Estimation of Noise Levels for Models of Chaotic Dynamical Systems

J. P. M. Heald and J. Stark

Centre for Nonlinear Dynamics, University College London, Gower Street, London WC1E 6BT, United Kingdom

(Received 19 January 1999)

We investigate how far it is possible to identify and separate dynamical noise from measurement noise in observed nonlinear time series. Using Bayesian methods, we derive estimates for the two noise levels, and find that, given a good model of the dynamics, these can give accurate results even if the dynamical noise level is orders of magnitude smaller than the measurement noise level, whereas a simple calculation of root mean square error badly understates the dynamical noise. We argue that this allows better estimates of the underlying dynamical time series, and so better predictions of its future and of its fundamental dynamical properties.

PACS numbers: 05.45.Tp, 05.40.Ca

Many complex physical phenomena can be generated by relatively simple nonlinear mechanisms. Often the precise nature of such mechanisms is not known and one is forced to build models directly from observed data, and in particular from observed time series. Deterministic nonlinear autoregressive (NAR) models have proved a convenient and popular structure for this purpose. Such models can in principle capture all of the coordinate free properties of a deterministic physical system. However, in practice no physical system is entirely free of noise, and no model is an exact representation of reality. We must therefore expect errors in the model dynamics (dynamical noise), and uncertainty in matching the model states to observations (measurement noise).

Such dynamical noise may represent genuine stochastic behavior in the physics, the effects of unmodeled phenomena, or other imperfections in the model dynamics. Measurement noise can reflect not only random instrumental errors in observables but also representativeness error that they may be fundamentally different quantities to the idealized model variables. A correct balance between the two forms of noise is particularly important for making good estimates of the recent state of the system, which is a prerequisite for predicting its future. This is well known in linear systems through their effect on the Kalman filter. In nonlinear systems changing the balance can also alter estimates of the deterministic skeleton of the dynamics, and modify system invariants such as Lyapunov exponents and the fractal dimension. It is thus important that our models accurately represent both forms of noise, and that we have reasonable estimates of their magnitudes (which may be far from clear, even in models with a strong physical motivation). The latter problem appears to have received relatively little attention in the physics literature. In particular, it has been usual to equate the root mean square (RMS) prediction error with the dynamical noise level in a system.

The aim of this Letter is to develop a more sophisticated analysis, which leads to more reliable estimates of the noise levels and demonstrates that in general the RMS prediction error is not a good indicator of the dynamical noise level (even after measurement noise reduction). Since the way in which the two types of noises interact with a nonlinear deterministic system is far from obvious, our analysis is based on the consistent use of a full probabilistic model M for the noise levels $\boldsymbol{\sigma} = \{\sigma_D, \sigma_M\}$, the underlying time series $\mathbf{x}$, and the observed data $\mathbf{y}$. A widely used form for this is the stochastic NAR structure

$$\begin{aligned} x_{n+1} &= f(x_n, x_{n-1}, x_{n-2}, \ldots) + \epsilon_{n+1}^{(\mathrm{D})}, \\ y_{n+1} &= x_{n+1} + \epsilon_{n+1}^{(\mathrm{M})}, \end{aligned} \tag{1}$$

where $\epsilon^{(\mathrm{D})}$ and $\epsilon^{(\mathrm{M})}$ are independent identically distributed Gaussian dynamical and measurement noise processes with mean 0 and variance σ_D^2 and σ_M^2, respectively.

The joint probability density $P(\mathbf{y}, \boldsymbol{\sigma} \mid M)$ for $\mathbf{y}$ and $\boldsymbol{\sigma}$ given the model M can be factorized in two ways, either as $P(\mathbf{y} \mid \boldsymbol{\sigma}, M)P(\boldsymbol{\sigma} \mid M)$ or as $P(\boldsymbol{\sigma} \mid \mathbf{y}, M)P(\mathbf{y} \mid M)$. The first of these represents the forward process by which one could imagine generating the data, taking a *priori* probability density for $\boldsymbol{\sigma}$ given only the model M, and then the *likelihood* that such a $\boldsymbol{\sigma}$ would generate a particular data set $\mathbf{y}$. The second represents the inverse process, where one considers first the possible data sets $\mathbf{y}$ given M, and then the *posteriori* probability density of $\boldsymbol{\sigma}$ given M and such a $\mathbf{y}$. The equivalence of the two implies

$$P(\boldsymbol{\sigma} \mid \mathbf{y}, M) = \frac{P(\mathbf{y} \mid \boldsymbol{\sigma}, M)P(\boldsymbol{\sigma} \mid M)}{P(\mathbf{y} \mid M)}. \tag{2}$$

This is the famous theorem of Bayes (1764). To be consistent, any assignment of probabilities must (at least implicitly) reflect its structure: probabilities that disagree with it cannot be coherent. A thorough discussion of this approach to uncertainty in physics can be found in the classic book of Jeffreys [1].

The denominator in (2) is effectively just a normalizing constant $P(\mathbf{y} \mid M) = \int P(\mathbf{y} \mid \boldsymbol{\sigma}, M)P(\boldsymbol{\sigma} \mid M)d\boldsymbol{\sigma}$. We shall consider only relative probabilities of the different noise levels in this Letter, so $P(\mathbf{y} \mid M)$ will be left unevaluated. In the numerator, $P(\boldsymbol{\sigma} \mid M)$ represents prior beliefs about possible values of $\boldsymbol{\sigma}$. Here, we shall assume no previous knowledge about the scale of σ_M or σ_D, so that all values of $\ln\sigma_M$ and $\ln\sigma_D$ are equally possible. This

 0031-9007/00/84(11)/2366(4)$15.00

leaves the likelihood $P(\mathbf{y}\,|\,\boldsymbol{\sigma}, M)$ to evaluate. The dynamical noise does not directly affect the likely observations $\mathbf{y}$: it acts on them only indirectly through the likely dynamical series $\mathbf{x}$. To obtain $P(\mathbf{y}\,|\,\boldsymbol{\sigma}, M)$ we must therefore consider the joint probability of $\mathbf{x}$ and $\mathbf{y}$, and then integrate out (marginalize) the possible realizations of $\mathbf{x}$:

$$P(\mathbf{y}\,|\,\boldsymbol{\sigma}, M) = \int P(\mathbf{y}\,|\,\mathbf{x}, \sigma_M, M) P(\mathbf{x}\,|\,\sigma_D, M)\, d^N\mathbf{x}\,. \tag{3}$$

Provided that there is a consistent splitting between the finite-time expanding and contracting directions, it is known [2] that at small noise levels σ_D the inferred trajectory $\mathbf{x}$ consistent with earlier and later measurements becomes sharply localized. The integral can therefore be estimated by a Laplace approximation, i.e., by replacing the integrand, which is proportional to $P(\mathbf{x}\,|\,\mathbf{y}, \boldsymbol{\sigma}, M)$, with an appropriate Gaussian distribution centered on the most probable series $\hat{\mathbf{x}}$.

The removal of the effects of measurement noise in $\mathbf{y}$ to estimate $\hat{\mathbf{x}}$ is known as noise reduction. A number of approaches have been suggested [2–5], but these usually seek an orbit which minimizes all deviations from determinism. This will typically *not* be the most probable orbit if $\sigma_D \neq 0$. We therefore first reconsider the noise reduction problem from a Bayesian perspective. The assumed Gaussian dynamical noise gives $\mathbf{x}$ the prior distribution

$$P(\mathbf{x}\,|\,\sigma_D, M) \propto P(x_{1,\ldots,d}) \exp\left[-\sum_{d+1}^{N} \left(\frac{(f_i - x_i)^2}{2\sigma_D^2} \right) \right], \tag{4}$$

where $P(x_{1,\ldots,d})$ is the probability distribution for the initial state-space point and $f_i = f(x_{i-1}, \ldots, x_{i-d})$. The Gaussian measurements give the likelihood

$$P(\mathbf{y}\,|\,\mathbf{x}, \sigma_M, M) \propto \exp\left[-\sum_{1}^{N} \left(\frac{(y_i - x_i)^2}{2\sigma_M^2} \right) \right], \tag{5}$$

and so, combining (4) and (5),

$$P(\mathbf{x}\,|\,\mathbf{y}, \boldsymbol{\sigma}, M) \propto P(x_{1,\ldots,d}) \exp\left[-\sum \left(\frac{(f_i - x_i)^2}{2\sigma_D^2} + \frac{(y_i - x_i)^2}{2\sigma_M^2} \right) \right]. \tag{6}$$

There are situations (e.g., the large-scale meteorological models in [6]) where a careful consideration of $P(x_{1,\ldots,d})$ might add useful information to the inferred distribution of possible histories $\mathbf{x}$ by ensuring that it reflects only physically plausible initial states. However, for simplicity we will take $P(x_{1,\ldots,d})$ to be a flat distribution over the state space. This reduces (6) to a nonlinear least squares problem, the minimization of

$$Q = \sum [(f_i - x_i)^2 + (\sigma_D^2/\sigma_M^2)(y_i - x_i)^2]. \tag{7}$$

As pointed out by Davies [5], the matrix $L = \partial f_i/\partial x_j - I$ is banded and triangular, so Newton-like methods can be used efficiently. The most probable $\mathbf{x}$ can be found rapidly by repeatedly solving (e.g., by QL decomposition [7]) the overdetermined linear system

$$\begin{pmatrix} -L(\mathbf{x}) \\ \sigma_D/\sigma_M \\ \sqrt{\delta_{\mathrm{LM}}} \end{pmatrix} \delta\mathbf{x} \approx \begin{pmatrix} \mathbf{f}(\mathbf{x}) - \mathbf{x} \\ (\sigma_D/\sigma_M)(\mathbf{y} - \mathbf{x}) \\ \mathbf{0} \end{pmatrix}. \tag{8}$$

Here $\delta_{\mathrm{LM}}(\delta\mathbf{x})^2$ is an additional Levenberg-Marquardt term, which limits the step-size $\delta\mathbf{x}$ and stabilizes the iteration if necessary. For best results δ_{LM} should not be set constant (as in [5]), but should be chosen adaptively, using either the simple approach of [7] or the algorithm of Moré which has become the standard [8]. This provides robustness in the initial stages of the optimization, without impeding final convergence to the optimum.

This procedure represents a balance between closeness to determinism and closeness to the observed data which echo the original noise reduction papers of Kostelich and Yorke [3]: their tradeoff parameter w can be identified from (7) as an *ad hoc* value set for σ_D^2/σ_M^2. We shall now use the resulting $\hat{\mathbf{x}}(\sigma_M, \sigma_D)$ to optimize the σ_M and σ_D. This closely parallels work by Gull and by MacKay [9,10], balancing analogously competing Gaussian influences in the contexts of image reconstruction and function fitting. We proceed similarly, combining the uniform priors on $\ln\sigma_M$ and $\ln\sigma_M$ with the Laplace approximation to (3) from (6) at $\hat{\mathbf{x}}$ to obtain the log posterior distribution,

$$\ln P(\ln\sigma_D, \ln\sigma_M | \mathbf{y}, M) = \text{const} - (N - d)\ln\sigma_D - N\ln\sigma_M - \frac{\sum_1^N (y_i - \hat{x}_i)^2}{2\sigma_M^2} - \frac{\sum_{d+1}^N (f_i - \hat{x}_i)^2}{2\sigma_D^2} - \frac{\ln\det A}{2}, \tag{9}$$

where A is the Hessian matrix of $\ln P(\mathbf{x}|\mathbf{y}, \boldsymbol{\sigma}, M)$ at $\hat{\mathbf{x}}$,

$$A = \nabla\nabla(Q/2\sigma_D^2) = (1/\sigma_D^2)\{L^T L + (\sigma_D^2/\sigma_M^2) I\}. \tag{10}$$

Differentiating (9) with respect to $\ln\sigma_M$ and $\ln\sigma_D$ (neglecting any changes in $L^T L$ caused by induced changes in $\hat{\mathbf{x}}$) now gives a pair of consistency relations for the most probable values of the log noise levels,

$$\sigma_M^2 = \frac{\sum_1^N (y_i - \hat{x}_i)^2}{N - \gamma_M}, \qquad \sigma_D^2 = \frac{\sum_{d+1}^N (f_i - \hat{x}_i)^2}{N - d - \gamma_D}. \tag{11}$$

These formulas and their application constitute the main result of this Letter. The quantities

$$\gamma_M = (1/\sigma_M^2)\mathrm{Tr}(A^{-1}), \qquad \gamma_D = N - \gamma_M \tag{12}$$

have an attractive and intuitive interpretation [9,10]: if Eq. (7) is decomposed into eigendirections of $L^T L$ corresponding to eigenvalues $\lambda_{(j)}^2$, the solution in the jth direction is given by $\hat{x}_{(j)} = \gamma_{M(j)} y_{(j)} + \gamma_{D(j)} x_{D(j)}$ where

$$\gamma_{M(j)} = \frac{(\sigma_D^2/\sigma_M^2)}{\lambda_{(j)}^2 + (\sigma_D^2/\sigma_M^2)}, \qquad \gamma_{D(j)} = 1 - \gamma_{M(j)}, \tag{13}$$

and $x_{D(j)}$ is the Newton estimate for a perfectly deterministic orbit calculated from $\hat{\mathbf{x}}$. Each $\gamma_{D(j)}$ therefore identifies the extent to which that eigendirection has been set by the dynamics, and the sum $\gamma_D = \sum_j \gamma_{D(j)}$ can be interpreted as the effective total number of degrees of freedom set by the dynamics rather than by the measurements. A similar effective number of degrees of freedom for linear scatterplot smoothers is discussed by Hastie and Tibshirani in [11].

Of course, $\hat{\mathbf{x}}$, γ_M, and γ_D depend strongly on the current values of σ_M and σ_D. The estimation was therefore practically carried out by first setting $\sigma_D = 0$ to find the most deterministic time series compatible with the data. Then, having initialized both γ_M and γ_D to zero, a consistent set of estimates for $\hat{\mathbf{x}}$, γ_M, γ_D, σ_M, and σ_D were found iteratively, by alternately using (11) to estimate σ_M and σ_D, then further minimizing (8) and applying (12) to reestimate $\hat{\mathbf{x}}$, γ_M, and γ_D. This was found to converge very rapidly, requiring only a handful of further least-squares steps after the initial minimization to pull $\hat{\mathbf{x}}$ back appropriately close to the observations.

As a detail, it should be noted that the matrix A^{-1} (which is full rank) need never be explicitly formed [12]. Instead (12) can be estimated using $\mathrm{Tr}(A^{-1}) = \sigma_D^2 \langle (G^{-1}\mathbf{r})^2 \rangle$ where $\mathbf{r}$ is a vector of unit variance Gaussian random numbers, and G is the L factor found in the QL decomposition used to solve (8) (so that $A = \sigma_D^{-2} G^T G$). Typically a good value for $\mathrm{Tr}(A^{-1})$ can be obtained in this way using only a few dozen random vectors $\mathbf{r}$. The exception is when (σ_D/σ_M) is very small, so that all but a few of the eigenvalues of A^{-1} are small, and the samples $G^{-1}\mathbf{r}$ acquire high kurtosis. In this case, the few contributing eigenvalues can be estimated directly using Lanczos methods.

To test the $\boldsymbol{\sigma}$ estimation algorithm, 400 different time series of 1000 points were generated from the Hénon system

$$x_{n+1} = 1 - 1.4x_n^2 + 0.3x_{n-1} + \epsilon_{n+1}^{(\mathrm{D})}. \qquad (14)$$

In each case the measurement noise level σ_M was fixed at 2.5×10^{-2}, and a different value between 10^{-7} and 10^{-2} was chosen for the dynamical noise level σ_D. This gave the results shown in Fig. 1. The simple Gaussian estimates for the uncertainties in $\ln\sigma_D$ and $\ln\sigma_M$ were obtained by differentiating (9) again, including induced changes in $\hat{\mathbf{x}}(\sigma_D, \sigma_M)$, to give

$$\begin{aligned} \Delta \ln\sigma_D &= \left(4\gamma_M - 2d - \frac{2}{\sigma_M^4}\,\mathrm{Tr}A^{-2} \right. \\ &\qquad \left. - 4\,\frac{\sigma_D^2}{\sigma_M^2} \left| G^{-1}\,\frac{(\mathbf{y}-\hat{\mathbf{x}})}{\sigma_M} \right|^2 \right)^{-1/2}, \\ \Delta \ln\sigma_M &= \left(2N - \frac{2}{\sigma_M^4}\,\mathrm{Tr}A^{-2} \right. \\ &\qquad \left. - 4\,\frac{\sigma_D^2}{\sigma_M^2} \left| G^{-1}\,\frac{(\mathbf{y}-\hat{\mathbf{x}})}{\sigma_M} \right|^2 \right)^{-1/2}. \end{aligned} \qquad (15)$$

The method has accurately revealed dynamical noise levels 1000 times smaller than the measurement noise, and given reasonable error estimates. Furthermore the estimate for σ_D is significantly different to the *ad hoc* value given by finding the RMS deviation from determinism after noise reduction:

$$\mathrm{RMS}(\boldsymbol{\epsilon}^{(\mathrm{D})}) = \left(\frac{\sum_{d+1}^{N} (f_i - \hat{x}_i)^2}{N-d} \right)^{1/2}. \qquad (16)$$

Compared to (11) this differs by an amount γ_D in the denominator. The significance of this term is shown in Fig. 2. While $N - d$ remains a constant 998, the number $N - d - \gamma_D$ of effective measurements of the dynamical noise level falls steadily, from about 150 at $\sigma_D = 10^{-2}$ to about 0.4 at $\sigma_D = 10^{-4}$ and 0.1 at $\sigma_D = 10^{-5}$. The estimate of σ_D from (11) in the latter case will thus be $\approx$100 times larger than from (16).

An extreme case occurs when the algorithm can find a perfectly deterministic or "shadowing" orbit, as it could for the 167 Hénon time series considered which appear as

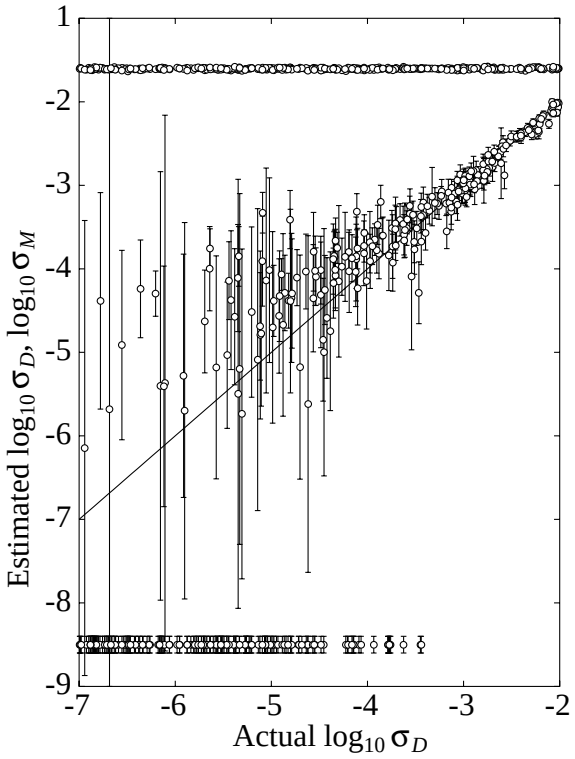

FIG. 1. Estimated $\log_{10}\sigma_M$ and σ_D against actual $\log_{10}\sigma_D$.

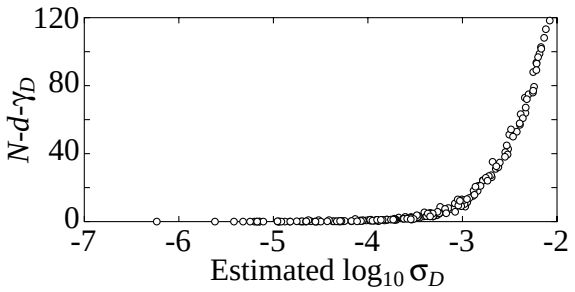

FIG. 2. $N - d - \gamma_D$ against estimated $\log_{10}\sigma_D$.

a band across the bottom of Fig. 1. The RMS dynamical error is then precisely zero, but, as is well known from the shadowing theorems of Anosov and Bowen [13], just because the model *can* deterministically shadow the observations to within $O(\sigma_M)$, this does *not* mean that σ_D is necessarily (or even probably) zero. Equation (11) deals quite simply with lengths of the orbit for which shadowing is indicated: it ignores them. Although the numerator $\sum_{d+1}^{N}(f_i - \hat{x}_i)^2$ becomes zero, so too does the denominator $N - d - \gamma_D$. All $N - d$ components of the time series have been set by the dynamics, so there are effectively *no* good measurements of the dynamical noise level and σ_D is essentially indeterminate.

A deeper understanding of this situation can be obtained from the full log probability distribution (9). When shadowing is possible and the true σ_D is very small, $\ln P(\log\sigma_D)$ is dominated by the terms $-(N - d)\ln\sigma_D - (1/2)\ln\det A$, giving the form shown in Fig. 3. This reflects a contribution of $-(1/2)\ln(\lambda_{(j)}^2 + \sigma_D^2/\sigma_M^2)$ from each nonzero eigendirection, which is insensitive to σ_D as long as $\lambda_{(j)}^{-1}\sigma_D \ll \sigma_M$. Thus the distance that the nonzero σ_D allows $\hat{x}_{(j)}$ to move from $x_{D(j)}$ is unobservably small in the measurement noise. Although there is no dynamical error, all that can be inferred when σ_D is *very* small is that $\sigma_D \lesssim \lambda_{(d+1)}\sigma_M$, where $\lambda_{(d+1)}$ is the smallest nonzero singular value of L.

However, in both hyperbolic and nonhyperbolic systems, shadowing may be possible at larger values of σ_D than this. For example, Hammel *et al.* [14] have conjectured that shadowing within a distance $\sigma_M < \sqrt{\sigma_D}$ is typically possible for up to $O(1/\sqrt{\sigma_D})$ iterates for 2D maps with a rectangular noise distribution of width σ_D. In such cases a shadowing orbit may still exist when the true $\sigma_D \gg \lambda_{(d+1)}\sigma_M$. The highly correlated effects of dynamical noise will then start to become statistically apparent above the isotropic measurement noise in the observations. This will be most noticeable at points in the orbit where stable and unstable manifolds are most nearly tangent (cf. [15]), corresponding to the smallest singular values $\lambda_{(j)}$ of L. As the deviations from isotropy become stronger, they cause the remaining terms in (9) to become significant, effectively adding a broad bump to the middle of the curve in Fig. 3. This is a second, more probable solution (σ_D, σ_M) to the consistency equations (11), with an orbit $\hat{\mathbf{x}}$ which is more probable than the shadowing orbit. A quick upward sweep through σ_D could be added to check for this possibility whenever the noise reduction algorithm finds a shadowing orbit.

In conclusion, therefore, even a perfectly deterministic orbit $\mathbf{x}$ may not be an optimal estimate for an underlying dynamical time series. Instead, we have shown that if a good model of the form (1) is available, accurate estimates of both the dynamical noise level σ_D and measurement noise level σ_M can be derived, with corresponding error bars and appropriate time-series estimates, even when σ_D is orders of magnitude smaller than σ_M. Our approach can also be applied to more general models, with appropriate modifications to the noise reduction step. In some physical situations, however, no *a priori* model will be available. We are currently extending our approach to cover this case by combining the model estimation with the noise reduction as in [16] within the overall probabilistic framework described here.

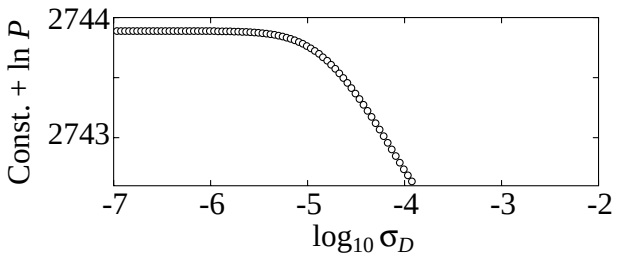

FIG. 3. Log probabilities for an orbit created with $\sigma_D = 0$.

[1] H. Jeffreys, *Theory of Probability* (Clarendon Press, Oxford, 1961), 3rd ed.
[2] J. D. Farmer and J. J. Sidorowich, Physica (Amsterdam) **47D**, 373 (1991); S. M. Hammel, Phys. Lett. A **148**, 421 (1990).
[3] E. J. Kostelich and J. A. Yorke, Physica (Amsterdam) **41D**, 183 (1990).
[4] P. Grassberger *et al.*, Chaos **3**, 127 (1993).
[5] M. E. Davies, Physica (Amsterdam) **79D**, 174 (1994).
[6] L. A. Smith, in *Past and Present Variability of the Solar-Terrestrial System: Measurement, Data Analysis and Theoretical Models*, edited by G. C. Castagnoli and A. Provenzale (IOS, Amsterdam, 1997).
[7] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C* (Cambridge University Press, Cambridge, 1992), 2nd ed.
[8] J. J. Moré, in *Numerical Analysis—Dundee 1977*, edited by G. A. Watson (Springer, Berlin, 1978), pp. 105–116.
[9] D. J. C. MacKay, Neural Comput. **4**, 415 (1992).
[10] S. F. Gull, in *Maximum Entropy and Bayesian Methods, Cambridge 1988*, edited by J. Skilling (Kluwer, Dordrecht, 1989), pp. 53–71.
[11] T. J. Hastie and R. J. Tibshirani, *Generalised Additive Models* (Chapman and Hall, London, 1990).
[12] J. Skilling, in *Maximum Entropy and Bayesian Methods, Cambridge 1988* (Ref. [10]), pp. 455–466.
[13] D. V. Anosov, *Geodesic Flows on Riemann Manifolds with Negative Curvature*, Proceedings of the Steklov Institute of Mathematics Vol. 90 (American Mathematical Society, Providence, RI, 1967); R. Bowen, Am. J. Math. **92**, 725 (1970).
[14] S. M. Hammel, J. A. Yorke, and C. Grebogi, Bull. Am. Math. Soc. **19**, 465 (1988).
[15] L. Jaeger and H. Kantz, Physica (Amsterdam) **105D**, 79 (1997).
[16] M. E. Davies and J. Stark, in *Nonlinearity and Chaos in Engineering Dynamics*, edited by J. M. T. Thompson and S. Bishop (Wiley, London, 1994).