Just do it?
Reductionism, Modelling and Black-box Forecasting

Patrick McSharry and Leonard A Smith
Mathematical Institute, University of Oxford,
24-29 St Giles', Oxford OX1 3LB, U.K.

July 7, 1998

Abstract-The reductionist approach has proven a powerful guide for scientific advancement over the last 300 years; constructing the simplest models consistent with the data remains a goal across the sciences. Yet are there instances where the blind pursuit of "simple" models is doomed from the start? Can we construct tests of internal consistency relating to the minimal duration of data from a given phenomenon, or between the data and a given model? In short, if the aim is to model a phenomena, should we just do it or first ponder the possible outcomes? This question is addressed in the context of the `datacomp.dat` data set.

I. Introduction

Nowadays it is common to attempt to reconstruct dynamics from data [12]. This can be approached in a variety of ways, depending upon what is known about the underlying *system*. One may build a *model* of the system from so-called first principles; or attempt to reconstruct a state-space model either explicitly (see [9] and references therein), or less explicitly via neural networks (see [12] and references therein); or adopt traditional statistical modelling approaches.

In the case of `datacomp.dat`, we know little of the system, having only 2000 observations, each of about 30 bits with no knowledge of the observational noise level. We are asked to extend the series by an additional 200 points, with the goal of minimizing the root mean square (RMS) error on the extension. We will consider the questions: What kinds of models should we consider? How should we choose a particular model structure? How might we decide if this model is consistent with our knowledge of the system? What kind of predictions should we make?

II. The Data

The time series `datacomp.dat` is shown in Figure 1a (here and below, time is measured in terms of the sampling interval). One is immediately struck by the range of scales over which the system shows changes in behaviour. The sampling rate is high enough so that the fast oscillations are well resolved, there being about 100 of them in the series; but the data also shows structured changes in behaviour on a time scale of about 200. This long term structure makes it doubtful that the available data is of sufficient duration to provide a good description of the range of behaviour typical of the underlying system. The lack of any significant quantization effects in the data indicate either that it has not been passed through any laboratory A/D converter or that significant filtering has taken place. In this sense, the data avoids some of the difficulties found in the Santa Fe test data (see [12], pp. 323-344). Yet it remains to be seen if the data set is long enough to yield a well-defined model within a particular model structure. We proceed by constructing a variety of model structures for this data set, and examining their data requirements.

III. Reconstruction and Prediction

In this section we consider four different approaches to modelling the observations. Our goal is to restrict the model class(es) considered in making predictions. First we illustrate that a class of stochastic processes is unlikely to provide a good model for this system. We then consider something of a "first principles" approach and construct a system of ordinary differential equations which mimic the general behaviours of the data. Third, simple analogue predictions are considered where the reconstruction is based on a time delay embedding. Finally, we deploy a different analogue technique, utilising the highly sampled data to determine local derivatives and thereby extract the dynamics.

A. Ruling out the easy options

Given the tremendous depth of research into properties of linear stochastic processes over the last 30+ years, we first wish to consider whether or not the data set might be easily modelled by such a process. Given the smoothness of the data, the abrupt changes in behaviour (on the 200 time scale), and the apparent lack of time-reversibility, we doubt that any simple linear stochastic process will provide a good model. One method to qualify this hunch is to generate "FFT surrogate" data sets with the same power spectra (and autocorrelation function) of the observations, but with random phases (see [8, 9, 11], and references thereof for details and improvements). Such surrogates are typical realizations of a linear stochastic process with the same autocorrelation function as the observations; one is shown in Figure 1b. The differences are apparent to the eye, although one could, of course, still quantify this by computing one of any number of test statistics which would reject the null hypothesis that the unknown system was a linear stochastic process at rather high levels of significance.

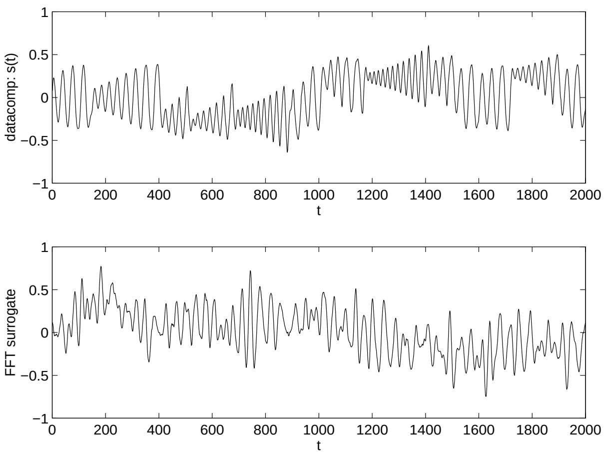

Figure 1: (a) The competition data `datacomp.dat` and (b) a linear FFT surrogate (b).

B. The Chair model

Both the fast oscillations and the longer structured behaviour of the observations may be modelled in a three-dimensional flow. The state space of one such model (Equations 1 and 2 below) is illustrated in Figure 2. This model is simply a collection of repelling harmonic oscillators centred at $\mathbf{x}_i = (x_i, y_i, z_i)$, $i = 1, 2, \ldots, n$, each centre is placed on the xy-plane (*i.e.* $z_i = 0, i = 1, 2, ..., n$). These oscillators dominate the dynamics when z is small, and are combined with a limit cycle centred at $\mathbf{x}_0 = (x_0, y_0, z_0)$ which is attracting when z is large. A typical trajectory near a $\mathbf{x}_i$ spirals outwards and upward (z increasing) about a line perpendicular to the xy-plane at $\mathbf{x}_i$. As the z co-ordinate increases there is a transition to the limit cycle mode, mimicking a structural change in `datacomp.dat`. Now in a large z mode, the trajectory is attracted toward the limit cycle, the radius of which is such that it nearly passes above the centres $\mathbf{x}_i$. Aperiodically, the trajectory strays above one of the centres, z then decreases exponentially as the trajectory falls down toward the particular centre $\mathbf{x}_j$; the small z motion is then dominated by the repeller at $\mathbf{x}_j$. And so on. The equations of motion are

$$
\begin{aligned}
\dot{x} = \quad & v(z) \sum_{i=1}^{n} s(r_i) \left[\alpha_i (x - x_i) - \omega_i (y - y_i)\right] \\
& + (1 - v(z)) \left[\alpha_0(x, y)(x - x_0) - \omega_0 (y - y_0)\right], \\
\dot{y} = \quad & v(z) \sum_{i=1}^{n} s(r_i) \left[\alpha_i (x - x_i) + \omega_i (y - y_i)\right] \\
& + (1 - v(z)) \left[\alpha_0(x, y)(y - y_0) + \omega_0 (x - x_0)\right], \\
\dot{z} = \quad & z \sum_{i=1}^{n} \left[a \exp(0.5/(r_i^2 + 0.1) - r_i^2) + 5 r_i\right] \\
& - 5z^2, \qquad (1)
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha_0(x, y) &= 1 - (1/r_0)\sqrt{x^2 + y^2}, \\
v(z) &= (1/2)(1 - \tanh((z - z_0)/0.5)), \\
s(r_i) &= (1/2)(1 - \tanh((r_i - 0.5)/0.1)). \\
r_i &= \sqrt{(x - x_i)^2 + (y - y_i)^2}, \qquad (2)
\end{aligned}
$$

and we take $\omega_0 = 20$, $r_0 = \sqrt{2}$, $n = 4$, $(x_0, y_0, z_0) = (0, 0, 5)$; $\omega_i = 40$, $\alpha_i = 1$ for $i = 1, .., 4$. The centres (x_i, y_i, z_i) are (1,1,0), (1,-1,0), (-1,1,0), (-1,-1,0). The coefficient a in the $\dot{z}$ equation determines whether or not the motion is periodic; a governs the damping of the downward force acting on the trajectory as it passes over the centres. For low values $1 \leq a \leq 2.5$, the motion is periodic. For $a = 2.9$, however, we observe aperiodic motion over time scales long relative to the length of `datacomp.dat`. By choosing an appropriate measurement function $h(x, y)$, a signal structurally similar to `datacomp.dat` may be obtained. Taking $h(x, y) = 0.3x + 0.1y$, yields the trace shown in Figure 3a.

We could, of course, significantly increase the superficial resemblance between Figure 3a and `datacomp.dat` by altering both the parameters of the model and $h(x, y)$. We have not done so for a variety of reasons; one of the more compelling is the length of `datacomp.dat`. Suppose for a moment that a chair model generated `datacomp.dat`, what duration of observations would be required to identify the particular values of the parameters, under the assumption that the general structure of the chair model was correct? The accuracy with which we must "identify the parameters" is clearly task dependent, in this case the instructions say our aim is to forecast 200 steps into the future. Preliminary tests show that the amount of data required is somewhat significantly greater than 2000 steps.

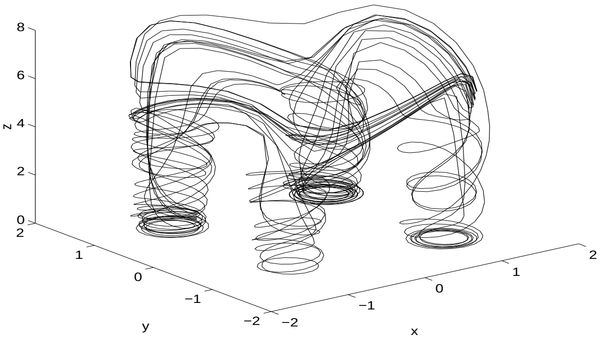

Figure 2: State space of the chair flow.

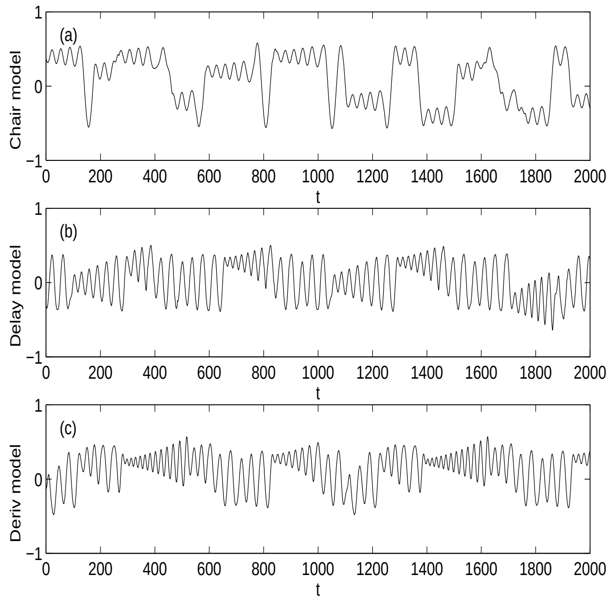

Figure 3: Three realistic models. (a) the signal observed from the chair flow using the measurement function $h(x, y) = 0.3x + 0.1y$, (b) the analogue delay model, and (c) the analogue derivative model.

C. Delay reconstructions

Analogue prediction was suggested as an antidote for no model (or no good model) by Lorenz [3]; this approach was successfully exploited in the Santa Fe time series prediction competition by Kostelich and Lathrop ([12] pp. 283-295). An analogue prediction is made by taking a reconstructed vector and searching for near neighbours in a library of such vectors. The image of this nearest neighbour is the prediction. In geometrical terms this corresponds to searching for the most similar pattern in the library of patterns and then forecasting the continuation of the most similar pattern.

Analogue forecasting provides a way to intelligently shuffle the original time series while conserving structure over some (random) time span of order τ_r. The technique involves making a time delay embedding [5] of the signal $\mathbf{s}_i = (s_{i-(m-1)}, ..., s_i)$. The delay vector $\mathbf{s}_i$ represents the pattern of length m in the data between times $t_{i-(m-1)}$ and t_i. In the delay space, near neighbours from the original signal provide analogues, ideally from the same region of the true state space. By patching together segments of data, each of length $\approx \tau_r \gg m$, we aim to capture the coherent long time structure of `datacomp.dat`. Beginning a new segment is only contemplated when there are "nearby" analogues, and once a new section of signal is selected, a minimal length on the order of τ_r is considered before another switch is contemplated. And so on. One such trajectory is illustrated in Figure 3b. In this example, $m = 10$ and the minimal period before searching for a new near neighbour considered $\tau_r = 100$ with variations chosen uniformly on the interval $[0, 200]$.

This method is particularly amenable to ensemble forecasting, discussed in Section IV. In this case multiple analogues are chosen, with different weights, as in Random Analogue Prediction (RAP, [6]).

D. Derivative reconstructions

The last model we consider is another analogue model, this time in a reconstructed state space based both on the data and its derivatives. The high signal to noise ratio of `datacomp.dat` allows us to exploit numerically estimated derivatives; these are computed via local polynomial fits to the time series itself (see [7]). This reconstruction sometimes holds an advantage over delay reconstructions by more effectively using information in high time resolution data (see [2]) while also avoiding the choice of a particular delay. In this reconstruction, the derivative vectors, $\mathbf{d}_i = (s_i, \dot{s}_i, \ddot{s}_i)$, can be used to make predictions, either by interpolative methods [5, 9, 12], or analogue methods. We will consider only the latter. Figure 4 contains a projection plot of s versus $\dot{s}$ for `datacomp.dat`. The dotted trajectory represents the first 1800 points of `datacomp.dat`, the dashed line represents the final 200 points and the circle marks the final point. The obvious coherent structure in this figure strongly suggests a deterministic approach to modelling `datacomp.dat` (at least between most time steps), yet the helter-skelter crossing of trajectories (some of which track each other for an extended time and thus would also yield near neighbours in a delay reconstruction) indicates a need either to consider significantly higher dimensional reconstructions or to invoke occasional stochastic perturbations. The duration of the data set restricts us to fairly low dimensional reconstructions, while the limited number of structural changes makes guessing a good stochastic formula for them unlikely, without additional information on the underlying system which generated the data.

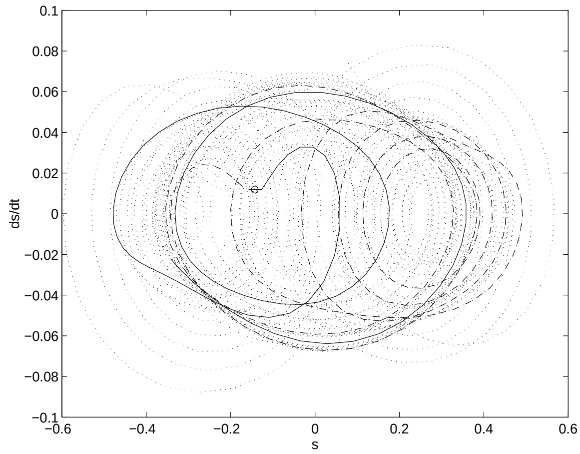

Figure 4: The reconstructed derivative state space, showing the evolution of $(s(t), \dot{s}(t))$ with time. The *dotted* trajectory represents the first 1800 points, the *dashed* line represents the final 200 points and the *circle* marks the final point. The continuation is the *solid* line.

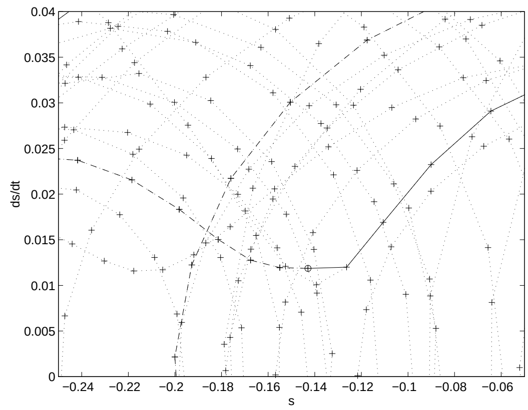

Figure 5: The zoomed in reconstructed derivative state space, showing the evolution of $(s(t), \dot{s}(t))$ with time. The *dotted* trajectory represents the first 1800 points, the *dashed* line represents the final 200 points and the *circle* marks the final point. The *crosses* mark the actual points. The continuation is the *solid* line.

Figure 5 is a zoom is about the final point (circle). Shown in 2-d projection, the solid line in Figure 5 represents a continuation of the series using a three-dimensional derivative reconstruction, *i.e.* $(s_i, \dot{s}_i, \ddot{s}_i)$; the corresponding time series is shown in Figure 6b. A typical model trajectory is given in Figure 3c.

Figure 5 brings into sharp focus the likely futility of deploying advanced interpolation methods on this data set. The dashed trajectory leading up to the final point has only one nearby trajectory. While most of the model state space is moderately well sampled, the particular region around the end point is sparse; the probability of accurately determining the future path over the next 200 time steps appears small.

IV. Forecast and Verification

Even if `datacomp.dat` originated in one of the relatively simple model structures considered above, data sets of duration 2000 fail to restrict the range of behaviours to the level such that good RMS forecasts over a period of 200 are likely: without additional information, how would one choose between the forecast of a suitably tuned chair-model and one of the data-driven models? And given that the forecasts of models which maintain realistic levels of uncertainty diverge rapidly (that is, over time scales much less than 200), the "optimal forecast" in the RMS error sense is unlikely to be a realistic trajectory of any of them [1, 10].

One option is to provide an ensemble forecast, where the ensemble is over both different models and also uncertainty in the initial condition, but this fails even to aim at the target of the competition. The mean of such an ensemble would provide a candidate, as would picking one model, based on intuition or luck and extending it 200 steps. Figure 6 shows extensions for three models. If the goal is truly a minimal RMS error trajectory over 200 steps, we would suggest a rapid decay to an estimated mean value of $\mu = 0.0146$. This suggestion is strengthened by the distinguishing behaviour near the end of `datacomp.dat` noted above.

Both delay and derivative reconstruction models offer reasonable trajectories; given the amount of data available, there is no reason to favour any one of these over the other. The winning predictor may win not because it has captured the most physics - but simply because it was fortunate enough to do the right thing. Figure 6c shows a combination of the derivative analogue prediction in Figure 6b but weighted so as to revert to the mean μ of the entire data set on a timescale t_α. For a forecast initiated at $t = t_0$,

$$F_{rms}(t) = \alpha(t) F_{deriv}(\mathbf{d}(t_0), t) + (1 - \alpha(t))\mu \qquad (3)$$

where the weighting function $\alpha(t)$ is

$$\alpha(t) = \exp(-(t - t_0)/t_\alpha) \qquad (4)$$

and $t_\alpha = 30$.

Can't someone do better than this? Yes. And no. It is doubtful that any purely data based approach to modelling `datacomp.dat` could yield a forecast model accurate in the RMS sense at 200 steps: given a variety of different model structures, any of which could be tuned to the existing data, how might one choose without recourse to serendipity? Without such serendipity, it seems clear that the best forecast model (in the RMS sense), is unlikely to prove the best model of the system in the reductionist sense. Each model structure will admit an "optimal model" with parameter values which provide the best fit for the combination of model structure, system, and cost function. In practice, every "optimal model" will fail to reproduce the exact state space dynamics of the system due to structural error. But why not at least use a model structure whose "optimal model" has a deep minima in the cost function of interest? (While admitting that the choice of model structure is based more on the background of the researcher than on the data.) One reason is simply that the available data is insufficient for one to expect the particular model obtained to be anywhere near that true minima.

Until a trajectory has explored the true state space, we can hardly expect to "reconstruct" a relevant model-state space: the data does not contain enough information to define what happens in soon to be visited regions of (either) state space. Out-of-sample, there is little to be gained from using an advanced interpolation model if the dynamics of the model when fit to these 2000 points would differ substantially from those of a model based on the next 2000 points. If the underlying system is a linear stochastic process, then there are many methods for estimating the appropriate complexity of the model; much less is known if the underlying system is a nonlinear deterministic process. And given only the data, we never know the process.

Given uncertainty in the initial condition or model structure, the model which minimizes the RMS forecast error at large lead times will not reflect the nonlinear dynamics of the system, thus it will fail to achieve the reductionist's goal. By looking for internal consistency between model and each individual forecast, given the observational uncertainty (see [4]), and attempting to construct models which shadow the data while **not** producing minimal RMS skill scores (see [1, 10]), we can attempt to construct boxes, be they black, grey, or transparent, which reproduce the observed dynamics to within the limits set by the frequency, duration, and accuracy of the available data. Without additional information about the system which produced the data, we prefer not to aim for more.

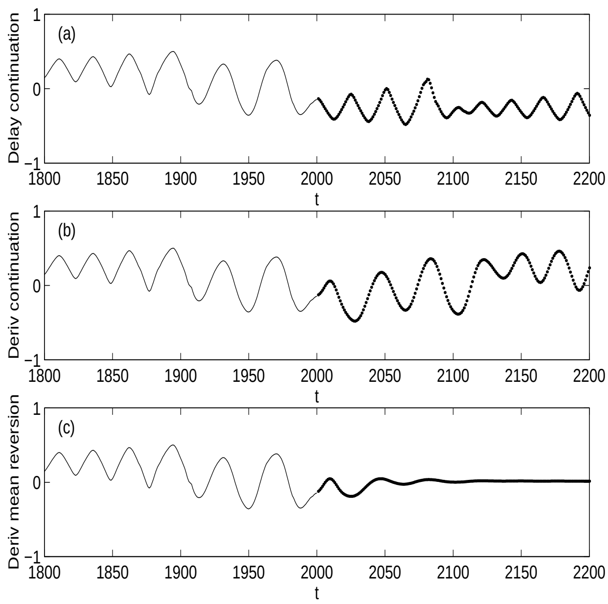

Figure 6: Continued time series of iterated analogue predictions of the time series for (a) delay reconstruction, (b) derivative reconstruction and (c) derivative reconstruction with mean reversion.

V. Acknowledgements

PM acknowledges support of a Marie Curie Research Fellowship from the European Union. LAS is supported by a Senior Research Fellowship at Pembroke College, Oxford. We gratefully acknowledge several rounds of discussion with A. Fowler on the philosophy of data-driven modelling.

References

[1] Gilmour, I. and Smith, L. A. "Enlightenment in shadows". In *Applied nonlinear dynamics and stochastic systems near the millennium*, edited by J. B. Kadtke and A. Bulsara, pp. 335–340, New York, 1997. AIP.

[2] Kantz, H. and Olbrich, E. "Scalar observations from a class of high-dimensional chaotic systems: Limitations of the time delay embedding". *Chaos*, **7**(3):423–429, 1997.

[3] Lorenz, E. N. "Deterministic nonperiodic flow". *J. Atmos. Sci.*, **20**:130–141, 1963.

[4] McSharry, P. *Uncertainty Propagation and Ensemble Selection in Nonlinear Systems.* University of Oxford, 1996. Transfer of status dissertation.

[5] Ott, E., Sauer, T. and Yorke, J. A. *Coping with Chaos.* John Wiley & Sons, Inc., 1994.

[6] Paparella, F.and Provenzale, A., Smith, L. A., Taricco, C. and Vio, R. "Local random analogue prediction of nonlinear processes". *Phys. Lett. A*, **235**:233–240, 1997.

[7] Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. *Numerical Recipes in C.* Cambridge University Press, Cambridge, 2nd edition, 1992.

[8] Provenzale, A., Smith, L. A., Vio, R. and Murante, G. "Distinguishing between low-dimensional dynamics and randomness in measured time series". *Physica D*, **58**:31–49, 1992.

[9] SMITH, L. A. "Identification and Prediction of Low-Dimensional Dynamics". *Physica D*, **58**:50–76, 1992.

[10] SMITH, L. A. "The Maintenance of Uncertainty". In *Nonlinearity in Geophysics and Astrophysics*, edited by G. Cini, volume CXXXIII of *International School of Physics "Enrico Fermi"*, Bologna, 1997. Il Nuovo Cimento.

[11] THEILER, J., EUBANK, S., LONGTIN, A., GALDRIKAN, B. AND FARMER, J. D. "Testing for nonlinearity in time series: the method of surrogate data". *Physica D*, **58**:77, 1992.

[12] WEIGEND, A. S. AND GERSHENFELD, N. A. *Time Series Prediction: Forecasting the Future and Understanding the Past*, volume XV of *SFI Studies in Complexity*. Addison-Wesley, New York, 1993.