Non-linear Multivariate Model Identification using Information Measures

Ivan Kojadinovic and Henri Ralambondrainy
IREMIA – University of Reunion Island
15, av. Rene Cassin, F-97715 Saint-Denis de la Reunion, France
Phone: +262-93-82-82, Fax: +262-93-82-60
email:{ikojadin,ralambon}@univ-reunion.fr

ABSTRACT: This paper proposes a model-free approach for ranking the relevancy of candidate input variables in non-linear multivariate problems. In order to measure the general dependence between a candidate input variable and the output variable, we employ the average mutual information. As the computation of the average mutual information implies the estimation of probability densities, we review kernel density estimation procedures. Finally, the approach is validated on the Box-Jenkins gas furnace data and the obtained models are compared with models proposed in the literature.

KEYWORDS: mutual information, non-linear model identification, kernel density estimation, fuzzy control

INTRODUCTION

Non-linear multivariate modelling is known to be a very difficult problem and its complexity increases as the number of possible input variables grows. When the problem to be modelled is strongly non-linear, classical linear techniques perform very badly. In the connectionist literature, several approaches to input variable selection, such as Sensitivity Based Pruning by Moody and Utans (1994), were proposed. Such approaches are neural-network-based and therefore should logically imply the use of a neural network in the modelling process.
In this paper, we investigate a model-free approach to input variable selection based on mutual information. Mutual information is an information-theoretic measure, which can be used to rank the relevancy of candidate input variables. Mutual information was used in Fraser and Swinney (1986) to establish an appropriate time delay in phase-space reconstruction from univariate non-linear time series data. In Moon et al. (1995), the authors used kernel density estimators for mutual information estimation and demonstrated the superiority of their approach over the recursive algorithm proposed by Fraser and Swinney.
This paper is organised as follows. First, we define the average mutual information. Then, after presenting kernel density estimators, we focus on the adaptive kernel method that we used for probability density estimation throughout this work. Finally, we validate this approach to input selection on the gas furnace data studied in Box and Jenkins (1970) and compare the generalisation ability of our models with models proposed in the literature.

MUTUAL INFORMATION

The average mutual information is a powerful statistic for measuring the general dependence between two variables. Let X and Y be two coordinates of a multivariate time series. We denote by $\{x_1,..,x_n\}$ (resp. $\{y_1,\ldots,y_n\}$) the time series of the variable X (resp. Y), where n is the record length. We assume that the sampling frequency is fixed. The average mutual information between the two time series is defined in bits as

$$I(X,Y) = \sum_{i,j} P_{X,Y}(x_i, y_j) \log_2 \left(\frac{P_{X,Y}(x_i, y_j)}{P_X(x_i) P_Y(y_j)} \right), \quad (1)$$

where $P_{X,Y}(x_i, y_j)$ is the joint probability density of X and Y evaluated at (x_i,y_j) and $P_X(x_i)$ and $P_Y(y_j)$ are the marginal probability densities of X and Y evaluated at x_i and y_j respectively.
When X and Y are statistically independent, $P_{X,Y}(x, y) = P_X(x) P_Y(y)$ which causes I(X,Y) to be equal to zero. On the contrary, the more X and Y are dependent, the higher their mutual information is.
A more rigorous introduction to mutual information can be found in Cover and Thomas (1991).

KERNEL DENSITY ESTIMATION

From (1), it can be seen that density estimation plays a key role in the computation of mutual information. As in Moon et al. (1995), we used kernel density estimators for the computation of the mutual information. A comprehensive introduction to kernel density estimation can be found in Silverman (1986). In the following, after some general definitions, we review the multivariate adaptive kernel method that we used throughout this work.

DEFINITION OF THE MULTIVARIATE KERNEL DENSITY ESTIMATOR

Given multivariate time series data with unit variance $\{x_1,...,x_n\} \in R^d \times ... \times R^d$, the multivariate kernel density estimator with kernel K and window width h is defined by

$$\hat{P}_X(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{1}{h}(x - x_i)\right) \tag{2}$$

A kernel is a continuous symmetric function K(x) defined on R^d that integrates to unity. In this paper, we have chosen to use the multivariate Epanechnikov kernel

$$K_e(x) = \begin{cases} \frac{1}{2c_d}(d+2)(1-x^T x) & if \quad x^T x < 1 \\ 0 & otherwise \end{cases} \tag{3}$$

where c_d is the volume of the unit d-dimensional sphere: $c_1 = 2, c_2 = \pi$, etc.
The kernel estimator can be seen as a sum of "bumps" centred at the observations.
In kernel density estimation, the choice of the window width h is crucial. If h is too small, the density may be "under"-smoothed and consequently spurious details may appear in the density estimate. If h is too large, the density may be "over"-smoothed and therefore some important details may be hidden. In Moon et al. (1995), the authors used the window width that minimises the mean integrated square error in $\hat{P}_X(x)$ if the underlying distribution is assumed to be multivariate gaussian. In Silverman (1986), the "optimal" Gaussian bandwidth corresponding to the multivariate Epanechnikov kernel is given by

$$h = \left\{ \frac{8d(d+2)(d+4)(2\sqrt{p})^d}{n(2d+1)c_d} \right\}^{\frac{1}{d+4}} \tag{4}$$

As the assumption underlying the choice of window width in (6) is rarely verified, we opted for the adaptive kernel method which allows the window width to vary at each observation. From a practical point of view, this method generates estimates that should be closer to the real underlying density and copes better with the tails of the distribution.

ADAPTIVE KERNEL ESTIMATES

Given multivariate time series data with unit variance $\{x_1,...,x_n\} \in R^d \times ... \times R^d$, the algorithm of the adaptive kernel method for estimating the density at the observations is given in Silverman (1986):

(i) Compute the pilot density estimate $\left(\tilde{P}_X(x_1),...,\tilde{P}_X(x_n)\right)$ using the window width defined in (4).

(ii) Define the local bandwidth factors λ_i by $l_i = \left(\tilde{P}_X(x_i)/g\right)^{-a}$ where g is the geometric mean of the $\tilde{P}_X(x_i)$ i.e. $\log g = \frac{1}{n}\sum_{i=1}^{n} \log \tilde{P}_X(x_i)$ and α is the sensitivity parameter ($0<\alpha<1$).

(iii) Compute the adaptive kernel estimates at $x_1, x_2, \ldots, x_n$ using $\hat{P}_X(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(hl_i)^d} K\left(\frac{t - xi}{hl_i}\right)$

From the previous, it can be seen that the local window width used at observation x_i is $h\lambda_i$.
We set $\alpha = 0.5$ thereby following the recommendations given in Silverman (1986) and in Abramson (1982).

APPLICATION TO FUZZY CONTROL

In order to assess the performance of our approach, we applied mutual information to non-linear multivariate model identification. As training data, we used the Box-Jenkins gas furnace data that are often studied and compared. The furnace input is the gas flow rate x(t) and the output is the CO_2 concentration. We consider at least 20 input candidates: x(t-10),x(t-9),…,x(t-1),y(t-10),y(t-9),…,y(t-1). As mentioned in Zhang and Knoll (1998), if all 20 inputs are used, building a fuzzy controller for predicting y(t) means solving a 20-input-1-output problem. If each input is defined using 4 linguistic terms, the number of possible rules is 4^{20}.

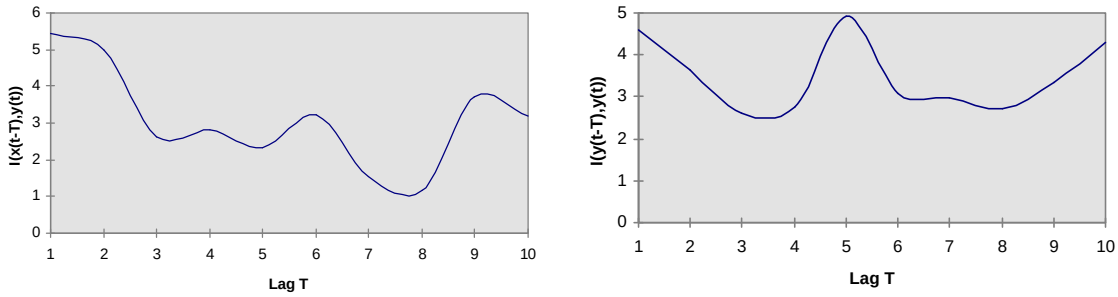

Figure 1: I((x(t-T),y(t)) and I((y(t-T),y(t)) plotted against the lag.

In Fig. 1, I((x(t-T),y(t)) and I((y(t-T),y(t)) were plotted against the lag. The most informative variables appear to be x(t-1), followed by x(t-2),y(t-5) and y(t-1). Furthermore, in order to understand the interdependencies between candidate variables, we computed the mutual information matrix (see Table II).

In order to validate the ranking of the candidate input variables obtained using mutual information, we tested the generalisation ability of different models, which were all implemented as fuzzy rule-based systems. The first 200 input-output pairs of the Box-Jenkins gas furnace data were used for learning and the last 80 input-output data pairs for testing. For all models, we used the method proposed by Nozaki et al. (1997) for generating fuzzy rules from numerical data. In order to ensure a fair comparison, the same number of linguistic terms was used for models with same number of inputs. The results in terms of generalisation ability are presented in Table I. The 2-input model with the lowest root mean square error (RMSE) and the lowest mean absolute percentage error (MAPE) appear to be the one with x(t-1) and y(t-5) as input variables. The best 3-input model is obtained by using the three variables with the highest mutual information. As one can see, the selection of the most informative variables in the mutual information sense seems to lead to the models that generalise the best. However, one should bear in mind that selecting the most informative variables does not necessarily lead to the most informative set of variables especially if the selected variables contain redundant information. As an example, compare the results obtained for Chiu's / Xu's / Lu's model (x(t-1), y(t-3) as input variables) and the Mutual Information 3 model (x(t-1), y(t-1) as input variables). Although, y(t-1) appears to be more informative then y(t-3) (see Table II), Chiu's / Xu's / Lu's model shows better generalisation ability. This may be due to the fact that I(x(t-1),y(t-3))is inferior to I(x(t-1),y(t-1)) which would mean that the couple (x(t-1), y(t-1)) is less informative then the couple (x(t-1), y(t-3)).

More details about the models that were used for the comparison can be found in Zhang and Knoll (1998).

Model	***Inputs***	***Number of linguistic terms per input variable***	***Test RMSE***	***Test MAPE***
Mutual Information 1	x(t-1), x(t-2)	20	3,23	5,10
Mutual Information 2	x(t-1), y(t-5)	20	**3,00**	**4,64**
Mutual Information 3	x(t-1), y(t-1)	20	3,19	4,73
Tong's / Pedrycz's	x(t-1), y(t-4)	20	3,69	5,57
Chiu's / Xu's / Lu's	x(t-1), y(t-3)	20	3,17	4,60
Mutual Information 4	x(t-1), x(t-2), y(t-5)	12	**2,74**	**3,94**
Mutual Information 5	x(t-1), x(t-2), y(t-1)	12	3,31	4,96
Mutual Information 6	x(t-1), y(t-1), y(t-5)	12	2,89	4,14
Chiu's	x(t-1), x(t-3), y(t-3)	12	3,57	5,11

Table I:. Comparison of different models in terms of generalisation ability.

	x(t-8)	x(t-7)	x(t-6)	x(t-5)	x(t-4)	x(t-3)	x(t-2)	x(t-1)	y(t-8)	y(t-7)	y(t-6)	y(t-5)	y(t-4)	y(t-3)	y(t-2)	y(t-1)	y(t)
x(t-8)	**NA**	2,43	4,57	3,64	1,47	3,05	2,55	1,97	7,02	5,67	5,14	2,91	2,72	2,23	3,29	1,49	1,17
x(t-7)	2,43	**NA**	2,41	4,57	3,69	1,59	3,06	2,52	3,21	7,08	5,62	4,94	2,89	2,73	2,31	3,20	1,53
x(t-6)	4,57	2,41	**NA**	2,39	4,62	3,67	1,56	3,06	2,72	3,16	7,00	5,60	4,97	2,78	2,70	2,31	3,22
x(t-5)	3,64	4,57	2,39	**NA**	2,40	4,67	3,61	1,61	3,12	2,77	3,07	7,06	5,58	5,01	2,72	2,81	2,34
x(t-4)	1,47	3,69	4,62	2,40	**NA**	2,31	4,84	3,66	2,59	3,11	2,77	3,18	7,00	5,56	5,14	2,72	2,82
x(t-3)	3,05	1,59	3,67	4,67	2,31	**NA**	2,25	4,82	3,87	2,58	3,05	2,72	3,06	6,82	5,52	5,20	2,62
x(t-2)	2,55	3,06	1,56	3,61	4,84	2,25	**NA**	2,17	4,04	3,84	2,56	3,02	2,90	3,09	7,09	5,49	4,99
x(t-1)	1,97	2,52	3,06	1,61	3,66	4,82	2,17	**NA**	2,89	4,00	3,88	2,57	2,94	**2,83**	3,12	**7,12**	5,45
y(t-8)	7,02	3,21	2,72	3,12	2,59	3,87	4,04	2,89	**NA**	4,35	3,79	2,64	2,69	5,07	3,11	3,02	2,70
y(t-7)	5,67	7,08	3,16	2,77	3,11	2,58	3,84	4,00	4,35	**NA**	4,46	3,71	2,69	2,75	5,00	3,12	2,97
y(t-6)	5,14	5,62	7,00	3,07	2,77	3,05	2,56	3,88	3,79	4,46	**NA**	4,48	3,57	2,62	2,72	4,99	3,11
y(t-5)	2,91	4,94	5,60	7,06	3,18	2,72	3,02	2,57	2,64	3,71	4,48	**NA**	4,50	3,60	2,62	2,77	4,91
y(t-4)	2,72	2,89	4,97	5,58	7,00	3,06	2,90	2,94	2,69	2,69	3,57	4,50	**NA**	4,57	3,51	2,59	2,74
y(t-3)	2,23	2,73	2,78	5,01	5,56	6,82	3,09	2,83	5,07	2,75	2,62	3,60	4,57	**NA**	4,65	3,59	2,60
y(t-2)	3,29	2,31	2,70	2,72	5,14	5,52	7,09	3,12	3,11	5,00	2,72	2,62	3,51	4,65	**NA**	4,76	3,65
y(t-1)	1,49	3,20	2,31	2,81	2,72	5,20	5,49	7,12	3,02	3,12	4,99	2,77	2,59	3,59	4,76	**NA**	4,61
y(t)	1,17	1,53	3,22	2,34	2,82	2,62	4,99	5,45	2,70	2,97	3,11	4,91	2,74	2,60	3,65	4,61	**NA**

Table II:. Mutual information matrix.

CONCLUSION AND FURTHER WORK

Although the approach to input selection presented in this paper led to the best models for the Box-Jenkins data, it suffers from a serious drawback. Indeed, selecting the most informative variables does not necessarily lead to the most informative set of variables. One possible way of improving the current approach would be to use the natural extension of mutual information for more than two variables, which is called redundancy. However, as this would imply the estimation of probability densities in higher dimensions, the sample size would have to be drastically increased. For instance, in order to compute the redundancy of a set of ten variables with a satisfactory accuracy, 10-dimensional probability densities would have to be estimated, which, according to Silverman, would require more then 800 000 points.

REFERENCES

Abramson, I.S., 1982, “On Bandwidth variation in kernel estimates – a square root law”, Ann. Statist., 10, 1217-1223

Box, G.E.P.; Jenkins, G.M., 1970, “Time series analysis, Forecasting and Control” , (3rd edition), Englewood Cliffs, NJ: Prentice-Hall

Cover T.M.; Thomas J.A., 1991, “Elements of Information Theory”, New York: Wiley

Fraser, A.; Swinney, H.L., 1986, Phys. Rev. A33,1134

Moon Y.;Rajagopalan B. ; Lall U., 1995, “Estimation of mutual information using kernel density estimators”, Phys. Rev. E, Vol. 33, No. 3

Silverman, B. W., 1986, “Density Estimation for Statistics and Data Analysis”, Chapman and Hall, New York

Moody J.; Utans J., 1994, “Architecture Selection Strategies for Neural Networks: Application to Corporate Bond Rating Prediction”, Neural Networks in the Capital Markets, John Wiley & Sons

Nozaki K.; Ishibuchi H.; Tanaka H., 1997, “A simple but powerful heuristic method for generating fuzzy rules from numerical data”, Fuzzy Sets and Systems 86 (1997) 251-270

Zhang, J.; Knoll, A., 1998, “Constructing Fuzzy Controllers for Multivariate Problems by Using Statistical Indices”, Proc. FUZZ’IEEE 98, 1619-1624