

Lag Selection for Regression Models Using High-Dimensional Mutual Information

Geoffroy Simon¹ * 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:

$$\hat{x}(t+1) = f(x(t), x(t-\tau), x(t-2*\tau), \dots, x(t-(d-1)*\tau)), \quad (1)$$

where f is the model (it can be linear or nonlinear) and $\hat{x}(t+1)$ is the prediction. Values $x(t), x(t-\tau), x(t-2*\tau), \dots, x(t-(d-1)*\tau)$ 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:

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

where $P[\cdot]$ 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).

