Improved false nearest neighbor method to detect determinism in time series data

Rainer Hegger and Holger Kantz
Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Strasse 38, 01187 Dresden, Germany
(Received 21 April 1999)

The false nearest neighbor method introduced by Kennel *et al.* [Phys. Rev. A **45**, 3403 (1992)] is revisited and modified in order to ensure a correct distinction between low-dimensional chaotic data and noise. Still, correlated noise processes can yield vanishing percentages of false nearest neighbors for rather low embedding dimensions and can be mistaken for deterministic signals. Therefore, the false nearest neighbors method has always to be combined with a surrogate data test. [S1063-651X(99)08510-4]

PACS number(s): 05.45.−a

The study of irregular time series data by means of nonlinear analysis methods has become a popular task within the last years [1,2]. The main idea is that the aperiodicity in the data is not due to stochasticity but due to nonlinearity. The justification for this point of view is that simple nonlinear processes can give rise to very complex dynamical behavior, though the underlying process is purely deterministic and might be quite low dimensional. The main visible difference between a stochastic process and a nonlinear deterministic process is that data from the latter are confined to a finite-dimensional manifold, whereas the former represents infinitely many degrees of freedom. Thus the most important and apparently easiest test for determinism is a dimensional analysis. For this reason the correlation dimension D_2 introduced by Grassberger and Procaccia [3] is one of the most often used methods in nonlinear time series analysis. Indeed, this tool gives reasonable results if the combination of dimensionality of the process, the length of the time series, and the noise level of the data is propitious. Especially, under well controlled laboratory conditions it is sometimes possible to estimate D_2 to a good approximation. Certainly, the usual problem is that often either the dimensionality of the system is too high to find a clear indication of a finite dimension from a finite time series, or noise on the data destroys the signature of determinism. Therefore, in most cases this dimension estimate does not yield conclusive results at all.

Another method to determine the dimensionality of the system is the false nearest neighbor method developed by Kennel *et al.* [5]. The main idea is that for deterministic systems, points which are close in the (reconstructed) phase space stay close under forward iteration. Translated into the concept of time delay embedding for scalar time series data [4], this statement is true if the dimension of the embedding space is high enough to fully resolve the determinism. If, on the other hand, the dimension is too small, points may appear as close neighbors purely by projection effects. Therefore, as argued in [5], these points are mapped randomly onto the whole attractor under forward iteration. Based on these ideas the algorithm works as follows: Given a point $\vec{x}_n$ in m dimensions, look for its nearest neighbor $\vec{x}_r$. Let the distance between these two points be ϵ. If the distance of the iterates of these two points is larger than $s\epsilon$, where s is an *a priori* fixed value (to be discussed below), then $\vec{x}_r$ is marked as a *false nearest neighbor* (FNN). Repeat the procedure for the whole time series. The fraction of false nearest neighbors (percentage FNN) then indicates whether the process is deterministic in m dimensions or not. If m is larger than the number of active degrees of freedom this fraction should be zero or at least very small, and nonzero otherwise. This method, which is also used to determine the minimal embedding dimension of scalar time series data, has gained popularity. However, as we shall show below, stochastic processes can also yield a vanishing fraction of false nearest neighbors for not too large m, when the method is used as a black box with inappropriate parameter settings. We therefore argue in favor of using a modified and more detailed neighbor statistics, to use only suitable embedding time lags also for stochastic data properly taking into account their linear correlations, and to complement each false nearest neighbors plot by a plot for suitable surrogates.

For a systematic approach we formulate the fraction of false nearest neighbors in a probabilistic way. For analytical ease, we will measure distances by the maximum norm, but there is no evidence that this is a restriction for the generality of our results. Assume that the distance between an m-dimensional vector $\vec{x}_n=(x_n,x_{n+1},\ldots,x_{n+m-1})^\dagger$ and its nearest neighbor $\vec{x}_r$ is ϵ. The conditional probability that these points are false nearest neighbors, which means that $|x_{n+m},x_{r+m}|>s\epsilon$, can be written as

$$P_f(s)=P(|x_{n+m},x_{r+m}|>s\epsilon \mid \|\vec{x}_n,\vec{x}_r\|=\epsilon), \tag{1}$$

or as

$$P_f(s)=\frac{P(|x_{n+m},x_{r+m}|>s\epsilon,\|\vec{x}_n,\vec{x}_r\|=\epsilon)}{P(\|\vec{x}_n,\vec{x}_r\|=\epsilon)}. \tag{2}$$

If we suppose we had infinitely many points this can be rewritten as

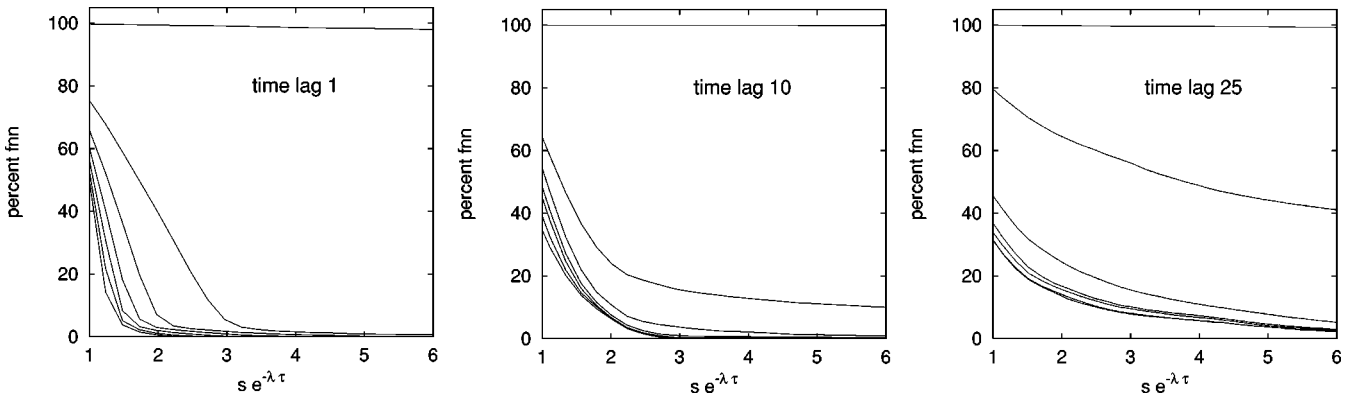

FIG. 1. Percentage of false nearest neighbors computed numerically for data from the Lorenz ordinary differential equations [7], as a function of $se^{-\lambda\tau}$ for embedding dimensions $m=1$ to 7, from top to bottom. The result depends on the time lag τ. Due to the sampling rate of our data $\tau=10$ is most suitable.

$$P_f(s)=\frac{\int dx_{n+m}\int dx_{r+m}\int d\vec{x}_n\hat{\mu}(x_{n+m},\vec{x}_n)\int d\vec{x}_r\hat{\mu}(x_{r+m},\vec{x}_r)\,\delta(\|\vec{x}_n,\vec{x}_r\|-\epsilon)\Theta(|x_{n+m},x_{r+m}|-s\epsilon)}{\int d\vec{x}_n\mu(\vec{x}_n)\int d\vec{x}_r\mu(\vec{x}_r)\,\delta(\|\vec{x}_n,\vec{x}_r\|-\epsilon)}, \tag{3}$$

where $\hat{\mu}(x)$ is the invariant measure in $\mathbf{R}^{m+1}$, $\mu(x)$ is the invariant measure in $\mathbf{R}^m$, and $\Theta(x)$ is the Heaviside step function. Equation (3) thus defines the average fraction of false nearest neighbors if the initial distance was ϵ. To get the fraction of false nearest neighbors for pairs having arbitrary initial distance, one has to average over all values of ϵ.

Let us now consider a scalar time series of a deterministic system. If the embedding dimension m is chosen high enough to fully resolve the determinism, we can write the joined measure as

$$\begin{aligned}\hat{\mu}(x_{n(r)+m},\vec{x}_{n(r)})&=\hat{\mu}(x_{n(r)+m}|\vec{x}_{n(r)})\mu(\vec{x}_{n(r)})\\&=\delta(x_{n(r)+m}-f(\vec{x}_{n(r)}))\mu(\vec{x}_{n(r)}),\end{aligned} \tag{4}$$

where $n(r)$ means that the index is either n or r. Thus we can perform the integration over $x_{n(r)+m}$ and the argument of the Θ function simply becomes

$$|f(\vec{x}_n),f(\vec{x}_r)|-s\epsilon, \tag{5}$$

which is negative if s is larger than the largest local expansion rate of the map f. Hence, for deterministic systems we see a fraction zero of false nearest neighbors only if s is sufficiently large (s was typically set to 10 in [5]). If we choose s too small, false neighbors are found and we are unable to identify a deterministic system as such. Figure 1 shows the fraction of false nearest neighbors for the Lorenz system as a function of s. We argue against replacing the study of the full s dependence by a calculation for a single fixed s value (which is done in most applications of the method). The minimal reasonable s is given by the maximum of the local deterministic expansion rate, which can be much larger than $e^{\lambda_{\max}\tau}$, where $\lambda_{\max}$ is the maximal Lyapunov exponent and τ the time lag. The time lag between successive measurements entering the delay vectors is a relevant embedding parameter which has gained much attention in the literature [6]. Also for the FNN statistics it is crucial for a sound detection of the correct dimension. Too short time lags enhance correlations and give rise to delay vectors close to the diagonal, such that deviations transverse to it are badly unfolded. Too large lags lead to complicated geometry of the reconstructed attractor. Both mechanisms introduce a distortion of the FNN statistics (see Fig. 1). The reasonably unfolded attractor can be embedded in three dimensions apart from points of measure zero, a result which we find only for $\tau\simeq 10$. For comparability we have normalized s to the expansion factor $\exp(\lambda_{\max}\tau)$.

A nonzero percentage of FNN is found for the ''correct'' m not only if s is too small compared to the lag τ, but also when the data are contaminated by measurement noise. Paradoxically, the effect of noise becomes the more severe the longer the time series is, since more data allow for smaller nearest neighbor distances. One can therefore test for additive noise by varying the time series length used, although this will yield a visible effect only when the supposed attractor dimension is small enough to also guarantee a considerable variation of the average nearest neighbor distance.

The main question we want to raise is whether a very low fraction of false nearest neighbors for some m and s is sufficient to characterize a system as being deterministic or even nonlinear. The simplest nondeterministic system to deal with is pure white noise. In this case the μ in Eq. (3) factorizes and the false nearest neighbor fraction is given by

$$P_f(s)=\int d\eta_n\mu(\eta_n)\int d\eta_r\mu(\eta_r)\Theta(|\eta_n,\eta_r|-s\epsilon). \tag{6}$$

For uniformly distributed random numbers $\mu(\eta)=\text{const}$ for $0\leqslant\eta\leqslant 1$ and zero otherwise; this is easily calculated to

$$P_f(s)=1-2s\epsilon+s^2\epsilon^2, \tag{7}$$

and after averaging over all ϵ, we have to replace ϵ and ϵ^2 by $\langle\epsilon\rangle$ and $\langle\epsilon^2\rangle$, respectively. $P_f(s)$ is unity if $\langle\epsilon\rangle$ is zero, which is fulfilled only for an infinite time series. I.e., only an infinite time series of white noise has 100% FNNs for arbitrary s. For finite time series length N, $P_f(s)$ decreases for increasing m, since the average interpoint distance $\langle\epsilon\rangle$ becomes larger and larger. In particular, Eq. (7) yields zero for $\epsilon=1/s$. For uncorrelated noise $\langle\epsilon\rangle\simeq 1/N^{1/m}$. One thus needs $N\geqslant s^m$ points to get a nonzero result, or one cannot trust the result for $m>\ln N/\ln s$. Of course, we already saw that we cannot counterbalance too small an N by decreasing s too much, since then we run into the problem that we cannot identify a deterministic chaotic system as being deterministic.

Summarizing, for $\langle\epsilon\rangle=1/s$ the fraction of false nearest neighbors is zero even for uncorrelated uniformly distributed random data. To avoid this bias, one needs a second reasoning for the nearest neighbor statistics. In the original work of Kennel *et al.* [5] this was solved in the following way: If $\sqrt{\sum_{i=0}^{m}(x_{n+i}-x_{r+i})^2}>A_{\rm tol}R_A$, where R_A is standard deviation of the data and represents the attractor size, and $A_{\rm tol}$ some factor (typically set to 2 in [5]), then $\vec{x}_r$ is also counted as a false nearest neighbor, independently of the distance of the images of the two points. This choice has two drawbacks: First, for an insufficient amount of data, this criterion introduces false neighbors in large m also for deterministic chaotic systems. Second, it underestimates the number of false neighbors for large s. Both aspects are a consequence of the fact that this criterion itself is not completely in the spirit of the false neighbor ideology: Pairs with too large distance are not really false neighbors, but are just inappropriate candidates to apply the method. We thus decide to disregard all points in the averaging procedure of Eq. (1) for which the distance towards its nearest neighbor (in m dimensions) is larger than or equal to $1/s$ times the standard deviation of the data. For uniform and Gaussian distributed white noise, this leads to 50–60% FNN for large s, independent of s and m. However, the number of points entering the average drops to zero very fast for large s. Simply setting $A_{\rm tol}=1/s$ cures the second problem mentioned above, but not the first one.

In Fig. 2, our implementation and the criterion of [5] are compared for uniformly distributed white noise. The continuous curves show the results of our numerics. The value of $\langle\epsilon\rangle$ thus obtained was inserted in Eq. (7), shown as dotted curves. One clearly observes statistical fluctuations due to poor statistics for large s and large m. The dashed curve shows the result for half of the data if we use the second condition of Kennel *et al.* instead of ours. One clearly sees that this condition gives many fewer false neighbors, and for (large s)/(large m) the FNN percentage drops farther when increasing the length of the time series (e.g., with 20 000 points it is less than 5% for $m=6$). Notice that the lowest dashed curve corresponds to $m=5$, and $m=6$ lies above.

So we argue in favor of computing the percentage of FNN for a large range of ''falseness values'' s, and to exclude all points from the statistics whose closest neigbor is already too far away to have a chance to become false. This allows us to clearly distinguish between white noise and low-dimensional chaos. However, the situation becomes worse also for our

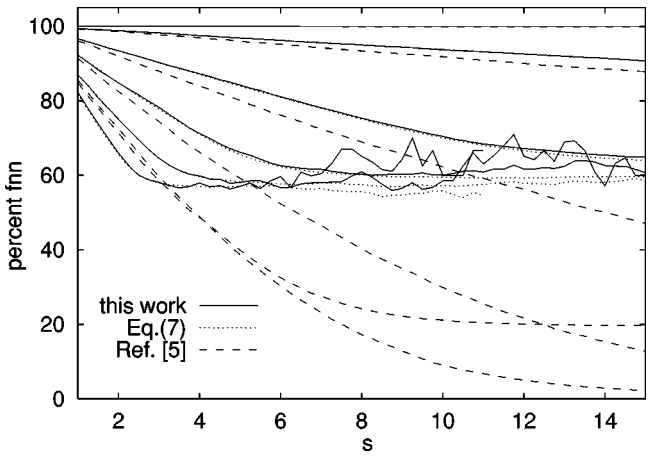

FIG. 2. Percentage of false nearest neighbors for 20 000 uniform white noise data for $m=1$ to 6 (from top to bottom). Our second criterion (continuous curve) fixes the percentage to $\approx 60\%$, whereas the original criterion (on only 10 000 points) leads to unsatisfactory results (broken curves). For details see text.

implementation if we deal with colored noise. As an example we consider an autoregressive process of order 2 [AR(2)] process (a discretized noise driven damped oscilator)

$$x_{n+1}=(2-\omega^2-\rho)x_n+(\rho-1)x_{n-1}+\eta_n, \qquad (8)$$

where η_n is white noise, and we fix $\omega=2\pi/20$, whereas ρ is varied to show the influence of the correlation time.

In Fig. 3 we see that, although the data stem from stochastic processes, the false nearest neighbor fraction tends to zero for rather small embedding dimensions, when ρ is small. The AR(2) process has two time scales, one set by the oscillation period, which we respected by a reasonable time lag τ (when the time lag of the time delay embedding is too small, one finds even fewer false nearest neighbors), and the second set by the damping. The latter cannot be accounted for, even though the results are obtained using a kind of Theiler window [9].

When speaking of determinism in time series data, one usually implicitly accepts some small amount of measurement noise. Dynamical noise, which interacts with the deter-

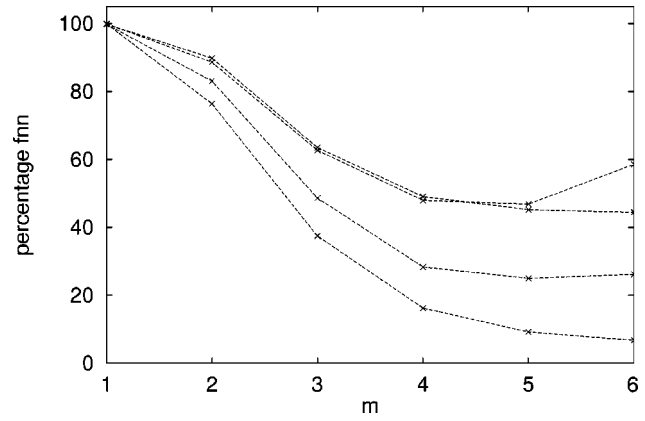

FIG. 3. Percentage of false nearest neighbors for colored noise processes at $s=8$ (saturation region for these data), for damping $\rho=0.02$ (lowest curve), $\rho=0.05$ (second from below), $\rho=0.2$, and $\rho=0.5$ (two uppermost curves).

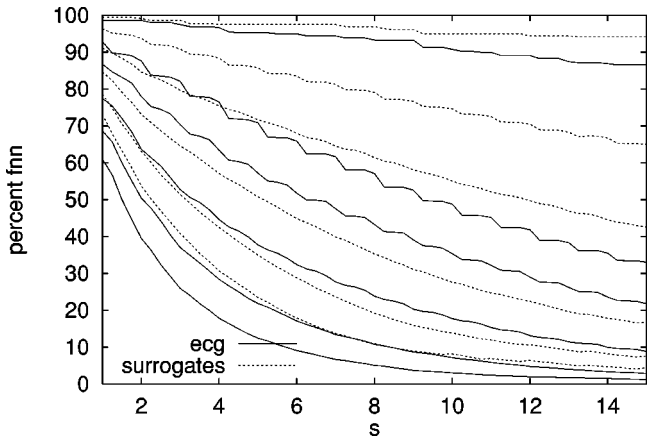

FIG. 4. Percentage of FNN for ECG data and their surrogates.

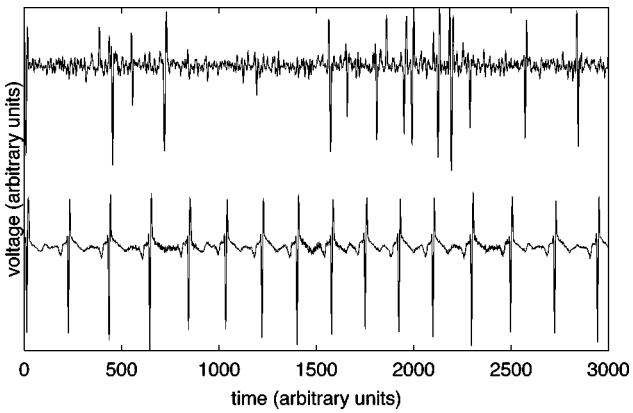

FIG. 5. Part of the human ECG data (lower trace) and their surrogates (upper trace).

ministic part of the dynamics, is equally likely to be present in almost all experimental data. Formally, such a process is then a stochastic process, since it is driven by noise. The difference between such a process, which one could call noisy chaos, and the processes depicted by Eq. (8) is that noise is essential for the latter, since purely deterministic solutions would be transients decaying to zero. Nevertheless, we see that a (linear) deterministic rule is responsible for the correlations in data from these models, and it is therefore not totally surprising that this deterministic component leads to a suppression of false nearest neighbors.

To clarify this point, we show in Fig. 4 the percentage of FNN for human ECG data and their surrogates [10,11]. Surrogates are numerically generated data sets which contain all the linear correlations and represent the correct marginal distribution of the original data but are otherwise random. The time lag for the FNN algorithm was chosen such that the data are reasonably unfolded in two dimensions. We show a part of the original and the surrogate time series (iterated amplitude adjusted Fourier transform surrogates [11]) in Fig. 5. As claimed above, the result is not very clear: ECG data and surrogates have very similar FNN statistics, and the randomized data lose their false nearest neighbors for moderate embedding dimensions. Although we do not claim that ECG data represent a deterministic process, the difference between the ECG and its surrogates is so striking that the insensitivity of the false nearest neighbor method is somewhat worrying.

To summarize, the false nearest neighbor method is not able to distinguish between deterministic and stochastic processes on an absolute level. If the correlation time of a stochastic process is large, the probability of identifying the data as a deterministic process is quite high. Only for white noise processes, i.e., uncorrelated random data, can about 50% of false nearest neighbors be assured by the additional requirement introduced in this paper. In order to find interpretable results for other data, one has to choose a reasonable time lag in the time delay embedding and to scan a whole range of s values. Together with the original data one should always study surrogate data, since it is otherwise hard to distinguish linear correlations from nonlinear deterministic rules. Even then the insight might remain incomplete as demonstrated for ECG data. Let us finally stress that despite our warning we think that the false nearest neighbor method is useful to determine in an intuitively convincing way the embedding parameters of a system of which one has good reason to assume that it is deterministic.

All numerical results relying on our modified implementation of the false nearest neighbor method are obtained with the algorithm contained in the TISEAN packet, which can be downloaded for free [8].

R.H. wants to acknowledge the support of the European Union under Grant number FMRX-CT96-0010 and to thank colleagues at the Istituto Nazionale di Ottica (Florence, Italy) for their kind hospitality.

[1] H.D.I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer-Verlag, New York, 1996).

[2] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, UK, 1997).

[3] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).

[4] F. Takens, *Detecting Strange Attractors in Turbulence*, edited by D. A. Rand and L. S. Young, Lecture Notes in Mathematics Vol. 898 (Springer-Verlag, New York, 1981); T. Sauer, M. Casdagli, and J.A. Yorke, J. Stat. Phys. **65**, 579 (1991).

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

[6] e.g., A.M. Fraser and H.L. Swinney, Phys. Rev. A **33**, 1134 (1986); T. Buzug and G. Pfister, Physica D **58**, 127 (1992).

[7] E.N. Lorenz, J. Atmos. Sci. **20**, 130 (1963).

[8] The TISEAN software packet of R. Hegger, H. Kantz, and T. Schreiber can be downloaded for free from http://www.mpipks-dresden.mpg.de/~tisean.

[9] J. Theiler, Phys. Rev. A **34**, 2427 (1986).

[10] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J.D. Farmer, Physica D **58**, 77 (1992).

[11] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).