Variable Selection Using Information Entropy in Time Series Prediction

T.Babaie+, C.Lucas++

+Control and Intelligent Processing Center of Excellence, Electrical and Computer Eng. Department, University of Tehran, Tehran, Iran

++School of Intelligent Systems, Institute for Studies in Theoretical Physics and Mathematics, Tehran, Iran

t.babaie@ut.ac.ir, lucas@ipm.ir

***Abstract*—** Non-linear prediction models have shown good performance in reflection of uncertainties and complexities in real world problems of decision making. Extending the predictor variables of predictions is the challenging and difficult task of prediction. In this paper, we investigate the problem of selecting non-contiguous input variables for usual prediction models in order to improve the prediction ability. In this paper, we discuss an Entropy-Based Approach to Time Series Analysis. The basic concept of entropy in information theory has to do with how much randomness in a signal or in a random event. In probability theory and information theory, and Statistics, a quantity called Relative entropy is often used to assess information content. We successfully test proposed algorithm with a chaotic time series by selecting in-puts on K.U.Leuven data set.

***Keywords*—** Time Series, Relative Entropy, Variable Selection

I. INTRODUCTION

FEATURE subset selection plays an important role in analysis of many problems of Time series prediction encountered in science and technology, for instance in climatology [1], and economics [2]. The objectives of variable selection is to improve the prediction performance of the predictors, to provide faster and more cost effective predictors, and to provide a better understanding of the underlying process that generated the data. Some of the independent variables may not contribute at all to the model. Thus we have to select from these variables to obtain a model which contains as little variables as possible while still being the "best" model. In principle, all possible combinations of independent variables should be tried for calculating a suitable model. This could turn out to be a formidable task, even if high performance computers are available. Important discussion would be about “criterion”, since that a simple criterion, like the goodness of fit, r^2, may lead to wrong conclusions if the number of selected variables approaches the number of observations.

In this study, we discuss an Entropy-Based Approach to Time Series Analysis. The basic concept of entropy [4], [8], [9] in information theory [5] has to do with how much randomness in a signal or in a random event. In *probability theory* and *information theory*, and *statistics*, a quantity called the Kullback-Leibler (KL) divergence or Relative entropy is often used to assess information content [3], [6], [7]. In other word the Kullback-Leibler divergence, or relative entropy, is a quantity which measures the difference between two probability distributions and often referred to as the discrimination gain. As mathematic point view KL is a measure of the distance between two probability density functions. In this paper, we use *Kullback-Leibler Information Criterion* to find desired variables. The new estimated distribution is chosen to be as close as possible to the original in the sense of minimizing the associated *Kullback-Leibler Information Criterion*, or *relative entropy*.

We utilize an algorithm for input variable selection in the spirit of stepwise selection, in which variables are progressively removed from the prediction model and, which variables are added to the prediction model. The removal and additional of variables is based on a median and a standard deviation of parameter distributions sampled with an entropy based approach. These statistics reflect the uncertainty for a variable to play an important role in the prediction task.

We apply the algorithm in a prediction setting, where input selection is performed for different one-step-ahead prediction models. Finally results used in the prediction of a prominent chaotic benchmark, K.U.Leuven time series data generated for an international competition on Advanced Black-Box Techniques for Nonlinear Modeling, held at the K.U.Leuven Belgium.

II. VARIABLE SELECTION: STEPWISE ALGORITHM

Stepwise selection is a method to find the "best" combination of variables by starting with a single variable, and alternative adding and eliminating the variables, step by step. Which variables to add or eliminate is decided according to desire criteria.

The method is started by first selecting the variable which results in the best fit for the dependent variable *Y*. Next, this

variable is used to test all combinations with the remaining variables in order to find the "best" pair of variables. All variables are tested if their contribution is significant after a new variable has been added. This may lead to the elimination of an already selected variable if this variable has become superfluous because of its relationship to the other variables. In all further steps, additional variables are added until either all variables are used up, or some stopping criterion is met (i.e. the criteria below a certain limit). The algorithm could be shown in these steps:

1) Calculate the partial correlations of all predictor variables, *Xi*, with the response variable *Y*. Use the variable with the highest partial correlation as the starting variable.
2) Add the variable with the highest criteria value.
3) Check all variables of the current model for their criteria values and remove any variable which falls below a predefined threshold.
4) Repeat the procedure with step 2 until some stopping criterion is met.

III. RELATIVE ENTROPY: KULLBACK-LEIBLER INFORMATION CERITERION

A. Information Entropy

Entropy is a concept in *thermodynamics*, *statistical mechanics* and *information theory*. The concepts of information and entropy have deep links with one another, although it took many years for the development of the theories of statistical mechanics and *information theory* to make this apparent.

In information theory, the Shannon entropy or *information entropy* is a measure of the uncertainty associated with a random variable. The concept was introduced by Claude E. Shannon in his 1948 paper "A Mathematical Theory of Communication" [10].

In physics, the word entropy has important physical implications as the amount of "disorder" of a system. In mathematics, a more abstract definition is used. Shannon defined a measure of entropy:

$$H(X) = -\sum_{x} P(X) \log_2 [P(x)]$$

bits, where P(x) is the probability that *X* is in the state *x*, and $Plog_2P$ is defined as 0 if $P=0$. The joint entropy of variables $X_1,..., X2$ is then defined by

$$H(X_1,..., X_n) = -\sum_{x_1} ... \sum_{x_n} P(x_1,..., x_n) \log_2 [P(x_1,..., x_n)]$$

B. Cross Entropy

In *information theory*, the *cross entropy* between two probability distributions measures the overall difference between the two distributions. *Cross entropy* is closely related to Kullback-Leibler divergence (which is also known as the *relative entropy*). The *cross entropy* for two distributions p and q over the same probability space is defined for discrete p and q this means as follows:

$$H(p,q) = -\sum_{x_1} P(x) \log q(x)$$

The situation for continuous distributions is analogous:

$$-\int_X p(x) \log q(x) dx.$$

C. Relative Entropy: Kullback-Leibler Divergence

In *probability theory* and *information theory*, the Kullback-Leibler divergence, or *relative entropy*, is a quantity which measures the difference between two probability distributions. It is named after Solomon Kullback and Richard Leibler, two NSA (National Security Agency) mathematicians [11]. The term "divergence" is a misnomer; it is not the same as divergence in calculus. One might be tempted to call it a "distance metric", but this would also be a misnomer as the Kullback-Leibler divergence is not symmetric and does not satisfy the triangle inequality.

The Kullback-Leibler divergence between two probability distributions p and q is defined as

$$KL(p,q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

for distributions of a discrete variable, and as

$$KL(p,q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}$$

for distributions of a continuous variable.

It can be seen from the definition that

$$KL(p,q) = -\sum_{x} p(x) \log q(x) + \sum_{x} p(x) \log p(x)$$
$$= H(p,q) - H(p)$$

denoting by $H(p,q)$ the *cross entropy* of p and q, and by $H(p)$ the entropy of p. As the *cross entropy* is always greater than or equal to the entropy, this shows that the Kullback-Leibler divergence is nonnegative, and furthermore $KL(p,q)$ is zero iff $p = q$.

Although $KL(p,q) \neq KL(q,p)$, so relative entropy is therefore not a true metric, it satisfies many important mathematical properties. For example, it is a convex function of p.

IV. K.U. LEUVEN TIME SERIES DATA

Within the framework of the International Workshop on Advanced Black-Box Techniques for Nonlinear Modeling, held at the K.U.Leuven Belgium July 8-10 1998, a new time series competition had been organized. The data are generated from a computer simulated 5-scroll attractor, resulting from a

generalized Chuas circuit, Fig.1. Chuas circuit is well known to be a paradigm for chaos [12, 13] being a simple nonlinear electrical circuit that reveals a rich variety of phenomena. The generalized Chuas circuit consists of nonlinearity with multiple breakpoints, leading to a family of n-scroll attractors [14, 15].

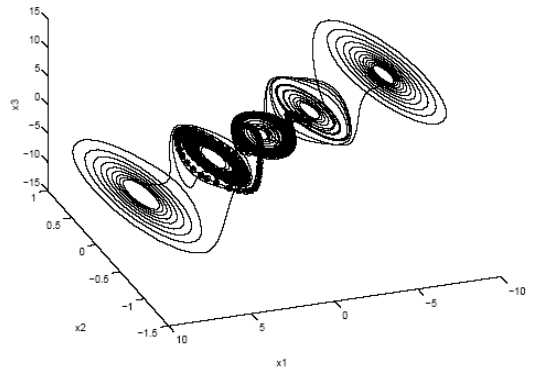

Fig.1. Computer simulated 5-scroll attractor from which the competition data have been generated.

The data were generated from the following computer simulated generalized Chua's circuit ([15], [16]):

$$\begin{cases} \dot{x}_1 &= \alpha\,[x_2 - h(x_1)] \\ \dot{x}_2 &= x_1 - x_2 + x_3 \\ \dot{x}_3 &= -\beta x_2 \end{cases}$$

with piecewise linear characteristic

$$h(x_1) = m_5 x_1 + \frac{1}{2}\sum_{i=1}^{5}(m_{i-1} - m_i)(|x_1 + c_i| - |x_1 - c_i|)$$

with parameters $\alpha = 9$, $\beta = 14{:}286$ and for the vectors $m=[m_0;m_1;:::;m_{2q-1}]$, $c=[c_1;c_2;:::;c_{2q-1}]$ one takes

$$m = [0.9/7; -3/7; 3.5/7; -2.7/7; 4/7; -2.4/7]$$
$$c = [1; 2.15; 3.6; 6.2; 9]$$

The generalized Chua's circuit has been simulated for initial state *[0:1; 0:2; 0:3]* with a Runge-Kutta integration rule (ode23 in Matlab). The competition data have been obtained then by taking a nonlinear combination of the 3 state variables:

$$y = W\tanh(Vx)$$

where $x = [x1; x2; x3]$ is the 3-dimensional state vector and the nonlinearity is a multilayer perceptron with 3 hidden units, interconnection matrices

$$W = [-0.0124 \quad 0.3267 \quad 1.2288]$$

$$V = \begin{bmatrix} -0.1004 & -0.1102 & -0.2784 \\ 0.0009 & 0.0792 & 0.6892 \\ 0.1063 & -0.0042 & 0.0943 \end{bmatrix}$$

and a zero bias vector. This multilayer perceptron is hiding the underlying structure of the attractor. The resulting time series is 2000 data points we have used, Fig.2.

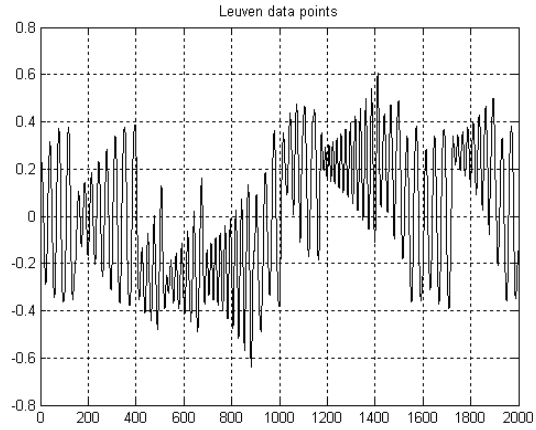

Fig.2. 2000 given data points

V. NONLINEAR MODEL AND VARIABLE SELECTION

Based on the results from the previous section, we train a non-linear model. Here, we have used a multi layer perceptron (MLP) network. The network is trained using optimization method by back propagating the error gradients. A partial correlation calculation is shown in Fig.3. The selected number of initial variables could be between 60 pervious values in time series.

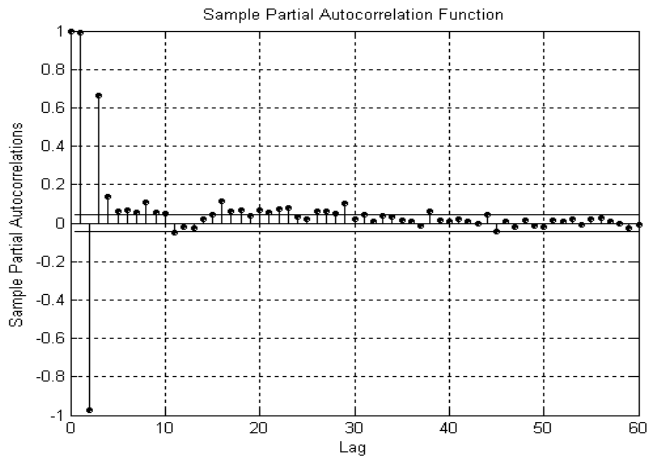

Fig.3. Partial correlation calculation for K.U.Leuven data set

Input selection algorithm has been used to yield input variables. Fig.4 presents an example of input selection in the case of one-step-ahead prediction. In this case, algorithm selected the model with 4 inputs. Numerical results for best

cases in final population are shown in Table1. Fig.5 and Fig.6 show resulted variable according to both minimum KLIC and RMSE criterion.

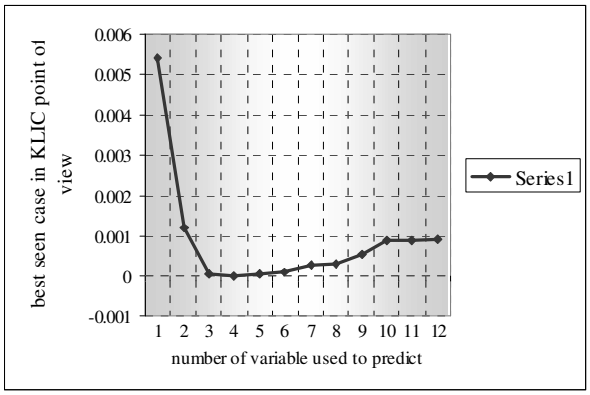

Fig.4. Input selection, Kullback-Leibler Entropy Criterion, the final model is chosen to be the least complex model 4 inputs.

TABLE1. SELECTED INPUTS FOR BEST MODELS ACCORDING TO THE MINIMUM KLIC AND THE CORRELATION COEFFICIENT AND ROOT MEAN SQUARE ERROR

Input variables	RMSE of prediction	Correlation criterion	KL criterion
(1,2,3)	0.005362	0.99959	5.22E-05
(1,2,3,8)	0.005155	0.99961	4.12E-05
(1,2,4)	0.008254	0.99931	2.22 E-04
(1,2,4,8)	0.004688	0.99967	7.91E-05

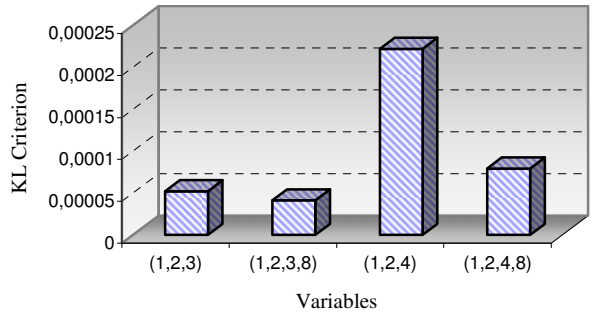

Fig.5. Selected inputs for best models according to the minimum KLIC

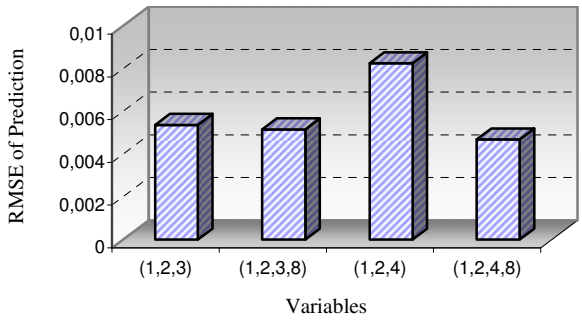

Fig.6. Selected inputs for best models according to the minimum RMSE

VI. CONCLUSION

The proposed algorithm selected parsimonious sets of inputs for all prediction models. The Structure built using the corresponding inputs led to good prediction performance. The main advantage of the proposed approach is that it combines fast input selection with accurate but computationally demanding non-linear prediction.

REFERENCES

[1] Dal Cin, C., Moens, L., Dierickx, P., Bastin, G., Zech, Y.: An Integrated Approach for Real-time Flood-map Forecasting on the Belgian Meuse River. (Natural Hazards), In press.

[2] Hamilton, J.D.: Analysis of Time Series Subject to Changes in Regime. Journal of Econometrics 45 (1990) 39-70

[3] Claude E. Shannon, Warren Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, 1963. ISBN 0252725484

[4] Thomas M. Cover, Joy A. Thomas. *Elements of information theory* New York: Wiley, 1991. ISBN 0471062596

[5] Maxwell's Demon: Entropy, Information, Computing, H. S. Leff and A. F. Rex, Editors, Princeton University Press, Princeton, NJ (1990). ISBN 069108727X

[6] Cover, T. M. and Thomas, J. A. *Elements of Information Theory.* New York: Wiley, 1991.

[7] Qian, H. "Relative Entropy: Free Energy Associated with Equilibrium Fluctuations and Nonequilibrium Deviations." 8 Jul 2000.

[8] Ellis, R. S. Entropy, Large Deviations, and Statistical Mechanics. New York: Springer-Verlag, 1985.

[9] Havil, J. "A Measure of Uncertainty." §14.1 in *Gamma: Exploring Euler's Constant*. Princeton, NJ: Princeton University Press, pp. 139-145, 2003.

[10] Shannon, C. E. "A Mathematical Theory of Communication." The Bell System Technical J. 27, 379-423 and 623-656, July and Oct. 1948.

[11] Kullback S. and Leibler R. A., On Information and Sufficiency. Annals of Mathematical Statistics 22(1):79–86, March 1951.

[12] Chua L.O., Komuro M. and Matsumoto T., "The Double Scroll Family," IEEE Trans. Circuits and Systems-I, 33, No.11 pp.1072-1118, 1986

[13] Madan R.N., (Ed.), Chua's Circuit: A Paradigm for Chaos. Singapore: World Scientific Publishing Co. Pte. Ltd, 1993.

[14] Suykens J.A.K., Vandewalle J., "Generation of n-double scrolls (n=1,2,3,4,…)," IEEE Transactions on Circuits and Systems-I (Special issue on chaos in nonlinear electronic Circuits), Vol.40, No.11, pp.861-867, 1993.

[15] Suykens J.A.K., Huang A. and Chua L.O., "A family of n-scroll attractors from a generalized Chua's circuit," Archiv fur Elektronik und Ubertragungstechnik (International Journal of Electronics and Communications) Vol.51, No.3, pp.131-138, 1997.

[16] Suykens J.A.K., Vandewalle J., "The K.U.Leuven competition data: a challenge for advanced neural network techniques", in *Proc. of the European Symposium on Artificial Neural Networks (ESANN'2000)*, Bruges, Belgium, 2000, pp. 299-304.