Statistical methods of parameter estimation for deterministically chaotic time series

V. F. Pisarenko[1] and D. Sornette[2,3,4,*]

[1]*International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Science, Warshavskoye sh., 79, kor. 2, Moscow 113556, Russia*

[2]*Institute of Geophysics and Planetary Physics, University of California, Los Angeles, California 90095, USA*

[3]*Department of Earth and Space Sciences, University of California, Los Angeles, California 90095, USA*

[4]*Laboratoire de Physique de la Matière Condensée, CNRS UMR 6622 and Université de Nice-Sophia Antipolis, 06108 Nice Cedex 2, France*

(Received 14 August 2003; revised manuscript received 13 November 2003; published 31 March 2004)

We discuss the possibility of applying some standard statistical methods (the least-square method, the maximum likelihood method, and the method of statistical moments for estimation of parameters) to deterministically chaotic low-dimensional dynamic system (the logistic map) containing an observational noise. A ''segmentation fitting'' maximum likelihood (ML) method is suggested to estimate the structural parameter of the logistic map along with the initial value x_1 considered as an additional unknown parameter. The segmentation fitting method, called ''piece-wise'' ML, is similar in spirit but simpler and has smaller bias than the ''multiple shooting'' previously proposed. Comparisons with different previously proposed techniques on simulated numerical examples give favorable results (at least, for the investigated combinations of sample size N and noise level). Besides, unlike some suggested techniques, our method does not require the *a priori* knowledge of the noise variance. We also clarify the nature of the inherent difficulties in the statistical analysis of deterministically chaotic time series and the status of previously proposed Bayesian approaches. We note the trade off between the need of using a large number of data points in the ML analysis to decrease the bias (to guarantee consistency of the estimation) and the unstable nature of dynamical trajectories with exponentially fast loss of memory of the initial condition. The method of statistical moments for the estimation of the parameter of the logistic map is discussed. This method seems to be the unique method whose consistency for deterministically chaotic time series is proved so far theoretically (not only numerically).

DOI: 10.1103/PhysRevE.69.036122 PACS number(s): 02.50.-r, 05.45.Tp, 02.60.Pn

The problem of characterizing and quantifying a noisy nonlinear dynamical chaotic system from a finite realization of a time series of measurements is full of difficulties. The first one is that one rarely has the luxury of knowing the underlying dynamics, i.e., one does not in general know the underlying equations of evolution. Techniques to reconstruct a parametric representation of the time series then may lead to so-called model errors.

Even in the rare situations where one can ascertain that the measurements correspond to a known set of equations with additive noise, the chaotic nature of the dynamics makes the estimation of the model parameters from time series surprisingly difficult. This is true even for low-dimensional systems, another even rarer instance in naturally occurring time series.

Here, we revisit the problem proposed by McSharry and Smith [1], who introduced an improved method over standard least-square fits to estimate the structural parameter of a low-dimensional deterministically chaotic system (the logistic map). We discuss the caveats underlying this problem. We suggest as well a ''piecewise'' likelihood method (called the segmentation method) that in fact is a simple modified version of the multiple shooting method proposed in Refs. [2–4]. Our conclusion stresses the inherent difficulties in formulating a bona fide statistical theory of structural parameter estimations for noisy deterministic chaos.

*Electronic address: sornette@moho.ess.ucla.edu

I. DEFINITION AND NATURE OF THE PROBLEM

Let us consider the supposedly simple problem considered by McSharry and Smith [1], in which one measures the sample $s_1, \ldots, s_N$ with

$$s_i = x_i + \eta_i, \tag{1}$$

where the underlying dynamical one-dimensional discrete recurrence equation

$$x_{i+1} = F(x_i, a) \equiv 1 - a x_i^2 \tag{2}$$

is known and the η_i's are Gaussian $N(0,\epsilon)$ independent and identically distributed (i.i.d.) random variables with zero mean and standard deviation ϵ. The problem is to determine the model parameter a from the measurements $s_1, \ldots, s_N$, knowing that Eq. (2) is the true dynamics.

At first sight, this problem looks like a statistical estimation of an unknown structural parameter, given observational data. Thus, it seems that such standard statistical approach as the maximum likelihood method can be used for this purpose. However, strictly speaking, it is not so. Indeed, the likelihood function $L(a, x_1 | s_1, \ldots, s_N)$ reads

$$\ln L(a, x_1 | s_1, \ldots, s_N) \propto -N \ln(\epsilon) - \frac{1}{2\epsilon^2} \sum_i [s_i - F^{(i)}(x_1, a)]^2, \tag{3}$$

where $F^{(i)}(x_1,a)$ is the ith iteration of the logistic map (2) with parameter a and initial value x_1. The key point of difficulty is that the ith iteration $F^{(i)}(x_1,a)$ depends on i, i.e., it is *nonstationary* although the dynamical system (2) has an invariant measure $\mu(x)$. For nonstationary dependence of the partial likelihood functions $L_i(\theta|s_i)$ on i, special additional (sufficient) conditions discussed below should be fulfilled guaranteeing optimal asymptotical properties of maximum likelihood estimates (MLE), such as consistency, asymptotical normality, efficiency. It should be stressed that no verification of these sufficient conditions has appeared so far in the literature devoted to this problem. Thus, application of the maximum likelihood method to unstable nonlinear systems distorted by noise has no mathematical ground so far. Then, widely used numerical simulations of examples are not sufficient to consider suggested methods as consistent and should be complemented with proofs of results showing under what conditions the ML or Bayesian methods continue to apply to nonstationary time series like Eq. (1).

A first taste of the difficulty of the problem is given by an analysis of the behavior of the ''one-step least-squares (LS) estimation'' and of the ''total least-squares'' method, given in the Appendix. The Appendix shows that least-squares methods are biased and should be corrected before comparing these to other methods, as done in Ref. [1]. In particular, the Appendix shows that it was *a priori* unfair or inappropriate to compare any estimate obtained with a given method (such as the one advocated by McSharry and Smith [1]) to uncorrected ML estimates due to the nonstationarity of the function; the appropriate corrections can be obtained from the standard statistical theory of confluence analysis [6–8].

We present a brief synthesis of known facts on the standard ML theory in the stationary case (for sample of i.i.d. random values) and its generalization to the nonstationary case. We expose these questions in a qualitative nonstrict manner to communicate with less-mathematically minded readers.

II. ON MAXIMUM LIKELIHOOD THEORY

A. The stationary case

Suppose that the probability density (PD) depends on a (multivariate in the general case) parameter θ. Suppose further that that the observed sample $s_1,\ldots,s_N$ is constituted by random values obeying the PD $f(x|\theta)$ with a true parameter value θ_0. Then, the derivative of the logarithm of the likelihood function $L(\theta|s_1,\ldots,s_N)=f(s_1|\theta)\cdot\ldots\theta f(s_N|\theta)$ can be expanded in the vicinity of θ_0 as follows:

$$\frac{1}{N}\frac{\partial \ln L(\theta)}{\partial\theta}\cong B_0+B_1(\theta-\theta_0)+\frac{1}{2}B_2(\theta-\theta_0)^2, \qquad (4)$$

where

$$B_0=\frac{1}{N}\sum_{i=1}^{N}\left.\frac{\partial\ \ln f(s_i|\theta)}{\partial\theta}\right|_{\theta=\theta_0}, \qquad (5)$$

$$B_1=\frac{1}{N}\sum_{i=1}^{N}\left.\frac{\partial^2\ \ln f(s_i|\theta)}{\partial\theta^2}\right|_{\theta=\theta_0}, \qquad (6)$$

$$B_2=\frac{1}{N}\sum_{i=1}^{N}\left.\frac{\partial^3\ \ln f(s_i|\theta)}{\partial\theta^3}\right|_{\theta=\theta_0}. \qquad (7)$$

Under some regularity conditions imposed on the PD $f(x|\theta)$ (see exact formulation in Ref. [9]), it can be shown that random value B_0 converges in probability to zero (which is its expectation), the random value B_1 converges in probability to some negative value $-I<0$, and the random value B_2 converges to some finite value C, as $N\to\infty$. Thus, taking θ sufficiently close to θ_0, we can make the third term in Eq. (4) much smaller in absolute value than B_1, whereas the first term in Eq. (4) can be made sufficiently small if N is sufficiently large. It follows that the likelihood equation

$$\frac{\partial \ln L(\theta)}{\partial\theta}=0 \qquad (8)$$

has a root θ^* (called the maximum likelihood estimate) in an arbitrary small vicinity of the true parameter value θ_0, which is to say that a consistent solution of the likelihood equation exists. In accordance with the law of the large numbers, the random value B_1 converges to the expectation

$$\mathrm{E}\left[\left.\frac{\partial^2\ \ln f(s_i|\theta)}{\partial\theta^2}\right|_{\theta=\theta_0}\right]=-\mathrm{E}\left[\left.\frac{\partial \ln f(s_i|\theta)}{\partial\theta}\right|_{\theta=\theta_0}\right]^2, \qquad (9)$$

the last equality following from the regularity conditions guaranteeing the possibility to differentiate under integrals including the PD. The positive expectations

$$I_i(\theta)=\mathrm{E}\left[\left.\frac{\partial \ln f(s_i|\theta)}{\partial\theta}\right|_{\theta=\theta_0}\right]^2=I(\theta) \qquad (10)$$

are called partial Fisher's amounts of information. We see that the cumulative Fisher's amount of information

$$J_N(\theta)=\sum_{i=1}^{N} I_i(\theta)=NI(\theta) \qquad (11)$$

grows linearly with N. It follows from the central limit theorem that the random variable

$$N^{1/2}B_0=N^{-1/2}\sum_{i=1}^{N}\left.\frac{\partial \ln f(s_i|\theta)}{\partial\theta}\right|_{\theta=\theta_0} \qquad (12)$$

converges in probability to a Gaussian random variable with zero mean and variance $1/I(\theta)$. For regular probability densities, the convergence of $N^{1/2}\ B_0$ to some Gaussian random variable holds true not only for the true parameter value θ_0 but for other values θ in the $1/N^{1/2}$ vicinity of the true value, in other words, in the $1/J_N(\theta)$ vicinity of the true value. This last property is called local asymptotic normality (LAN) of the PD in question. It is very important, and all generaliza-

tions of maximum likelihood methods for nonstationary cases are based on the LAN property.

B. The nonstationary case

In the nonstationary case, the role of the normalizing factor is played by the cumulative Fisher's information J_N. But unlike the stationary case, J_N can grow nonlinearly. We refer to Theorem 4 of Ref. [9] giving sufficient conditions for consistency, asymptotical normality and efficiency of MLE for nonstationary case. Sufficient conditions should provide, roughly speaking, following three main properties:

(i) Unlimited growth of the cumulative Fisher's information with N: $J_N \to \infty$, as $N \to \infty$.

(ii) Local convergence of the derivative of the likelihood (12) to a Gaussian law (LAN).

(iii) Relative smallness of the third term in the likelihood expansion (4).

There are some other additional conditions but they are not so evident and explicit (see Ref. [9] for an exact formulation). In the following, we restrict the discussion to the particular case of observations representing nonstationary signals $x_i(\theta)$ on a noisy background (1) (additive noise).

In addition to the signal generated by the logistic map (2), it is useful to compare with two other examples serving as benchmarks:

(i) The standard harmonic map with unknown frequency a

$$x_k(a) = \sin(ak). \tag{13}$$

(ii) The sine signal with exponentially nonlinear frequency

$$x_k(a) = \sin[\exp(ak)]. \tag{14}$$

The parameter a in both cases is taken to belong to a compact parametric space on the real axis. The signal (14) has an important property in common with the logistic map, namely, it is exponentially sensitive to parameter variations (see Refs. [10–12] for an early related discussion on the sensitive dependence on parameters).

Figure 1 shows the log-likelihood as a function of a for the map (14). It looks like a Brownian trajectory. A similar effect is observed for the logistic map. Figure 2 shows the log-likelihood obtained for the logistic map as a function of the structural parameter a, for a trajectory of 50 data points generated with $a = 1.85$ and a noise standard deviation of 0.2. We have fixed $x_1 = 0.3$ to a known value, thus deliberately facilitating the estimation problem. The log-likelihood has been sampled with steps in a equal to 10^{-7}. Rather than the smooth function generally obtained in standard ML problems, as well as for the generalization (13) involving the estimation of the frequency of a harmonic oscillation in the presence of Gaussian noise, we observe that the log-likelihood function for the logistic map is like a white noise.

From the similarity of these two figures, one can hope that the MLE problem for these two cases are similar. But, unlike the situation for the logistic map, the cumulative Fisher's information and other characteristics can be evaluated analytically for the signal (14). We now use Theorem 4 of Ref.

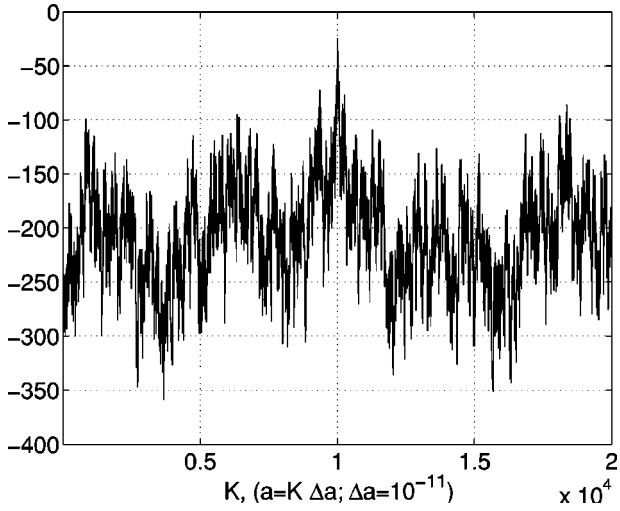

FIG. 1. Log-likelihood as a function of a for the map (14), for a trajectory of $N = 50$ data points generated with $a = 0.4$ and a noise standard deviation of 0.2.

[9] giving sufficient conditions for consistency, asymptotical normality and efficiency of MLE for nonstationary cases to check if they are fulfilled. Among the four conditions of this theorem, we take for checking the following two (see Eqs. (4.11) and (4.12) in Chap. 3, Sec. 4, Theorem 4.2 of Ref. [9]):

$$\text{Condition 1: } \sup_a \left\{ \frac{U(a,N)}{[J_N(a)]^2} \right\} \to 0, \quad N \to \infty, \tag{15}$$

$$\text{Condition 2: } \frac{\sup_a \{[J_N(a)]^2\}}{\mathrm{Inf}_a \{[J_N(a)]^2\}} < \infty, \quad N \to \infty, \tag{16}$$

where the cumulative Fisher's information for the sum of the signal plus noise equals

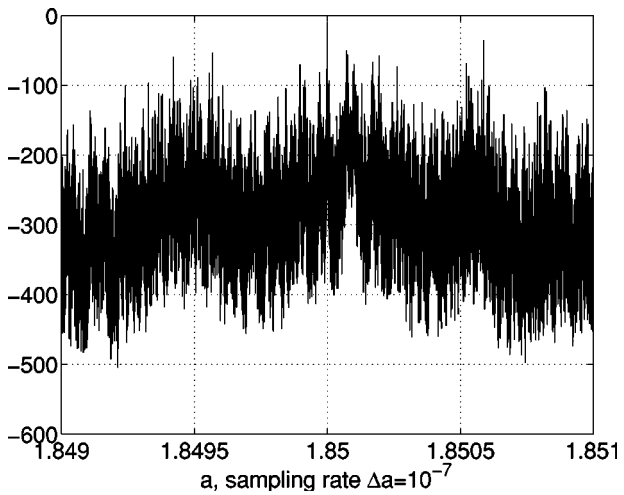

FIG. 2. Log-likelihood obtained for the logistic map as a function of the structural parameter a, for a trajectory of $N = 50$ data points generated with $a = 1.85$ and a noise standard deviation of 0.2. The initial value is fixed equal to $x_1 = 0.3$.

$$J_N(a)=\sum_{i=1}^{N}\left(\frac{\partial x_i}{\partial a}\right)^2, \tag{17}$$

and

$$U_N(a)=\sum_{i=1}^{N}\left(\frac{\partial x_i}{\partial a}\right)^4. \tag{18}$$

Supremum and infimum in Eqs. (15) and (16) are taken over an open bounded convex set of the parameter space. Loosely speaking, condition (16) provides the possibility of normalizing the first derivative of the log-likelihood [the first term in the right-hand side of Eq. (4)] in a vicinity $|a-a_0|<\delta$ of the true parameter value a_0 by the square root of the Fisher information $J_N(a_0)$, so that it can converge to a standard Gaussian random variable.

It can be shown that Eqs. (15) and (16) are fulfilled for the standard harmonic map (13) (see p. 267 in Ref. [9]), with

$$J_N(a)\sim N^3 \quad \text{and} \quad U_N(a)\sim N^5. \tag{19}$$

The cumulative Fisher's information $J_N(a)$ grow in this case as N^3 unlike the standard case of i.i.d. where it grows as N. As a consequence, the standard deviation of the limit Gaussian distribution of the MLE decreases as $1/N^{3/2}$ instead of the standard $1/N^{1/2}$.

For the sine exponential (14), we have

$$J_N(a)\sim N^2 e^{2aN} \quad \text{and} \quad U_N(a)\sim N^4 e^{4aN}, \tag{20}$$

leading to the violation of conditions, Eqs. (15) and (16). Of course, the violation of sufficient conditions does not automatically imply that the MLE are inconsistent. But still, there is no mathematical proof of their consistency, and one can suspect that the MLE of parameter a is inconsistent (see below).

For the logistic map (2), we cannot explicitly put down the derivatives of the signal $x_k(a,x_1)$ and are not able to check conditions (15) and (16) (and other conditions of the Theorem 4 mentioned above) analytically. Instead, we are forced to use numerical experiments to explore the consistency and other useful properties of ML estimates. Of course, the numerical experiments (computer simulations) cannot provide complete evidence confirming such or such assertion. There is no guarantee that under different parameter values, sample sizes, noise distribution, accuracy of computations, etc., the results of numerical experiments will confirm earlier conclusions. But still, such experiments are useful research tools for studying nonlinear dynamics, and sometimes they are quite illustrative and informative. We thus tried to evaluate numerically the cumulative Fisher information for the logistic map, using numerical experiments in which the standard deviation of the noise was taken equal to 1. We estimated numerically the following elements of the Fisher information matrix J:

$$J=\begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix},$$

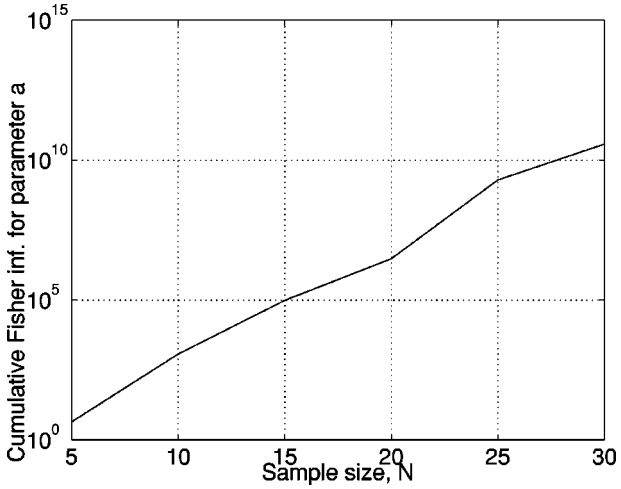

FIG. 3. Exponential growth of the cumulative Fisher information J_{11} as a function of N for the logistic map with $a=1.85$ and $x_1=0.3$.

where

$$J_{11}=\sum_{i=1}^{N}(\partial x_i/\partial a)^2, \tag{21}$$

$$J_{12}=J_{21}=\sum_{i=1}^{N}(\partial x_i/\partial a)(\partial x_i/\partial x_1), \tag{22}$$

$$J_{22}=\sum_{i=1}^{N}(\partial x_i/\partial x_1)^2. \tag{23}$$

We have taken $a=1.85$ and $x_1=0.3$. The increments for the evaluation of derivatives were taken equal to 10^{-9} for sample sizes $N=5$–15, and equal to 10^{-12} for $N=20$–30. The accuracy of the computation of the derivatives was controlled by decreasing the increments by a factor of 2 and checking that we get the same values to a precision with three figures.

Figure 3 shows the exponential growth of the diagonal elements of the Fisher matrix (with a slight nonmonotonicity at $N=20$) as a function of N. It turned out that the Fisher matrix becomes more and more degenerate with growing N, as shown in Fig. 4 which plots the ratio of smallest eigenvalue λ_1 of the Fisher matrix to its largest value λ_2. This ratio λ_1/λ_2 decreases exponentially which leads to a degeneracy of the inverse matrix $B=J^{-1}$. Similar effect were discovered by Horbelt [2], but only qualitatively without its explanation based on the Fisher information. In standard regular situations, the matrix B in the limit $N\to\infty$ gives the covariance matrix of the limit two-dimensional Gaussian distribution of ML estimates. For the logistic map, as we pointed out above, there is no theorem justifying the use of B as a measure of the standard deviations of the estimated parameters. Nevertheless, we shall use $\sqrt{B(1,1)}$ and $\sqrt{B(2,2)}$, shown in Fig. 5, as the "standard deviations" of the estimates of a and x_1 if the use of B was justified. These "standard deviations" decrease approximately as $1/N^{1.6}$, which

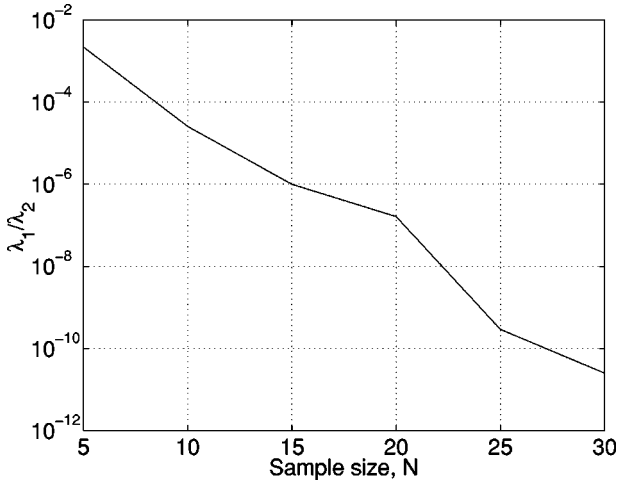

FIG. 4. Ratio of smallest eigenvalue λ_1 of the Fisher matrix to its largest value λ_2 for the logistic map with $a=1.85$ and $x_1=0.3$.

differs somewhat from the dependence $\sim 1/N$ given by Horbelt [2]. It should be noted that, if one of the two parameters a,x_1 was known then from the exponential growth of the Fisher partial information coefficients J_{11},J_{22}, the ''standard deviation'' for the remaining estimated parameter would decrease exponentially, in agreement with the numerical tests performed by Horbelt [2]. These results allow us to pinpoint the origin of the much slower decrease of the ''standard deviations'' for the estimation of two parameters as being due to the growing degeneracy of the Fisher matrix (B matrix). Figure 6 confirms this explanation by plotting $1-\rho$, where ρ is the ''coefficient of correlation'' calculated from the matrix B as a function of N. $1-\rho$ is found to decrease exponentially, reflecting a stronger and stronger correlation between the estimations of a and x_1 as N increases.

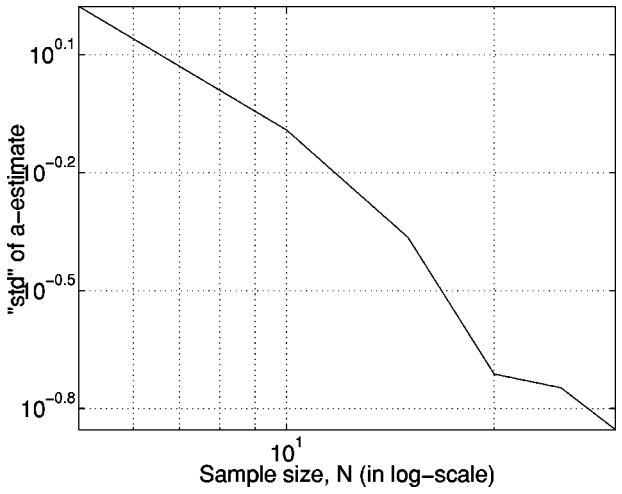

FIG. 5. Pseudo ''standard deviations'' $\sqrt{B(1,1)}$ of the estimate of a for the logistic map with $a=1.85$ and $x_1=0.3$.

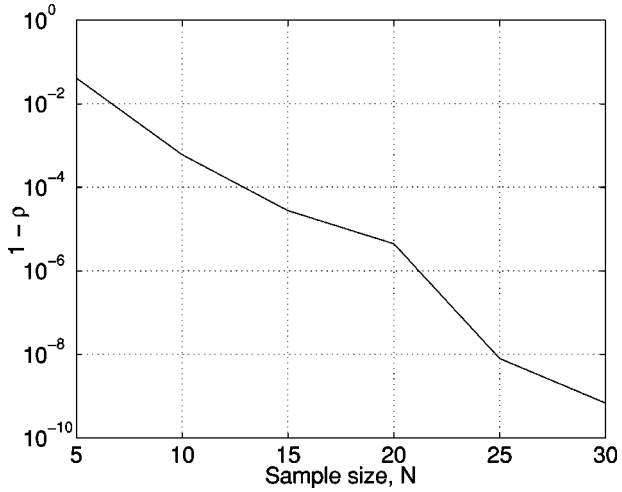

FIG. 6. $1-\rho$, where ρ is the ''coefficient of correlation'' calculated from the matrix B, as a function of the sample size N for the logistic map with $a=1.85$ and $x_1=0.3$.

III. A "PIECEWISE" MAXIMUM LIKELIHOOD APPROACH IN TERMS OF (a,x_1)

A. Constraints due to sensitive dependence on the initial value x_1 and on the parameter a

The majority of the conclusions drawn in Refs. [1–5,13,14)] are based on numerical experiments. But, in many cases, the information on these experiments is not complete. Therefore, we suggest to provide the following standard information (which can be debated) for each numerical experiment:

(i) Sample size.

(ii) Parameter set (range, grid step, separately for each parameter).

(iii) Noise distribution.

(iv) Number of simulations, or bootstrap number (if any).

(v) Accuracy of the computation.

(vi) Additional useful relevant information.

We feel that, without such information, it is difficult to judge the meaning of the corresponding numerical experiment. Other forms of the simulation information are quite possible.

In this spirit, we have performed a series of numerical experiments with the logistic map on the sensitivity of this map to variations of parameters (a,x_1). We have taken $a=1.85$ and $x_1=0.3$ as the reference values and have perturbed them by random noise uniformly distributed in the interval $[-\delta,+\delta]$, δ taking the values $10^{-3};10^{-7};10^{-11};10^{-15}$. The accuracy of the computer computation was 10^{-15}. We generated 1000 simulations for each δ and for each sample size N, where N was varied from $N=2$ up to $N=100$. The maximum absolute deviation between the perturbed and reference trajectories is found to increase with disturbed (a,x_1). The results are shown in Fig. 7, which show that small initial deviations increase exponentially so that deviations of the order 10^{-15} (corresponding to our computer accuracy) leads to a divergence of order 1 after about 80 iterations. An approximate empirical formula describing the exponential growth of $\epsilon(n)$ shown in Fig. 7 is

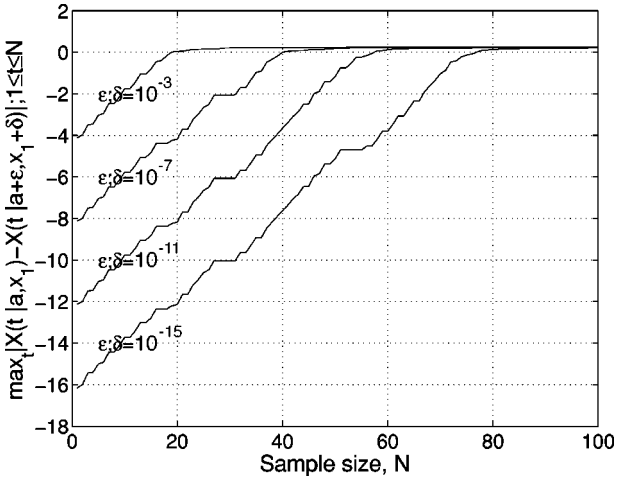

FIG. 7. Maximum absolute deviation $\epsilon(n)$ between the perturbed and reference trajectories as a function of sample size N for different amplitude δ of the initial deviation of (a,x_1) with respect to the reference values $a=1.85$ and $x_1=0.3$ for the logistic map.

$$\epsilon(n)=\epsilon(1)\times 10^{0.2N-1}. \tag{24}$$

Several other parameter values were tried for (a,x_1) leading to dependencies quite similar to Eq. (24). If we accept that Eq. (24) is approximately true for all parameter values, then it means that we cannot use samples with size $N>70$ for discriminating signals when the available computer accuracy is 10^{-15}. Using double precision shifts out this limit very little as seen from Eq. (24). As we are going to see, such high accuracy (10^{-15}) is needed for computing the logistic map likelihood. Thus, to speak about consistency of ML (or other) estimates solely on the basis of numerical experiments, as it is often done, is very dangerous. We know that ML estimates for nonlinear problems are biased for small samples. In our case, the bias for $N=70$ can be small (for high signal-over-noise ratio) but is nevertheless finite. Some modifications of the ML method such as multiple shooting [2–5] or the segmentation method we propose here amount to use subsamples of moderate size taken as sequential segments of the initial sample. We can affirm that such moderate size should not be more than 70 and, in fact, much less due to some other reasons, perhaps, no more than 25–30. Larger subsamples would be used inefficiently. Further combination of segments may not eliminate the finite bias. One can suspect that such finite bias is retained in these modifications and thus, these modifications are not consistent, in contradiction with what has been claimed in the literature. Of course, this is only a suspicion and not a strict proof of inconsistency, but in absence of strict proof of consistency, it has its own right for existence and is justified by the elements given above.

Let us now consider the MLE of (a,x_1) for the logistic map, which requires to minimize the sum

$$\sum_i [s_i-F^{(i)}(x_1,a)]^2, \tag{25}$$

which looks superficially as a standard nonlinear least-square sum. There is however one very important distinction, as we already pointed out above: the nonlinear function depends on the index i whereas, in the standard least-square method, one has a sum of the type

$$\sum_i [s_i-F(x_i,a)]^2, \tag{26}$$

where the $x_1,\ldots,x_N$ are assumed to be known. The crucial difference between Eq. (25) and Eq. (26) is that, in the former, the mean value of s_i is expressed in terms of a function which is dependent on the index i, with sensitive dependence on the initial conditions. As we noted above, this leads to an exponential growth of the cumulative Fisher information [this was verified analytically for the sine exponential map (14) and numerically for the logistic map]. Although the growth of the cumulative Fisher information is, generally speaking, useful for estimating the parameters (the faster the growth, the smaller is the variance of the limit distribution of ML estimates), such an exponential flow of information for the logistic map turns out to be too much to ''be digested by estimation algorithms.'' Figure 2 shows the log-likelihood as a function of a for the map (14) with true parameters $a=1.85$ and $x_1=0.3$, noise standard deviation 0.2, with $N=50$ with a sampling interval of a equal to 10^{-7}.

Figure 2 gives a sense of the nature of the problem in a pictorial way. Rather than the smooth function generally obtained in standard ML problems, as well as for the sine harmonic map (13), we observe that the log-likelihood function for the logistic map is like a white noise. Our tests show that, if one wishes to use a grid method to search the maximum of the likelihood, one has to take a sampling step in a which decreases exponentially with the sample size N as

$$\Delta a=10^{-0.2N}. \tag{27}$$

This is nothing but the inverse of the dependence with N of the average distance between two trajectories with slightly different values of a reported in Eq. (24). For $N=70$, this gives $\Delta a=10^{-14}$, which reaches the precision limits of our computer. It is clear that Eq. (27) prevents us from using time series of size $n=70$ or larger. In addition, if the parameter x_1 is unknown (which is the real situation), then the grid is two dimensional, which worsens considerably the computation problem. In Fig. 2, the true value $a=1.85$ is the maximum of the likelihood function. But this is no more necessarily true in a two-dimensional situation where x_1 is unknown. To account for this fact, we shifted the (a,x_1) grid in our ''piecewise'' ML method for each simulated trajectory by a random two-dimensional value distributed uniformly over the grid mesh. Thus, the true parameter values were so-to-say ''distributed'' uniformly over the unit mesh, which is more realistic than their coincidence with grid axes resulting in a high likelihood value. Note also that any local search for a maximum will be trapped in local maxima. Even improved nonlocal searches such as simulated annealing or genetic algorithms (which are not exhaustive as the grid method) may be trapped far from the true value by the hierarchy of height structures visible in Fig. 2. A natural idea would be to attempt to smooth the likelihood function over a.

A direct smoothing approach would however require that the likelihood be calculated on a dense grid, which brings us back to square one. Any other smoothing method should be strictly justified analytically, otherwise one faces the risk of unexpected surprise arising from the exponential instability of the logistic map.

The lack of a rigorous theoretical basis should not prevent us from using the ML method, especially since other solutions previously proposed in the literature turns out to be controversial. But some additional precaution must be pointed out. The exponential sensitivity to variations of the initial value x_1 and of a of the logistic map is actually the more stringent limiting factor for the usable size N of time series. For the parameters a for which the logistic map exhibits the phenomenon of sensitivity upon the parameter [10], the direct minimization of Eq. (25) is not feasible directly even for time series as short at $n=70$.

B. Addressing the sensitivity to small variations in x_1 and a: multiple shooting versus segmentation ("piecewise")

The following remarks on the parametrization of the signal of the logistic map $x_{i+1}=F^{(i)}(x_1,a)$ are helpful to motivate our proposed solution. Formally, this parametrization using two parameters (x_1,a) is the most parsimonious. But, as we saw above, it assumes unlimited accuracy of all related computations and is not robust, not realistic and practically useless. Instead, we can look upon a trajectory of the logistic map as a randomlike function forgetting its initial condition after some (not quite explicitly defined) time interval of K iterations. Thus, if we have a realization of length N, we could look at it as if consisting of $m=N/K$ different subrealizations of length K. Each subrealization, in addition to a common parameter a, contains its own parameter corresponding to its initial value x_{Kj+1}, for $j=1,\ldots,m-1$. One can consider these initial values as unknown parameters, in addition to (x_1,a). The corresponding parametrization of the signal of the logistic map by the set of parameters $a,x_1,x_{K+1},\ldots,x_{K(m-1)+1}$ is now robust. This robust parametrization has a very characteristic feature: the number of parameters increases with the sample size N: as we are going to see, this property has an important consequence for the ML estimation. This leads to the natural idea of cutting the sample $(s_1,\ldots,s_N)$ into K portions of some size K, and to treat each portion separately. If we minimize the total sum of the residuals

$$\sum_{i=1}^{K}[s_i-F^{(i)}(a,x_1)]^2+\sum_{i=K+1}^{2K}[s_i-F^{(i)}(a,x_{K+1})]^2+\cdots+\sum_{i=(m-1)K+1}^{mK}[s_i-F^{(i)}(a,x_{K(m-1)+1})]^2 \tag{28}$$

over the parameters $a,x_1,x_{K+1},\ldots,x_{K(m-1)+1}$, we arrive at a scheme similar to multiple shooting (see Refs. [2–5]). It should be noted that, in this scheme, the parameter a is kept the same in all terms of the sum (28), whereas the initial values x_{Kj+1} are fitted separately in each term. In most of the implementation of Refs. [2–5], the constraint of continuity of model trajectory is added. Thus, initial conditions for a model on different segments become mutually dependent variables and only x_1 remains a free parameter. In this sense, multiple shooting resembles the shadowing technique of Judd [14] (the difference being that Judd does not impose the condition of minimal square distance between the model and the observed time series). In other instances, the authors of Refs. [2–4] also use different segments with independent initial conditions applying multiple shooting inside each segment separately [5].

Here, we suggest another version consisting of fitting separately the parameters in each subinterval. In our "segmentation fitting" scheme, the parameter a is allowed to take different values in different subintervals. Our segmentation fitting scheme treats each subinterval absolutely independently of other intervals, and the resulting a estimates are averaged.

It is difficult to decide *a priori* what version of these two methods gives the most efficient estimation of the parameter a. We are going to show several examples in which the segmentation fitting method results in smaller sample mean-square deviations of the a estimates. However, we do not exclude the possibility that there are situations depending on the choice of the sample size N, the size K chosen to partition the time series into $m=N/K$ subseries, and the noise level ϵ, for which the multiple shooting method has a smaller sample mean-square deviation. We think that while these two methods have probably comparable efficiency, the segmentation fitting method is somewhat simpler and has smaller bias. Both of them may be inconsistent as $N\to\infty$, or differently worded, their consistency has not been proved analytically, although the bias can be very small if N is large and the noise level is small. Advocates of the multiple shooting method often assert that their method is unbiased as $N\to\infty$, referring to asymptotical consistency of ML estimates, and to arguments of type: "Multiple shooting makes use of all available information in a robust, reasonable way." As we noted above, there is no strict mathematical ground for such affirmations. We showed that robust parametrization leads to unlimited growth of the number of fitted parameters. This in turn creates a situation where ML estimates can be biased. We give a simple example of such inconsistency of ML estimates in the Appendix.

C. Practical implementation and tests of the "piecewise" method

Let us cut the sample $s_1,\ldots,s_N$ into n_1 portions of size no more than $n_2=20$, and treat each portion separately. As we said, this amounts to reestimating a different initial condition for each such subseries, which is a natural step since the sensitivity upon initial conditions amounts to losing the information on the specific value of the initial condition.

Our numerical tests show that our MLE works well by considering subseries of size in the range $n_2=4$–25 (for the true value of a equal to the value 1.85 considered by McSharry and Smith [1] that we take as our benchmark for the sake of comparison). For larger samples (say, $N=100$), we recommend to cut this sample into n_1 subsamples of size

TABLE I. Comparison of sample MSE (mean-square error with respect to true value $a=1.85$) of a estimates obtained by two methods: multiple shooting and ''piecewise'' ML; initial value $x_1=0.3$; number of simulations $m=100$.

N	n_1	n_2	Noise std	a-mesh	x_1-mesh	Multiple shooting MSE	''Piecewise'' ML method MSE
14	7	2	0.1	0.005	0.005	0.086	0.090
35	7	5	0.1	0.005	0.005	0.058	0.050
70	7	10	0.1	0.0025	0.0025	0.040	0.039
40	20	2	0.1	0.0025	0.0025	0.051	0.035
60	20	3	0.1	0.0025	0.0025	0.042	0.027

$n_2=4$–25, and treat them separately, and then to average the resulting n_1 a estimates. In order to determine the optimal value of n_1 for a fixed N (say $N=100$) and for the value $a=1.85$ investigated here, we calculate the standard deviation a over the n_1 subsamples as a function of n_1. We find that, basically independently of the noise level ϵ, the pair $n_1=25, n_2=4$ gives the smallest standard deviation a.

We have implemented this approach and compared it first with the shooting method in Table I. One can observe that the two methods give comparable mean-square errors, with a slight advantage to the ''piecewise'' ML for larger N. We also observe in general (not shown) that the multiple shooting method has a larger bias than the ''piece-wise'' ML.

We now turn to the comparison between the ''piecewise'' ML and the method proposed by McSharry and Smith [1].

IV. ML VERSION OF McSHARRY AND SMITH AND COMPARISONS

The main result of McSharry and Smith's paper [1] consists in their formulas (13) and (14) for their proposed ML cost function. Their idea is to substitute in the ML cost function the unknown invariant measure $\mu_a(x)$ of the dynamical system (2), for a given value of the parameter a, for what should be a realization of the latent variables x_i's. Note that a should be varied in order to determine the maximum likelihood. In practice, the integral over the unknown invariant measure $\mu_a(x)$ is replaced by a sum over a model trajectory (which can be calculated since the model is assumed to be known) of length $\tau \gg N$. Unfortunately, this most important step is not confirmed by any numerical results (see below).

Before continuing, let us note that there is a mistake in the probability density function PDF and likelihood given by their Eqs. (7)–(9). Using the intuition that pairs (s_i, s_{i+1}) should be used in their Eqs. (5) and (6) to track the deterministic relation between x_i and $x_{i+1}=F(x_i,a)$, we see that a single latent variable x_i is associated with each pair (s_i, s_{i+1}) since s_i is compared with x_i and s_{i+1} with $F(x_i,a)$. Thus, each x_i is used only once when scanning all possible pairs (s_i, s_{i+1}), for $i=1,\ldots,N-1$ and in their ML cost function (13) and (14). Actually, the correct likelihood should use only once each *observed* random variable s_i, not the latent variable x_i. Therefore, using pairs (s_i, s_{i+1}), McSharry and Smith take into account each $s_i, i=2,\ldots,N-1$ twice, and the end values s_1, s_N once. For $N \gg 2$, their expression (7) is approximately equals (up to the end terms) to the square of the correct likelihood. Taking the logarithm in their Eq. (13) gives approximately twice the correct likelihood, which gives almost the same estimate as the exact likelihood.

While this mistake has no serious consequences for the numerical accuracy of their calculation for long time series $N \gg 2$, it illustrates the difference between their construction of the likelihood and our direct approach presented in the preceding section. By writing the conditional likelihood for a pair (s_i, s_{i+1}) under a latent variable x_i, and by averaging this conditional likelihood weighted by the invariant measure $\mu(x|a)$, McSharry and Smith suggest that, by doing so, they incorporate additional information on the system in question. If we had a usual probability space, then such averaging would provide the unconditional likelihood of the pair (s_i, s_{i+1}) but, for deterministically chaotic time series, the exact meaning of this averaging is not clear. Another questionable step of McSharry and Smith is to multiply these pairwise likelihoods as if the pairs (s_i, s_{i+1}) were independent. If this was so, this would indeed give the unconditional likelihood for the data sample $s_1,\ldots,s_N$. McSharry and Smith avoid the maximization with respect to x_1 in their likelihood (13) and (14) and replace it by an averaging over a proxy of the invariant measure. It is doubtful that such a step is warranted, not speaking of optimality, and we can hope that our approach would lead to a more efficient estimate of a.

We now compare our ''piecewise'' maximum likelihood approach in terms of (a, x_1) proposed in Sec. III with the ML method of McSharry and Smith using numerical tests. We consider 1000 time series with $N=100$ data points and subdivide each of them into $n_1=25$ subseries of $n_2=4$ data points. We fix the true a equal to 1.85 as in Ref. [1] and study noises with standard deviations equal to 0.5 and 1.0. Table II shows a significant improvement offered by our ''piecewise'' ML method over the average ML of McSharry and Smith, at least for the set of parameters studied here. It is not possible to guarantee that this will be the case for all possible parameter values but we believe our method cannot be worse than the average ML of McSharry and Smith. A difficulty that should be mentioned is that the chaotic nature of the dynamics and in particular the sensitivity of the invariant measure with respect to the control parameter a is reflected in an ugly looking log-likelihood landscape shown in Fig. 8, with

TABLE II. Comparison between the ML method of McSharry and Smith [1] and our ''piecewise'' ML method described in Sec. III over 1000 realizations of the system (2) with true value $a=1.85$ giving 1000 time series of length $N=100$, each of them decorated with Gaussian noise with two different standard deviations (SD) (0.5 and 1). q_1 and q_2 are the sample quantiles at the 2.5% and 97.5% probability level, so that q_2-q_1 gives the width of the 95% confidence intervals. Our ''piecewise'' ML method provides us with an estimation $\hat{\epsilon}$ of the standard deviation of the noise given in the last column.

Noise		Mean (a)	SD (a)	q_1	q_2	q_2-q_1	$\hat{\epsilon}$
SD 0.5	Ref. [1]	1.816	0.0714	1.630	1.925	0.295	
	''piecewise'' ML	1.841	0.0390	1.762	1.913	0.151	0.459
SD 1	Ref. [1]	1.764	0.123	1.510	1.975	0.465	
	''piecewise'' ML	1.885	0.0467	1.781	1.959	0.178	0.766

many competing valleys. Standard numerical methods like gradient or simplex are unapplicable. We have used a systematic 2D grid search. Other methods in the field of computational intelligence, such as stimulated annealing and genetic algorithms, could also be used. The sensitivity of the invariant measure with respect to the control parameter a means that the invariant distribution can bifurcate from an almost uniform distribution on the interval $[-a,1]$ to a distribution consisting of three δ functions (this happens around $a\approx 1.75$).

In addition to performing better, our ''piece-wise'' ML approach does not depend on the noise level, in contrast with the ML cost function (13) and (14) proposed by McSharry and Smith [1]. This is an important advantage when the true level of noise is not known (noise error). Our method is insensitive to such noise error while we have found examples where the optimal estimation of the structure parameter a with the method of McSharry and Smith is obtained for a value of the noise standard deviation different from the true value. In general, the true noise level is not known and the method of McSharry and Smith does not apply in such situation. Our ''piecewise'' ML method actually provides us with

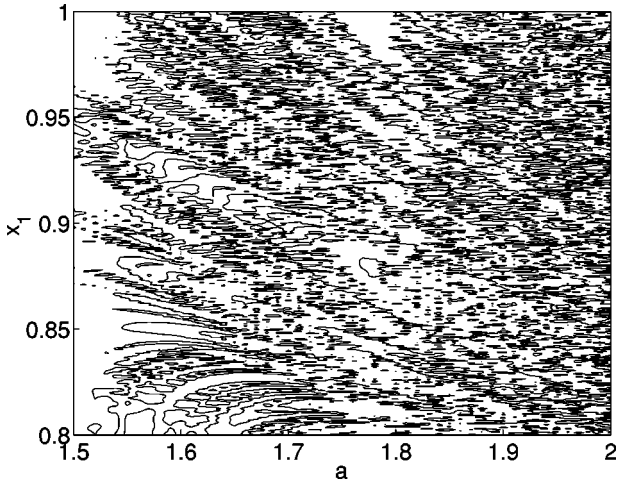

FIG. 8. Contour lines of the ''piecewise'' log-likelihood given by expression (3) for a given realization of $N=20$ data points generated with a starting value $x_1=0.9$, $a=1.85$, and noise standard deviation equal to 1. The log-likelihood landscape is similar to a two-dimensional Brownian sheet (two-dimensional generalization of a random walk).

an estimation $\hat{\epsilon}$ of the standard deviation of the noise given in the last column of Table II. These estimates have a small bias down (two fitted parameters were taken into account), which may be due to the fact that n_1 is not sufficiently large ($n_1=25$; $n_2=4$; $N=n_1n_2=100$).

V. DISCUSSION OF OTHER APPROACHES

Meyer and Christensen [13] have proposed to replace the *ad hoc* construction of the ML cost function of McSharry and Smith by a Bayesian approach, assuming noninformative priors for the structural parameter a, for the initial value x_1 and for the standard deviation of the noise. Their approach improves significantly on McSharry and Smith [1] by recognizing the role of x_1 but turns out to be incorrect, as shown by Judd [14], because their approach amounts to assuming a stochastic model, thus referring to quite another problem.

Based on the formulation of Ref. [15], Judd [14] develops a formulation which is almost identical to the likelihood (3) but there are important distinctions. Similarly to us, Judd introduces x_1 but he does not employ it. He prefers to eliminate the dependency on x_1 by averaging this parameter with a fiducial distribution (see, e.g., Ref. [8], Chap. 21, Interval Estimation, Fiducial Intervals). Judd incorrectly calls the method based on his Eqs. (4) and (5) a ML method. In fact, his Eqs. (4) and (5) gives a a hybrid of ML, Bayesian and so-called fiducial methods. It is a ML method with respect to the structural parameter a. It is Bayesian with respect to the initial value x_1. It is fiducial since it does not assume any a *priori* density for x_1, but uses a prior density function $\rho(s_1-w)$ (using the notation of Ref. [14]) that is in fact a Gaussian density of the noise with mean value equal to the unknown initial value s_1. Using such density is equivalent to weighting a two-parameter likelihood by weights corresponding to different values of noise disturbances. Thus, the averaged likelihood (5) in Ref. [14] describes an ensemble of different noise disturbances of an unknown initial value s_1. This provides a (reasonable but not optimal) method of elimination of the second parameter x_1 from the maximization procedure. It is neither a pure Bayesian method that would assume explicitly some a *priori* density for x_1 which could be arbitrary, and not necessarily equal to $\rho(x_1-w)$, nor a ML method for two unknown parameters as we suggested above in Sec. III.

In this context in view of the emphasis on Bayesian meth-

ods to solve this problem [13,14], it is perhaps useful to stress that the probability theory rule $P\{A,B\}=P\{A|B\}P\{B\}$ is often freely called "the Bayes rule." This is why the averaging of likelihoods over conditional state variables can be called Bayesian approaches, although this is not quite correct since the latent (state) variables are not random values in the standard meaning of this notion (as it is assumed by McSharry and Smith), although the state variables have a limit invariant measure, as we said above. The Bayesian approach assumes that parameters are random values. We stressed already that the series of state variables can be considered as a degenerate set of random values that are determined by one single random variable, namely, x_1. What is more natural? To consider x_1 as a random variable with a distribution determined by the invariant measure, or to consider x_1 as an unknown parameter to be estimated? The answer, in our opinion, is dictated by consideration of efficiency: the different examples that we have explored suggest that the latter is as a rule more efficient (has smaller mean-square error), at least for some combinations of sample size N and noise level.

As all the above has shown, the major obstacle is the loss of information on the initial value x_1 by the unstable logistic map beyond some limited number of iterations. We proposed the simple recipe of cutting the time series in short pieces and of averaging the estimations. Judd proposes a shadowing method [14]. It is not obvious that this will result in a consistent estimation and that this will overcome the intrinsic difficulty in treating long realizations (which is a necessary condition for unbiased estimations).

In sum, there is no analytical proof of consistency for all the estimation methods discussed until now (including the suggestions performed by the most convincing work to date [14] and our "piecewise" ML). It is useful to analyze the only method to our knowledge for which one can derive a proof of consistency in the present context, that is, the method of statistical moments.

VI. THE METHOD OF STATISTICAL MOMENTS

The method of statistical moments provides a consistent estimate of the parameters for nonlinear maps with ergodic properties. The method of statistical moments is the unique theoretically proven consistent estimator among all methods suggested so far by other authors. Although the moment estimates are known to have little efficiency, they are consistent. Consistency of all estimates suggested earlier including ours above were confirmed only numerically, which is very dangerous for instable nonlinear maps.

We consider four moment of the observed time series: $\langle s\rangle_N$, $\langle s^2\rangle_N$, $\langle s^3\rangle_N$, and $\langle s_i s_{i+1}\rangle_N$, where the brackets stand for time averaging over some time interval N. Building on the knowledge that the series $\{x_i\}$ is ergodic [16] and using Eqs. (1) and (2), we obtain the following relations

$$\langle s\rangle_N \to \langle x\rangle_\infty , \tag{29}$$

$$\langle s^2\rangle_N \to \langle x^2\rangle_\infty + \epsilon^2, \tag{30}$$

TABLE III. Estimation of the structural parameter a by the method of statistical moments [expression (34)] for the logistic map $x_{i+1}=1-ax_i^2, a=1.85$; the observations are $s_i=x_i+\eta_i$; η_i is a Gaussian random variable $N(0,\epsilon)$. As in Table II, q_1 and q_2 are the sample quantiles at the 2.5% and 97.5% probability level, so that q_2-q_1 gives the width of the 95% confidence intervals. Each estimate for a and standard deviation (SD) is based on 1000 simulated samples.

Sample size N	Noise SD ϵ	Estimate $(a)\pm$SD	q_1	q_2	q_2-q_1
100	0.05	1.8768±0.0926	1.684	2.000	0.316
1000	0.05	1.8544±0.0418	1.774	1.936	0.162
10 000	0.05	1.8503±0.0136	1.824	1.878	0.054
100 000	0.05	1.8499±0.0044	1.842	1.858	0.016
100	0.1	1.8456±0.1546	1.499	2.000	0.501
1000	0.1	1.8532±0.0815	1.693	2.000	0.307
10 000	0.1	1.8505±0.0279	1.795	1.908	0.113
100 000	0.1	1.8497±0.0089	1.833	1.867	0.034
100	0.5	1.2411±0.7331	0	2.000	2.000
1000	0.5	1.6907±0.3496	0.903	2.000	1.097
10 000	0.5	1.8244±0.1659	1.467	2.000	0.533
100 000	0.5	1.8554±0.0741	1.715	2.000	0.285

$$\langle s^3\rangle_N \to \langle x^3\rangle_\infty + 3\langle x\rangle_\infty \epsilon^2, \tag{31}$$

$$\langle s_i s_{i+1}\rangle_N \to \langle x\rangle_\infty - a\langle x^3\rangle_\infty . \tag{32}$$

Besides, averaging Eq. (2), we get

$$\langle x\rangle_\infty = 1 - a\langle x^2\rangle_\infty . \tag{33}$$

This provides us with five limit relations (29)–(33) with five unknown parameters: $a, \langle x\rangle_\infty$, $\langle x^2\rangle_\infty$, $\langle x^3\rangle_\infty$, and ϵ. Solving these five relations with respect to the unknown parameters, we get the so-called estimates of the method of moments:

$$\hat{a} = \frac{\langle s_i s_{i+1}\rangle_N + 2\langle s\rangle_N - 3(\langle s\rangle_N)^2}{3\langle s\rangle_N\langle s^2\rangle_N - \langle s^3\rangle_N}, \tag{34}$$

$$\langle \hat{x}\rangle_\infty = \langle s\rangle_N , \tag{35}$$

$$\langle \hat{x}^2\rangle_\infty = \langle s^2\rangle_N - \hat{\epsilon}^2, \tag{36}$$

$$\langle \hat{x}^3\rangle_\infty = \frac{1}{\hat{a}}[\langle s\rangle_N - \langle s_i s_{i+1}\rangle_N], \tag{37}$$

$$\hat{\epsilon}^2 = \frac{\langle s^3\rangle_N - \langle x^3\rangle_\infty}{3\langle s\rangle_N}. \tag{38}$$

Because of the limit relations (29)–(32) (which are valid because of the ergodicity of the time series $\{x_i\}$ [16]), the estimates (34)–(38) are consistent if $N\to\infty$.

We present in Table III the estimates of the parameter a given by expression (34). The consistency of the method of statistical moments is clearly suggested by the numerical re-

sults, as seen from the bracketing of the true value by $(a) \pm$ SD and by q_1 and q_2. However, as we already pointed out, the method of statistical moments is rather inefficient: the ratio of its standard deviation for a to that of the ''piecewise'' ML is about 4 for $N=100$ and $\epsilon=0.1$ for instance.

VII. CONCLUDING REMARKS

We have proposed a ''piecewise'' ML method to estimate the structural parameter of a deterministically chaotic low-dimensional system (the logistic map), which adds the initial value x_1 to the structural parameter to be determined. We have compared quantitatively this method with the ML method proposed by McSharry and Smith [1] based on an averaging over the unknown invariant measure of the dynamical system. A key aspect of the implementation of our approach lies in the compromise between the need to use a large number of data points for the ML to decrease a systematic bias and the unstable nature of dynamical trajectories which loses exponentially fast the memory of the initial condition. This second aspect prevents using our ''piecewise'' ML for systems larger than 10–25 data points. For larger time series, we have found it convenient to divide them into subsystems of very small lengths and then to average over their estimations. Numerical tests suggest that this ''piecewise'' ML method provides often significantly better estimates than previously proposed approaches.

The difference between the averaging over the invariant measure and our ''piecewise'' ML of McSharry and Smith is reminiscent of the distinction between ''annealed'' versus ''quenched'' averaging in the statistical physics of random systems, such as spin glasses [17,18]. It has indeed been shown that the correct theory of strongly heterogeneous media is obtained by performing the thermal Gibbs-Boltzmann averaging over fixed structural disorder realizations, similarly to our use of a specific trajectory of the latent variables x_i's. In contrast, performing the thermal Gibbs-Boltzmann averaging together with an averaging over different realization of the structural disorder describes another type of physics, which is not that of fixed heterogeneity. This second incorrect type of averaging is similar to the averaging of the ML over the invariant measure performed by McSharry and Smith.

There are several ways to improve our approach. One simple implementation is to use overlapping running windows. Another method is to reestimate the realized trajectory by using the extended Kalman filter method (however, difficulties may arise due to the existence of a maximum in the logistic map). Using shadowing methods as proposed in Ref. [14] in our context would also be interesting to investigate.

Let us end with a cautionary note. As we just said, the ML approach for two parameters (a,x_1) that we suggest here evidently works only for a limited sample size N (perhaps, $N<25$ or so) due to the sensitivity upon initial conditions of the chaotic logistic map. As is well-known in classical statistics, ML estimates have a bias that can be considerable if N is not large (say, $N<100$ or so). The ML estimates are usually only asymptotically unbiased. Thus, for $N=25$ (and all the more for $N=4$), ML estimates can exhibit a considerable bias. Thus, averaging biased estimates as we proposed may not result in a consistent estimation. Therefore, we cannot assert that our ML method (as well as any other suggested methods) is consistent. We can only observe, for particular combinations of the considered parameters, the numerically determined mean-square error of our suggested estimates with respect to the true parameter value. We are pleased if these errors are not too high, although our estimates can be biased (though, with small bias). But we are not able to make such bias arbitrarily small by increasing the sample size N, due to the instability under the iterations of the logistic map which leads to a loss of information about the initial value x_1. Thus, the situation is rather hopeless for the establishment of a meaningful statistical theory of estimation using the continuous theory of classical statistics to such discontinuous objects as the invariant measures of chaotic dynamical systems.

ACKNOWLEDGMENTS

We are grateful to K. Ide for useful discussions. This work was partially supported by a LDRD–Los Alamos grant and by the James S. McDonnell Foundation.

APPENDIX: ONE-STEP AND TOTAL LEAST-SQUARES ESTIMATIONS

McSharry and Smith noticed that the one-step least-squares method gives strongly biased results for the estimation of a [1]. Indeed, the method of estimation of the parameter a by the one-step least-squares method is evidently inconsistent, since the deviations (of the random variables) to be minimized in a least-squares sense are

$$\begin{aligned} s_{i+1}-F(s_i,a) &= x_{i+1}+\eta_{i+1}-F(x_i+\eta_i,a) \\ &= \eta_{i+1}+2ax_i\eta_i+a\eta_i^2, \end{aligned} \tag{A1}$$

which has nonzero expectation equal to $a\epsilon^2$. But, the fundamental least-squares principle consists in the minimization of deviations with zero mean. There are no least-squares schemes that would suggest to minimize random deviations with nonzero mean depending on an unknown parameter. Thus, it is not reasonable to include the least-squares method in any reasonable comparison.

The method called by McSharry and Smith as ''total least squares'' (TLS) is applied in situation when the variables x_i are known only with some errors η_i. This situation is called in statistics a ''confluence analysis,'' or ''estimation of a structural relation between two (or more) variables in the presence of errors on both variables'' [6–8]. In such a situation of confluence analysis, since the x_i's are in fact unknown (nuisance) parameters whose number grows with sample size, there is no guarantee of consistency of the ML estimates of the structural parameter a.

As an example, let us consider the very simple confluent scheme:

$$Y_i=X_i+\eta_i, \tag{A2}$$

$$Z_i = X_i + \zeta_i . \tag{A3}$$

Suppose we observe a sample of N pairs $(Y_i, Z_i), i = 1, \ldots, N$, where X_i are unknown arbitrary values and η_i, ζ_i are i.i.d. Gaussian random variables with standard deviation ϵ. The problem consists in estimating the parameter ϵ. Similarly to the situation with Eqs. (1) and (2) studied in Ref. [1], no restrictions are placed on the X_i's. The likelihood $L(\epsilon, X_1, \ldots, X_N | (Y_i, Z_i), i = 1, \ldots, N)$ is

$$\begin{aligned} L(\epsilon, X_1, \ldots, X_N | (Y_i, Z_i), \quad i = 1, \ldots, N) \\ \propto \epsilon^{-2N} \exp\left[-(1/2\epsilon^2) \sum_{i=1}^{N} (Y_i - X_i)^2 \right. \\ \left. -(1/2\epsilon^2) \sum_{i=1}^{N} (Z_i - X_i)^2 \right]. \end{aligned} \tag{A4}$$

The MLE $\hat{X}_i$'s of the X_i's (that coincide in this case with the least-squares estimates) are

$$\hat{X}_i = \frac{Y_i + Z_i}{2}. \tag{A5}$$

Inserting Eq. (A5) into Eq. (A4), we get

$$\begin{aligned} \hat{L}(\epsilon | (Y_i, Z_i), \quad i = 1, \ldots, N) \\ \propto \epsilon^{-2N} \exp\left[-(1/4\epsilon^2) \sum_{i=1}^{N} (Y_i - Z_i)^2 \right]. \end{aligned} \tag{A6}$$

Thus, the MLE of the parameter ϵ obtained from Eq. (A6) satisfies

$$\epsilon^2 = \frac{1}{4N} \sum_{i=1}^{N} (Y_i - Z_i)^2. \tag{A7}$$

Since $\mathrm{E}[(Y_i - Z_i)^2] = 2\epsilon^2$, the estimate (A7) is inconsistent. A consistent (corrected) estimate is

$$\epsilon^2 = \frac{1}{2N} \sum_{i=1}^{N} (Y_i - Z_i)^2. \tag{A8}$$

Thus, we see that the MLE of the structural parameter ϵ is inconsistent due to the increasing number of nuisance parameters. Thus, the direct use of the least-square (or TLS) in the confluent situation is not justified, and was not recommended in any statistical textbook. Instead, standard statistical works recommend a ''corrected'' ML estimates (see, for instance Refs. [7,8]).

We should stress in addition that there is a significant difference between the standard confluent analysis and the problem addressed in Ref. [1]. Confluent analysis deals with arbitrary unknown (distorted) arguments x_i, whereas in Ref. [1], the latent variables x_i are related by the nonlinear map (2). The information on the structure of the x_i's is not used in confluence analysis while it can really help in the estimation procedure as shown in Ref. [1] and in the present work.

[1] P.E. Mcsharry and L.A. Smith, Phys. Rev. Lett. **83**, 4285 (1999).

[2] W. Horbelt, dissertation, University of Freiburg, 2001, http://webber.physik.uni-freiburg.de/~horbelt/diss/

[3] J. Timmer, H. Rust, W. Horbelt, and H.U. Voss, Phys. Lett. A **274**, 123 (2000).

[4] W. Horbelt, J. Timmer, and H.U. Voss, Phys. Lett. A **299**, 513 (2002).

[5] W. Horbelt, J. Timmer, M.J. Bunner, R. Meucci, and M. Ciofini, Phys. Rev. E **64**, 016222 (2001).

[6] R. Frisch, Statistical Confluence Analysis by Means of Complete Regression Systems, Oslo, 1934.

[7] R.C. Geary, J. Am. Stat. Assoc. **48**, 94 (1953).

[8] M. Kendall and A. Stuart, *The Advanced Theory of Statistics, Curvilinear Dependencies*, 2nd ed. (Hafner, New York, 1961), Chap. 29, Sec. 29.50.

[9] I.A. Ibragimov and R.Z. Hasminskii, *Statistical Estimation: Asymptotic Theory* (Springer-Verlag, New York, 1981), Chap. 3, Sec. 5, Theorem 5.1.

[10] J.D. Farmer, Phys. Rev. Lett. **55**, 351 (1985).

[11] D.K. Umberger and J.D. Farmer, Phys. Rev. Lett. **55**, 661 (1985).

[12] R. Ecke, J.D. Farmer, and D.K. Umberger, Nonlinearity **2**, 175 (1989).

[13] R. Meyer and N. Christensen, Phys. Rev. E **62**, 3535 (2000).

[14] CK. Judd, Phys. Rev. E **67**, 026212 (2003).

[15] M.L. Berliner, J. Am. Stat. Assoc. **86**, 939 (1991).

[16] P. Collet and J.-P. Eckmann, *Iterated Maps on The Interval as Dynamical Systems* (Birkhäuser, Basel, 1980).

[17] M. Mézard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[18] D. Sornette, *Critical Phenomena in Natural Sciences* (Springer, Heidelberg, 2000), Chap. 16.