Method of false nearest neighbors: A cautionary note

Donald R. Fredkin
Department of Physics, University of California, San Diego, La Jolla, California 92093-0319

John A. Rice
Department of Statistics, University of California, Berkeley, Berkeley, California 94720
(Received 2 December 1994)

The method of false nearest neighbors [M. B. Kennel, R. Brown, and H. D. I. Abarbanel, Phys. Rev. A **45**, 3403 (1992)] has been proposed for detecting deterministic structure in empirical time series and for estimating the embedding dimension if the series is deterministic. We show that the method can falsely indicate that a stationary random process is deterministic. Remedial modifications of the method are proposed.

PACS number(s): 05.45.+b, 05.40.+j

I. INTRODUCTION

Let $x(1), x(2), \ldots, x(N)$ be a scalar time series. Kennel *et al.* [1], hence referred to as KBA, proposed the method of false near neighbors for the determination of a minimal embedding dimension. In this paper we discuss certain shortcomings of this procedure as a method for distinguishing stochastic processes (which do not have finite embedding dimensions) from dynamical systems and propose and examine possible remedies.

For a given time delay T and proposed embedding dimension d, one forms the collection of d-vectors

$$y(k) = [x(k), x(k+T), \ldots, x(k+(d-1)T)]$$

and finds the nearest neighbor in the Euclidean metric of each such vector. We denote the nearest neighbor of $y(k)$ by $y(n(k))$. If the series is a projection of a dynamical system with an attractor of dimension d_A and $d > 2d_A$, then the embedding is sufficient to unfold the geometry of attractor [2] so that points that are close together in the collection of $y(k)$ are also close together in phase space. On the other hand, if d is too small, members of the collection of $y(k)$ that are quite separated in phase space may be neighbors in R^d because the attractor has been projected onto a low dimensional space.

The method of false near neighbors makes the determination of whether d is sufficiently large by comparing the $(d+1)$st coordinates of $y(k)$ and $y(n(k))$: $x(k+dT)$ and $x(n(k)+dT)$. If many of the distances $|x(k+dT) - x(n(k)+dT)|$ are large, many of the nearest neighbors are "false" and have been pulled apart by increasing the dimension from d to $d+1$, and implying that d is too small. On the other hand, if the distances are predominantly small, only a small proportion of the neighbors are false, and d is deemed to be a sufficient embedding dimension. A common value of d found consistently while varying T is taken as evidence that the series is a projection of a deterministic dynamical system with an attractor of dimension $d_A < 2d$. Further analysis is then conducted to examine this hypothesis and to elucidate the geometry of the attractor.

To make this idea operational the criteria under which a neighbor is declared false must be specified. KBA propose that a neighbor be declared false if

$$\frac{[x(k+dT) - x(n(k)+dT)]^2}{\| y(k) - y(n(k)) \|^2} > R_{\text{tol}}^2 \tag{1}$$

or if

$$\frac{\| y(k) - y(n(k)) \|^2 + [x(k+dT) - x(n(k)+dT)]^2}{R_A^2} > A_{\text{tol}}^2 \tag{2}$$

where

$$R_A = \frac{1}{N}\sum_{k=1}^{N}[x(k) - \bar{x}]^2.$$

The second criterion was proposed in order to provide correct diagnostics for noise and KBA demonstrated its effectiveness in correctly identifying white noise as nondeterministic. On the basis of considerable numerical experimentation, they found that the procedure was robust with respect to the choice of R_{tol} and A_{tol}; in our computations below we used $R_{\text{tol}} = 17.3$ and $A_{\text{tol}} = 1.81$, following their recommendations.

Our discussion is motivated and illustrated by the analyses of three time series each of length $N = 30\,000$, which we will refer to as series A, B, and C. For each series we used several time delays T and computed the percentage of false near neighbors. The results, shown in Fig. 1, suggest that the series are deterministic, not stochastic, with embedding dimensions of about 6, 3, and 10, respectively.

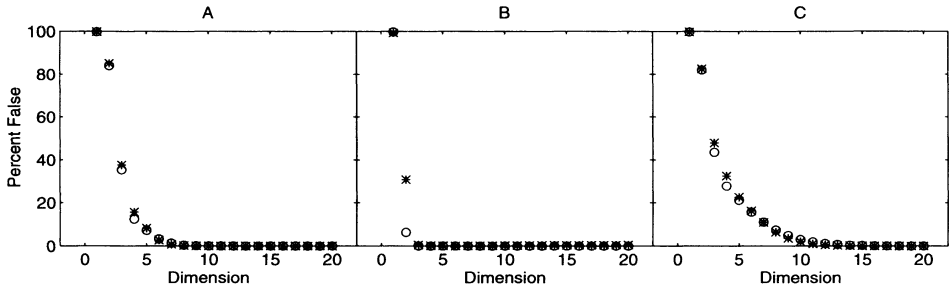

FIG. 1. Percent false nearest neighbors as a function of embedding dimension d for each of the three series. The results for time lags (T) of 10, 10, and 100 sample units are plotted as open circles and those for T=20, 20, 200 are shown as asterisks. These values were chosen based on the mutual information functions shown in Fig. 2.

The remainder of the paper is organized as follows: In Sec. II we explain why stochastic processes can and demonstrably do give rise to results such as those shown in these figures. In Secs. III and IV we propose and investigate two possible remedies. Some concluding remarks are made in Sec. V.

II. THE EFFECTS OF AUTOCORRELATION

Figure 2 shows autocorrelation and mutual information functions [3] for the three series. By either measure there is considerable short term memory in each of the series. In this section we argue that high autocorrelation can cause the method of false near neighbors to incorrectly indicate that a stationary random time series is deterministic.

Suppose that $x(n)$ is obtained by sampling from a smooth trajectory, either stochastic or deterministic, and hence has substantial autocorrelation. The nearest neighbor of $y(k)$ will tend to be adjacent in time, since for small r, $y(k+r)$ will be close to $y(k)$. For a stochastic process, it can be argued that this tendency becomes stronger as d increases. Since $|x(k+dT)-x(k+r+dT)|$ will be small, such a nearest neighbor will not unfold as the embedding dimension is increased and will be accepted as "true" by the method of false near neighbors. Such neighbors are not useful for discriminating between

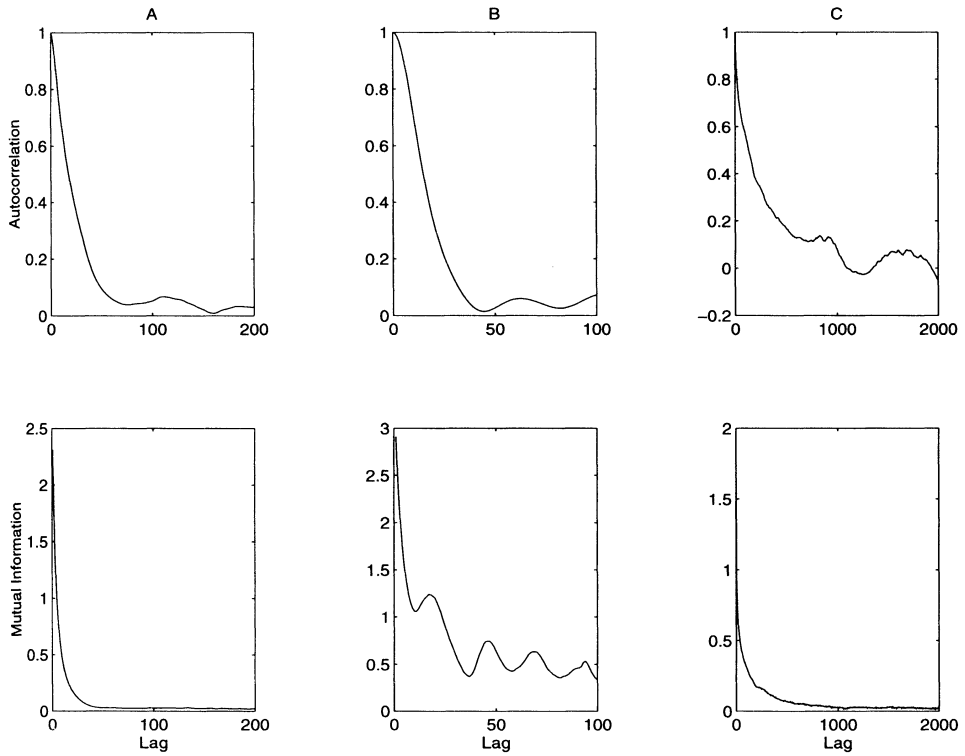

FIG. 2. Autocorrelation functions (top row) and mutual information functions (bottom row) of the three series. The smallest lag (in sampling units) is zero for autocorrelation and one for mutual information.

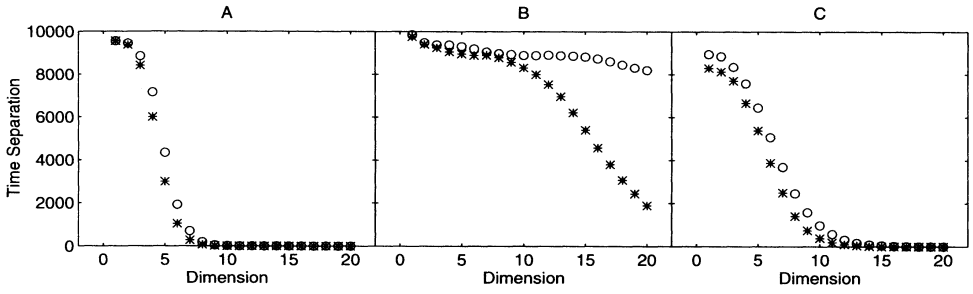

FIG. 3. Average time separations (in sampling units) of nearest neighbors for the three series as a function of embedding dimension. The results for time lags (T) of 10, 10, and 100 sample units are plotted as open circles and those for T=20, 20, 200 are shown as asterisks.

stochastic and deterministic processes.

Since a stochastic process cannot be embedded in a finite dimensional space, and a d dimensional projection will occupy a d dimensional volume, a finite sample will be ever thinner as d increases and the chance of seeing nearest neighbors which are not adjacent in time (and would hence unfold) diminishes. A smooth stochastic process may consequently appear to have a finite embedding dimension and be accepted as deterministic.

For a deterministic process there exist many nearest neighbors that are well separated in time because the orbit returns many times to any neighborhood of any of its points, which lie on a low dimensional manifold. These time separated neighbors do not unfold when the embedding dimension is increased above the correct value.

A smooth stochastic process may consequently appear to have a finite embedding dimension and be accepted as deterministic if time adjacent near neighbors are allowed.

To demonstrate that this is not a purely hypothetical concern, Fig. 3 shows the average time separation $|k - n(k)|$ as a function of d for each of the series. For series A and C, our fear is realized, while a substantial number of the nearest neighbors for series B are well separated in time.

III. IMPOSING A MINIMUM TIME SEPARATION

The analysis of the preceding section suggests that in order to avoid the spurious effects of autocorrelation, the search for the nearest neighbor of each point $y(k)$ should be restricted to m such that $|k - m| > \tau_{\min}$, where $\tau_{\min}$ is greater than the correlation time of the series. Figure 4 shows the effects of imposing this constraint on each of the three series. With this modification, the method gives no evidence that series A and C can be embedded in a low dimension. For series B with $T = 10$, the results are consistent with a low dimensional deterministic system. However, when $T = 20$, the percentage of false nearest neighbors falls below 1% at $d = 3$ but increases for higher dimensions, becoming greater than 1% for $d \geq 11$. One's conclusions will be affected by moderate variation in T.

IV. PREWHITENING

We have seen that a stationary random process can appear to be embeddable in a low dimension if it is strongly autocorrelated. This suggests that a combination of linear filtering and decimation be used to transform the series into one which has little autocorrelation. If a finite impulse response filter is used, these operations will not affect the dimension of the attractor of a deterministic system. We thus seek filter coefficients $a(k)$ such that

$$u(t) = x(t) + \sum_{k=1}^{M} a(k)x(t-k) \tag{3}$$

has little autocorrelation. If necessary, filtering can be followed by decimation to achieve this end. If the coef-

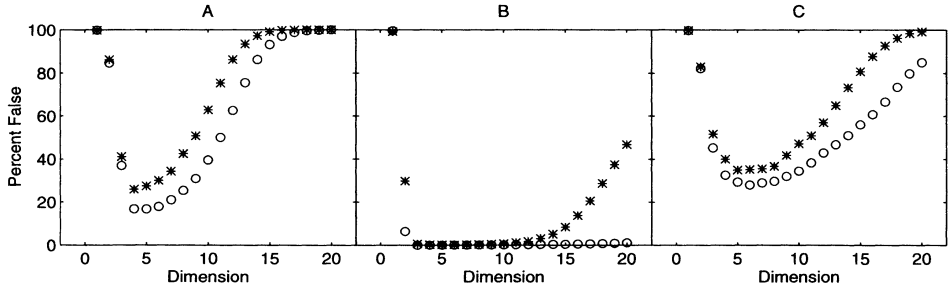

FIG. 4. Percent false nearest neighbors as a function of embedding dimension d for each of the three series with the constraint that nearest neighbors be separated in time by at least $\tau_{\min} = 100, 50, 1000$ respectively. The results for time lags (T) of 10, 10, and 100 sample units are plotted as open circles and those for T=20, 20, 200 are shown as asterisks.

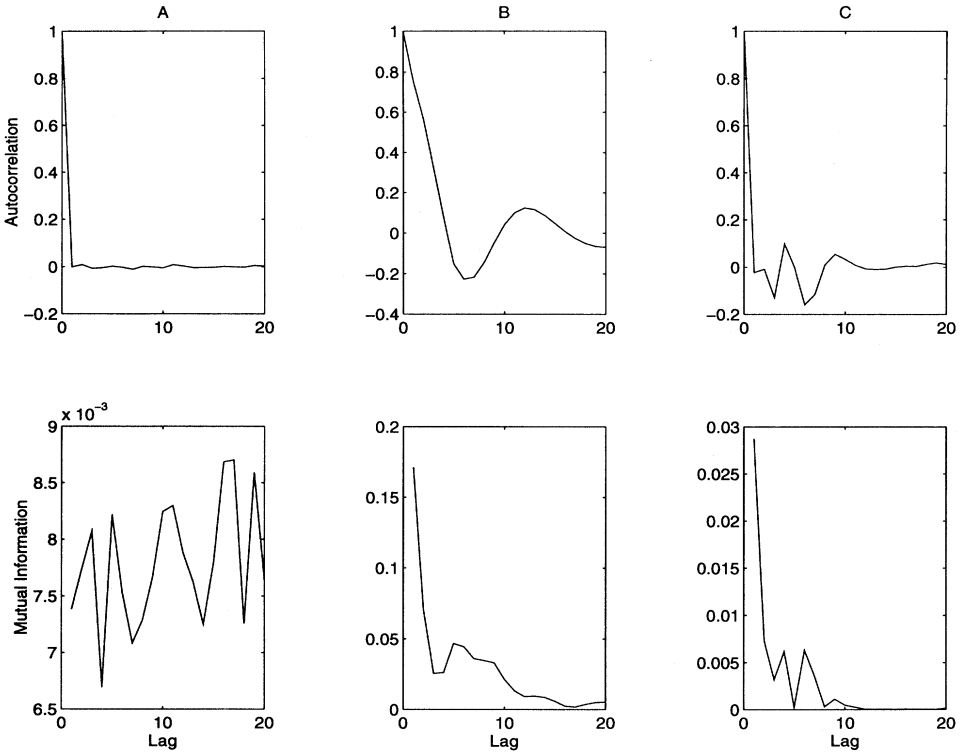

FIG. 5. Autocorrelation functions (top row) and mutual information functions (bottom row) of the three prewhitened series. The prewhitening filters were of lengths 2, 6, and 4, respectively; longer filters made little improvement. The smallest lag (in sampling units) is zero for autocorrelation and one for mutual information.

ficients $a(1), a(2), \ldots, a(M)$ are such that the sequence $u(t)$ is white noise, then the coefficients $a(k)$ are related to the autocovariance function $r(k)$ of the sequence $x(t)$ by the Yule-Walker equations

$$r(n) + \sum_{k=1}^{M} a(k) r(n+k) = 0, \quad n = 1, 2, \ldots, M. \qquad (4)$$

We determine the coefficients from (4) and then use them in (3) to generate the sequence $u(t)$. In general, we do not expect $x(t)$ to be an autoregressive process of order $\leq M$, so $u(t)$ will not be exactly white.

Figure 5 shows the resulting autocorrelation and mutual information functions. For series A, prewhitening decorrelated the series quite effectively. For series B, it was found necessary to follow filtering by decimation (every fourth point) in order to construct a decorrelated series. Prewhitening was reasonably successful for series C.

Figure 6 shows the results of the method of false near neighbors on the decorrelated series. As in the preceding section, there is little evidence that series A and C can be embedded.

V. SUMMARY AND CONCLUSIONS

Series A is an autoregressive process satisfying the equations

$$x(t) = 1.59x(t-1) - 0.60x(t-2) + u(t),$$

where $u(t)$ is a white noise sequence. We have seen that if no precautions are taken the method of false near neighbors erroneously suggests that it can be embedded in a space of dimension $d \approx 6$. When appropriate safeguards against autocorrelation are taken, one is not led to this conclusion.

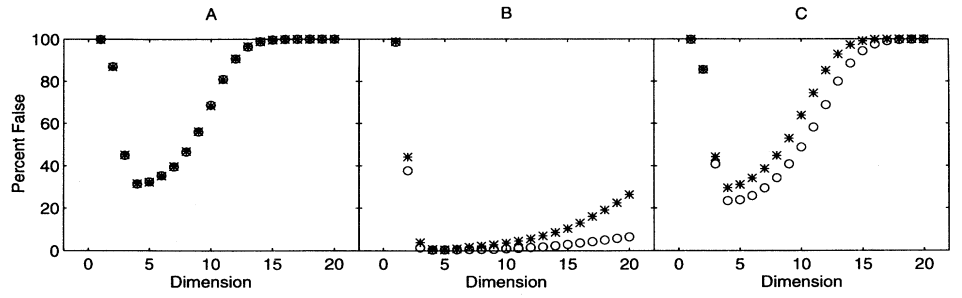

FIG. 6. Percent false nearest neighbors as a function of embedding dimension d for each of the three prewhitened series. The prewhitened series B was further decimated by a factor of 4. The results for time lags (T) of 1, 1, and 2 sample units are plotted as open circles and those for T=2, 2, and 4 are shown as asterisks.

Series B is generated by the Lorenz system

$$\dot{x} = \sigma(y - x),$$
$$\dot{y} = -xz + rx - y,$$
$$\dot{z} = xy - bz,$$

with $r = 45.92$, $b = 4.0$, $\sigma = 16.0$ and initial conditions $x(0) = y(0) = z(0) = 1.0$. The differential equations were solved using RKSUITE [4] with nominal relative accuracy 10^{-6} and the series consists of the values $x(k\Delta t)$, $k = 0, \ldots, 30\,000$ with $\Delta t = 0.01$. We have seen the effects of safeguards against autocorrelation. The percentage of false nearest neighbors is quite small (less than 1%) at dimension 3 and well beyond, but eventually increases. The amount of increase is quite sensitive to the precise value of the time lag T.

Series C is a recording of the current passing through an NMDA [(*N*-methy)-*D*-aspartate] receptor from a rat hippocampal slice [5]. Figure 1 suggests that the dynamics of this channel are deterministic rather than stochastic, but after the effects of autocorrelation are removed, no evidence for this remains.

The method of false nearest neighbors can serve two functions. First, it can provide an estimate of the embedding dimension for a process that is known to be deterministic and in this function it is robust against the addition of noise, as shown in KBA. Second, the method can be used to assess whether an empirical time series is deterministic. We have seen that in this role it is necessary to safeguard against the effects of autocorrelation. However, the price that is paid is that the results for even one of the most classical low dimensional chaotic systems are ambiguous.

Figure 3 suggests that deterministic and stochastic series differ in the dependence of time separation of nearest-neighbors on dimension. This phenomenon may thus provide a useful and simple diagnostic.

ACKNOWLEDGMENTS

We thank Henry Abarbanel for making available a program for the false nearest-neighbor method and for helpful conversations. This research was supported by grants from the Office of Naval Research and the National Science Foundation.

[1] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, Phys. Rev. A **45**, 3403 (1992).
[2] T. Sauer, J. A. Yorke, and M. Casdagli, J. Stat. Phys. **65**, 579 (1991).
[3] H. D. I. Abarbanel, R. Brown, J. J. Sidorwich, and L. S. Tsimring, Rev. Mod. Phys. **65**, 1331 (1993).
[4] R. W. Brankin, I. Gladwell, and L. F. Shampine, Soft-report 92-S1, Department of Mathematics, Southern Methodist University, Dallas, Texas, 1992 (unpublished).
[5] A. J. Gibb and D. Colquhoun, J. Physiol. **456**, 143 (1992).