to test for their presence. Such investigations do also provide new insight into the origin of artifacts and spurious results of dimension analysis and into new ways to avoid or at least reduce their effects. In particular we have discussed the problematic case of quasiperiodic time series, i.e. N-tori, and demonstrated that in order to create an embedding without dynamical correlations we need as many independent time series as vectors to be employed for sampling the torus. Otherwise a correct dimension analysis is impossible. By this result we provide an answer to the observations of Jedynak *et al.* [1994] on the difficulty of estimating correlation dimensions from N-tori. We have shown that this difficulty is not a result of lack of sufficiently large data sets (as assumed by Jedynak *et al.* [1994]), but of dynamical correlations, hence it will persist for arbitrarily large data sets.

The situation is more favourable in the case of chaotic attractors, since due to positive Lyapunov exponents the dynamical correlations will gradually decay with time; still within local neighbourhoods of state space they pose severe problems.

Periodical sampling of trajectories will typically lead to an oscillation of the scale-dependent correlation dimension estimate, which will be particularly pronounced for rather low-dimensional point sets in high-dimensional embedding spaces; but it should be stressed that these oscillations are not an artifact of time-delay embedding, since we have also succeeded in detecting them in completely known Lorenz attractors (i.e. in sets of vectors (x, y, z) instead of time-delay reconstructed vectors from Eq. (2)) [Galka, 2000].

We have proposed random sampling (or random resampling, in the case of time-delay reconstructed attractors) as a simple and efficient way to remove this oscillation; this approach provides an improved sampling scheme of invariant densities in state spaces, and may therefore prove to be useful also for other purposes. It can also be regarded as a straightforward way to remove the sampling frequency from the state space representation of the data; it may be desirable to remove this frequency, since it does not correspond to any inherent properties of the underlying dynamical system.

By random (re)sampling the second level of sampling is completely removed, and only the effects of the first level dominate the structure of dynamical correlations; also the opposite case is possible, e.g. in attractors generated by discrete maps, such as the Hénon attractor, we would expect to find similar spurious correlations from sampling the invariant density by isolated points *with finite precision.* These points would then be positioned on a grid corresponding to the finite precision and thereby introduce spurious correlations, although no trajectories would be involved.

Random (re)sampling has the disadvantage of intensifying the artifacts of the second level of sampling, i.e. enhancing *trajectory bias*. A simple remedy against this effect has been introduced by Theiler already in 1986 [Theiler, 1986], and his correction should be applied together with random (re)sampling. This combined approach will provide satisfactory results for chaotic attractors, whereas it fails in the case of N-tori.

We have shown in this paper that the common view that the Theiler correction alone was able to remove all artifacts of dynamical correlations from the results of correlation dimension estimation, provided the parameter W is chosen sufficiently large, is incorrect. We would like to mention that the only way to remove reliably all dynamical correlations by the Theiler correction is given by choosing W as half the number of vectors, i.e. omitting *all* distances, which would obviously make no sense. Choosing a very large W will approach this limit, and we have indeed observed that the quality of the scaling regions deteriorates if W is increased beyond a certain threshold. The best way to avoid any dynamical correlations is certainly given by creating independent vectors, as demonstrated for the 3-torus and the Lorenz attractor; but clearly this approach will be impossible in the usual situation that the data is available only as a time series of limited length. In such situations it is also much more difficult to provide confidence profiles for dimension estimates, since frequently not sufficient data is available for applying bootstrap techniques. When analysing single realisations the artifacts investigated in this paper will typically be buried by statistical errors of larger amplitude.

Therefore we presume that new investigations into dynamical correlations on reconstructed attractors, as provided by this paper, and new approaches to reducing correlations, such as random (re)sampling, may prove to be useful for correlation dimension estimation and related analyses of numerical data from experiments and observations. We intend to present applications to experimental data from Taylor-Couette flow in a forthcoming paper.

Acknowledgements

AG gratefully acknowledges support by the Deutsche Forschungsgemeinschaft (DFG) under contract GA 673/1-1. This work has benefitted considerably from discussions with Leonard Smith and from the comments and suggestions of an anonymous referee.

References

Buzug, T. & Pfister, G. [1992] "Optimal delay time and embedding dimension for delay-time coordinates by analysis of the global and local dynamical behavior of strange attractors," *Phys. Rev. A* **45**, 7073–7084.

Cutler, C. D. [1994] "A theory of correlation dimension for stationary time series," *Phil. Trans. R. Soc. London A* **348**, 343–355.

Denker, M. & Keller, G. [1986] "Rigorous statistical procedures for data from dynamical systems," *J. Stat. Phys.* **44**, 67–93.

Frank, M., Keller, G. & Sporer, R. [1996] "Practical implementation of error estimation for the correlation dimension," *Phys. Rev. E* **53**, 5831–5836.

Galka, A., Maaß, T. & Pfister, G. [1998] "Estimating the dimension of high-dimensional attractors: A comparison between two algorithms," *Physica D* **121**, 237–251.

Galka, A. [2000] *Topics in Nonlinear Time Series Analysis - With Implications for EEG Analysis* (World Scientific Publishing Company, Singapore)

Grassberger, P. [1990] "An optimized box-assisted algorithm for fractal dimensions," *Phys. Lett. A* **148**, 63–68.

Grassberger, P. & Procaccia, I. [1983] "Measuring the strangeness of strange attractors," *Physica D* **9**, 189–208.

Grassberger, P., Schreiber, T. & Schaffrath, C. [1991] "Nonlinear time sequence analysis," *Int. J. Bifurcation Chaos* **1**, 521–547.

Jedynak, A., Bach, M. & Timmer, J. [1994] "Failure of dimension analysis in a simple five-dimensional system," *Phys. Rev. E* **50**, 1770–1780.

Judd, K. [1994] "Estimating dimension from small samples," *Physica D* **71**, 421–429.

Kantz, H. & Schreiber, T. [1997] *Nonlinear time series analysis* (Cambridge University Press, Cambridge).

Mikosch, T. & Wang, Q. [1995] "A Monte Carlo method for estimating the correlation exponent," *J. Stat. Phys.* **78**, 799–813.

Provenzale, A., Smith, L. A., Vio, R. & Murante, G. [1992] "Distinguishing between low-dimensional dynamics and randomness in measured time series," *Physica D* **58**, 31–49.

Smith, L. A. [1997] "The maintenance of uncertainty," In: Castagnoli, C. & Provenzale, A. (eds.), *Proceedings of the International School of Physics "Enrico Fermi"*, vol. CXXXIII, pp. 177–246.

Takens, F. [1981] "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*, eds. Rand, D. A. & Young, L.-S., Lecture Notes in Mathematics (Springer, Berlin) **898**, pp. 366–381.

Theiler, J. [1986] "Spurious dimensions from correlation algorithms applied to limited time series data," *Phys. Rev. A* **34**, 2427–2432.

Theiler, J. [1988a] *Quantifying chaos: practical estimation of the correlation dimension*, Ph.D. thesis, California Inst. of Technology, Pasadena.

Theiler, J. [1988b] "Lacunarity in a best estimator of fractal dimension," *Phys. Lett. A* **133**, 195–200.

Theiler, J. [1990a] "Estimating fractal dimension," *J. Opt. Soc. Am. A* **7**, 1055–1073.

Theiler, J. [1990b] "Statistical precision of dimension estimators," *Phys. Rev. A* **41**, 3038–3051.

Figure captions

Fig. 1: Correlation dimension estimate vs. radius for a circle sampled by 10^4 vectors by periodic sampling (solid line) and by random sampling (dashed line). The dotted line denotes the correct dimension $d_c = 1$.

Fig. 2: Correlation dimension estimate vs. radius for a 3-torus (left panel) sampled by 10^5 vectors by periodic sampling (solid line) and by random sampling (dashed line); and for 100 realisations of the same 3-torus, each composed of 10^5 completely independent vectors (right panel). The right panel shows average estimates (solid line) and statistical errors of the average (approximate 95% confidence profiles; thin dash-dotted lines). Radius units correspond to a 16bit data format; the dotted lines denote the correct dimension $d_c = 3$.

Fig. 3: Correlation dimension estimate vs. radius for 100 realisations of the Lorenz attractor, reconstructed by time-delay embedding from the x-component, using time series of 10^5 points length. Average estimates for periodic sampling (solid lines) and for random resampling (dashed lines) are shown; statistical errors of the average (approximate 95% confidence profiles) are denoted by thin dash-dotted lines. Theiler correction parameter: $W = 1$ (upper left panel), $W = 100$ (upper right panel) and $W = 1000$ (lower right panel). Lower left panel shows the result for 100 realisations of 10^5 completely independent vectors. Radius units correspond to a 16bit data format; the dotted lines denote the (supposedly) correct dimension $d_c = 2.06$. Note that the vertical axis does not start at a dimension estimate of zero, but of 1.95.

Figure 1

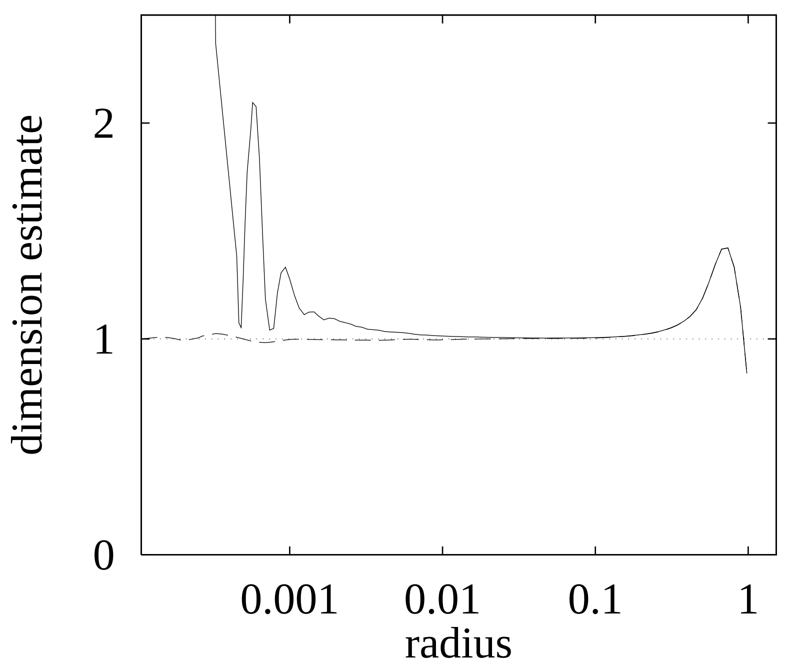


Figure 2

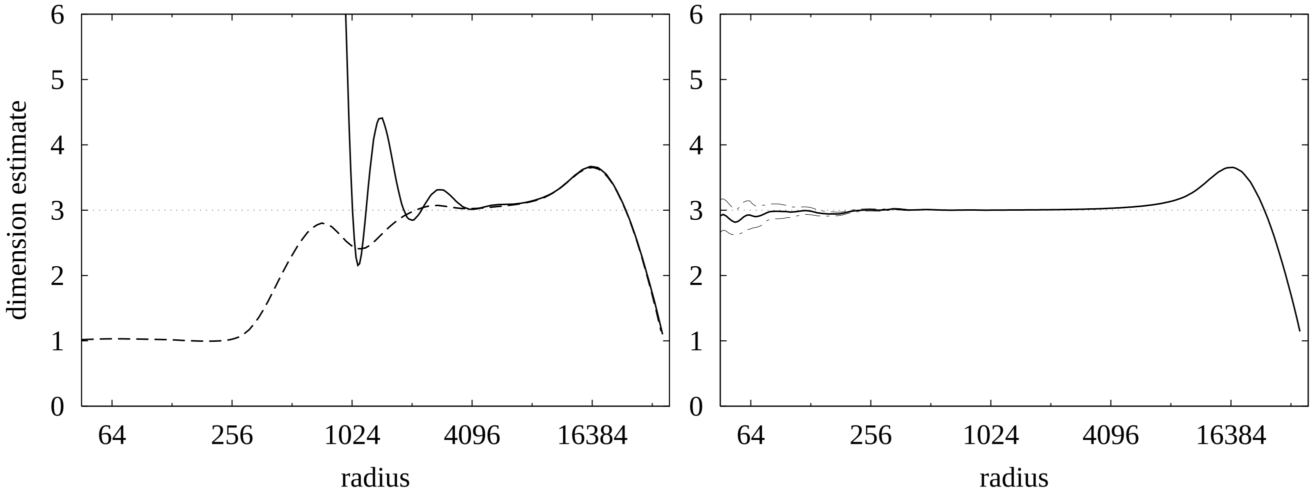
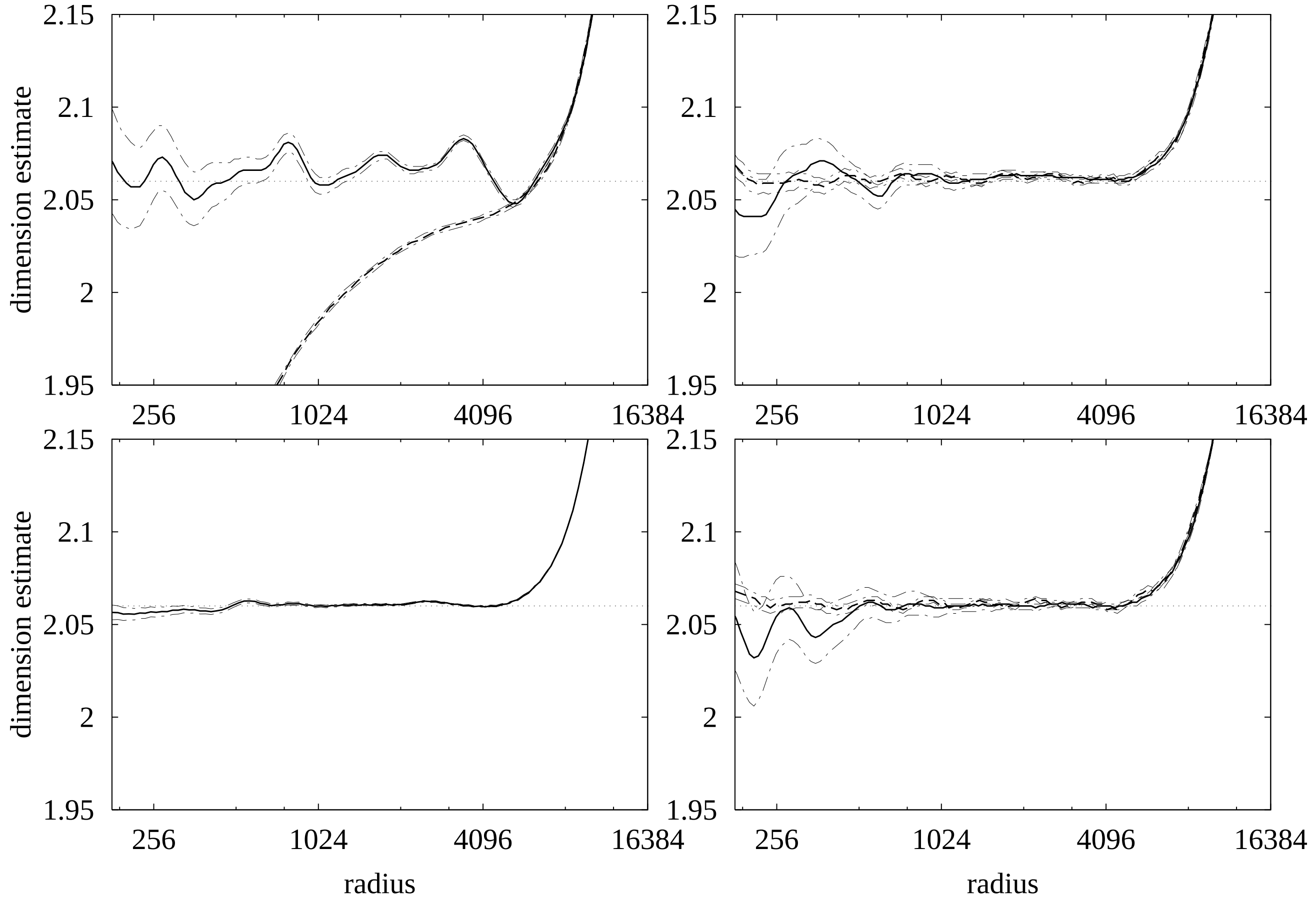

dimension estimate
0
1
2
3
4
5
6
64
256
1024
4096
16384
radius
0
1
2
3
4
5
6
64
256
1024
4096
16384
radius

Figure 3


2.15
2.1
2.05
2
1.95
dimension estimate
256
1024
4096
16384
radius