2005 International Symposium on Nonlinear
Theory and its Applications (NOLTA2005)
Bruges, Belgium, October 18-21, 2005

Prediction, Behaviour, and Ignorance

Devin Kilminster[†,‡] and Reason Machete[†]

†Oxford Centre for Industrial and Applied Mathematics,
Mathematical Institute, 24–29 St. Giles', Oxford OX1 3LB, UK
‡Centre for the Analysis of Time Series,
London School of Economics, Houghton Street, London WC2A 2AE, UK
Email: devin@27720.net, machete@maths.ox.ac.uk

Abstract—In an "imperfect model scenario", the choice of values for the model's parameters is not merely a matter of "estimation", rather we must "fit" the parameters according to some criterion for the performance of the model. For time-series, one such criterion is that of prediction. A "predictionist" model is fit so as to optimise for short term prediction. Predictionist models, however can fall short in other ways — a model that makes very good short term prediction might fail to have reasonable long term behaviour when "free-run", for example. Interestingly, though, "behaviourist" models (ones which optimise a free-running behavioural criterion) often perform quite robustly when considered as predictors.

We investigate these issues in the context of "ignorance" — a score for measuring the performance of predictors. We develop a behavioural analog of ignorance and derive a few interesting connections between the two scores.

1. Introduction

When we build models from time-series data, we are faced with the problem of selecting a particular model (often by choosing values for a number of *parameters*) from amongst a larger *model class*. The problem of making this choice is sometimes phrased in terms of statistical estimation in which the aim is to *estimate* the "true" values for the parameters of the model class. Such an approach runs in to the problem of "model-imperfection", however. Our model-classes are almost never perfect — "truth" is almost never contained within the model class. Furthermore, within many model classes, the parameters don't even have an obvious correspondence with quantities in the real world. The philosophical implications of procedures that attempt to "estimate" true values of non-existent quantities are worrying to say the least!

A more philosophically sound stance is to accept that the problem of choosing the values for the parameters is really a problem of *fitting* rather than estimation. That is, the parameters are chosen so as to be fit for some particular purpose. In a *perfect-model scenario*, that is, when truth lies within the model class, for most reasonable purposes the best fit model will in fact be the truth, however outside of this, one would expect models fit for different purposes to differ.

In Figure 1, we show time-delay plots for data taken from a non-linear circuit, and two free-running realisations of different models of that circuit, both within the same model class. *Model 1* was fit so as to optimise its performance at making short-range predictions, while *Model 2* was fit so as to optimise its long-term "behaviour". Specifically, both models are taken from a class of *ellipsoidal basis-function models*, where given the time-series, $y_1, y_2, \ldots$, the model is:

$$y_{t+1} = f(y_t, y_{t-5}, y_{t-10}) + \epsilon_t,$$

where $f = \sum_{i=1}^{k} \omega_i \Phi_i$ is a linear sum (with parameters ω_i) of *basis functions*, either *ellipsoidal*,

$$\Phi_i(x) = \mathrm{e}^{-\left((x-c_i)^T R_i{}^T R_i (x-c_i)\right)^{p_i}},$$

(with parameters c_i, p_i and diagonal R_i) or members of the standard affine basis. The $\epsilon_t \sim N(0, \sigma^2)$ are i.i.d with parameter, σ. Both models have $k = 30$ basis functions.

Model 1 was built using methods similar to those described in [3, 4] — f is chosen to minimise the average one step squared error, $(f(y_t, y_{t-5}, y_{t-10}) - y_{t+1})^2$, that is to maximise its likelihood. The variance, σ^2, of ϵ_t is also chosen to maximise likelihood. Thus from the standpoint of statistical estimation, Model 1, is the maximum likelihood estimate. From the standpoint of fitting for a purpose, however, one notes that it attempts to place (on average[1]) the greatest possible probability on the one-step outcome, y_{t+1} — that is it optimises short-term (1-step) predictions. Model 1 can be said to be a *predictionist* model.

Even a casual examination of the delay plots in Figure 1 reveals that the maximum likelihood, predictionist, Model 1 fails to reflect important dynamical properties of the original system — Model 1 is essentially a slightly noisy periodic orbit; occasionally the noise forces it sufficiently far that it explores other parts of the attractor, but generally it fails to do so with a frequency matching that suggested by the data. Model 1 possesses poor *behaviour*. Roughly, behaviour can be defined as the property that the model exhibits the same sorts of phenomena with similar frequencies to those exhibited by the original system.

Model 2 was created by adjusting the parameters of Model 1 so as to minimise a behaviourist criterion. We

[1] Strictly speaking, the geometric average.

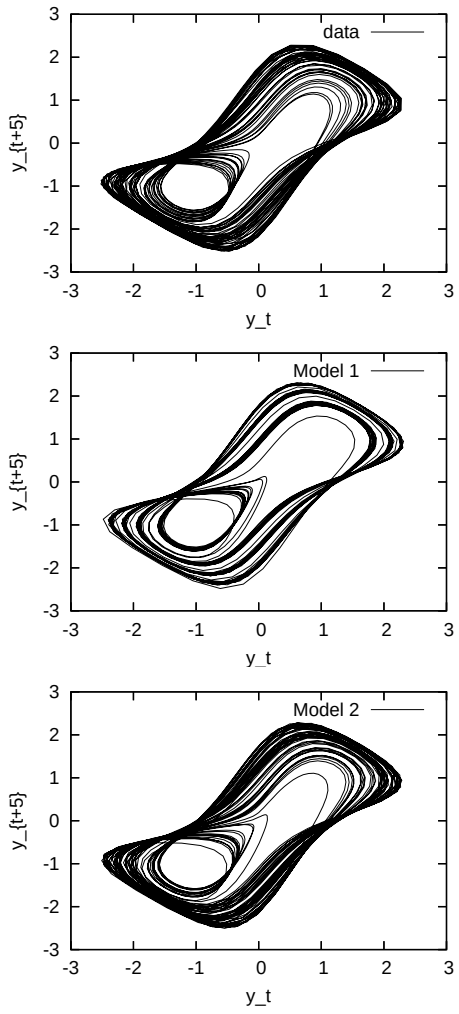

Figure 1: Delay plots for data from a non-linear circuit and two free-running models. Model 2 exhibits better "behaviour", although Model 1 makes better 1-step predictions.

shall defer the discussion as to the exact criterion used until Section 3, however we will note that the procedure is similar to those presented in [7, 5, 6]. The criterion attempts to minimise the difference in the distributions of strands of trajectories found in the data and the corresponding strands of trajectories generated by the model. In Figure 1, it can be seen that, at least in terms of exploration of the attractor, Model 2 better captures the dynamical features of the data. It should be emphasised that Model 2 is still, of course, imperfect — close examination of the delay plots reveal differences between model and data.

The plots of Figure 1 demonstrate that in the imperfect model case, there is a tradeoff between the different criteria we would like our models to have; specifically the predictionist ability to make good short term predictions and the behavioural ability to reflect longer range dynamical properties of the original system. Which model is better? The answer to that question must depend upon what the model will be *used* for — outside the perfect model scenario, there is no such thing as a single "best" model. If one wished to use the model solely to make short term predictions, then clearly the predictionist Model 1 is better. If one would like to capture more of the long-term dynamics of the system, then Model 2 seems a better choice.

It is important to note, however, that behaviourist models seem to enjoy a certain robustness over predictionist models, and this is explored in some detail in [7]. If we assume that the free-running long-term distribution[2] of observations exhibited by the original system is given by the family of densities, p, and those for the model given by $\tilde{p}$, we can make the following definitions: First, the *average prediction error*,

$$\mathrm{APE}_{P,F} = \int_{\mathbb{R}^m} p(P) \int_{\mathbb{R}^n} |\tilde{p}(F|P) - p(F|P)| \, \mathrm{d}F \mathrm{d}P,$$

where $P = y_{t-m+1}, \ldots, y_t$ is thought of as the *past* on which a prediction is to be made and $F = y_{t+1}, \ldots, y_{t+n}$ is thought of as the *future* to be predicted. It measures the average distance in 1-norm between $\tilde{p}(F|P)$, the predicted distribution of the future given P, and $p(F|P)$, the actual distribution of the future given P. Secondly, the *behaviour error*,

$$\mathrm{BE}_H = \int_{\mathbb{R}^l} |\tilde{p}(H) - p(H)| \, \mathrm{d}H,$$

where $H = y_t, \ldots, y_{t+l-1}$, is the *partial history* of length l. The behaviour error measures the distance in 1-norm between the distribution of partial histories exhibited by the model and the original system. The following theorem can be easily proved:

$$\mathrm{APE}_{P,F} \leq \mathrm{BE}_{P,F} + \mathrm{BE}_P \leq 2\mathrm{BE}_{P,F} \tag{1}$$

The point is that the cost to prediction of pursuing good behaviour is bounded. That is, a well behaved model will be not too bad as a predictor. The converse is not true. In [7], examples are constructed making APE arbitrarily small while BE remains large — the pursuit of prediction is not guaranteed to produce good behaviour.

2. Ignorance and Prediction

Whilst the theorem, (1), provides some reassurance for preferring, in general, a behaviourist approach to modelling, it leaves something to be desired. The problem is that the quantities, APE and BE, are not directly accessible. For example, given our model, we have the probabilistic prediction, $\tilde{p}(F|P)$, but generally we don't know $p(F|P)$. This mismatch is the central problem for any attempt to verify probabilistic forecasts — we only ever have the point outcome, never its "true" distribution given the data.

[2] We need to assume that these distributions have densities which, at least, our models do.

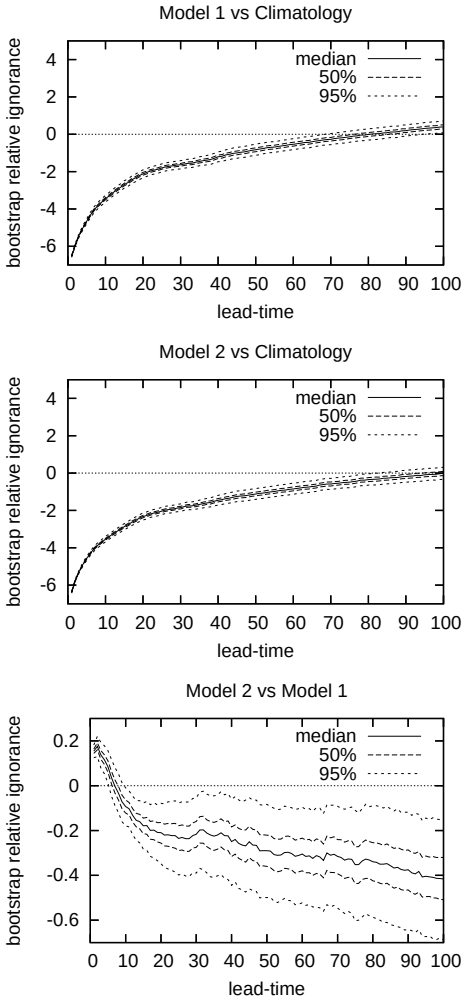

Figure 2: Bootstrap averages of the relative ignorance of Model 1 vs "Climatology", Model 2 vs "Climatology", and Model 2 vs Model 1 at various lead times. A relative ignorance for A vs B of less than zero indicates that A has better predictive performance.

There are, none the less, a wide range of scores designed to solve the problem of evaluating probabilistic prediction. One of the most appealing is the *ignorance* [2, 8]. Given the model distribution, $\tilde{p}$, the information, P, on which the prediction is to be made, and the point outcome, F, the ignorance is defined:

$$\text{ign}_P(\tilde{p}; F) = -\log \tilde{p}(F|P)$$

The smaller the ignorance, the "better" the prediction. Of course, the ignorance of a single prediction is not very informative, in practice, one will be interested in the average ignorance over a number of test cases. This is, of course, an estimate of the expected ignorance.

Other than its simplicity, the ignorance has a number of pleasing properties: It is *proper*, that is, given beliefs about the true distribution, i.e. $p(F|P) = q(F)$, one expects to minimise ignorance by the prediction, $\tilde{p}(F|P) = q(F)$. Also it is *local*, meaning that its value does not depend upon parts of the predicted distribution far from the actual outcome, F, in fact, ignorance is essentially the *only* score that is proper, local, and smooth. Ignorance has an information theoretic interpretation in that it measures the information needed to describe the outcome to a receiver in possession of the prediction. (The degree to which the receiver remains "ignorant" of the outcome.) The relative ignorance between two models is invariant of the co-ordinate system used to describe the outcome. The relative ignorance also has an interpretation in the theory of betting. If a bettor and the house each rationally bet or set odds on the outcome, then the bettor expects a log-return equal to the relative ignorance of the house.

In Figure 2, we illustrate the use of ignorance by showing plots of average relative ignorance between the two models and a "climatology" which predicts by sampling from the data without taking any account of dynamics. We show the ignorance for predictions at "lead-times" between 1 and 100 steps. The actual application of ignorance can be difficult — in principle we have $\tilde{p}(F|P)$, however in practice, this density can be difficult to compute. One solution is to approximate by using an *ensemble*, E, — a finite sample from $\tilde{p}(F|P)$. Letting the size of the ensemble be N, the dimension of F to be d and r the Euclidean metric, it can be shown that the *distance score*,

$$D(E; F) = K_d + \log N + d \log r(F, E),$$

becomes an unbiased estimator of the ignorance as N increases, for some constant K_d. We have used ensembles of size $N = 100$. Because only a finite amount of out-of-sample data is available, we have plotted bootstrap [1] intervals for the averages. As should be expected, Model 1 outperforms Model 2 for the shortest lead-times, however, for longer lead-times, Model 2 makes better predictions. For the improved dynamical behaviour, the slightly poorer predictions for the shortest lead-times seems generally to be a reasonable trade-off. If we care about longer lead-times, the behaviourist model is definitely to be preferred.

3. Ignorance and Behaviour

If we wish to develop an analog of the theorem, (1), in which predictive performance is measured by the ignorance, how should we measure behavioural performance? Noting that behaviour aims to match the unconditioned distributions of partial histories, H, an obvious possibility is:

$$\text{ign}(\tilde{p}; H) = -\log \tilde{p}(H)$$

By the Kullback-Leibler inequality, this is expected to be minimised when $\tilde{p} = p$. The desired theorem is easy to derive. Since $\tilde{p}(F|P) = \tilde{p}(P, F)/\tilde{p}(P)$, $-\log \tilde{p}(P|F) = -\log \tilde{p}(P, F) + \log \tilde{p}(P)$. So,

$$\mathbb{E}\left(\text{ign}_P(\tilde{p}; F)\right) = \mathbb{E}\left(\text{ign}(\tilde{p}; P, F)\right) - \mathbb{E}\left(\text{ign}(\tilde{p}; P)\right).$$

The term on the left is an expected predictive ignorance, and the terms on the right are expected behavioural ignorances. The terms on the right form a difference, suggesting that the predictive ignorance can be small, even if the behavioural terms are large.

$$\begin{aligned}\mathbb{E}\left(\mathrm{ign}(\tilde{p};P)\right) &= -\int_{\mathbb{R}^n} p(P)\log\tilde{p}(P)\,\mathrm{d}P \\ &\geq -\int_{\mathbb{R}^n} p(P)\log p(P)\,\mathrm{d}P \\ &= h(P),\end{aligned}$$

the entropy. So we have our bound (letting $H = P, F$):

$$\mathbb{E}\left(\mathrm{ign}_P(\tilde{p};F)\right) + h(P) \leq \mathbb{E}\left(\mathrm{ign}(\tilde{p};H)\right)$$

That is, the behavioural ignorance bounds the sum of the predictive ignorance and the entropy. The entropy isn't directly accessible, but at least it is independent of the model. Estimates of the other two quantities can be made by computing averages.

The parameters of Model 2 were fit by minimising a distance score estimate of such a behavioural ignorance.

4. Combining Predictions or Models

The densities we have been using up to now actually have to be defined with respect to some reference measure space, $(\mathbb{R}^n, \mathcal{B}, \mu)$; the density, $\tilde{p}$, gives probability $\int_B \tilde{p}(x)\,\mathrm{d}\mu(x)$ to the *event* $B \in \mathcal{B}$. By using a restriction, $\mathcal{A} \subset \mathcal{B}$, of the σ-algebra of events, we can create versions of ignorance which ignore certain details of the outcome. The trick is to write $\tilde{p}|_{\mathcal{A}}(\cdot|P) = \mathbb{E}\left(\tilde{p}(\cdot|P)|\mathcal{A}\right)$[3], and define the restricted ignorance,

$$\mathrm{ign}_{\mathcal{A}P}(\tilde{p};F) = -\log\tilde{p}|_{\mathcal{A}}(F|P).$$

It can happen that we have two models, with densities, $\tilde{p}$ and $\tilde{q}$, with the property that,

$$\mathbb{E}\left(\mathrm{ign}_P(\tilde{p};F)\right) < \mathbb{E}\left(\mathrm{ign}_P(\tilde{q};F)\right),$$

but,

$$\mathbb{E}\left(\mathrm{ign}_{\mathcal{A}P}(\tilde{p};F)\right) > \mathbb{E}\left(\mathrm{ign}_{\mathcal{A}P}(\tilde{q};F)\right).$$

That is, $\tilde{p}$ makes better predictions than $\tilde{q}$, but we have found a strength, encoded by $\mathcal{A}$, which $\tilde{q}$ has over $\tilde{p}$. It is then possible to define combined *predictions*,

$$\tilde{r}(F|P) = \tilde{p}(F|P)\frac{\tilde{q}|_{\mathcal{A}}(F|P)}{\tilde{p}|_{\mathcal{A}}(F|P)},$$

and,

$$\begin{aligned}\mathbb{E}\left(\mathrm{ign}_P(\tilde{r};F)\right) &= \mathbb{E}\left(\mathrm{ign}_P(\tilde{p};F)\right) + \\ &\quad \mathbb{E}\left(\mathrm{ign}_{\mathcal{A}P}(\tilde{q};F)\right) - \mathbb{E}\left(\mathrm{ign}_{\mathcal{A}P}(\tilde{p};F)\right) \\ &< \mathbb{E}\left(\mathrm{ign}_P(\tilde{p};F)\right).\end{aligned}$$

[3] Too much expectation notation is confusing.

This procedure is reasonably straight-forward when dealing with *predictions*, the analogous situation of combining *models* (to reduce behavioural ignorance) is more complicated. The problem is that the combination, $\tilde{r}$, does not necessarily correspond to a long term distribution.[4]

5. Conclusions

We have illustrated some of the tradeoffs to be considered when fitting models to time-series in the imperfect model scenario. In particular, we considered building predictionist and behaviourist models of data from a non-linear circuit, showing the relative robustness of the behaviourist model over the predictionist. We showed how the predictionist score of ignorance can be extended into the behavioural setting and finally offered some hint as to the possible advantages of doing so.

Acknowledgements

Support for parts of this work was provided by US WRP funds administered by NOAA. We also acknowledge fruitful discussions had with members of CATS and OCIAM.

References

[1] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals an d other measures of statistical accuracy. *Statistical Science*, 1:54–77, 1986.

[2] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society: SERIES B (METHODOLOG ICAL)*, XIV(1):107–114, 1952.

[3] Kevin Judd and Alistair Mees. Embedding as a modeling problem. *Physica D*, 120:273–286, 1998.

[4] Kevin Judd and Alistair I. Mees. On selecting models for nonlinear time series. *Physica D*, 82:426–444, 1995.

[5] Devin Kilminster. Reconstructing noisy dynamical systems. In *NOLTA '98*, volume 2, pages 627–630, Crans-Montana, Switzerland, 1998.

[6] Devin Kilminster. The benefits of complicated embeddings. In *NOLTA '99*, volume 1, pages 383–386, Waikoloa, Hawaii, 1999.

[7] Devin Kilminster. *Modelling dynamical systems via behaviour criteria*. PhD thesis, University of Western Australia, Perth, 2002.

[8] Mark S. Roulston and Leonard A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653–1660, 2002.

[4] A necessary condition is *invariance*, that $\tilde{r}(y,H) = \tilde{r}(H,y)$.