ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Lag Selection for Regression Models Using High-Dimensional Mutual Information

Geoffroy Simon[1] * and Michel Verleysen[1,2] †

1- Université catholique de Louvain, Machine Learning Group - DICE,
Place du Levant 3, B-1348 Louvain-la-Neuve, BELGIUM

2- Université Paris I - Panthéon Sorbonne, SAMOS-MATISSE, UMR CNRS 8595,
Rue de Tolbiac 90, F-75634 Paris Cedex 13, France

Abstract. Mutual information may be used to select the embedding lag of a time series. However, this lag selection is usually limited to the analysis of the mutual information between a pair of lagged values in the series. In this paper, generalized mutual information estimators are proposed to take into account more than two variables in the lag selection. Experimental results show that lag selection using mutual information should also take into account the output of the regression model.

1 Introduction

While working with time series, one has first to analyze them. Such analysis leads to a characterization of the series (stationary, periodic, chaotic, ... [1, 2]), as well as to the computation of some invariants (dimension, lag, Lyapunov exponents, ... [3, 4, 5, 1, 2]). The lag is an important value for the embedding of the series i.e. for its reconstruction in a state space [5].

Two approaches are usually used to estimate the lag. The first one consists in selecting the first value that corresponds to a zero of the autocorrelation function [1, 2]. The second one selects a value corresponding to a minimum of the mutual information (MI) [6, 1, 2]. However both approaches have the same goal: to select variables that are as much independent (or uncorrelated) as possible in order to reconstruct a trajectory in the state space that approaches at best the true dynamics of the time series.

In both these approaches one usually estimates the first lag and then uses multiples of it for the other lags. The lag selection problem is thus reduced to a particular case using only two variables: one at time t and one at time $t-\tau$, where τ is the lag. Autocorrelation or MI with three (or more) variables is not computed although the reconstructed state space is often 3-dimensional (or higher). There is thus a need to estimate the lag with more than two variables.

Furthermore, in a time series prediction context, the common approach is to estimate the lag by computing a criterion (autocorrelation or MI) between the inputs of the model regardless of the desired model output(s). The usual hypothesis behind this method is that a good reconstruction in a state space leads to a good prediction accuracy. However, this common belief is usually

*G. Simon is funded by the Belgian F.R.I.A.
†M. Verleysen is a Research Director of the Belgian F.N.R.S.

not quantitatively measured, even in the lag selection step where it could be measured easily.

In this paper it is first suggested selecting the lag using MI between more than two variables at the model input side, and then taking into account the desired model output. This approach is recommended in the context of time series prediction as it allows selecting a set of past values in the series that contains as much information as possible with respect to the output to be predicted. The MI will be estimated in any dimensional space using a k-nearest neighbour based MI estimator introduced recently [7]. Experimental results based on this MI estimator show the positive impact of taking into account the model output in the lag selection.

The paper is organized as follows. Section 2 presents the MI approach for lag selection. Section 3 then introduces the lag selection using high-dimensional k-nearest neighbour based MI estimator, and introduces a way to take into account the desired model output. Experimental results in section 4 show the usefulness of taking into account the model output while selecting the lag of a time series.

2 Embedding in a state space: the lag selection

A time series is defined as a series of values x_t measured from a varying process. The x_t values are usually ordered according to the time index t.

In time series prediction context one has to build a model of the series that can be denoted as:

$$\hat{x}(t+1) = f\left(x(t), x(t-\tau), x(t-2*\tau), ..., x(t-(d-1)*\tau)\right), \qquad (1)$$

where f is the model (it can be linear or nonlinear) and $\hat{x}(t+1)$ is the prediction. Values $x(t), x(t-\tau), x(t-2*\tau), ..., x(t-(d-1)*\tau)$ are often grouped in vectors called state vectors [5]. Notation d is the dimension of the time series and τ is the lag [5, 1, 2]. The question regarding how to choose the dimension d is decisive for the model [3, 4, 1, 2] but is quite independent from the goal of this work which is to choose an adequate lag τ. Dimension d will therefore be deemed to be fixed a priori throughout the rest of this paper.

In practice lag τ is often selected as the first zero value of the autocorrelation function [1, 2], but this criterion only measures the linear dependencies between a variable $x(t)$ and the lagged one $x(t-\tau)$. A nonlinear alternative proposed to select the lag is the mutual information (MI) [6, 1, 2]. This lag selection approach will be developed in this paper.

An estimator of the MI can be defined as:

$$\mathrm{MI}(x(t), x(t+\tau)) = \sum_{x(t),x(t+\tau)} \mathrm{P}\left[x(t), x(t+\tau)\right] \log\left(\frac{\mathrm{P}\left[x(t), x(t+\tau)\right]}{\mathrm{P}\left[x(t)\right]\mathrm{P}\left[x(t+\tau)\right]}\right), \qquad (2)$$

where P[.] denotes the probability. This criterion is a nonlinear measure of how much information on $x(t)$ can be deduced from the knowledge of $x(t+\tau)$. The lag to select is the one minimizing (2).

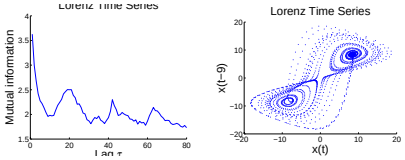

Fig. 1: Left: MI for the Lorenz time series. Right: reconstructed 2-dimensional state space (the selected lag is $\tau = 9$).

Figure 1 shows the use of the MI criterion on a time series obtained from the Lorenz equations [1, 2], in the $d = 2$ case. The left part shows the MI with respect to the lag τ. According to this plot, $\tau = 9$ is chosen. The reconstructed 2-dimensional state space is shown on the right.

3 Lag selection using high-dimensional MI

The main limitations of the criterion in Equation (2) are first its definition for two variables $x(t)$ and $x(t-\tau)$ only and second the non-use of the desired predicted value. The following methodology is therefore introduced to get rid of both these drawbacks. A similar generalization could be provided for the autocorrelation function, but this paper focuses on MI as mentioned in section 2.

The 2-dimensional definition in (2) can be generalized to any dimension:

$$\mathrm{MI}(X_1, ..., X_m) = \sum_{X_1,...,X_m} \mathrm{P}\left[X_1, ..., X_m\right] \log\left(\frac{\mathrm{P}\left[X_1, ..., X_m\right]}{\mathrm{P}\left[X_1\right] \ ... \ \mathrm{P}\left[X_m\right]}\right), \tag{3}$$

where X_1, ..., X_m are variables that can be $x(t)$, ..., $x(t-(m-1)*\tau)$ for example. In practice, high-dimensional MI is very difficult to estimate, but an estimator of such high-dimensional MI, based on k-nearest neighbours, has been introduced recently [7]. This estimator has been used in the experiments presented in Section 4.

Shortly, the estimator of $\mathrm{MI}(X, Y)$ between X and Y is based on counting the number of points in the X and Y spaces respectively, up to a limit distance fixed according to the k^{th} nearest-neighbour in the combined XY space. The existence of a relation between the points in subspaces X and Y, obtained by projection of the points in the XY space, will give rise to a relation between the two counts as soon as the limit distance tends to be small. A variant of this estimator is also given in [7] with distinct bounds in each space X and Y. As expected from [7] experimental results confirm that both estimators behave similarly (except for a small bias without consequence on the search for minima and maxima); the first estimator is used in this paper. The most interesting property of the estimator in [7] is that X and Y are not restricted to be scalar variables; they may be vector as well, what provides an estimator for (3).

Now, taking into account the output $x(t+1)$, Equation (3) becomes:

$$\mathrm{MI}(X_0; (X_1, ..., X_m)) = \sum_{X_0,X_1,...,X_m} \mathrm{P}\left[X_0, (X_1, ..., X_m)\right] \log\left(\frac{\mathrm{P}\left[X_0, (X_1, ..., X_m)\right]}{\mathrm{P}\left[X_0\right] \mathrm{P}\left[X_1, ..., X_m\right]}\right), \tag{4}$$

where X_0 denotes the model output $x(t+1)$ and variables X_1, ..., X_m can be $x(t)$, ..., $x(t-(m-1)*\tau)$ as above. Note that any prediction horizon $h \geq 1$ can be considered as output; $h = 1$ is chosen here for the ease of illustration.

A k-nearest neighbour estimator for the MI defined in (4) is also proposed in [7]. Indeed variables X and Y can be replaced by vectors in the computation of $\mathrm{MI}(X, Y)$ once a norm has been defined in the X and Y spaces.

The two above criteria can be used together to select the embedding lag. Intuitively it is expected that the MI estimated with (3) should be as low as possible as the goal is to have state vector components as much independent as possible. This approach is coherent with the 2-dimensional case usually applied to select the lag [6]. On the contrary the MI estimated with (4) should be as high as possible since in this case the state vector components should provide as much information as possible to explain the output. Since high-dimensional MI estimators are available, it is now possible to observe the influence of the model output in the lag selection.

Finally note that definitions (3) and (4) are independent of the regression model that is used for prediction. They thus provide a general framework for the lag selection, in the form of an unsupervised method.

4 Experimental results

This section illustrates the use of the two MI criteria (3) and (4) in the lag selection problem on artificial and real-world time series.

A real time series is the Santa Fe A time series [8] that has been proposed in the Santa Fe Time Series Prediction and Analysis Competition held in 1991. The 1000 data were collected from a Far-Infrared-Laser in a chaotic state. The artificial time series were generated from the Henon map and the Lorenz system:

- Henon map [1, 9]:

$$\begin{aligned} x_{t+1} &= 1 - 1.4x_t^2 + y_t, \\ y_{t+1} &= 0.3x_t. \end{aligned} \tag{5}$$

  2000 data have been generated using these finite difference equations.

- Lorenz system [1, 9, 2]:

$$\begin{aligned} \dot{x}_t &= 10(y_t - x_t), \\ \dot{y}_t &= 28x_t - y_t - x_t z_t, \\ \dot{z}_t &= x_t y_t - \tfrac{8}{3} z_t. \end{aligned} \tag{6}$$

  4000 data have been generated using an integration step of 0.015. The first 2000 data have been discarded to remove any transient state and let the trajectory fall to the attractor.

The real-world and the two artificial time series are shown in Figure 2.

From Figure 3 it can be observed that MI (3) presents some local minima for the 2-dimensional case. Though these minima appear more or less at the same values that the maxima of (4) as expected, they are obviously much less

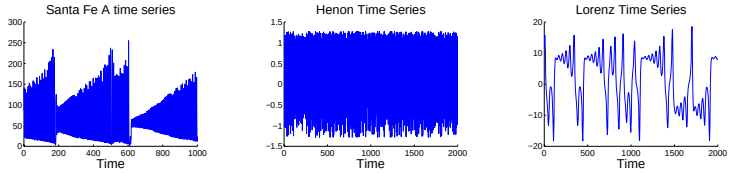

Fig. 2: From left to right: Santa Fe A, Henon and Lorenz time series.

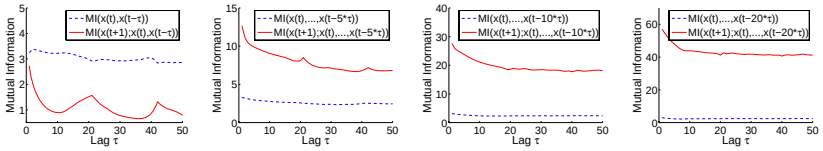

Fig. 3: Lorenz time series: comparisons of the MI values obtained according to criteria (3) and (4) with respect to the lag. From left to right: 2-dimensional, 5-dimensional, 10-dimensional and 20-dimensional state vectors.

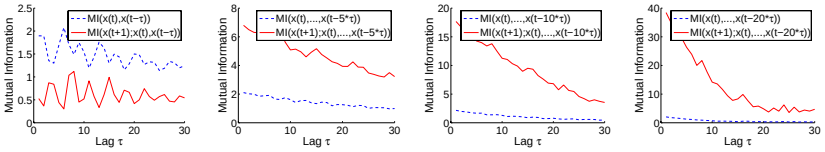

Fig. 4: Same results than Figure 3 for the Santa Fe A time series.

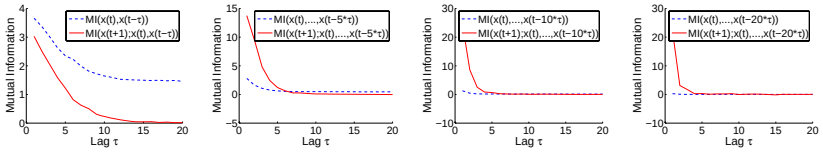

Fig. 5: Same results than Figures 3 and 4 for the Henon time series.

distinguishable. This tends to prove that using only the model inputs to select the lag as in Equation (3) is not sufficient. Furthermore the minima that appear in the 2-dimensional case are no longer observable in any other dimension. What is much more problematic is the fact that for dimensions higher than two the MI computed with (3) is not longer low while MI (4) is high. On the contrary, both the criteria have high values together. Similar comments can be formulated from Figure 4. In Figure 5, no clear minimum can be found for MI (3) while MI (4) is high, even in two dimensions.

These figures confirm two facts. First, when more than two consecutive values (one lag) are taken into account, it may become difficult to select the adequate value of the lag with criterion (3). This means that unfolding in a state space become difficult, and that different unfoldings (corresponding to different lags) may become indiscernible (from the criterion point of view). Furthermore, what is even more important is that the minimum of criterion (3), even when it exists, may not correspond anymore to a maximum of criterion (4). Although, it is clear that a maximum of (4) is the goal to be reached: one tries to select a set of past values in the series that contains as much information as possible with respect to the output $x(t+1)$ to be predicted; this goal corresponds exactly to (4). This result questions the optimality of lag selection performed with a state space unfolding goal. As using the desired output value is possible through criterion (4) (or equivalent), it is strongly suggested to use the latter instead of criterion (3) to select an optimal lag.

5 Conclusion

In this paper, mutual information estimators generalized to the high-dimensional case and taking into account the output of a regression model have been used to select the embedding lag.

Experimental results show that using a methodology that simply generalizes the usual lag selection approach based on two lagged values becomes problematic in spaces of dimension higher than two. It is thus recommended to use a mutual information estimator that takes into account the output of the regression model. With such approach the selected lagged variables are those that indeed provide the most useful information in the context of time series prediction.

Further work will include a study of the influence of the prediction horizon on the methodology that should be used to select the most adequate lag.

References

[1] H. Kantz, T. Schreiber, *Nonlinear Time Series Analysis*, Cambridge Nonlinear Science Series 7, Cambridge University Press, Cambridge, 1997.

[2] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* Springer-Verlag, New-York, 1997.

[3] F. Takens, Detecting strange attractors in turbulence, D. A. Rand and L. S. Young, eds, *Dynamical Systems and turbulence* Lecture Notes in Mathematics 898, Springer, 1981.

[4] P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, *Physica D* 9:189-208, Elsevier, 1983.

[5] T. Sauer, J. Yorke, and M. Casdagli, Embeddology, *Journal of Statistical Physics*, 65:579-616, Springer, 1991.

[6] A. M. Fraser and H. L. Swinney, Independent coordinates for strange attractors from mutual information, *Phys. Rev. A*, 33:1134–1140, 1986.

[7] A. Kraskov, H. Stögbauer, and P. Grassberger, Estimating mutual information *Phys. Rev. E* 69(6):066138, 2004.

[8] A. Weigend, N. Gershenfeld, *Time Series Prediction: Forecasting the future and Understanding the Past*, Santa Fe Institute, MA, Addison-Wesley Publishing Company, 1994.

[9] D.T. Kaplan, L. Glass, *Understanding Nonlinear Dynamics*, Springer, New York, 1995.