Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 70 (2007) 1265–1275
NEUROCOMPUTING
www.elsevier.com/locate/neucom

High-dimensional delay selection for regression models with mutual information and distance-to-diagonal criteria

Geoffroy Simon[a,*,1], Michel Verleysen[a,b,2]

[a]*Machine Learning Group - DICE, Université catholique de Louvain,Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium*
[b]*SAMOS-MATISSE, UMR CNRS 8595, Université Paris I - Panthéon Sorbonne Rue de Tolbiac 90, F-75634 Paris Cedex 13, France*

Available online 11 January 2007

Abstract

Delay selection for time series phase space reconstruction may be performed using a mutual information (MI) criterion. However, the delay selection is in that case limited to the estimation of a single delay using MI between two variables only. A high-dimensional estimator of the MI may be used to select more than one delay between more than two variables but this approach is rather time consuming. In this paper, an alternative fast criterion is proposed to optimize all delays for a high-dimensional phase space reconstruction: the distance-to-diagonal (DD) criterion, based on a geometrical heuristic. The use of the distance to diagonal criterion is illustrated and compared to MI on artificial and benchmark time series.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Delay and multiple delay selection; Regressor; Phase space reconstruction; Mutual information; Distance-to-diagonal; Regression model

1. Introduction

Time series are encountered in many fields of application. In some situations the application goal consists in forecasting future values; in other ones, extracting a model of the series without forecasting goal is searched. In both cases, a preliminary analysis of the time series may provide useful information for the design of the forecasting or regression model. Many techniques have been developed in physics, statistics, mathematics and econometrics in order to characterize a series as periodic (or not), stationary (or not), chaotic (or deterministic), homoscedastic (or heteroscedastic), etc. Analysing a time series also gives other useful information as the presence and the amount of noise, the intrinsic dimension of the series and the delay between successive values that are best inserted as inputs to the model. The last two characteristics can greatly influence the quality of the model under design.

The dimension and the delay are mandatory information for the phase space reconstruction, or *embedding* [17], of a series. Knowing or estimating these two values allows the construction of state vectors. The state vectors, also called the *regressors*, contain all the information available about the state of the process at a given time. They are also considered as the vectors that provide the most useful information in order to forecast the next state of the process. Since the publication of theoretical results about the dimension needed to reconstruct the structure of a series through an embedding in its phase space [22,17] many methods have been developed for estimating this dimension: correlation dimension [9], false neighbours [12], Box–Couting [17], minimum dimension [5], etc. Concerning the selection of the delay common approaches are the use of the autocorrelation of the series, or the mutual information (MI) [8]. The goal is to select variables as regressor components that are as uncorrelated or independent as possible. Consequently each variable in the regressor will provide useful information not provided by the other variables. Wrapper methods, i.e. methods using the outputs of e.g. a prediction model, can also be used to select regressor components; nevertheless, wrapper methods are supervised, and usually result in a huge

*Corresponding author. Tel.: +32 10 47 80 61; fax: +32 10 47 25 98.
E-mail address: simon@dice.ucl.ac.be (G. Simon).
URL: http://www.dice.ucl.ac.be/~gsimon/.
[1]G. Simon is funded by the Belgian F.R.I.A.
[2]M. Verleysen is a Research Director of the Belgian F.N.R.S.

0925-2312/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.10.150

computational load when nonlinear models are used. In this paper, we are interested in unsupervised methods that do not make use of a prediction model; phase state reconstruction is the goal, for which no supervision is available. In addition, we concentrate on nonlinear approaches: in [8], it was shown that the MI is superior to other filter methods like the autocorrelation, as further confirmed e.g. in [1,14]. This paper focuses on filter approaches that do not make any assumption (for example linearity) about a data model.

The correlation is a linear measure of dependency. It is adapted when a linear model is built, but does not provide the required information in a nonlinear context. The MI is a non-parametric measure of dependency between variables, able to detect any relation, whether the latter is linear or not. Though MI is a criterion commonly used for delay selection it suffers from two limitations in that context. The first one is that MI is usually computed between two variables only: it detects the relations between two regressor values, but fails in that context to provide the information about high-dimensional relationships between all regressor values. The second limitation arises from the fact that a unique delay is selected. The regressors are then constructed using variables that are equally distributed in time (at multiples of the chosen delay). This approach artificially constrains the regressors: the delay is selected according to two-dimensional information while the regressors themselves can be multi-dimensional.

A more general approach would be to select different delays for the various variables taken into account in the regressors. The selection of these delays should ideally be performed in a space which has the same dimensionality as the regressors. A first attempt to solve this more general problem is to use a high-dimensional MI estimator, as the one provided recently in [13]. This specific MI estimator has already been used in some previous works. In [21,20,10,16] it is used for feature selection in a supervised context where the model output is taken into account. In [18] it is used to observe the effect of the dimension on the delay selection, using a unique delay in an unsupervised context. The high-dimensional MI estimator is used in this paper as a generalization of the unsupervised two-dimensional case. However, one of the main limitations of this estimator is its computation time in $O(N^2)$, where N is the number of values in the series. As the number of possible combinations of variables in a regressor increases exponentially with the regressor size, testing all the possibilities with an algorithm in $O(N^2)$ becomes rapidly unfeasible in practice.

In this paper, an alternative to the unsupervised approach based on high-dimensional MI is also proposed for delay and multiple delay selection: the distance-to-diagonal criterion (DD for short). The DD criterion is a geometrical heuristic based on the regressor distribution in the state space: the DD criterion measures the portion of the space filled by the regressor distribution. The DD criterion is justified in a geometrical way: it is shown that if the regressor distribution does not fill a large portion of the state space, it means that their components are highly correlated, therefore that their global information content is low. Maximizing the portion of the space filled by the regressor distribution is thus a way to maximize the information content of the regressors, before using the latter e.g. in a prediction model. It will be shown that the MI and DD criterion provide qualitatively the same information, with an obvious advantage to the DD criterion in terms of computation load, making the selection of delays in high-dimensional regressors feasible in practice.

In addition, to its use to the selection of a single delay in high-dimensional regressors, it is also shown that the DD criterion may be used for multiple delay selection. Finally, it is shown that all these properties are obtained within an algorithm with a computation time in $O(p * N)$ where p is the regressor size. These properties will be illustrated on artificial and benchmark time series.

This paper is organized as follows. Section 2 explains the regressor reconstruction in a phase space in relation with the embedding theory. Section 3 recalls the approach of delay selection using MI in the usual two-dimensional unsupervised case. Then, the general framework of multiple delay selection using high-dimensional MI is introduced. The high-dimensional MI estimator used for this unsupervised multiple delay selection is briefly explained. Section 4 then introduces the DD criterion and compares it to the high-dimensional MI. Section 5 presents the time series, the regression model, the methodology and the experimental results. Section 6 concludes this paper and discusses future developments of the proposed method.

2. Embedding, phase space reconstruction and regression

A time series S is a series of scalar values $x(t)$, $1 \leqslant t \leqslant N$, measured at a (usually) constant sampling rate from a time varying process. The temporal data $x(t)$ are ordered according to the time index.

From a theoretical point of view the embedding [17] of a time series S is a one-to-one relation between the original (temporal) structure of the time series and the reconstruction of the structure in a phase space. Theoretical results ensure the existence of this one-to-one relation provided that the dimension of the reconstruction space is strictly larger than two times the intrinsic dimension of the original series [22,17]. Note that in this paper the term structure of a time series is preferred to attractor as the latter is more specifically related to chaotic systems; in practice, although all systems are not chaotic, they all show a characteristic structure once reconstructed in their phase space.

As mentioned in the introduction, many methods designed for estimating the dimension of the original structure have been proposed in the literature: correlation dimension [9], false neighbours [12], Box–Couting [17], minimum dimension [5], etc. The estimation of the dimension of the structure is a necessary step for determining the dimension of the phase

space or embedding space [17]. The estimation of the structure dimension is not considered in this paper; in the experimental section several possible dimensions will be considered for each time series.

In practice, the embedding of a time series is achieved by reconstructing the series in its phase space. So-called *delayed vectors* in the phase space are defined as

$$x_t = \{x(t), x(t-\tau), x(t-2*\tau), \ldots, x(t-(p+1)*\tau)\}, \quad (1)$$

where τ is the delay and p is the dimension of the reconstructed phase space. Consequently a forecasting or *regression* model f of a time series S can be built on these delayed vectors as:

$$\hat{x}(t+1) = f(x(t), x(t-\tau), x(t-2*\tau), \ldots, x(t-(p+1)*\tau)). \quad (2)$$

Note that model f can be either linear, as simple ARX models, or nonlinear, as neural networks for example. Obviously the vectors of model inputs needed for the regression in Eq. (2), called the *regressors*, correspond in fact to the state vectors. In the remaining of this paper, the terms regressors, state vectors and delayed vectors will be used indifferently.

Once the dimension p of the regressors is known or estimated for instance through the intrinsic dimension estimators listed above, the value τ of the delay has to be estimated. This can be performed in an unsupervised way using MI as described in Section 3.

It should be noted that according to Taken's theorem [22], only the dimension of the regressors is important to reconstruct them in the phase space. Therefore, in theory, a delay value of one could be adopted in all situations. However, a theoretical delay of one may be too restrictive for some practical aspects of time series analysis and forecasting. In practice, time series values (therefore regressors) are sampled using an a priori chosen frequency. Furthermore, they are often noisy and their number is limited by the finite size of the series. Therefore, one has to face the problem of using as much information as possible contained in a limited number of regressors (whose size has been fixed in advance). It is then obvious that any methodology able to extract the highest content of information in these vectors is desirable. An example of such methodology is to select an adequate delay value in order to maximize the information content in the regressors [8]. But this approach is limited to the case of two-dimensional regressors. More general methodologies suited for high-dimensional regressors are presented in this paper in Sections 3.2 and 4.

3. Unsupervised delay selection using MI

In this section the common unsupervised two-dimensional approach for single delay selection is first recalled. The methodology, first described in [8], is based on the MI between two variables. Then, the methodology is generalized in the framework of multiple delay selection in high-dimensional space. This methodological generalization is based on a k-nearest neighbours (k-NNs) based high-dimensional MI estimator proposed recently [13]. The MI estimator is briefly described and its use for unsupervised multiple delay selection is introduced.

3.1. Single delay selection using two-dimensional MI

The goal of delay selection using the MI criterion is to select variables $x(t-i*\tau)$, with $0 \leqslant i \leqslant p-1$, that are as independent as possible for the regressor reconstruction. The traditional two-dimensional approach consists in selecting a delay τ such that the MI between variable $x(t)$ and the delayed variable $x(t-\tau)$ is minimum. As several local minima may be found, the delay corresponding to the first local minimum is usually considered. This approach is similar to its linear counterpart where the delay τ is usually selected as the first minimum of the autocorrelation function. However, the delay selection using MI has been proved to be more efficient than using the autocorrelation function [8]. Indeed, the delay selected by MI is smaller than the one obtained using the autocorrelation. This last property leads to regressors that are more compact in time, providing more information content for a subsequent prediction. In the following of this paper the MI will be used as criterion to select the delay τ.

The MI between two variables X_1 and X_2 can be defined as [6]

$$\mathrm{MI}(X_1, X_2) = \int \mathrm{P}[X_1, X_2] \log\left(\frac{\mathrm{P}[X_1, X_2]}{\mathrm{P}[X_1]\mathrm{P}[X_2]}\right) \mathrm{d}X_1 \, \mathrm{d}X_2, \quad (3)$$

where P[.] denotes the probability. This criterion is a nonlinear measure of how much information on X_1 can be deduced from the knowledge of X_2 and vice versa. In the context of regressor reconstruction, variables X_1 and X_2 are $x(t)$ and $x(t-\tau)$, respectively. In practice, the MI is computed for several delays τ. The graph of the MI versus the delay is then plotted and the delay corresponding to the first minimum is selected. Fig. 1 shows the use of the MI on a time series obtained from the Lorenz equations [11,1], in the $p=2$ case. The left part shows the MI with respect to the lag τ. According to this plot, $\tau = 11$ is chosen. The reconstructed two-dimensional phase space is shown on the right of Fig. 1.

The main limitation of this approach is that the MI is computed between two variables $x(t)$ and $x(t-\tau)$ only. This fact leads to constrained regressors in Eq. (1); indeed, in the case of multi-dimensional regressors, variables $x(t-2*\tau)$, $x(t-3*\tau)$, …, etc. will be automatically selected, without care of the information content added by these variables to the information already contained in $x(t)$ and $x(t-\tau)$. Two more interesting paths could be followed: either to evaluate the information contained as a whole by all variables in (1), or to allow more freedom by removing the constraint that delays are multiple one from

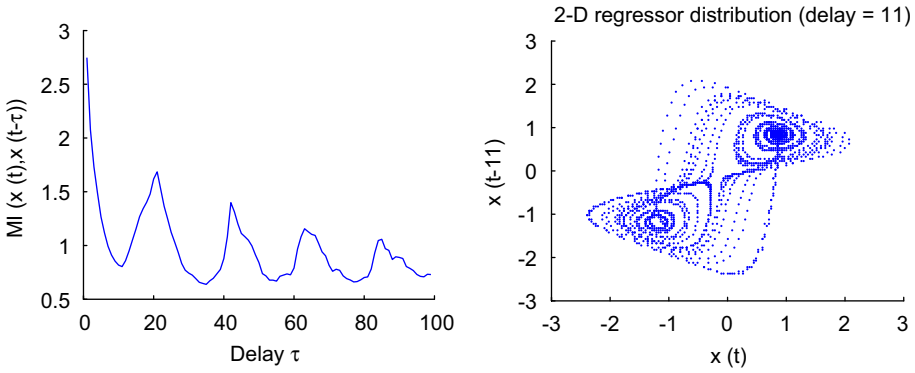

Fig. 1. Left: MI for the Lorenz time series. Right: reconstructed two-dimensional phase space (the selected lag is $\tau = 11$).

another

$$x_t = \{x(t-\tau_0), x(t-\tau_1), x(t-\tau_2), \ldots, x(t-\tau_{p-1})\}. \quad (4)$$

In both cases there is a need for a multi-dimensional criterion. The high-dimensional MI estimator based on k-NN, presented in Section 3.2, may be used for this purpose. An alternative is the DD criterion introduced in Section 4. After presenting these two criteria it is shown that the DD one is an advantageous alternative to the high-dimensional MI in a regressor reconstruction context.

3.2. Multiple delay selection using high-dimensional MI

Eq. (3) can easily be generalized to any dimension. One can indeed define the MI for m variables $X_1, \ldots, X_m$ as

$$\mathrm{MI}(X_1, \ldots, X_m) = \int \mathrm{P}[X_1, \ldots, X_m] \log\left(\frac{\mathrm{P}[X_1, \ldots, X_m]}{\mathrm{P}[X_1] \ldots \mathrm{P}[X_m]}\right) \mathrm{d}X_1, \ldots, \mathrm{d}X_m, \quad (5)$$

where variables $X_1, \ldots, X_m$ are, respectively, $x(t)$, …, $x(t-(p-1)*\tau)$, for example, in the context of regressor reconstruction.

High-dimensional MI, as defined in Eq. (5), is however difficult to estimate in practice; the unknown probability densities cannot be estimated accurately in a high-dimensional space. Estimating the MI in a high-dimensional space should thus be performed using another approach. Recently, Kraskov et al. have proposed a suitable estimator based on k-NN [13]. This estimator has been used in [21,20,10,16] for supervised feature selection: the group of features that maximizes the MI with the model desired output is selected. It has also been used in a preliminary study in [18] for the unsupervised selection of a single delay in high-dimensional regressors.

Shortly, the idea of the high-dimensional MI estimator is to rely on the fact that the MI is a measure of the discrepancy between the probability density of the joint variables and the product of the marginal densities (see Eq. (5)). Local densities around each vector (here regressors) can be estimated by counting the number of its neighbours up to a predefined distance or, conversely, by measuring the distance to its k-NN. Marginal densities can be estimated in the same way, provided that marginal variables are used instead of the joined ones. A way to estimate the MI is thus to estimate the joint density by a k-NN technique, and relate this measure to a similar k-NN estimation of the marginal densities. In the estimator [13], the relation is built by estimating the distance to the k-NN in the joint space, and counting the number of neighbours in the marginal spaces, up to a projection in these spaces of the previously estimated distance. If a relation exists between the variables, the numbers of neighbours estimated in this way will be related too. This reasoning is extended to the limit of small distances. Technical details about two variants of this principle may be found in [13], where it is explained that the two variants behave similarly except for different biases without consequence on the search for minima. The first version with a smaller bias in [13] is used in this paper. In practice, the most interesting property of the k-NN based MI estimator is that X and Y are not restricted to be scalar variables. As the estimator is based on distance measures variables X and Y may be multi-dimensional vectors. An implementation of this method is available online [2].

The high-dimensional MI estimator can be used as unsupervised criterion to select the delay(s). Generalizing the two-dimensional unsupervised methodology in [8], the delays for regressor reconstruction are selected such that they minimize the high-dimensional MI between the variables in (4). The variables that are selected are thus deemed to carry, as a whole, the highest possible information content for reconstructing the series in the phase space.

4. Regressor reconstruction using the distance-to-diagonal criterion

The theoretical complexity of the above detailed high-dimensional MI estimator is $O(N^2)$ where N is the number of data [13]. Indeed the operation with the highest computational cost is the computation of all distances in

order to find the nearest neighbours. Implementations more efficient than $O(N^2)$ may be used (see for example [13]); even in this case though, the estimation remains costly, especially in a feature selection context when many estimations of the MI for different sets of variables have to be performed. Furthermore, the MI estimator in [13] remains an estimator, loosing its accuracy when the dimension p of the regressors increases. A much simpler and faster criterion can be found to perform regressor reconstruction with comparable results but reduced computational cost. This criterion is DD.

As a reminder concerning the motivation of this work, the sampling frequency of a time series has a large impact on the information contained in the regressors. Indeed a too high sampling frequency leads to successive values that are highly dependent. This fact can be seen in Fig. 2 where two- and three-dimensional regressor distributions for a time series obtained from the Lorenz equations [11,1] are plotted. In both left and right figures it can be seen that the regressor distributions are limited to a portion of the space that is concentrated around its main diagonal. This problem is well known; delaying the values in the regressors by τ instead of 1 is a way to decrease the dependencies between variables in the regressor. As dependencies are decreased, it is reasonable to think that the information content in the regressor is increased (for a fixed regressor dimension). Delaying the variables in order to avoid the concentration of the regressors around the main diagonal is the geometrical concept at the basis of the DD heuristic detailed below.

Intuitively, maximizing the information content of a regressor should lead to a regressor that is far from the main diagonal of the phase space. Indeed, for regressors lying on or near the diagonal, it would mean that the regressor components are highly correlated, therefore that their global information content is low. Considering now all regressors in the phase space, their distribution should occupy the largest possible region and not be concentrated around the diagonal as illustrated in Fig. 2. In practice, measuring the part of the space filled by the regressor distribution can be performed by measuring how far each regressor is from the main diagonal of the phase space. The DD criterion is then the sum of all the distances between the regressors and the main diagonal.

More formally, for a time series S, the p-dimensional x_t regressors are constructed using either Eq. (1) or (4). Then, the sum of distances $\mathscr{D}(S)$ of the regressors to the main diagonal of the state space, i.e. to a line defined by the p-dimensional origin $(0,0,\ldots,0)$ and the unit vector $(1,1,\ldots,1)=\mathbf{1}^{\mathrm{T}}$, is given by:

$$\mathscr{D}(S)=\sum_{t=1}^{N-p+1}\|x_t\|^2-((x_t)^{\mathrm{T}}\cdot\mathbf{1})^2. \tag{6}$$

In this equation, each term simply measures the difference between the norm of x_t and the norm of the projection of x_t on the line passing through the unit vector $\mathbf{1}$.

The DD criterion is illustrated in Fig. 3 in the two-dimensional case. The left figure presents the $\mathscr{D}(S)$ values for the two-dimensional regressor distributions obtained using several delays τ. An adequate delay corresponds to a maximum in this graph. The value selected for the regressor reconstruction in the right figure is $\tau=31$. This value is the first local maximum of the DD criterion. Indeed, reaching the global maximum of the graph should not be considered as a unique goal. If two maxima are close, it is suggested to select the first one, corresponding to the lowest value of the delay. Too large delays could indeed lead to regressors that are useless e.g. when a further goal is to use the regressors in a time series prediction context: the advantage given by a small increase of the DD criterion could be largely balanced by the fact that values further away in time would be used for the prediction. Therefore, a good practice is to select local maxima or smaller lags in a plateau (for example a delay between $\tau=22$ and $\tau=30$ could have been selected).

The main advantage of the proposed DD criterion is its ability to select directly multiple delays for the regressor reconstruction. Furthermore, regressors may be constructed using either a single delay as in Eq. (1) or multiple delays as in Eq. (4).

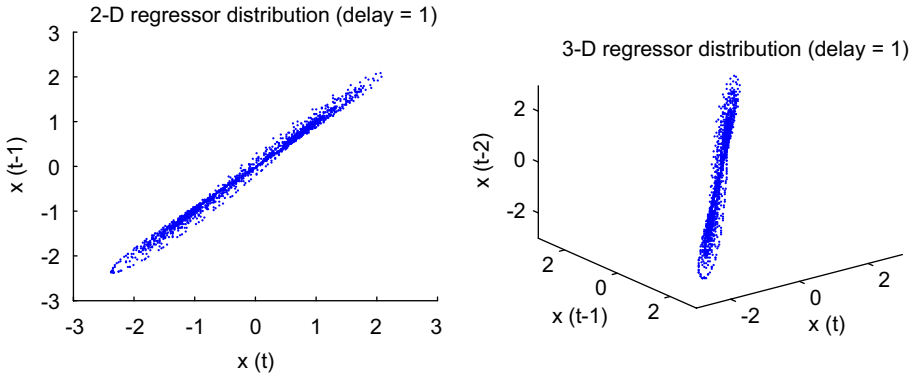

Fig. 2. Left: two-dimensional regressor distribution for the Lorenz time series. Right: three-dimensional regressor distribution for the Lorenz time series. In both cases, a delay $\tau=1$ has been used.

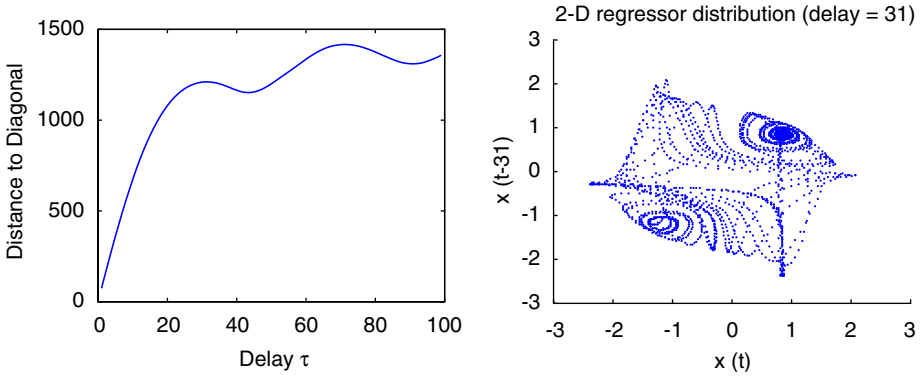

Fig. 3. Left: DD for the Lorenz time series. Right: reconstructed two-dimensional phase space (the selected lag is $\tau = 31$).

In comparison to the high-dimensional MI estimator, the DD criterion has a reduced computational complexity. Indeed the computation of the distance of a point to a line is of order $O(p)$ with p the space dimension. As this operation is repeated $N - p + 1$ times for the regressors obtained from a time series S of N values, the theoretical complexity is strictly bounded by $O(p * N)$. Furthermore, as p increases, there are more and more components in the regressors. As a consequence, there are more and more possible combinations of past values to construct regressors. All these combinations have to be considered, for the DD criterion as well as for any other one, in order to select the regressor with delay(s) that reconstruct adequately the time series structure. A reduced computation time of the criterion corresponding to a single combination is thus even more advantageous as the number of possible regressors increases dramatically while the regressor size grows large.

As a summary, the selection of optimal delays in time series regressors may be performed in two ways. First, the traditional two-dimensional MI criterion may be extended to high-dimensional regressors, thanks to recently published estimators. The computational complexity of this approach is in $O(N^2)$; even if this complexity may be reduced by efficient neighbour search algorithms, the computational load remains high; furthermore, the MI estimator may loose its accuracy when the dimension of the regressors increases. The other approach is the use of the DD criterion, a geometrical heuristic introduced in this paper, which has a reduced computation time in $O(p * N)$. As it will be illustrated in Section 5.3 on artificial and benchmark time series, the regressors reconstructed using the DD criterion are comparable to those obtained using high-dimensional MI.

5. Experimental results

In this section, both the high-dimensional MI and the DD criteria are used to select multiple delays for regressor reconstruction. In addition, to prove empirically the adequateness of the DD criterion in this context with respect to the MI approach, radial basis function networks (RBFN [4]) are used to forecast several times series. The aim is to show that, although the regressors obtained using the DD criterion are not exactly identical to the ones obtained using high-dimensional MI, the RBFN models built using the regressors selected by both criteria lead to comparable prediction errors. The regressors obtained using the DD criterion are therefore as useful for prediction as those obtained using the high-dimensional MI criterion in terms of information content, with the advantage of being obtained with a reduced computational cost.

5.1. The time series

In all experiments artificial and benchmark time series have been used. Here is a brief description of these series:

- *Autoregressive model*, AR(3):

$$x(t) = a_1 * x(t-1) + a_2 * x(t-2) + a_3 * x(t-3) + \varepsilon(t). \quad (7)$$

  2500 data have been generated; the first 500 data have been discarded to remove any bias due to initial conditions. The values of the a_i coefficients are $a_1 = 0.302$, $a_2 = 0.898$ and $a_3 = 0.251$. ε_t has been generated using a normal distribution with zero mean and 0.1 variance.
- *Artificial*:

$$x(t) = \sin(x(t-1)) + 2 * x(t-4) - 4 * x(t-8) + \varepsilon(t). \quad (8)$$

  As above, ε_t has been generated using a normal distribution with zero mean and 0.1 variance.
- *Lorenz system* [11,1]:

$$\dot{x}(t) = 10(y(t) - x(t)),$$

$$\dot{y}(t) = 28x(t) - y(t) - x(t) * z(t),$$

$$\dot{z}(t) = x(t) * y(t) - \tfrac{8}{3} z(t). \quad (9)$$

  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 Santa Fe A time series [23] was proposed in the Santa Fe time series prediction and analysis competition in 1991. The data were collected from a Far-Infrared-Laser in a chaotic state. The data set proposed for the competition, with 1000 data, has been used here.

5.2. The RBFN as regression model

RBFN [4] are used here as a regression model in order to forecast the time series. Details concerning the learning of the RBFN model can be found in [15,3]. In this work, the centers of the Gaussian kernels have been determined by a vector quantization algorithm. The width of the kernels is the width of the clusters (obtained by the vector quantization algorithm) multiplied by a common Width Scaling Factor optimized on the learning set. The weights of the kernels are determined by solving a system of linear equation.

The model accuracy in one-step-ahead prediction is evaluated using the usual generalization error E_{gen} defined as

$$E_{\text{gen}} = \lim_{M->\infty} \frac{\sum_{i=1}^{M} (x(t+1) - \hat{x}(t+1))^2}{M}, \tag{10}$$

where M is the size of the test set.

In order to select the best model for a given time series, a model selection strategy has been used to optimize the number of Gaussian kernels used in the RBFN models. The model selection strategy used here is the Bootstrap [7]. This procedure gives an estimation of the generalization error E_{gen} using a data set of finite size M. In all the experiments below, the estimate of the generalization error E_{gen} is obtained as the mean of the generalization errors obtained on 100 bootstrap samples.

5.3. Experiments, results and comments

The section begins with the results obtained on the Santa Fe A time series. The experiments with this time series are detailed in depth. Such a study is aimed at discussing practical considerations about the use of the high-dimensional MI and the DD criteria. The Santa Fe A time series has been chosen as a first example due to the illustrative power of its characteristic structure.

The two-dimensional case is first illustrated. The left plot in Fig. 4 presents the MI between $x(t)$ and the delayed variable $x(t-\tau)$. The two-dimensional regressor distribution with $\tau = 2$ is also shown in Fig. 4. This approach corresponds to the two-dimensional methodology as in [8] where the first local minimum is selected as delay. The results obtained using the DD are presented in the two right parts of Fig. 4. Here, the selected delay corresponds to the first local maximum of the DD. It can be noticed that the selected delays are very close (respectively, $\tau = 2$ and $\tau = 3$) although the regressor distributions slightly differ.

For the three-dimensional case, the results are presented in Figs. 5 and 6. In Fig. 5, the three-dimensional MI is plotted for the various regressors. The regressors are ordered according to the variables they contain: the left figure presents the MI for regressors $(x(t), x(t-1), x(t-2))$; $(x(t), x(t-1), x(t-3))$; $(x(t), x(t-1), x(t-4))$; …; $(x(t), x(t-2), x(t-3))$; $(x(t), x(t-2), x(t-4))$; etc. All the regressors containing three delayed variables with delay $\tau \in [1, 50]$ have been tested. A zoom of the MI for the 200 first regressors is provided. The same has been done using the DD criterion in the two last plots of Fig. 5. For the MI criterion, the plot reflects the contents of the regressors: all regressors containing $x(t)$ and $x(t-1)$ are in the first peak; all regressors containing $x(t)$ and $x(t-2)$ are in the second one, and so one. In each peak, the first local minimum has been selected, leading to the regressors in Table 1. In the same table, the four regressors with the largest values of DD have also been reported. Many similarities can be observed in this table. The regressor distributions obtained using the best combination of delays, for each criterion, have been plotted in Fig. 6. For comparison purposes, the regressor distribution obtained using the methodology of [8] is also presented in the left part. It can be noticed that the regressor distributions are very similar. In the following, we only consider high-dimensional MI and DD criteria.

In higher dimensions, the figures obtained using the high-dimensional MI and DD criteria are comparable to the ones obtained in three-dimensions and presented in Fig. 6. The only difference is that there are many more

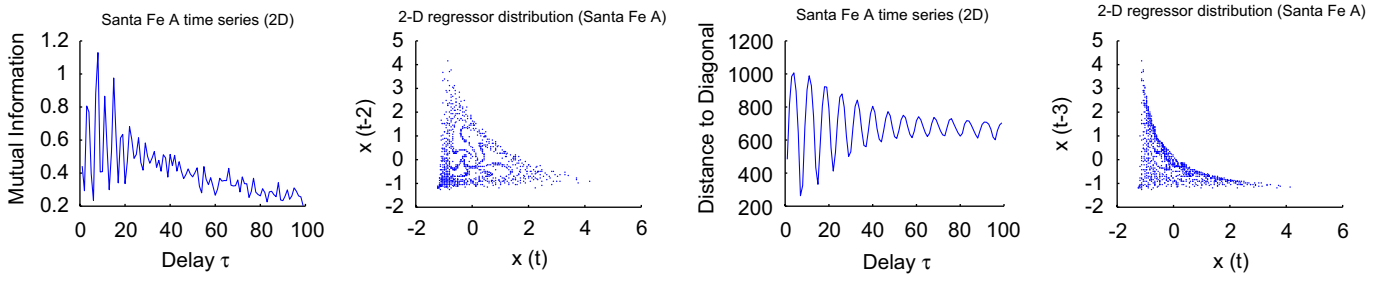

Fig. 4. From left to right: MI with respect to delay and two-dimensional regressor distribution with delay selected using MI ($\tau = 2$); DD with respect to the delay and two-dimensional regressor distribution with delay selected using DD ($\tau = 3$).

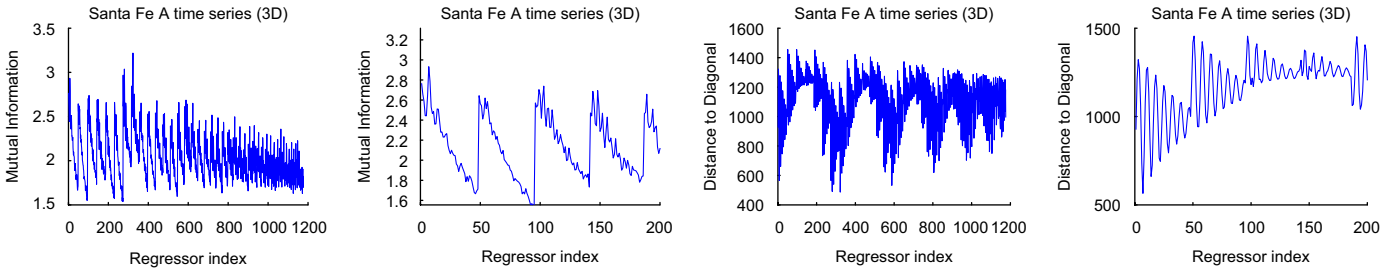

Fig. 5. From left to right: three-dimensional MI for various regressors and a zoom of the MI values for the first 200 regressors; three-dimensional DD for various regressors and a zoom of the DD values for the first 200 regressors. See text for details.

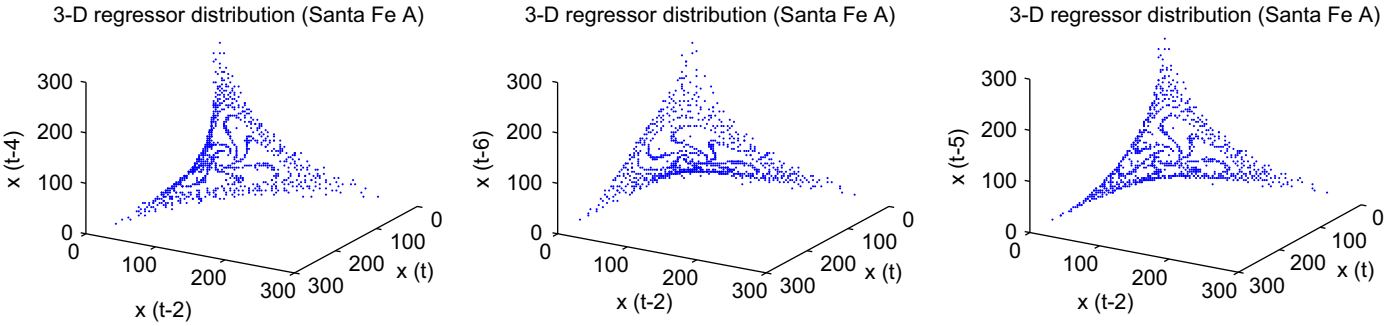

Fig. 6. Three-dimensional regressor distributions reconstructed using, from left to right, the two-dimensional MI with multiple of the selected delay, the high-dimensional MI and the DD criteria, respectively.

Table 1
Four best regressors obtained using the high-dimensional MI and the DD criteria

Pos.	High-dimensional MI	Distance-to-diagonal
1.	$(x(t), x(t-2), x(t-6))$	$(x(t), x(t-2), x(t-5))$
2.	$(x(t), x(t-3), x(t-9))$	$(x(t), x(t-3), x(t-5))$
3.	$(x(t), x(t-4), x(t-10))$	$(x(t), x(t-4), x(t-9))$
4.	$(x(t), x(t-1), x(t-5))$	$(x(t), x(t-1), x(t-4))$

possible combinations for the regressor contents, as it can be seen from Fig. 7 where high-dimensional MI and DD for a few thousands of regressor contents are provided. Analysing these figures in dimensions 4–6 lead to the selection of the optimal regressor contents according to each criterion. These regressors are given in Table 2. It can be observed from Table 2 that the variables chosen by the criteria for the regressor reconstruction are very similar. In this example, it can be noted that the regressors obtained using the DD criterion are always more compact in time than the ones obtained with the high-dimensional MI (the only exception is the two-dimensional case).

Using the best regressors as in Table 2 prediction experiments can be performed with a one-step-ahead goal. The RBFN model complexity, corresponding to the number of Gaussian kernels in the model, has been optimized using a fast-Bootstrap strategy. The plots of the estimations of the generalization error obtained using the best regressor for each criterion are provided in Fig. 8 with, from left to right, top to down, two- to six-dimensional regressors, respectively. Table 3 summarizes the results obtained for the best models in each experiment. Each line summarizes the model complexity and the corresponding estimate of the generalization error obtained with the Fast-Bootstrap. The model structures differ slightly, a fact obviously expected as the variables in the regressors are not exactly the same. However, in terms of generalization error, the high-dimensional MI and the DD criteria perform equivalently in one-step-ahead prediction.

Note that, as mentioned earlier, there are more and more possible combinations to construct regressors as the regressor size p increases. This fact can be seen from Figs. 4, 5 and 7. As an extensive search has to be performed through all possible regressor combinations, for all criterions, it is obvious that reasonable values of p have to be considered.

The same methodology has been applied to the three other time series. Tables 4–6 summarizes the best regressors obtained for each time series, using the various criteria. The results presented here are limited to dimensions two to four. For the three times series it can be noticed once again that the regressors reconstructed using both criteria are similar. The one-step-ahead prediction has also been performed for these series with the best regressors presented in Tables 4–6. As for the Santa Fe A time series, the estimates of the generalization error are summarized in Tables 7–9 for the AR(3), artificial and Lorenz time series, respectively. Note that, for the artificial time series, the

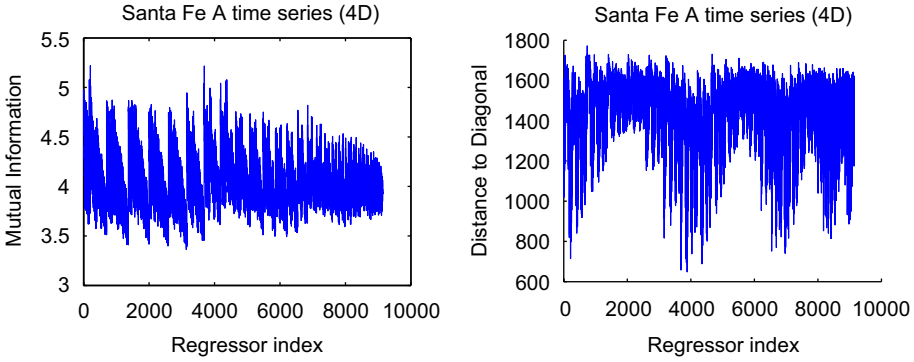

Fig. 7. Examples of the high-dimensional MI and the DD criteria obtained with four-dimensional regressors.

Table 2
Best regressors for each criterion in two- to six-dimensional space for the Santa Fe A time series

p	High-dimensional MI	Distance-to-diagonal
2	$(x(t), x(t-2))$	$(x(t), x(t-3))$
3	$(x(t), x(t-2), x(t-6))$	$(x(t), x(t-2), x(t-5))$
4	$(x(t), x(t-3), x(t-5), x(t-9))$	$(x(t), x(t-2), x(t-4), x(t-6))$
5	$(x(t), x(t-1), x(t-5), x(t-6), x(t-11))$	$(x(t), x(t-3), x(t-4), x(t-6), x(t-9))$
6	$(x(t), x(t-1), x(t-2), x(t-5), x(t-6), x(t-11))$	$(x(t), x(t-2), x(t-4), x(t-5), x(t-6), x(t-9))$

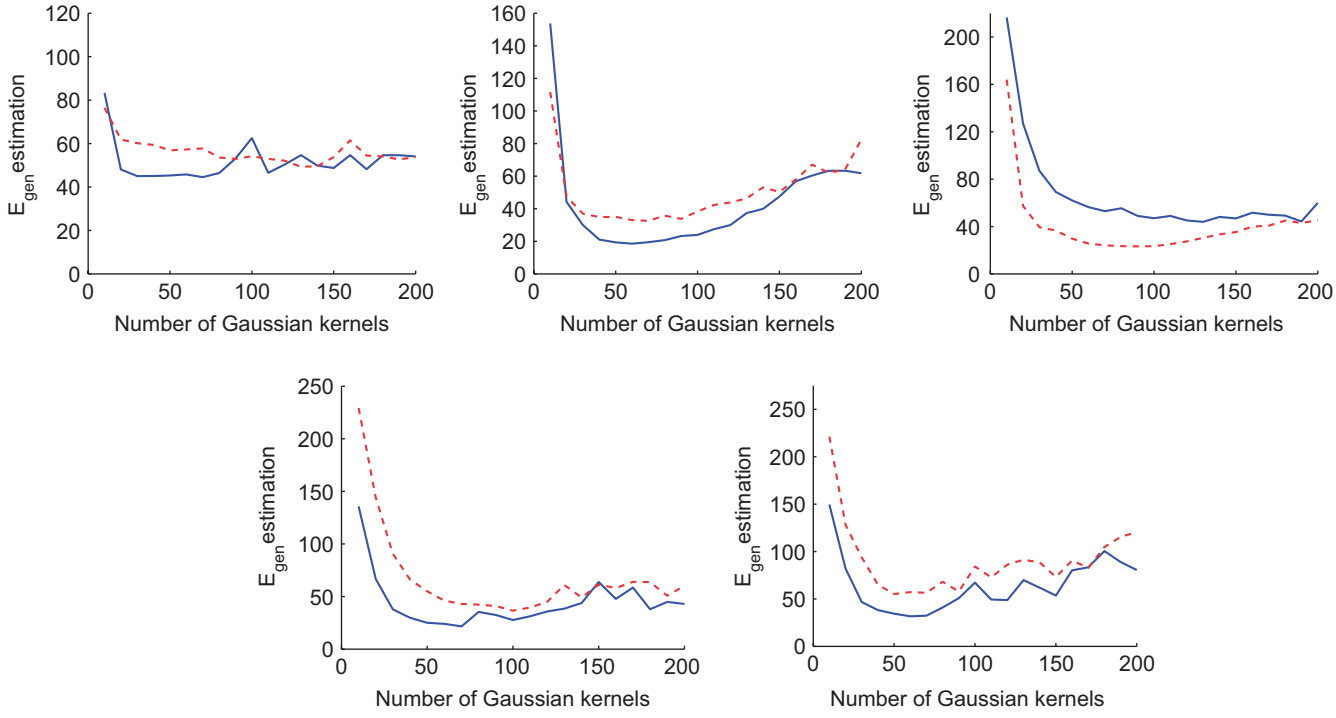

Fig. 8. Estimate of the generalization error using a bootstrap procedure for RBFN models with increasing number of kernels. From left to right, top to down: two- to six-dimensional regressors. Plain lines: results obtained with the best regressors selected using high-dimensional MI; dashed lines: results obtained with the best regressors selected using DD.

high-dimensional MI sometimes gives negative values. This fact does not penalize the search for a minimum as the estimator is known to be biased, as mentioned earlier in Section 3.2. From these last three tables, it can still be observed that the high-dimensional MI and the DD criteria perform equivalently: no criterion is always better than the

other; there is sometimes an advantage for one or the other approach depending on the used time series and the considered dimension.

From Fig. 8 and Tables 7–9, it can be observed that the high-dimensional MI and DD criteria lead to comparable performances. Indeed, it can be seen that both the generalization error estimates E_{gen} and the numbers of Gaussian kernels are roughly identical. Non-significant differences are observed though, due to the fact that the regressors obtained after applying the MI and DD criteria are not exactly identical. The same comments can be deduced from Tables 7–9.

6. Conclusions

In this paper, an extension of the usual approach of single delay selection for regressor reconstruction using mutual information (MI) is proposed. A high-dimensional mutual information estimator, based on the k-nearest neighbours (k-NN) algorithm, is used to select different delays in higher than two-dimensional spaces.

Furthermore the Distance-to-Diagonal (DD) criterion is introduced as an alternative to the high-dimensional mutual information. This alternative is based on a geometrical heuristics.

Table 3
Summary of the generalization error estimates for the best RBFN models in each dimension and for each criterion

p	High-dimensional MI		Distance-to-diagonal	
	Number of kernels	E_{gen}	Number of kernels	E_{gen}
2	70	44.537	130	49.292
3	60	18.545	70	32.590
4	130	43.912	90	23.251
5	70	21.652	100	36.483
6	60	31.732	50	55.160

See text for details.

Table 4
Best regressors for each criterion in two- to four-dimensional space for the AR(3) time series

p	High-dimensional MI	Distance-to-diagonal
2	$(x(t), x(t-3))$	$(x(t), x(t-1))$
3	$(x(t), x(t-3), x(t-8))$	$(x(t), x(t-5), x(t-8))$
4	$(x(t), x(t-1), x(t-4), x(t-9))$	$(x(t), x(t-1), x(t-4), x(t-9))$

Table 5
Best regressors for each criterion in two- to four-dimensional space for the artificial time series

p	High-dimensional MI	Distance-to-diagonal
2	$(x(t), x(t-5))$	$(x(t), x(t-3))$
3	$(x(t), x(t-1), x(t-8))$	$(x(t), x(t-2), x(t-4))$
4	$(x(t), x(t-1), x(t-8), x(t-10))$	$(x(t), x(t-2), x(t-3), x(t-5))$

Table 6
Best regressors for each criterion in two- to four-dimensional space for the Lorenz time series

Dimension	High-dimensional MI	Distance-to-diagonal
2	$(x(t), x(t-11))$	$(x(t), x(t-31))$
3	$(x(t), x(t-4), x(t-15))$	$(x(t), x(t-3), x(t-33))$
4	$(x(t), x(t-6), x(t-11), x(t-20))$	$(x(t), x(t-13), x(t-15), x(t-30))$

Table 7
Summary of the generalization error estimates for the best RBFN models in each dimension and for each criterion, AR(3) time series

Dimension	High-dimensional MI		Distance-to-diagonal	
	Number of kernels	E_{gen}	Number of kernels	E_{gen}
2	4	7.323	4	3.701
3	7	7.538	7	9.497
4	10	3.758	10	3.758

Table 8
Summary of the generalization error estimates for the best RBFN models in each dimension and for each criterion, artificial time series

Dimension	High-dimensional MI		Distance-to-diagonal	
	Number of kernels	E_{gen}	Number of kernels	E_{gen}
2	10	878.99	10	959.86
3	10	851.89	20	871.78
4	10	856.66	20	824.04

Table 9
Summary of the generalization error estimates for the best RBFN models in each dimension and for each criterion, Lorenz time series

Dimension	High-dimensional MI		Distance-to-diagonal	
	Number of kernels	E_{gen}	Number of kernels	E_{gen}
2	200	153.35	130	253.93
3	160	32.181	200	40.144
4	180	22.173	200	57.165

Both criteria have been used to select delays for regressor reconstruction. The methods have been compared in a one-step-ahead prediction context on various artificial and benchmark time series. It is shown that the high-dimensional mutual information and the DD approaches are comparable in the delays they select. They also lead to comparable performances in a one-step-ahead prediction context, with an advantageous to the DD criterion with respect to its lighter computational cost.

Further work includes a theoretical study of the DD criterion as well as the use of these high-dimensional mutual information and DD criteria in a multiple step-ahead prediction framework.

References

[1] H.D.I. Abarbanel, Analysis of Observed Chaotic Data, Springer, New York, 1997.

[2] S. Astakhov, P. Grassberger, A. Kraskov, H. Stögbauer, MILCA—Mutual Information Least-dependent Component Analysis, software implementation package available from ⟨http://www.klab.caltech.edu/~kraskov/MILCA/⟩.

[3] N. Benoudjit, M. Verleysen, On the kernel widths in radial-basis function networks, Neural Processing Letters, Kluwer Academic Publishers, Dordrecht, MA, vol. 18(2), 2003, pp. 139–154.

[4] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, Oxford, 1995.

[5] L. Cao, Practical method for determining the minimum embedding dimension of a scalar time series, Physica D 110 (1997) 43–50.

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

[7] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, London, 1993.

[8] A.M. Fraser, H.L. Swinney, Independent coordinates for strange attractors from mutual information, Phys. Rev. A 33 (1986) 1134–1140.

[9] P. Grassberger, I. Procaccia, Measuring the strangeness of strange attractors, Physica D 9 (1983) 189–208.

[10] Y. Ji, J. Hao, N. Reyhani, A. Lendasse, Direct and Recursive Prediction of Time Series Using Mutual Information Selection, in: J. Cabestany, A. Prieto, F. Sandoval (Eds.), Proceedings of International Workshop on Artificial Neural Networks, IWANN 2005, Barcelona, Spain, June 8–10, Computational Intelligence and Bioinspired Systems, Lecture Notes in Computer Science, vol. 3512, Springer, Berlin, pp. 1010–1017.

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

[12] M.B. Kennel, R. Brown, H.D.I. Abarbanel, Determining minimum embedding dimension using a geometrical construction, Phys. Rev. A 45 (1992) 3403–3411.

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

[14] J.M. Nichols, J.D. Nicholsb, Attractor reconstruction for non-linear systems: a methodological note, Math. Biosci. 171 (1) (2001) 21–32.

[15] M.J. Orr, Optimising the widths of radial basis functions, in: Proceedings of Fifth Brazilian Symposium on Neural Networks, Belo Horizonte, Brazil, December, 1998.

[16] N. Reyhani, J. Hao, Y. Ji, A. Lendasse, Mutual information and gamma test for input selection, in: Proceedings of European Symposium on Artificial Neural Networks, ESANN 2005, Bruges, Belgium, April 26–28, 2005, pp. 503–504.

[17] T. Sauer, J. Yorke, M. Casdagli, Embeddology, J. Statist. Phys. 65 (1991) 579–616.

[18] G. Simon, M. Verleysen, Lag selection for regression models using high-dimensional mutual information, in: Proceedings of European Symposium on Artificial Neural Networks, ESANN 2006, Bruges, Belgium, April 26–28, 2006, pp. 395–400.

[20] A. Sorjamaa, J. Hao, A. Lendasse, Mutual information and k-nearest neighbors approximator for time series predictions, in: W. Duch, J. Kacprzyk, E. Oja, S. Zadrozny (Eds.), Proceedings of International Conference on Artificial Neural Networks, ICANN'05, Warsaw, Poland, September 11–15, Artificial Neural Networks: Formal Models and Their Applications, Lecture Notes in Computer Science, vol. 3697, Springer, Berlin, 2005, pp. 553–558.

[21] A. Sorjamaa, N. Reyhani, A. Lendasse, Input and structure selection for k-NN approximator, in: J. Cabestany, A. Prieto, F. Sandoval (Eds.), Proceedings of International Workshop on Artificial Neural Networks, IWANN 2005, Barcelona, Spain, June 8–10, Computational Intelligence and Bioinspired Systems, Lecture Notes in Computer Science, vol. 3512, Springer, Berlin, 2005, pp. 985–991.

[22] F. Takens, Detecting strange attractors in turbulence, in: D.A. Rand, L.S. Young (Eds.), Dynamical Systems and Turbulence, Lecture Notes in Mathematics, vol. 898, Springer, Berlin, 1981.

[23] A. Weigend, N. Gershenfeld, Time Series Prediction: Forecasting the Future and Understanding the Past, Santa Fe Institute, MA, Addison-Wesley Publishing Company, New York, 1994.

Geoffroy Simon was born in 1978 in Dinant, Belgium. He received the M.Sc. degree in Computer Sciences in 2002 from the Facultés Universitaires Notre Dame de la Paix (Namur, Belgium). He is now working as Ph.D. student at the Microelectronic Laboratory of the Electrical Engineering Department of the Université catholique de Louvain (UCL). His research topics cover nonlinear time series analysis, artificial neural networks, nonlinear statistics and self-organization applied to time-series forecasting problems. He is currently member of the UCL Machine Learning Group. His work is funded by a grant from the Belgian F.R.I.A. (Fonds pour la formation à la Recherche dans l'Industrie et dans l'Agriculture).

Michel Verleysen was born in 1965 in Belgium. He received the M.S. and Ph.D. degrees in electrical engineering from the Université catholique de Louvain (Belgium) in 1987 and 1992, respectively. He was an Invited Professor at the Swiss E.P.F.L. (Ecole Polytechnique Fédérale de Lausanne, Switzerland) in 1992, at the Université d'Evry Val d'Essonne (France) in 2001, and at the Université Paris 1—Panthéon-Sorbonne in 2002–2004. He is now a Research Director of the Belgian F.N.R.S. (Fonds National de la Recherche Scientique) and Lecturer at the Université catholique de Louvain. He is editor-in-chief of the Neural Processing Letters journal, chairman of the annual ESANN conference (European Symposium on Artificial Neural Networks), associate editor of the IEEE Transactions on Neural Networks journal, and member of the editorial board and program committee of several journals and conferences on neural networks and learning. He is author or co-author of about 200 scientific papers in international journals and books or communications to conferences with reviewing committee. He is the co-author of the scientific popularization book on artificial neural networks in the series "Que Sais-Je?", in French. His research interests include machine learning, artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, adaptive signal processing, and high-dimensional data analysis.