Better Nonlinear Models from Noisy Data: Attractors with Maximum Likelihood

Patrick E. McSharry* and Leonard A. Smith†
Mathematical Institute, University of Oxford, Oxford OX1 3LB, England
(Received 19 March 1999)

A new approach to nonlinear modeling is presented which, by incorporating the global behavior of the model, lifts shortcomings of both least squares and total least squares parameter estimates. Although ubiquitous in practice, a least squares approach is fundamentally flawed in that it assumes independent, normally distributed (IND) forecast errors: nonlinear models will not yield IND errors even if the noise is IND. A new cost function is obtained via the maximum likelihood principle; superior results are illustrated both for small data sets and infinitely long data streams.

PACS numbers: 05.45.Tp, 02.60.Pn

A nonlinear model must be tuned via parameter estimation, ideally forcing it to mimic the observations. Typically, tuning aims for parameters which yield the least squared error [1–3] (or total least squared error [4,5]) between the one-step forecasts and the data. After proving that even for the simplest nonlinear models both least squares and total least squares systematically reject the correct parameter values (i.e., those that generated the data), a new and more robust method is derived which incorporates the global dynamics of the model (and hence its attractor). Failure to recognize the effects of imperfect observations will lead to biased parameter estimates, whereas an inability to reflect the data indicates model error. The present Letter focuses on the first issue: using the fact that one can always estimate the probability density function (PDF) of the model-state variable for different values of the unknown parameters, the maximum likelihood principle is employed to derive a new cost function which incorporates this information. This global approach has the potential to outperform all one-step (or few-step) methods, whether they are based on least squares criteria or some future improvement. Note that even with an infinite amount of data the optimal least squares solution is simply incorrect; see Fig. 1 and the discussion below. The new cost function is shown to yield results consistent with the correct answer even for relatively small data sets and large noise levels in a variety of chaotic systems. It is applicable to high dimensional systems and may also be applied to nonlinear stochastic systems.

Suppose the evolution of a system's state variable, $\mathbf{x}_i \in \mathbb{R}^m$, is governed by the map

$$\mathbf{x}_{i+1} = F(\mathbf{x}_i, \mathbf{a}), \quad (1)$$

where the model's parameters are contained in the vector $\mathbf{a} \in \mathbb{R}^l$. For $m = 1$, the system state x_i is a scalar; assuming additive measurement noise η_i yields observations $s_i = x_i + \eta_i$. In a noise free setting (i.e., $\eta_i = 0 \; \forall \; i$), $l + 1$ sequential measurements $s_i, s_{i+1}, \ldots, s_{i+l}$ would, in general, be sufficient to determine $\mathbf{a}$. With noise, the PDF of both the measurement noise and the model-state variables are required to estimate $\mathbf{a}$ properly.

Given the correct model structure $F(\mathbf{x}, \mathbf{a})$ and data generated by particular parameters $\mathbf{a}_0$ (the "true" parameter values), a cost function is often used to obtain an estimate of the unknown parameters $\mathbf{a}_0$. Ideally, this estimate converges to $\mathbf{a}_0$ in the limit of an infinite number of observations. Complications of nonlinearity combined with the randomness of unavoidable measurement errors suggests a likelihood analysis for parameter estimation. Indeed, both least squares and total least squares are special cases of the likelihood method.

The one-step *least squares* (LS) estimate, $\mathbf{a}_{\text{LS}}$, is the value of $\mathbf{a}$ which minimizes the least squares cost function

$$C_{\text{LS}}(\mathbf{a}) = \sum_{i=1}^{N-1} E_i^2, \quad (2)$$

where $E_i = s_{i+1} - F(s_i, \mathbf{a})$, the one-step prediction error. Figure 1 shows the failure of this well-known technique when applied to the well-known logistic map [6]: $a_0 = 2$,

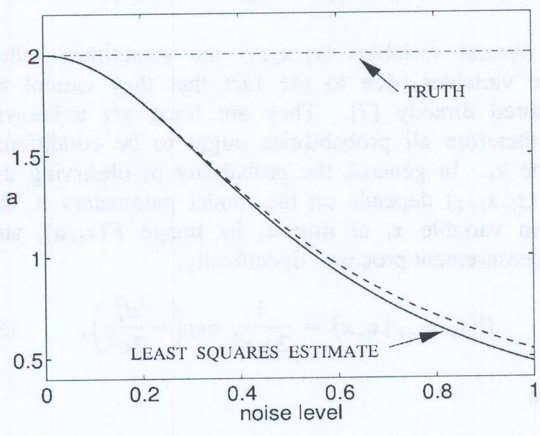

FIG. 1. Least squares estimate of a as a function of noise level using the analytic result (3) corresponding to an infinite data set. The underlying system is the logistic map with $a_0 = 2$. Parameter a is systematically underestimated for both normally distributed noise (solid), and uniformly distributed noise (dashed). Both deviate significantly from the correct value (dot-dashed).

0031-9007/99/83(21)/4285(4)$15.00

yet $a_{\rm LS} < 2$ at all nonzero noise levels, even for an infinite data set. To see this, first recall this one-dimensional map $F(x,a) = 1 - ax^2$. Writing the observed prediction error, $s_{i+1} - F(s_i, a)$, explicitly in terms of the underlying system state and a realization of the noise process yields $E_i = \eta_{i+1} - a_0 x_i^2 + a(x_i + \eta_i)^2$. When the least squares cost function (2) is a minima $\sum_{i=1}^{N} [E_i \partial E_i / \partial a] = 0$, and in the limit of an infinitely long data set, this sum converges to an integral over x taken with respect to the system's invariant measure, $\mu(x,a)$. For $a = 2$ in the logistic map, $\mu(x,2) = 1/\pi\sqrt{1-x^2}$ for $-1 \le x \le 1$ and zero otherwise. Thus $\langle x^2 \rangle = 1/2$ and $\langle x^4 \rangle = 3/8$. Since $\langle x^n \rangle = 0$ for odd n [$\mu(x,2)$ is an even function], if the distribution of the noise process is also even, then

$$a_{\rm LS} = \left(\frac{4\langle \eta^2 \rangle + 3}{8\langle \eta^4 \rangle + 24\langle \eta^2 \rangle + 3} \right) a_0 . \qquad (3)$$

Equation (3) yields the parameter estimate corresponding to an infinitely long data set as a function of noise level. For uniformly distributed noise [i.e., $\eta \overset{\rm d}{=} U(-\epsilon, \epsilon)$], $\langle \eta^n \rangle = \epsilon^n/(n+1)$ for even n, while for normally distributed noise [i.e., $\eta \overset{\rm d}{=} N(0, \epsilon^2)$], $\langle \eta^2 \rangle = \epsilon^2$ and $\langle \eta^4 \rangle = 3\epsilon^4$, the noise level is defined as $\sigma_{\rm noise}/\sigma_{\rm signal}$, where $\sigma_{\rm noise}^2$ and $\sigma_{\rm signal}^2$ are the variances of the noise and the signal, respectively; for the uniform case $\sigma_{\rm noise}^2 = \epsilon^2/3$, whereas $\sigma_{\rm noise}^2 = \epsilon^2$ for the normal case. Despite having a complete knowledge of the measurement process and data of infinite duration, the parameter estimates are biased.

Let (x_i, x_{i+1}) denote a successive pair of system states corresponding to observations (s_i, s_{i+1}) where

$$s_i = x_i + \eta_i, \qquad \eta_i \overset{\rm d}{=} N(0, \epsilon^2) . \qquad (4)$$

The system variables (x_i, x_{i+1}) are sometimes called *latent* variables, due to the fact that they cannot be measured directly [7]. They are fixed yet unknown, and therefore all probabilities ought to be conditional on the x_i. In general, the probability of observing the pair (s_i, s_{i+1}) depends on the model parameters $\mathbf{a}$, the system variable x_i at time i, its image $F(x_i, \mathbf{a})$, and the measurement process. Specifically,

$$P(s_i, s_{i+1} \,|\, \mathbf{a}, x) = \frac{1}{2\pi\epsilon^2} \exp\!\left(-\frac{d_i^2}{2\epsilon^2} \right), \qquad (5)$$

where

$$d_i^2(s_i, s_{i+1}, x, \mathbf{a}) = (s_i - x)^2 + [s_{i+1} - F(x, \mathbf{a})]^2 . \qquad (6)$$

Assuming the s_i and s_{i+1} are independent, the probability of observing a sequence of $N-1$ pairs, $\mathbf{S} = \{(s_i, s_{i+1})\}_{i=1}^{N-1}$, corresponding with a particular set of model states $\tilde{\mathbf{X}} = \{\tilde{x}_i\}_{i=1}^{N-1}$, is given by the joint PDF:

$$P(\mathbf{S} \,|\, \mathbf{a}, \tilde{\mathbf{X}}) = \prod_{i=1}^{N-1} P(s_i, s_{i+1} \,|\, \mathbf{a}, \tilde{x}_i) . \qquad (7)$$

Identifying the *likelihood* of parameters $\mathbf{a}$ generating the data $\mathbf{S}$ with the probability of observing data $\mathbf{S}$ given that the model has parameters $\mathbf{a}$ (see [8]) yields

$$L(\mathbf{a}, \tilde{\mathbf{X}} \,|\, \mathbf{S}) = P(\mathbf{S} \,|\, \mathbf{a}, \tilde{\mathbf{X}}) , \qquad (8)$$

where the conditional status of the model-state variables is explicit. Substituting (5) and (7) in (8) yields

$$L(\mathbf{a}, \tilde{\mathbf{X}} \,|\, \mathbf{S}) = \frac{1}{(2\pi\epsilon^2)^{N-1}} \exp\!\left(-\frac{1}{2\epsilon^2} \sum_{i=1}^{N-1} d_i^2 \right). \qquad (9)$$

$L(\mathbf{a}, \tilde{\mathbf{X}} \,|\, \mathbf{S})$ depends on the PDF of the likely model states and the parameters which maximize (9) will vary with the assumptions made regarding this distribution. These assumptions are paramount to this Letter; ignoring information from the distribution will lead to *total least squares* (TLS), whereas requiring consistency between the data and the PDF of the model-state variables yields the new cost function below. Casella and Berger [7] compare these assumptions for linear systems.

Ignoring the PDF of the model-state variables, total least squares resolves the dependence on $\tilde{x}_i$ by substituting any values $\tilde{x}_i$ which maximize $L(\mathbf{a} \,|\, \mathbf{S})$, that is

$$L(\mathbf{a} \,|\, \mathbf{S}) = \frac{1}{(2\pi\epsilon^2)^{N-1}} \exp\!\left(-\frac{1}{2\epsilon^2} \sum_{i=1}^{N-1} \min_{x \in \mathbb{R}} d_i^2 \right). \qquad (10)$$

The maximum of $L(\mathbf{a}, \mathbf{S})$ then corresponds to the minimum of the associated TLS cost function

$$C_{\rm TLS}(\mathbf{a}) = \sum_{i=1}^{N-1} \min_{x \in \mathbb{R}} d_i^2 . \qquad (11)$$

Thus, while the least squares cost function (2) minimizes the squared vertical distances $d_i^2 = [s_{i+1} - F(s_i, \mathbf{a})]^2$, the TLS solution minimizes the squared perpendicular distances (6) between the measured point (s_i, s_{i+1}) and a point on the hypersurface $[\tilde{x}, F(\tilde{x}, \mathbf{a})]$. No restrictions are placed on the values $\tilde{x}_i, i = 1, N$: the $\tilde{x}_i$ are not a trajectory of $F(x, \mathbf{a})$ nor do they reflect $\mu(x, \mathbf{a})$. Using the particular $\tilde{x}_i$ which minimizes each d_i^2 reflects the decision to ignore any knowledge of the PDF of the model-state variables. But, given that the model is always in hand, the PDF of the model, $\mu(x, \mathbf{a})$, is always obtainable (not that of the system, but of the model). This additional information may be incorporated by integrating the dependence on x_i out of the likelihood function in (9), yielding

$$L(\mathbf{a} \,|\, \mathbf{S}) = \prod_{i=1}^{N-1} \int_x P(s_i, s_{i+1} \,|\, \mathbf{a}, x) d\mu(x, \mathbf{a}) , \qquad (12)$$

and associated *maximum likelihood* (ML) cost function

$$C_{\rm ML}(\mathbf{a}) = -\sum_{i=1}^{N-1} \log \int_x \exp\left(-\frac{d_i^2}{2\epsilon^2}\right) d\mu(x,\mathbf{a}). \quad (13)$$

Equation (13) is the main result of this Letter; it is superior to both $C_{\rm LS}$ and $C_{\rm TLS}$. In practice, the integral in (13) is usually replaced by a sum over a model trajectory whose length, $\tau \gg N$, is limited only by computational constraints. Thus

$$C_{\rm ML}(\mathbf{a}) \approx -\sum_{i=1}^{N-1} \log\left(\sum_{k=1}^{\tau} \exp\left\{-\frac{1}{2\epsilon^2}[(s_i - x_k)^2 + (s_{i+1} - x_{k+1})^2]\right\}\right), \quad (14)$$

where $x_k = F^k(x_0, \mathbf{a})$ and x_0 is any posttransient value $F(x, \mathbf{a})$ in the relevant basis of attraction. For fixed N this can be computed efficiently [9]. As $\tau \to \infty$, the particular value of x_0 is irrelevant since the sum is dominated by those values of k for which $s_i \approx x_k$ and $s_{i+1} \approx x_{k+1}$, the particular values of k being irrelevant.

The logistic map's invariant measure varies drastically for different values of a. The bifurcation diagram (see Fig. 2a) illustrates this behavior in the range $1.5 \le a \le 2$. With $a_0 = 2$ and a noise level of 0.19, $a_{\rm LS} = 1.7$ (see Fig. 1); note from Fig. 2a that this corresponds to a model with a period three orbit. Obviously, values of a corresponding with periodic windows are unlikely to be responsible for a data set with wildly aperiodic behavior.

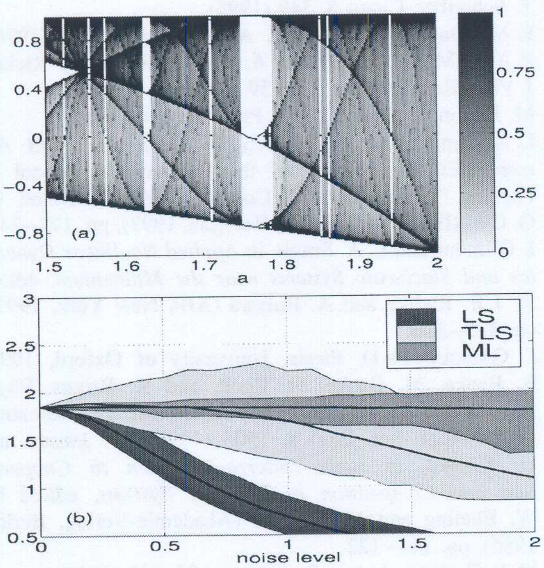

FIG. 2. Logistic map: (a) bifurcation diagram illustrating the variation in $\mu(x, a)$, (b) distribution of estimates contrasting the LS,TLS, and ML cost functions from 1000 realizations where $a_0 = 1.85$, $N = 100$, with normally distributed noise. The shading reflects the 95% limits, the solid line the mean.

Equation (14) allows this visually obvious result to be included implicitly into a cost function.

The TLS cost function for the logistic map may be obtained analytically by solving the cubic equation

$$\frac{\partial d_i^2}{\partial x_i} = 4a^2 x_i^3 + [2 + 4(s_{i+1} - 1)a]x_i - 2s_i = 0, \quad (15)$$

and taking the root satisfying $\partial^2 d_i^2/\partial x_i^2 > 0$. Results for three different cost functions are shown in Fig. 2. While the TLS cost function is much better than that of LS, the ML cost function is better still for all cases considered. Results are shown for data sets $N = 100$; for larger N the TLS solution improves, but in all cases tested the spread of the distribution remains smaller for the ML estimate. The simplicity of the logistic map makes it a weak test case and motivates further trials.

The Moran-Ricker map [10] has a functional form $F(x, a) = x \exp[a(1 - x)]$; its invariant measure (not shown) allows larger values of x with larger parameter values a. Figure 3 illustrates the results for each cost function where $a_0 = 3.7$ and $N = 100$. The ML cost function consistently yields the best estimates for all noise levels considered. While it is common to claim an algorithm generalizes to higher dimensional cases, this algorithm is easily generalized to $m > 1$ by substituting the term $[||\mathbf{s}_i - \mathbf{H}(\mathbf{x}_k)||]$ for the term $[(s_i - x_k)^2]$ in Eq. (14). Here the function $\mathbf{H}$ projects the system state vector $\mathbf{x}$ into the space of observations $\mathbf{s}$. In delay coordinates, this corresponds to taking the "last" component of $\mathbf{x}_k$. Figure 4 shows results for the two-dimensional Hénon map [11] in delay coordinates, $F(x_i, x_{i-1}) = 1 - ax_i^2 + bx_{i-1}$, where $a_0 = 1.4$ and $b_0 = 0.3$. While the ML cost function surface is more highly structured due to sensitivity to the parameters, its minima are in the relevant regions as opposed to the smooth but incorrect LS minimum. The LS cost function has a biased minima,

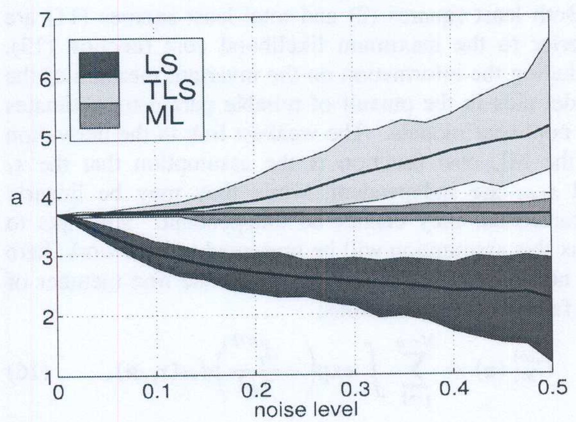

FIG. 3. Moran-Ricker map: a comparison of LS,TLS, and ML cost functions for 200 realizations with $a_0 = 3.7$, $N = 100$ with normally distributed measurement errors. The shading is as in Fig. 2.

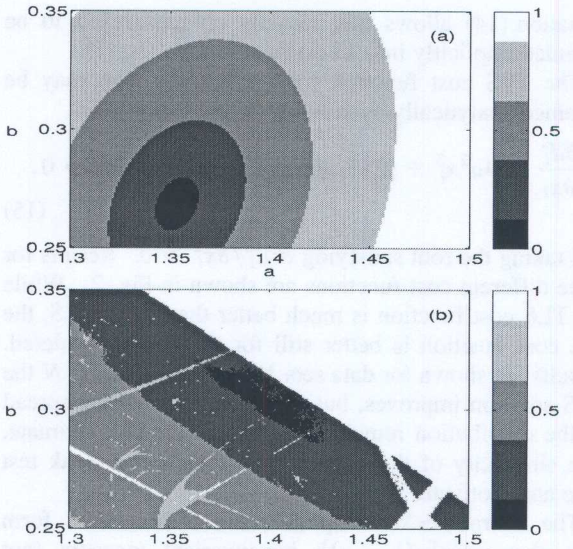

FIG. 4. Value of cost function in parameter space for a 2D delay reconstruction of the Hénon map for $a_0 = 1.4$, $b_0 = 0.3$, $N = 500$, and a noise level of 0.05: (a) C_{LS} and (b) C_{ML}.

while ML is consistent with (a_0, b_0). Whether this consistency is worth the computation depends on the problem at hand. Certainly the fact that ensemble forecasts of chaotic models using the ML estimates will relax naturally to a distribution consistent with $\mu(\mathbf{x}, \mathbf{a})$ is of value [12]. Better estimates of $\mathbf{a}$ also allow improved long term deterministic forecasts. The fact that higher dimensional models may require much larger data sets is a problem of uniqueness under the observations and cannot be laid at the door of the cost function. In Fig. 4, $N = 500$ and the noise level is 0.05. Equation (14) also allows an estimate of the magnitude of dynamical noise in a stochastic system [12] when the shape of the distribution of the noise is correctly specified.

Both least squares (2) and total least squares (11) are inferior to the maximum likelihood cost function (13). Including the information on the invariant measure of the model aids in the pursuit of reliable parameter estimates for nonlinear models. The weakest link in the derivation of the ML cost function is the assumption that the s_i and s_{i+1} are independent; while they may be linearly uncorrelated, they cannot be independent. Attempts to relax this assumption will be presented in later work; here we note that (13) may be viewed as the first member of the family of cost functions:

$$C_{\mathrm{ML}}^{(n)}(\mathbf{a}) = \sum_{i=1}^{N-n} \int \exp\left(-\frac{d_i^{(n)2}}{2\epsilon^2}\right) d\mu(x_i, \mathbf{a}), \qquad (16)$$

where

$$d_i^{(n)2} = \sum_{j=0}^{n} [s_{i+j} - F^j(x_i, \mathbf{a})]^2. \qquad (17)$$

For $n > 0$, $C_{\mathrm{ML}}^{(n)}(\mathbf{a})$ is a multistep cost function [3,13] moderating the assumption that the s_i are independent, the aim being to find parameter values which both have the correct PDF and shadow the observations [12,13]. The ultimate $n = N - 1$ multistep least squares approach, solving simultaneously for x_0 and a, may prove intractable for $N = 500$ (see [14]). Future work will also focus upon the choice of optimal model order in local polynomial prediction, and the interpretation of cost functions when the underlying model structure is unknown [15].

This work was supported by EC Grant No. ERBFMBICT950203, ONR Grant No. N00014-99-1-0056, and Pembroke College, Oxford.

*Email address: mcsharry@maths.ox.ac.uk
†Email address: lenny@maths.ox.ac.uk

[1] Å. Björck, *Numerical Methods for Least Squares Problems* (SIAM, Philadelphia, 1996).
[2] P. Grassberger, T. Schreiber, and C. Schaffrath, Int. J. Bifurcation Chaos Appl. Sci. Eng. **1**, 521 (1991).
[3] L. Jaeger and H. Kantz, Chaos **6**, 440 (1996); H. Kantz and L. Jaeger, Physica (Amsterdam) **109D**, 59 (1997).
[4] S. Van Huffel and J. Vandewalle, *The Total Least Squares Problem* (SIAM, Philadelphia, 1991).
[5] E. J. Kostelich, Physica (Amsterdam) **58D**, 138 (1992).
[6] R. M. May, Nature (London) **261**, 459 (1976).
[7] G. Casella and R. L. Berger, *Statistical Inference* (Wadsworth & Brooks/Cole, California, 1990).
[8] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C* (Cambridge University Press, Cambridge, 1992), 2nd ed.
[9] T. Schreiber, Chaos **5**, 349 (1995).
[10] R. M. May and G. Oyster, Am. Nat. **110**, 573 (1976); P. A. P. Moran, Biometrics **6**, 250 (1950); W. E. Ricker, J. Fish. Res. Bd. Can. **11**, 559 (1954).
[11] M. Hénon, Commun. Math. Phys. **50**, 69 (1976).
[12] L. A. Smith, in *Nonlinearity in Geophysics and Astrophysics,* Proceedings of the International School of Physics "Enrico Fermi", Course CXXXIII, edited by G. Cini (Il Nuovo Cimento, Bologna, 1997), pp. 177–246; I. Gilmour and L. A. Smith, in *Applied Nonlinear Dynamics and Stochastic Systems near the Millennium,* edited by J. B. Kadtke and A. Bulsara (AIP, New York, 1997), pp. 335–340.
[13] I. Gilmour, Ph.D. thesis, University of Oxford, 1998; E. Baake, M. Baake, H. Bock, and K. Briggs, Phys. Rev. A **45**, 5524 (1992); J. Timmer, Int. J. Bifurcation Chaos Appl. Sci. Eng. **8**, 1505 (1998); W. Jansen and U. Kriegel, in *Lotka-Volterra-Approach to Cooperation and Competition in Dynamic Systems,* edited by W. Ebeling and M. Peschel (Akademie-Verlag, Berlin, 1985), pp. 114–122.
[14] M. L. Berliner, J. Am. Stat. Assoc. **86**, 938 (1991).
[15] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson, Physica (Amsterdam) **51D**, 52 (1991); L. A. Smith, Physica (Amsterdam) **58D**, 50 (1992); L. A. Smith, Philos. Trans. R. Soc. London A **348**, 371 (1994).