Improving the false nearest neighbors method with graphical analysis

T. Aittokallio,[1,2] M. Gyllenberg,[1] J. Hietarinta,[3] T. Kuusela,[1,3] and T. Multamäki[3]
[1]*Department of Applied Mathematics, University of Turku, FIN-20014 Turku, Finland*
[2]*Turku Centre for Computer Science, Lemminkäisenkatu 14 A, FIN-20520 Turku, Finland*
[3]*Department of Physics, University of Turku, FIN-20014 Turku, Finland*

(Received 21 December 1998)

We introduce a graphical presentation for the false nearest neighbors (FNN) method. In the original method only the percentage of false neighbors is computed without regard to the distribution of neighboring points in the time-delay coordinates. With this presentation it is much easier to distinguish deterministic chaos from noise. The graphical approach also serves as a tool to determine better conditions for detecting low-dimensional chaos, and to get a better understanding on the applicability of the FNN method.
[S1063-651X(99)07507-8]

PACS number(s): 05.45.Tp, 07.05.Kf

I. INTRODUCTION

One of the main tasks of time series analysis is to determine from a given time series the basic properties of the underlying process, such as nonlinearity, complexity, chaos, etc. Among the most widely used approaches is state space reconstruction by time delay embedding [1]. After this step has been taken one can calculate correlation dimensions, various entropy quantities and estimates for Lyapunov exponents. The crucial problem is how to select a minimal embedding dimension for the pseudo-phase-space. If the embedding dimension is too small, one cannot unfold the geometry of the (possible strange) attractor, and if one uses a too high embedding dimension, most numerical methods characterizing the basic dynamical properties can produce unreliable or spurious results.

The false-nearest-neighbors (FNN) algorithm [2–4] is one of the tools that can be used to determine the number of time-delay coordinates needed to reconstruct the dynamics. In this method one forms a collection

$$\mathbf{y}(k)=[x(k),x(k+1),\ldots,x(k+d-1)] \tag{1.1}$$

of d-dimensional vectors for a given time delay (here normalized to 1), $x(1),x(2),\ldots,x(N)$ is a scalar time series. If the number d of time-delay coordinates in Eq. (1.1) is too small, then two time-delay vectors $\mathbf{y}(k)$ and $\mathbf{y}(l)$ may be close to each other due to the projection rather than to the inherent dynamics of the system. When this is the case, points close to each other may have very different time evolution, and actually belong to different parts of the underlying attractor.

In order to determine the sufficient number d of time-delay coordinates one next looks at the nearest neighbor of each vector (1.1) with respect to the Euclidean metric. We denote the nearest neighbor of $\mathbf{y}(k)$ by $\mathbf{y}(n(k))$. We then compare the "$(d+1)$"st coordinates of $\mathbf{y}(k)$ and $\mathbf{y}(n(k))$, e.g., $x(k+d)$ and $x(n(k)+d)$. If the distance $|x(k+d)-x(n(k)+d)|$ is large the points $\mathbf{y}(k)$ and $\mathbf{y}(n(k))$ are close just by projection. They are false nearest neighbors and they will be pulled apart by increasing the dimension d. If the distances $|x(k+d)-x(n(k)+d)|$ are predominantly small, then only a small portion of the neighbors are false and d can be considered a sufficient embedding dimension.

In the FNN algorithm [2–4] the neighbor is declared false if

$$\frac{|x(k+d)-x(n(k)+d)|}{\|\mathbf{y}(k)-\mathbf{y}(n(k))\|}>R_{tol}, \tag{1.2}$$

or if

$$\frac{\|\mathbf{y}(k)-\mathbf{y}(n(k))\|^2+\{x(k+d)-x(n(k)+d)\}^2}{R_A^2}>A_{tol}^2, \tag{1.3}$$

where

$$R_A^2=\frac{1}{N}\sum_{k=1}^{N}[x(k)-\overline{x}]^2, \tag{1.4}$$

and $\overline{x}$ is the mean of all points. The parameter R_{tol} in the first threshold test (1.1) is fixed beforehand, and in most studies it has been set to $10-20$. The second criterion (1.3) was proposed in order to provide correct diagnostics for noise and usually one takes $A_{tol}\approx 2$. If this test fails, then even the ($d+1$-dimensional) nearest neighbors themselves are far apart in the extended $d+1$ dimensional space and should be considered false neighbors.

Using tests (1.2) and (1.3) one can check all d-dimensional vectors in the data set, and compute the percentage of false nearest neighbors. By increasing the dimension d this percentage should drop to zero or to some acceptable small number. In that case the embedding dimension is large enough to represent the dynamics.

This method works quite well with noise-free data, and the percentage of false neighbors does not depend on the number of data points if it is sufficient. However, if data is corrupted with noise, the percentage of false nearest neighbors for a given embedding dimension increases as the amount of data is increased, and therefore a longer time se-

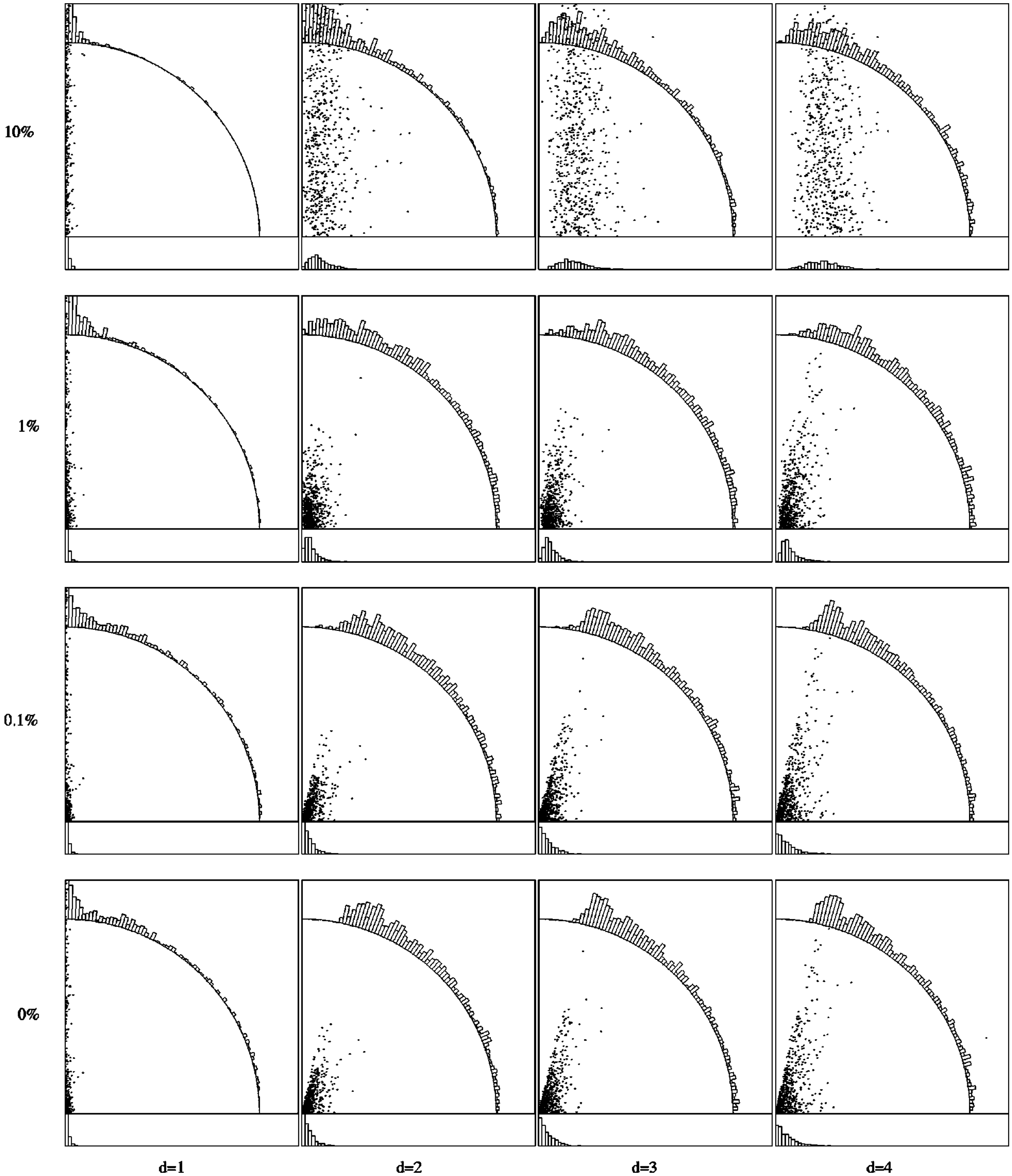

FIG. 1. In each subfigure we have plotted the target distance R_Δ (vertical axis, range 0 to 0.072) as a function of the nearest neighbor distance $\tilde{R}_d$ (horizontal axis, range 0 to 0.072) for the Henon system (the dimension of the attractor is 1.26). In each subfigure we have also given two distributions in the form of histograms: R_d distribution on the bottom part of the graph, and the radial distribution on the quarter arc. The total number of data points is 1000. The rows correspond to indicated noise levels, the columns to indicated embedding dimensions.

ries leads to erroneous false nearest neighbors as a result of noise corruption rather than of an incorrect embedding dimension. One possible solution to this problem is to modify the threshold test (1.2) to account for additional noise effects. For example, instead of test (1.2) the threshold could be determined by [5]

$$\frac{|x(k+d)-x(n(k)+d)|}{\|\mathbf{y}(k)-\mathbf{y}(n(k))\|}>R_{tol}+\frac{2\epsilon R_{tol}\sqrt{d}+2\epsilon}{\|\mathbf{y}(k)-\mathbf{y}(n(k))\|}. \tag{1.5}$$

Here the new parameter ϵ must be chosen properly. Obviously the optimal value for ϵ should be determined by the

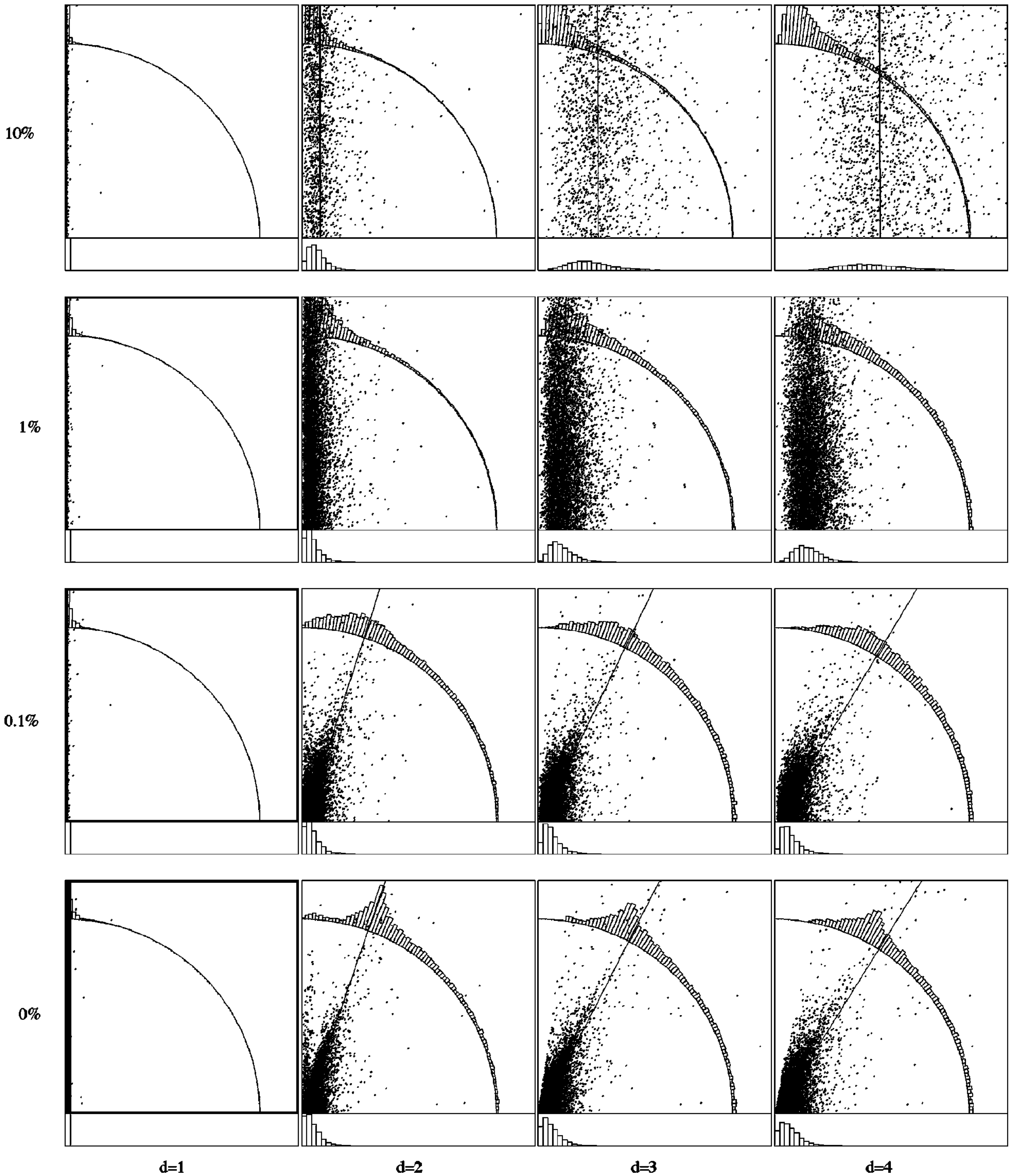

FIG. 2. Same as in Fig. 1 but for the Lorenz system (the dimension of the attractor is 2.06). The box size is 0.024×0.024 and the total number of data points is 10 000. The regression lines are also plotted on each graph.

noise level, but unfortunately, we have usually very limited information on the amplitude of the noise in a given time series.

II. GRAPHICAL REPRESENTATION OF NEAREST NEIGHBOR DISTRIBUTIONS

Without a clear understanding of the *distribution* of neighboring points in the time-delay coordinates we cannot really estimate the applicability or limitations of the FNN method. Indeed, the original test (1.2) or the modified test (1.5) cannot guarantee that we have reached a sufficient embedding dimension, even if the percentage of false nearest neighbors is low.

We have therefore constructed a simple graphical presentation that simultaneously displays all essential features. The basic idea is that we show the target distance $R_\Delta = |x(k+d) - x(n(k)+d)|$ as a function of the original distance $R_d = \|\mathbf{y}(k) - \mathbf{y}(n(k))\|$ for all d-dimensional vectors in the data set. (The variable R_d should be scaled with the normalization coefficient $\sqrt{d}$ in order to remove unessential changes in the graphs due to changes in the embedding dimension; see the Appendix).

As the first example we have chosen the Henon system

$$X_{n+1}=1-1.4X_n^2+Y_n, \quad Y_{n+1}=0.3X_n. \tag{2.1}$$

The parameters of this system were selected from the chaotic region (the dimension of the attractor is 1.26), and the total number of data points is 1000. In Fig. 1 we have plotted $(\tilde{R}_d, R_\Delta)$ pairs $(\tilde{R}_d = R_d/\sqrt{d})$ for all vectors $\mathbf{y}$. The displayed box size is 0.072×0.072 units. Two distributions have also been presented in each graph: the $\tilde{R}_d$ distribution on the bottom part of the graphs, and the radial distribution plotted on the quarter arc. The embedding dimension d is scanned from 1 to 4, and each set of four graphs is presented in four different cases where the amplitude of the additional uniformly distributed (measurement) noise is 0%, 0.1%, 1%, and 10% of the total amplitude.

According to Eq. (1.2) a neighbor is false if it lies above the straight line going through the origin with slope R_{tol}. If we use the test (1.5) the line has the same slope but there is an intercept equal to the noise correction term (scaled with $\sqrt{d}$). Normally we must know the slope *a priori* but using these graphs it is not necessary. If there is no noise we clearly see that with the embedding dimension >1 all points lie in the sector determined by the x axis and a line with slope angle well below 90°. This important feature can be understood if we assume that the dynamics is given by

$$x(k+dT)=f(x(k),x(k+1),\ldots,x(k+d-1)). \tag{2.2}$$

Then we can write

$$|x(k+d)-x(l+d)|\leqslant\|\nabla f(\xi)\|\|\mathbf{y}(k)-\mathbf{y}(l)\| \tag{2.3}$$

for some ξ, which implies that

$$\frac{R_\Delta}{R_d}\leqslant\|\nabla f(\xi)\|. \tag{2.4}$$

Therefore all points in the $(\tilde{R}_d, R_\Delta)$ plots must lie under a line that depends on the specific system. The limit (2.4) is true only when the embedding dimension is sufficient, and for noise it is never possible. If the time series includes only a small amount of additional noise we see its effect as a blurred boundary.

If the embedding dimension is too low the points cumulate close to the y axis. The radial distribution plot confirms this result. If $d=1$ the distribution has significant values only with angles close to 90° but if $d>1$ the distribution is almost zero within a distinct range at high angles. The $\tilde{R}_d$ distribution is high only in the vicinity of zero. A small amount of noise (0.1%, the second row from the bottom in Fig. 1) does not change the picture much.

If the level of additional noise is increased to 1% the points do not show as well formed a pattern. Also the radial distribution is quite broad but it nevertheless has a clear zero range at high angles if the embedding dimension is 3, which can be regarded as an indication of underlying chaotic (or at least deterministic) dynamics. The maximum of the $\tilde{R}_d$ distribution has clearly shifted towards large values which is typical for pure noise.

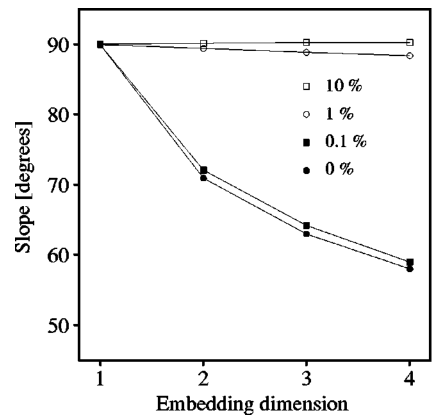

FIG. 3. Slope of the regression line as a function of the embedding dimension for different percentage of noise taken from the graphs in Fig. 2.

In the case of more noisy data (10% on the top row of Fig. 1) the distribution of points is totally different. Increasing the embedding dimension does not really change the overall shape of the point distribution. The radial distribution is fairly even, and the $\tilde{R}_d$ distribution is well centered and its maximum shifts toward higher values when the embedding dimension is increased. [In this case the modified test (1.5) does not really take noise effects into account.]

In Fig. 2 we have presented corresponding graphs for the Lorenz system

$$\begin{aligned}\dot{X}&=16(Y-X),\\ \dot{Y}&=X(45.92-Z)-Y,\\ \dot{Z}&=XY-4Z,\end{aligned} \tag{2.5}$$

using 10 000 data points and the sampling delay of 0.05. For these parameter values the dimension of the attractor is 2.07. Here we observe similar kind of behavior for various distributions, as in the case of the Henon system. Since the true dimension of the attractor is greater than 2, a clearly bounded sector pattern of points can only be seen in the graphs with embedding dimension $\geqslant 3$. For $d=2$ most of the points lie under a line with slope under 90°, which is also reflected in the noticeable maximum of the radial distribution, and since there is only a small portion of points between this maximum and the y axis we can estimate that the true dimension of the attractor is not much greater than 2.

The effect of even a small amount of noise can be clearly seen in Fig. 2. Already with 1% of noise the sector pattern has changed to a vertical one. This is shown clearly in the regression lines [corresponding to the first principal component of the points $(\tilde{R}_d, R_\Delta)$] plotted in Fig. 2. In the two bottom rows the regression lines have a slope well below 90°, and this can be taken as evidence of deterministic dynamics. For the two top rows the regression line is almost vertical (see also Fig. 3), indicating noise contamination. Furthermore, we see that the $\tilde{R}_d$ distribution shows approxi-

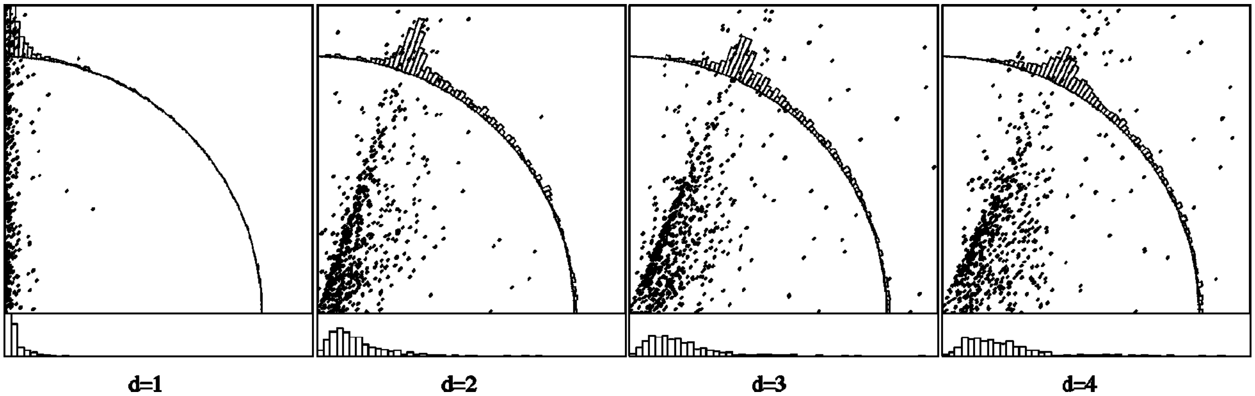

FIG. 4. Same graphs as in the bottom row of Fig. 2 but the total number of data points is only 1000.

mately Gaussian shape, which spreads out and moves further and further away from the origin as the noise level or embedding dimension increases. The radial distribution, on the other hand, moves closer to the 90° line as noise contamination increases, which means that the height-width ratio of the point distribution increases, and therefore that it is more and more difficult to predict the next point.

In the standard procedure noise effects are taken into account by the condition (1.3), which means that points outside a circle of radius $A_{tol}R_A$ are counted false (actually it is an ellipse, due to the scaling of $\tilde{R}_d$.) For Figs. 2 and 4 this radius is 500 times the box size (and for Figs. 1 and 5 the factor is about 20). Although the boundary is quite far away one can imagine that higher levels of noise and higher embedding dimensions both increase the number of false neighbors, as has been reported [3,4].

If the total number of data points of the preceeding system is decreased to 1000 the graphs are not so simple to interpret (Fig. 4). There is no significant difference between graphs with embedding dimension 2 and 3. As usual, reliable estimation of the underlying dynamical dimension requires a sufficient number of data points. However, by using this graphical representation we can nevertheless make a rough estimate on dimension, even when only relatively few data points are available.

As a final example we have analyzed the Mackey-Glass system

$$\dot{X}=\frac{0.2X(t+31.8)}{1+[X(t+31.8)]^{10}}-0.1X(t), \tag{2.6}$$

using the sampling delay of 2. As the dimension of the attractor with these parameter values is about 3.6, the embedding dimension must be at least 4. This can be seen in Fig. 5: only in the rightmost graph is there a clear sector type of pattern, and the radial distribution is zero over a nonzero range of angles near 90°.

III. CONCLUSIONS

In this paper we have presented a graphical method by which one can better understand the predictions and limitations of the false-nearest-neighbor analysis. This tool consists of a $(\tilde{R}_d, R_\Delta)$ plot augmented with two distributions. The slope of the regression line of points in the $(\tilde{R}_d, R_\Delta)$ graphs is a further tool in recognizing noise in deterministic systems.

The advantage of the present method is that we can notice even small amounts of noise contamination. At the same time we now see that determining the correct embedding dimension can become difficult, even with a small amount of noise. Although the criteria (1.2) and (1.5) always produce a yes or no answer, our recommendation is to first check whether noise contamination is too high for determining the embedding dimension.

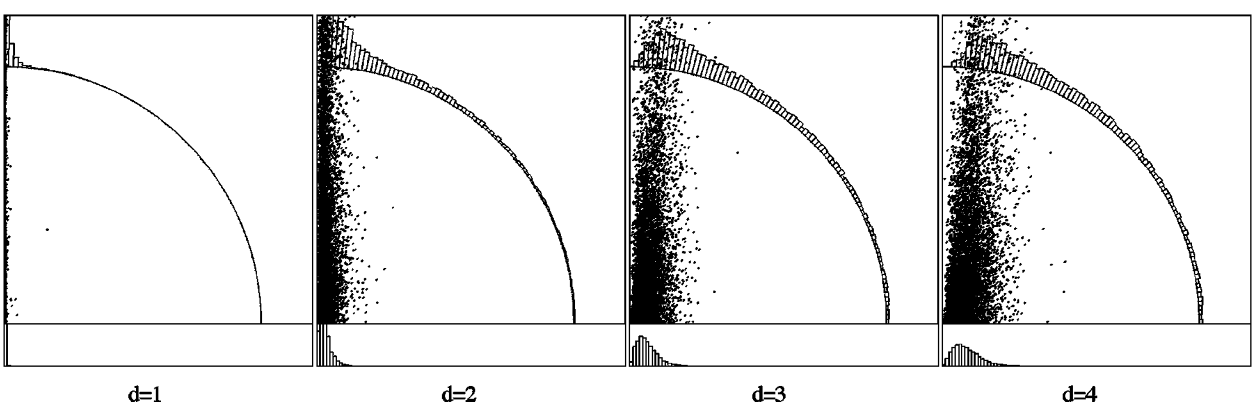

FIG. 5. Target distance R_Δ as a function of the nearest neighbor distance $\tilde{R}_d$ for the Mackey-Glass system (the dimension of the attractor is $\sim$3.6). The box size is 0.048×0.048, the total number of data points 10 000.

Thus, when faced with experimental data, one should first determine whether the FNN method really is applicable. We would say that the noise level in the time series is too high if the following properties are evident:

1. The radial distribution is spread out over the whole range from 0 to 90°,

2. the R_d distribution has a clear maximum far away from zero;

3. the slope of the regression line is close to 90°.

For example, the data studied in the two top rows of Fig. 2 has too much noise for any embedding dimension analysis by FNN method.

On the contrary, the time series is produced by a deterministic system without excessive noise contamination if both of the following conditions hold:

1. the slope of the regression line is well below 90;

2. the R_d distribution is centered close to zero.

In such a case one can next try to find the embedding dimension, the criteria for that is: The embedding dimension is sufficient for unfolding the dynamics if the points in the $(\tilde{R}_d, R_\Delta)$ plot form a clear sector pattern with a zero radial distribution over a distinct range below 90°,

The appearance of an empty wedge at $d=3$ in the two bottom rows of Figs. 2 and 4 is clear, although noise effects can already be seen at the 0.1% level of Fig. 2.

APPENDIX

Let f be a function that has been sampled very densely. Then we can assume that the nearest neighbor of the d-dimensional vector is the vector that starts at the next (or previous) sample point

$$[f(t_0+\delta),\ f(t_0+2\delta),\ f(t_0+3\delta),\ldots,f(t_0+d\delta)]. \tag{A1}$$

The distance between these two points is therefore

$$R_d=\sqrt{\sum_{i=1}^{d}\{f(t_0+i\delta)-f(t_0+(i-1)\delta)\}^2}$$
$$\approx\sqrt{\sum_{i=1}^{d}\delta^2 f'(t_0+i\delta)^2}\approx\delta\sqrt{d}|f'(t_0)|, \tag{A2}$$

where we have assumed that the function f changes relatively slowly (or that it is linear). The distance between the targets is

$$R_\Delta=|f(t_0+\delta+1)-f(t_0+\delta)|\approx\delta|f'(t_0)|, \tag{A3}$$

and by combining the results (A2) and (A3) we conclude that the ratio of R_Δ/R_d is $1/\sqrt{d}$, and therefore it is reasonable in all cases to normalize this ratio with $\sqrt{d}$.

[1] N.H. Packard, J.P. Crutchfield, J.D. Farmer, and R.S. Shaw, Phys. Rev. Lett. **45**, 712 (1980).

[2] M.B. Kennel, R. Brown, and H.D.I. Abarbanel, Phys. Rev. A **45**, 3403 (1992).

[3] H.D.I. Abarbanel, R. Brown, J.J. Sidorowich, and L.S. Tsimring, Rev. Mod. Phys. **65**, 1331 (1993).

[4] H.D.I. Abarbanel and M.B. Kennel, Phys. Rev. E **47**, 3057 (1993).

[5] C. Rhodes and M. Morari, Phys. Rev. E **55**, 6162 (1997).