Local optimal prediction: exploiting strangeness and the variation of sensitivity to initial condition†

By Leonard A. Smith

Mathematical Institute, University of Oxford, Oxford OX1 3LB, U.K.

Accurate prediction of a nonlinear system from limited data requires sensitivity to the variation of the system's properties in state space. Two aspects of this variability are examined, throwing new light on the 'limits of predictability' as well as individual predictions. A prediction scheme which embraces the variability both of dynamics and geometry is outlined and illustrated. The paper concludes with a discussion of residual predictability, proposing a simple test to detect systematic prediction error, which indicates that further improvement in prediction accuracy is possible.

1. Structure in the sensitivity to initial condition

Variability and unpredictability are two of the hallmarks of deterministic chaos; it is our aim here to show how the first can be exploited to reduce the second. Figure 1*a* shows the error-doubling time for a variety of initial conditions on the Lorenz attractor (Ziehmann-Schlumbohm 1994; Smith *et al.* 1994). The colours reflect the minimum time required for an infinitesimal uncertainty to double. The red points double within one Lorenz second, the orange within two, the yellow three and so on. This illustrates the main points of this paper: that the predictability of the flow is both variable and highly organized. By exploiting this variability, we can significantly improve our predictions. Moreover, arguments based on uniform error growth are misleading. The argument linking the largest Lyapunov exponent to the 'prediction horizon' is a good example; Lyapunov exponents need not reflect practical limits of prediction.

To evaluate a set of predictors, we must choose a criteria for comparison. For nonlinear systems, the results will be much more sensitive to the particular statistic chosen than an intuition based on independent, identically distributed (IID) gaussian residuals would suggest. Prediction errors from chaotic systems are neither gaussian nor independent; they are correlated both in time and state space, and are usually chaotic themselves. We also note that short term error growth will depend on the nature of the initial uncertainty; the structure observed when this uncertainty is oriented by the largest (global) Lyapunov exponent (as in figure 1*a*) will differ from cases where it is determined either by the locally fastest grow-

† This paper was produced from the author's disk by using the TEX typesetting system.

Phil. Trans. R. Soc. Lond. A (1994) **348**, 371–381
Printed in Great Britain

ing direction, by a random displacement, the most likely observed displacement, or from pre-processing of the data.

2. Nonlinear prediction as interpolation in state space

The recent success of prediction methods for chaotic dynamical systems (Abarbanel *et al.* 1993; Eubank & Farmer 1990; Tong 1990) is due, in large part, to the successful translation of an extrapolation problem to an interpolation problem (Eckmann & Ruelle 1985). Ideally, these methods consider a point in the state space of a deterministic physical system. If the equations of motion are known, the future of an initial condition may be determined by integration. Alternatively, if sufficient data are available, then we can interpolate the future trajectory given only the observations. Yet the functions involved in chaotic dynamical systems are, at their simplest, nonlinear, and the data are typically distributed on a strange attractor. We are faced with a high dimensional (> 2) interpolation of a complicated function sampled on an inhomogeneous distribution. With noise.

The two basic approaches to this problem consider either global interpolation functions for the entire state space, or restrict attention to local regions. The global approach requires a complicated interpolation scheme (e.g. radial basis functions or neural nets). Here we will consider the 'local' approach (Farmer & Sidorowich 1987; Sugihara & May 1990), which allows much simpler interpolation schemes. We can test for nonlinearity (Casdagli 1992; Casdagli *et al.* 1992) by evaluating a series of local predictors based on the k nearest neighbours and determining the value, k_c, at which the observed prediction error is smallest. For linear stochastic systems, k_c should correspond to the largest k, while for noise free deterministic systems, k_c should be of the order of the dimension of the system, given enough data. For deterministic systems with noise and nonlinear stochastic systems, a minimum at 'moderate' values of k is expected.

As a concrete example, consider the chaotic, two-dimensional Ikeda map,

$$\boldsymbol{D}_{\text{Ikeda}}(x, y) = (1.0 + \mu[x\cos(t) - y\sin(t)], \mu[x\sin(t) + y\cos(t)]), \tag{2.1}$$

where $t = 0.4 - 6.0/(x^2 + y^2 + 1)$ and $\mu = 0.90$. We will make one step ahead predictions of the value of x, given 1024 observations and base points (x, y) with gaussian noise ($\sigma = 0.125$) added to the observations. The solid line in figure 2 shows the out-of-sample, normalized average absolute error, $E(k)$ indicating $k_c \approx 32$. Yet if we consider only the five predictors with $k = 8$, 16, 32, 64 and 128, then we find the $k = 32$ predictor is the most accurate just 27% of the time (the distribution being 19, 23, 27, 23 and 8%, respectively). This illustrates a shortcoming of selecting a predictor with this approach: a global choice of k fails to account for the variation in the length-scales of either the dynamics or the data distribution. These two effects are shown schematically in figure 3. Figure 3*a* shows a one-dimensional local linear approximation to a polynomial curve. The circle shows the radius at which the expected value of the noise is equal to the error introduced by the linear approximation of the true curve. The optimal local radius within which data should be used, r_{opt}, depends upon the statistics of the noise, the data density and the local curvature, and this is our point: that even with uniformly distributed data, r_{opt} will change with the local curvature of the function estimated. Figure 3*b* illustrates additional complications due to the non-uniform distribution of data, these effects change not only with location,

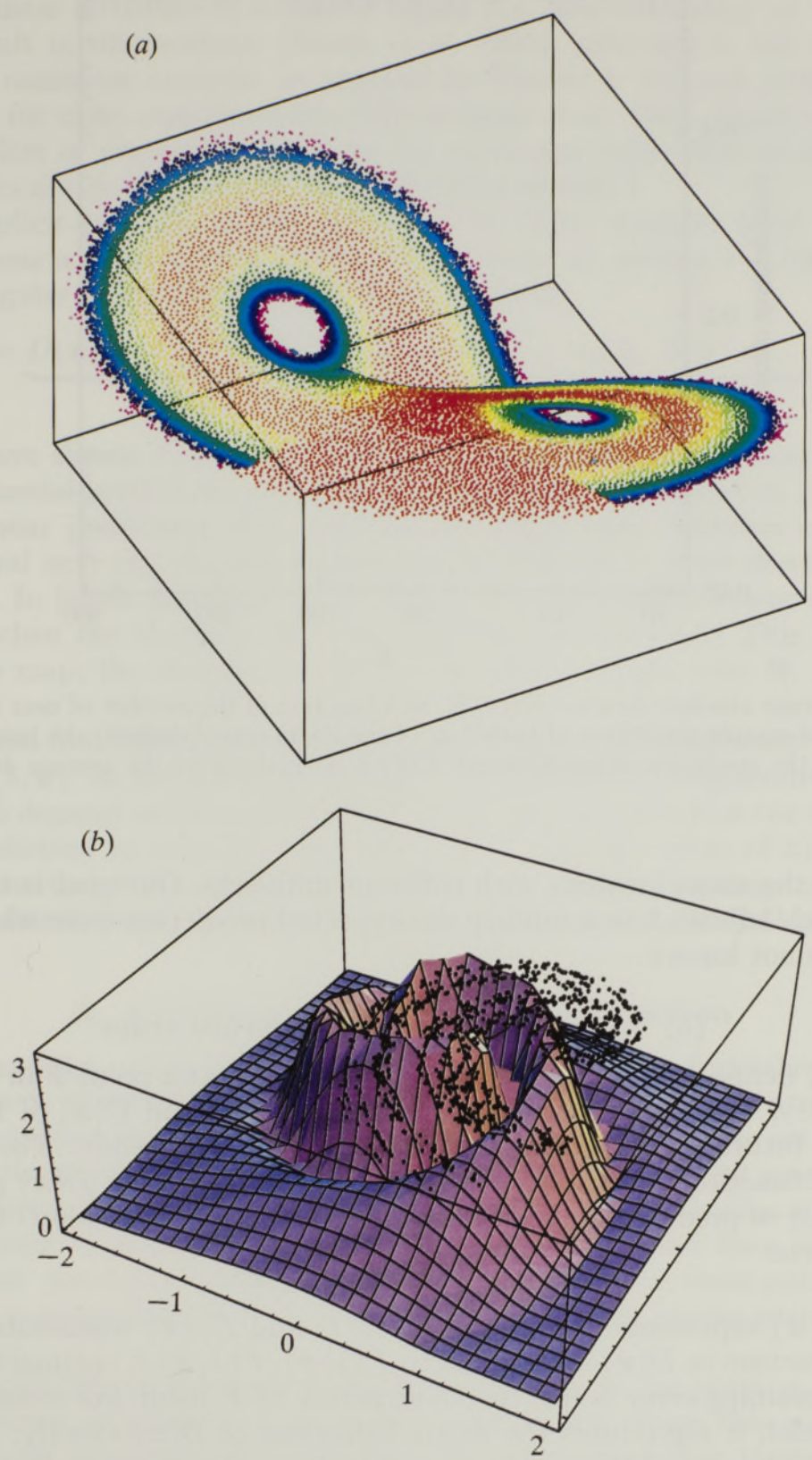

Figure 1. (*a*) Error doubling time on the Lorenz attractor. The colour code reflects the number of Lorenz seconds before which doubling occurs (< 1) red, (2) orange, (3) yellow, (4) light green, (5) dark green, (6) blue, (7) purple, (> 8) lavender. The number of rapidly doubling points is suppressed for clarity. (*b*) Variation in the absolute value of the quadratic term of the x component of the Ikeda map, averaged over angle. The dots show the location of the attractor.

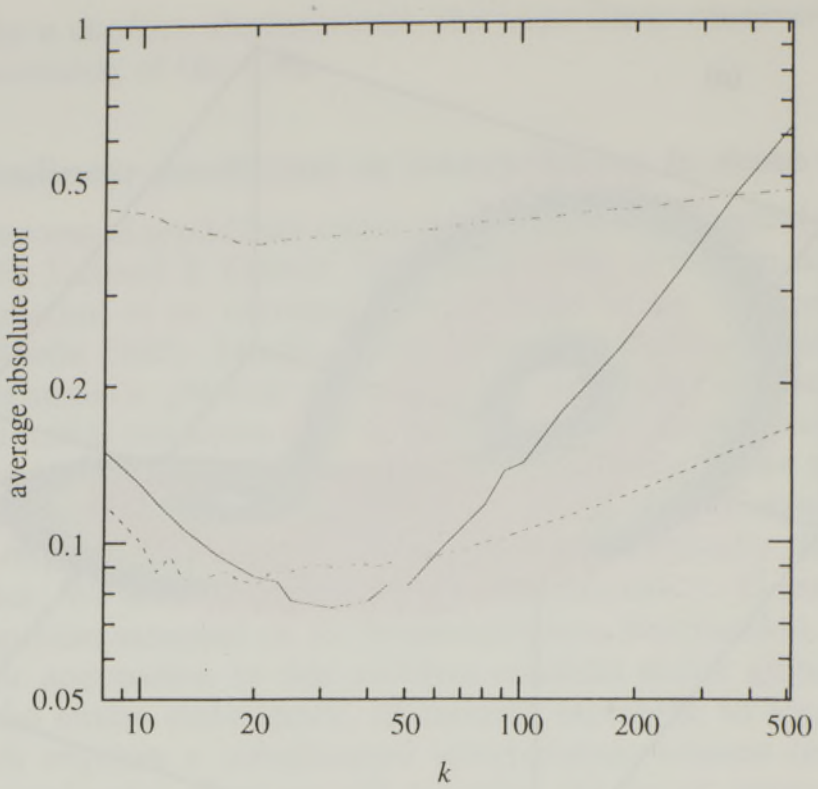

Figure 2. Average absolute forecast error, E, as a function of the number of near neighbours, k, used for out-of-sample prediction of (solid) x_i of the Ikeda map, (dashed) the laser system, and (dot–dashed) the stochastic sunspot model. $E(k)$ is normalized by the average deviation. Note the log scale.

but also at the same location with different data-sets. Our goal is to develop a scheme which adjusts k to minimize the expected prediction error when the local curvature is not known.

(*a*) *Imperfect predictions: colourfast chaos*

Let $\boldsymbol{D}(\boldsymbol{x})$ define a deterministic dynamical system at a point $\boldsymbol{x}$ in state space. For simplicity, we consider a scalar measurement function $D(\boldsymbol{x})$ of $\boldsymbol{D}(\boldsymbol{x})$ representing the future property of $\boldsymbol{D}(\boldsymbol{x})$ which we wish to predict. The problem is then one of function approximation; we wish to approximate $D(\boldsymbol{x})$ given a particular family of predictors, F, with parameters λ. We may divide D heuristically into two parts:

$$D(\boldsymbol{x}) = F(\lambda, \boldsymbol{x}) + E_F(\boldsymbol{x}), \tag{2.2}$$

where $F(\lambda, \boldsymbol{x})$ represents an optimal fit to D and $E_F(\boldsymbol{x})$ represents the deterministic structure in $D(\boldsymbol{x})$ orthogonal to $F(\lambda, \boldsymbol{x})$. $F(\lambda, \boldsymbol{x})$ is optimal in the sense that the remaining error is due to the structure of F itself. For example, if F is a linear model, it reproduces the linear behaviour of $D(\boldsymbol{x})$ exactly; in this case $E_F(\boldsymbol{x})$ would consist of the quadratic and higher order terms in $D(\boldsymbol{x})$.

In general, we expect $E_F(\boldsymbol{x})$ to be a good measurement function: when Taken's theorem (Takens 1981; Sauer *et al.* 1991) applies to a chaotic data stream, it will also apply to the time series of residuals from our prediction scheme. This implies that the residuals will be chaotic, albeit with more complicated macroscopic structure and a smaller signal to noise ratio, and indicates the impossibility of truly bleaching chaotic data (Theiler & Eubank 1994); chaos is colourfast in the sense that the residuals of non-perfect predictors will, in general, be chaotic. Removing

the global linear structure of a chaotic signal (i.e. 'pre-whitening' or 'bleaching') will not result in IID residuals (Brock *et al.* 1991), although it will obscure the results of a nonlinear analysis, as stressed by Theiler & Eubank (1993). Similar results hold for more complex filters (Broomhead *et al.* 1992; Sauer *et al.* 1991), where the filter or predictor influences the particular value that we observe, but the dynamics are determined by the underlying system.†

As an explicit example of the separation in (2.2), consider local polynomial prediction near a point $\boldsymbol{x}_0$. We expand $D(\boldsymbol{x})$ about $\boldsymbol{x}_0$, setting $r = ||\boldsymbol{x} - \boldsymbol{x}_0||$ and denoting angular orientation by the vector $\boldsymbol{\Theta}$ yields

$$D(\boldsymbol{x}) = \underbrace{D(\boldsymbol{x}_0, 0) + a_1(\boldsymbol{x}_0, \boldsymbol{\Theta})r}_{F(\lambda, \boldsymbol{x})} + \underbrace{a_2(\boldsymbol{x}_0, \boldsymbol{\Theta})r^2 + a_3(\boldsymbol{x}_0, \boldsymbol{\Theta})r^3 + \dots}_{E_F(\boldsymbol{x})}, \tag{2.3}$$

where we have shown $F(\lambda, \boldsymbol{x})$ and $E_F(\boldsymbol{x})$ for local linear prediction. For higher order polynomial predictors, additional terms would be shifted to $F(\lambda, \boldsymbol{x})$. For other nonlinear predictors (e.g. a particular radial basis function scheme or a specific neural net) this decomposition may be difficult to write down, but holds in principle. In linear predictors, r_{opt} will tend to be smaller where $a_2(\boldsymbol{x}_0, \boldsymbol{\Theta})$ is large even when the learning data are uniformly distributed.‡ This is observed in the Ikeda map; the absolute value of $a_2(\boldsymbol{x}, \boldsymbol{\Theta})$, averaged over $\boldsymbol{\Theta}$, is shown in figure 1*b*.

In the noise-free case, k_c is the minimum number of neighbours required to solve for $F(\lambda, \boldsymbol{x})$. In the presence of noise, the number of neighbours (and thus the r_{opt}) will depend on the nature of the noise level as well. It is the aim of local optimal prediction to vary the parameters of F (e.g. the value of k) to balance the local structure of E_F against that of the noise, while adapting to the details of the data distribution, in cases where the analytic structure of D is unknown.

3. A method of local optimal prediction

We now consider an algorithm to determine the optimal k from the data alone. The data-set is divided into a learning set from which the predictors are constructed, and a test set for out-of-sample evaluation. The learning set is analysed to estimate the global parameters k_c and $k_{\max}$ ($\gg k_c$), the largest neighbourhood to be considered. To predict a point $\boldsymbol{x}_0$ in the test set, the basic idea is to consider a small number of points in the learning set close to $\boldsymbol{x}$ and then employ local 'drop-one-out' predictors with various k to predict each of these points in turn. The local k is determined from the prediction error of the known points from the learning set.

More specifically, to determine the best value of k at the point $\boldsymbol{x}_0$, determine the $k_{\max}$ nearest neighbours of $\boldsymbol{x}_0$ in the learning set. From this subset, select the N_{drop} (≈ 8) points nearest to $\boldsymbol{x}_0$ with the requirement that these test points are well separated in time.¶ In figure 3*b*, these points lie within the smallest circle. For several values of k (corresponding to the larger circles in figure 3*b*),

† In contrast to the series obtained by repeatedly iterating a non-perfect predictor.

‡ In regions of very low data density, reverting to the 'zero-order' method of simply taking the image of the nearest neighbour can significantly improve the prediction error.

¶ This requirement is crucial to avoid picking consecutive points from the same segment of the trajectory which leads to highly correlated (and misleading) estimated errors.

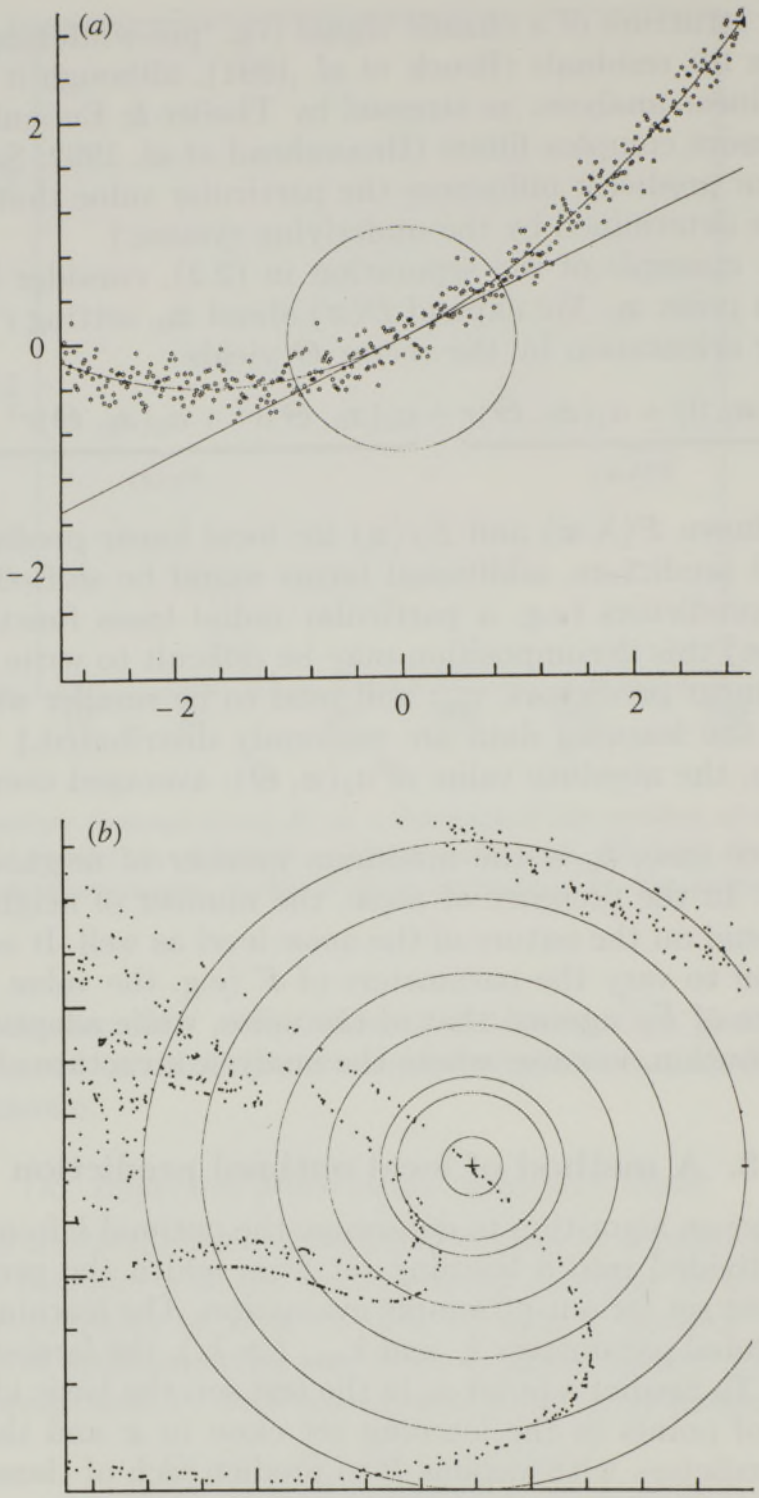

Figure 3. (*a*) Schematic diagram showing the need to balance the noise level against the local curvature in a linear fit to a polynomial function. The circle shows the radius at which the deviation due to the linear approximation equals the expected value of the noise. (*b*) Circles containing 9, 17, 33, 65, 129 and 257 points near the point can be predicted (+). The data are not evenly distributed within the circles, tending to be almost linear for $k = 8$ and remaining completely skewed to the one side of the prediction point until $k = 256$.

construct N_{drop} distinct predictors by dropping out, in turn, each of the N_{drop} test points and predicting the point omitted. Finally, combine these results for each k to obtain an estimated prediction error at $\boldsymbol{x}_0$ for a k neighbour predictor, and determine the optimal radius at $\boldsymbol{x}_0$.

The choice of optimal radius is non-trivial, and detailed results will be presented

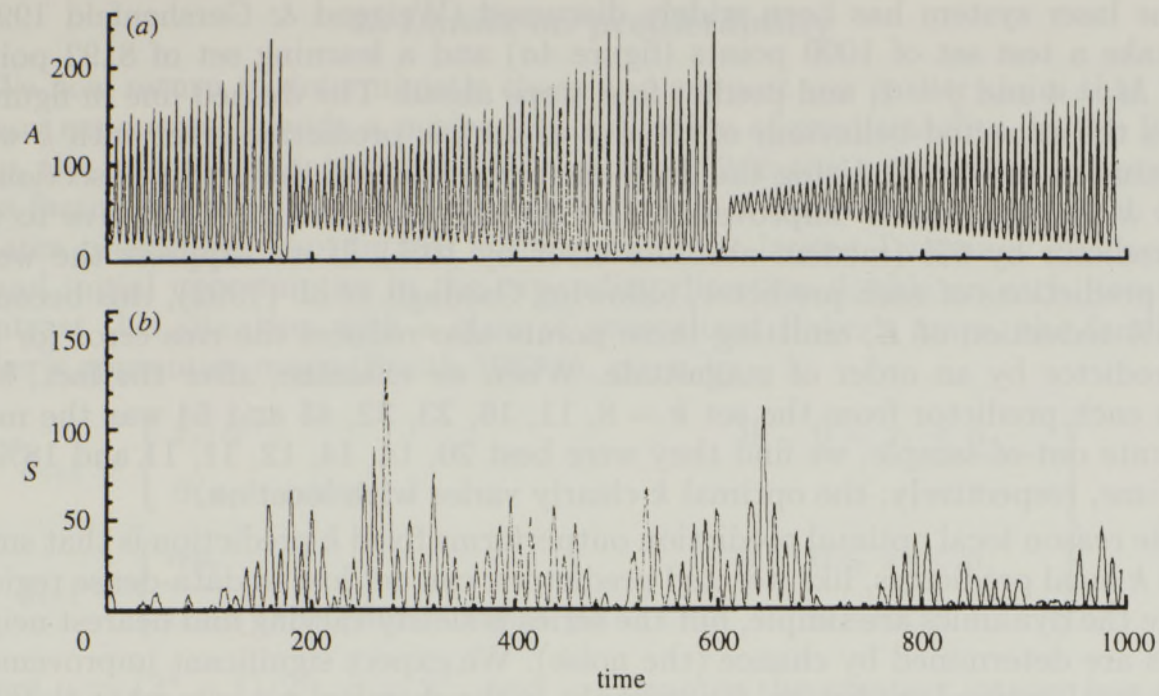

Figure 4. Time series data from (*a*) laser experiment and (*b*) the stochastic sunspot model.

elsewhere (Smith & Drysdale 1994). Picking the k with the minimum estimated error frequently improves the predictor a large fraction of the time, but *increases* the overall rms error by making a few bad choices. A second approach is to consider the estimated error as a function of k and find the local minimum nearest k_c. Alternatively, we can partition the entire space, and use the partition element to define the initial guess for k (see Smith 1992, 1993*b*). When considering local minima, it is useful to check that there is not a significantly (in terms of the standard error of the estimated prediction error) deeper minima elsewhere. In the results presented below, the second method is used.

We consider data from a deterministic laser system (Weigend & Gershenfeld 1993) and a nonlinear, stochastic model (Barnes *et al.* 1980). To apply these ideas when only a single time series, s_i, is available, we first form a reconstruction space via the method of delays (Sauer *et al.* 1991). This yields the series of M-dimensional vectors, $\boldsymbol{x}_i$

$$\boldsymbol{x}_i = (s_i, s_{i-j}, \ldots, s_{i-j(M-1)}), \tag{3.1}$$

where j is the delay time. Under ideal circumstances, Takens's theorem (Takens 1981; Sauer *et al.* 1991) assures us that many properties from the true state space dynamics are preserved by this reconstruction. In each case, we use previously published reconstruction parameters M and j to make direct, fixed step, local linear predictions, and evaluate the models out-of-sample. For delay reconstructions, this model is

$$F(\lambda, \boldsymbol{x}) = \lambda_0 + \sum_{\ell=0}^{M-1} \lambda_{\ell+1} s_{i-j\ell}, \tag{3.2}$$

where the λ_i are determined by least squares fit to the k nearest neighbours of $\boldsymbol{x}$. The variation of the predictor with $\boldsymbol{x}$ can be made smooth by weighting the contributions of points as a function of their distance from $\boldsymbol{x}$.

The laser system has been widely discussed (Weigend & Gershenfeld 1993). We take a test set of 1000 points (figure 4*a*) and a learning set of 8192 points with $M = 4$ and $j = 1$, and predict four steps ahead. The dashed line in figure 2 shows the expected behaviour of the out-of-sample prediction error with k with a minimum at $k = 20$. Using the drop-one-out scheme of the previous section to allow k to vary with $\boldsymbol{x}$ improves the average absolute error, E, relative to the k_c predictor by 6% (median absolute error by 10%). If we suppress the worst 10% predictions of each predictor, following Casdagli *et al.* (1992), this becomes a 6.5% reduction of E; omitting these points also reduces the rms error for the k_c predictor by an order of magnitude. When we examine, after the fact, how often each predictor from the set $k = 8$, 11, 16, 23, 32, 45 and 64 was the most accurate out-of-sample, we find they were best 20, 14, 14, 12, 11, 11 and 18% of the time, respectively; the optimal k clearly varies with location.

One reason local optimal prediction outperforms fixed k prediction is that small fixed k local predictors, like iterated predictors, can get lost in data-dense regions where the dynamics are simple, but the series is slowly varying and nearest neighbours are determined by chance (the noise). We expect significant improvement using local optimal prediction, for example, in the chemical experiments presented by Professor Olsen.

Successful nonlinear prediction has also been interpreted as evidence of deterministic dynamics. This interpretation is misleading for a class of nonlinear, but fundamentally stochastic systems which do not meet the criteria of Laplacian determinism. We call such systems 'aleatoric' since, while the underlying driving mechanism appears not to be deterministic, the dynamics are governed in large part by deterministic laws, as with a roll of the dice. Consider, for instance, the stochastic Barnes model (Barnes *et al.* 1980) for annual mean sunspot numbers, Y (figure 4*b*). Based on an ARMA(2,2) model with nonlinear modifications to ensure that Y remains positive and tends to increase more rapidly than it decreases, the model is

$$Z_n = \phi_1 Z_{n-1} + \phi_2 Z_{n-2} + a_n - \theta_1 a_{n-1} - \theta_2 a_{n-2}, \tag{3.3}$$

$$Y_n = Z_n^2 + \alpha(Z_n^2 - Z_{n-1}^2)^2, \tag{3.4}$$

where $\phi_1 = 1.90693$, $\phi_2 = -0.98751$, $\theta_1 = 0.78512$, $\theta_2 = -0.40662$, $\alpha = 0.03$ and the a_n are IID gaussian random variables with zero mean and $\sigma = 0.4$.

To avoid any bias from our knowledge of the underlying system, the reconstruction parameters of Casdagli *et al.* (1992) ($M = 3$, $j = 1$) for the observed sunspot series were used to make one year ahead predictions. The dot–dashed curve in figure 2 traces the out-of-sample average absolute error as a function of k showing a minimum at $k = 20$, behaviour similar to that expected from a deterministic system. Further, the data density and sensitivity of the dynamics vary tremendously in different regions of reconstruction space, so local optimal prediction improves prediction and the time series of prediction errors is far from IID. Distinguishing stochastic aleatoric dynamics from deterministic chaotic dynamics in practice may require huge data-sets, either to get good statistics on very near returns in reconstruction space or to quantify the average decay of predictability more directly (Sugihara & May 1990; Casdagli *et al.* 1992).

4. Limits on predictability

We now return to deterministic chaos and support our earlier claim that Lyapunov exponents provide a misleading indication of predictability. In the Baker map, the unit square is stretched by a factor of 2 in the x direction, compressed by a factor of 2 in y, and the resulting rectangle is then cut and stacked to form an area preserving map. In this uniform case, the largest Lyapunov exponent is 1, and initial uncertainties in the expanding direction double on each iteration. Contrast this situation with a class of generalized Baker's maps, the family of Baker's apprentice maps (Smith 1993*a*), given by

$$\left.\begin{aligned} x_{i+1} &= \begin{cases} x_i/\alpha & \text{if} \quad 0 \leqslant x_i < \alpha, \\ \beta(x_i - \alpha) \bmod 1 & \qquad\; \alpha \leqslant x_i < 1, \end{cases} \\ y_{i+1} &= \begin{cases} \alpha y_i & \text{if} \quad 0 \leqslant x_i < \alpha, \\ \alpha + (1/\beta)(\lfloor \beta(x_i - \alpha) \rfloor + y_i) & \qquad\; \alpha \leqslant x_i < 1, \end{cases} \end{aligned}\right\} \tag{4.1}$$

where $\alpha = (2^n - 1)/2^n$, $\beta = 2^{2^n}$ and $\lfloor z \rfloor$ denotes the greatest integer less than or equal to z. In this case a small fraction $(1 - \alpha)$ of the 'dough' is stretched a great deal (2^{2^n}) before being cut and stacked, while the majority of the initial conditions are displaced only slightly $(1/\alpha)$. For each n, equations (4.1) define an area preserving map whose positive Lyapunov exponent, λ_1 $(= 1 - \alpha \log_2 \alpha)$ is greater than 1 bit per iteration. Thus the Lyapunov exponents of each of these maps is *greater* than that of the Baker map, yet for the majority of initial conditions, the apprentice maps are much more predictable. This is reflected in the error doubling time, but not by the Lyapunov exponents. For these maps, points of equal doubling time fall in vertical bands reminiscent of those seen in the Lorenz system in figure 1*a*. Local optimal prediction will adapt to this variability automatically. Nevertheless, characterizing the decay of predictability by the growth of errors will be complicated by the lack of independence between consecutive errors.

5. Residual predictability

We began by examining the structure of error doubling times on a strange attractor and observed organization in this structure. To conclude, we suggest adopting the same approach to look for residual predictability, using a 'colour code' based on the (signed) prediction error rather than the doubling time. Organization in these errors means improved prediction is possible. In higher dimensions, where visualization is more difficult, we test for structure in the residuals by considering nearest neighbour pairs in the reconstruction space, and simply count the number of pairs where both of the associated prediction errors have the same sign. If the residuals are independent and identically distributed (IID), then the expected number of pairs with like signs is easily determined. As an IID sequence should remain IID under any general regrouping (see, for example, Dawid 1984), contrasting the expected and observed number of pairs with like signs provides an immediate, quantitative evaluation of the prediction scheme. There are, of course, many more powerful tests, like the BDS test which considers delay reconstructions of the series of the residuals themselves (Brock *et al.*

1991); the advantages of our test include simplicity and the use of the delay space coordinates of the original data to examine small length scales where traces of predictability are most likely to be found.

The difficulty of predicting the evolution of a nonlinear system varies with the state of the system. Fortunately, this variation is highly organized, and we may exploit it both to improve our predictions and our estimate of the degree to which we believe them. This organized variability also requires careful consideration when interpreting global bounds on predictability. The goal of nonlinear prediction is to remove all recognizable structure from the time series, and it is useful to look for structure in the residuals within the framework of the original reconstruction space of the observations. Ultimately, our goal is to exploit the structure in the dynamics to eliminate structure in the residuals.

It is a pleasure to acknowledge both technical assistance and mathematical discussions with D. Drysdale, K. Fraedrich, M. Muldoon, and C. Ziehmann-Schlumbohm. I am also grateful to S. Ellner, B. LeBaron, B. Seifert, R. Smith and D. Wolpert for suggestions and directions through the statistical literature. Issues related to residual predictability have been enlightened by conversations with D. Broomhead and J. Huke and the observations of R. Jones. This work was supported by a Senior Research Fellowship at Pembroke College.

References

Abarbanel, H. D. I., Brown, R., Sidorowich, J. J. & Tsimring, L. S. 1993 *Rev. mod. Phys.* **65**, 1331–1392.

Barnes, J. A., Sargent, H. H. & Tryon, P. V. 1980 Sunspot cycle simulation using random noise. In *The ancient sun* (ed. R. O. Pepin, J. A. Eddy & R. B. Merrill), pp. 159–163. New York: Pergamon.

Brock, W. A., Hsieh, D. & LeBaron, B. 1991 *Nonlinear dynamics, chaos, and instability: statistical theory and economic evidence.* Cambridge, MA: MIT Press.

Broomhead, D. S., Huke, J. P. & Muldoon, M. R. 1992 *Jl R. statist. Soc.* B **54**, 373–382.

Casdagli, M. 1992 *Jl R. statist. Soc.* B **54**, 303–328.

Casdagli, M. *et al.* 1992 Nonlinear modeling of chaotic time series. In *Applied chaos* (ed. J. H. Kim & J. Stringer), pp. 335–380. New York: John Wiley.

Dawid, A. P. 1984 *Jl R. statist. Soc.* A **147**, 278–292.

Eckmann, J. P. & Ruelle, D. 1985 *Rev. mod. Phys.* **57**, 617–656.

Eubank, S. & Farmer, J. D. 1990 An introduction to chaos and randomness. In *Proc. SFI Summer School* (ed. E. Jen). Addison-Wesley.

Farmer, J. D. & J. Sidorowich, J. 1987 *Phys. Rev. Lett.* **59**, 8.

Sauer, T., Yorke, J. A. & Casdagli, M. 1991 *J. statist. Phys.* **65**, 579–616.

Smith, L. A. 1992 *Physica* D **58**, 50–76.

Smith, L. A. 1993 *a* Do Lyapunov exponents limit predictabilty? Preprint.

Smith, L. A. 1993 *b* Does a meeting in Santa Fe imply chaos? In Weigend & Gershenfeld (1994, pp. 323–344).

Smith, L. A. & Drysdale, D. 1994 Local optimal prediction of low dimensional dynamics. (In preparation.)

Smith, L. A., Ziehmann-Schlumbohm, C. & Fraedrich, K. 1994 Structure within the sensitivity to initial condition. Preprint.

Sugihara, G. & May, R. M. 1990 *Nature, Lond.* **344**, 734–741.

Takens, F. 1981 Detecting strange attractors in fluid turbulence. In *Dynamical systems and turbulence* (ed. D. Rand & L.-S. Young), vol. 898, p. 366. New York: Springer-Verlag.

Theiler, J. & Eubank, S. 1994 *Chaos* **4**, 1–12.

Tong, H. 1990 *Nonlinear time series analysis.* Oxford University Press.

Weigend, A. & Gershenfeld, N. (eds) 1993 *Predicting the future and understanding the past.* New York: Addison-Wesley.

Ziehmann-Schlumbohm, C. 1994 Vorhersagestudien in chaotischen Systemen und in der Praxis – Anwendung von Methoden der nichtlinearen Systemanalyse. Ph.D. thesis, Freie Universität Berlin, Meteorologische Abhandlungen N.F. Serie A Monographien.

Discussion

R. J. Bhansali (*Department of Statistics and Computational Mathematics, University of Liverpool, U.K.*). Dr Smith stated talk that the residuals obtained after fitting the 'local' model will not be 'white noice' with probability one. This led me to wonder why he cannot iterate further and attempt to forecast the residuals themselves from the past? In a sense, the main reason why a time-series forecaster seeks to have 'white' residuals is to ensure that what remains after fitting an appropriate model is unforecastable from the past. I can see some theoretical as well as practical difficulties in ensuring that this is so for the approach he takes for nonlinear prediction. I feel, however, that, at least in principle, it should be possible to improve over a forecasting procedure which does not yield residual which are 'white' in the sense of being purely random.

L. A. Smith. First, let me clarify that this is not a property particular to local predictors; global models tend to have even more structure in their residuals than local models. In general, we expect that non-perfect predictors of any sort will yield non-'white' residuals when applied to chaotic systems.

Whether or not one can detect this from the in-sample residuals is a separate question. If so, then adopting a more general prediction scheme is in order; iterating the prediction procedure is a good example. Even when the in-sample residuals are not distinguishable from an IID series, we expect that, given enough data, the out-of-sample residuals are. Here again one may use this new data in an iterative manner to predict their residuals; but in this case the fair comparison is with the original model trained on the entire data-set (including the previously 'out-of-sample' observations). In both cases, the prefered approach will depend on whether the parameters in the predictor have their optimal values, in a least squares sense. Even with the parameter values which minimize the out-of-sample r.m.s. error, the predictor will not yield white (IID) residuals; these are two distinct goals when forecasting nonlinear systems. If the parameters are not optimal, then further refinement of the initial model with additional data is profitable (i.e. it can reduce the r.m.s. error). If the optimal values have been obtained then a more flexible prediction scheme is called for; again iterated prediction provides an example. In short, I agree with Dr Bhansali that it should be possible to improve over a forecasting system in which residual predictability is observed.