Influence of Data Dimensionality on the Quality of Forecasts Given by a Multilayer Perceptron

Krzysztof Michalak Halina Kwasnicka

Wroclaw University of Technology
Institute of Applied Informatics

Abstract

One of the phenomena that can be observed when using neural networks for time series prediction is that the quality of forecasts is correlated to the dimensionality of data. Higher data dimensionality leads in most cases to higher prediction errors. This phenomenon is connected by some authors to the decrease of the variance of the distance between data points which occurs when the length of predicted vectors increases.

In this paper a proof is given that the variance of distance between data points also decreases with the correlation dimension of data. Therefore, the drop in forecast quality might be expected not only when the length of data vectors is increased but also when using vectors of the same length to represent data of increasing dimensionality. We also present some experimental results that illustrate the dependence between data dimensionality, variance of the distance between data points and the forecast error obtained when using a multilayer perceptron to predict future values of some time series.

Key words: correlation dimension, data dimensionality, neural networks, multilayer perceptron, time series prediction
PACS: 05.45.Tp, 07.05.Mh

1 Introduction

Neural networks are commonly used in non-linear modeling and forecasting. Neural models are often regarded as an alternative to linear regression models

Email addresses: michalak@zacisze.wroc.pl (Krzysztof Michalak), Halina.Kwasnicka@pwr.wroc.pl (Halina Kwasnicka).

(for example VAR methods) and frequency domain methods [7]. One of the issues that arise when using neural models for time series prediction is that the quality od the forecasts deteriorates as the data dimensionality increases. This might be an important issue because in many forecasting scenarios the dimensionality of data is high. In our previous work [4] we have shown experimental results that illustrate this phenomenon. In this paper some theoretical background is given.

This paper is structured as follows. Section 2 contains an overview of one of the possible approaches to neural forecasting of time series. Section 3 discusses data dimensionality issues. In Section 4 a proof is given that the variance of the distance between data points decreases with the correlation dimension of data. In Section 5 some experimental results are given. Section 6 concludes the paper.

2 Neural Networks in Time-Series Prediction

One of the most popular network architectures is a multilayer perceptron [1]. An example of this architecture is shown in Figure 1.

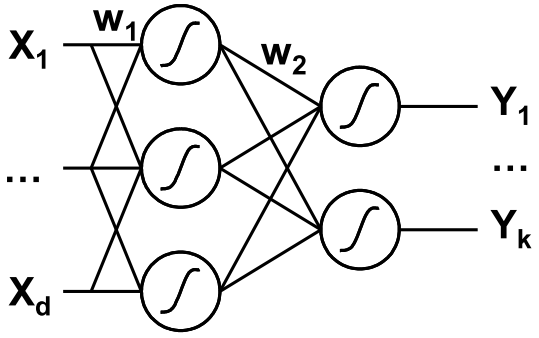

Fig. 1. Typical multilayer perceptron architecture

Such neural network may be used to model functions defined on R^d with values in R^k. To use a multilayer perceptron for time-series prediction it is necessary to prepare data so that it forms a set of vectors in R^d. Consider a time series $\{p_t\}_{t\in\mathrm{N}}$. From this time series a set of vectors

$$x_i = \{p_i, p_{i+1}, ..., p_{i+d-1}\} \subset \mathrm{R}^d \tag{1}$$

can be constructed using the sliding window technique.

The most common approach is to train the network using vectors $x_1, x_2, ..., x_{t_0-d+1}$ which contain all time series values up to the time instant t_0 and then by feed-

ing the network with vectors $x_{t_0-d+2}, x_{t_0-d+3}, ...$ to read predictions of the future values of the time series $p_{t_0+1}, p_{t_0+2}, ...$ from the network output. The number k of output neurons is in this application called the forecast horizon. The most straightforward approach is to set $k = 1$, that is to predict only one future value of the time series, but longer forecast horizons are also used.

The quality of the forecasts given by a multilayer perceptron can be measured using the mean squared error (MSE). For a number m of predictions made by the network the value of the MSE can be calculated as follows:

$$MSE = \frac{\sum_{i=1}^{m} \sum_{j=1}^{k} (Y_{ij} - T_{ij})^2}{mk} \ , \tag{2}$$

where Y and T are predicted and actual values respectively. Obviously, the lower the MSE is the better the predictions are.

3 Data Dimensionality

Having transformed the time series to a set of vectors one can measure the dimensionality of data. The most straightforward approach would be to use the embedding dimension d. However, the embedding dimension is chosen arbitrarily at the proprocessing stage and therefore does not provide any valuable information concerning the behaviour of the time series.

One of the invariants that measure data dimensionality taking into account actual behaviour of the time series is the correlation dimension [3]. To calculate the value of the correlation dimension for a set of data points x_i let's first define a set of all distances between data points:

$$\Delta = \{|x_i - x_j| : i < j\} \ . \tag{3}$$

The correlation dimension is defined using the correlation integral:

$$C(r) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(r - \delta_i) \ . \tag{4}$$

where $\delta_i \in \Delta$, $n = |\Delta|$ is the number of distances and $H(z) = 1$ for positive z, and 0 otherwise. The correlation dimension D_{corr} is then defined as:

$$D_{\text{corr}} = \lim_{r\to 0} \frac{log(C(r))}{log(r)} \ . \tag{5}$$

In numerical computations, especially when the Grassberger-Procaccia algorithm [3] is used, the correlation dimension is calculated by computing the value of the correlation dimension for various values of r and by approximating the value of D_{corr} from the equation:

$$D_{\text{corr}} = \frac{log(\sum_{i=1}^{n} H(r - \delta_i)) - c}{log(r)} \ . \tag{6}$$

using linear approximation. The value of the constant $c \in \mathrm{R}$ is also determined in this process.

It is proven that if the dimensionality of data increases in terms of the embedding dimension the variance of the distances between data points $S(\Delta)$ decreases. This in turn is connected by some authors to the deterioration of forecasts quality [2,8]. In Section 4 we show that such decrease in the variance also occurs when the correlation dimension of data increases. Therefore with the higher correlation dimension of data one can expect higher prediction MSE.

4 Correlation Dimension and the Variance of Distances Between Data Points

In this section we will prove that the variance of distances between data points $S(\Delta)$ decreases with the correlation dimension of data.

Lemma 1 *For every $n \in \mathrm{N}, n \geq 2$ the function:*

$$f_n(x) = n \sum_{i=1}^{n} i^{2x} - \left(\sum_{i=1}^{n} i^x \right)^2 \tag{7}$$

is increasing with $x \in (0, 1)$.

Proof

By induction on n.

For $n = 1$ we have:

$$f_1(x) = 1 \sum_{i=1}^{1} i^{2x} - \left(\sum_{i=1}^{1} i^x \right)^2 = 1^{2x} - (1^x)^2 = 0 \ . \tag{8}$$

We will show that if $f_n(x)$ is non-decreasing with $x \in (0,1)$ then $f_{n+1}(x)$ is increasing with $x \in (0,1)$.

$$f_{n+1}(x) = (n+1)\sum_{i=1}^{n+1} i^{2x} - \left(\sum_{i=1}^{n+1} i^x\right)^2 =$$

$$= n\sum_{i=1}^{n+1} i^{2x} + \sum_{i=1}^{n+1} i^{2x} - \left(\sum_{i=1}^{n} i^x + (n+1)^x\right)^2$$

$$= n\sum_{i=1}^{n} i^{2x} + n(n+1)^{2x} + \sum_{i=1}^{n+1} i^{2x} - \tag{9}$$

$$- \left(\left(\sum_{i=1}^{n} i^x\right)^2 + 2(n+1)^x \sum_{i=1}^{n} i^x + (n+1)^{2x}\right) =$$

$$= f_n(x) + n(n+1)^{2x} + \sum_{i=1}^{n+1} i^{2x} - 2(n+1)^x \sum_{i=1}^{n} i^x - (n+1)^{2x}$$

Therefore, to show that f_{n+1} is increasing with $x \in (0,1)$ it suffices to show that:

$$g_n(x) = n(n+1)^{2x} + \sum_{i=1}^{n} i^{2x} - 2(n+1)^x \sum_{i=1}^{n} i^x \tag{10}$$

is increasing with $x \in (0,1)$.

Notice that:

$$g_n(x) = \sum_{i=1}^{n} (n+1)^{2x} + i^{2x} - 2(n+1)^x i^x = \tag{11}$$

$$= \sum_{i=1}^{n} \left((n+1)^x - i^x\right)^2 \ .$$

Considering that for $i \leq n$:

$$\left((n+1)^x - i^x\right)' = (n+1)^x \ln(n+1) - i^x \ln i > 0 \tag{12}$$

every term in (11) is increasing with $x \in (0,1)$. Therefore if $f_n(x)$ is non-decreasing with $x \in (0,1)$ then $f_{n+1}(x)$ is increasing with $x \in (0,1)$. This, by induction, proves Lemma 1.

Theorem 2 *Let X_1, X_2 be two sets of data points, Δ_1, Δ_2 sets of distaces between data points in X_1 and X_2 respectively. Assume that:*

$$1 < D_{corr}(X_1) < D_{corr}(X_2) \tag{13}$$

$$|\Delta_1| = |\Delta_2| = n \tag{14}$$

$$\delta_{j1} < \delta_{j2} < ... < \delta_{jn} \text{ for } j = 1, 2 \ \delta_{ji} \in \Delta_j \tag{15}$$

$$\delta_{11} = \delta_{21} \tag{16}$$

Then the variances of distances between data points satisfy:

$$S(\Delta_1) > S(\Delta_2) \ . \tag{17}$$

Proof

From (6) it follows that:

$$\sum_{i=1}^{n} H(r - \delta_{ji}) = e^c r^{D_{corr}(X_j)} \tag{18}$$

and from (15) by setting $r = \delta_{jk}$:

$$k = e^c \delta_{jk}^{D_{corr}(X_j)} \tag{19}$$

$$k + 1 = e^c \delta_{jk+1}^{D_{corr}(X_j)} \ . \tag{20}$$

Therefore:

$$\frac{\delta_{jk+1}}{\delta_{jk}} = \left(\frac{k+1}{k}\right)^{1/D_{corr}(X_j)} \tag{21}$$

$$\delta_{jk} = \delta_{1k} k^{1/D_{corr}(X_j)} \ . \tag{22}$$

The variance $S(\Delta_j)$ is given by the equation:

$$S(\Delta_j) = E(\Delta_j^2) - E^2(\Delta_j) = \tag{23}$$

$$= \frac{\sum_{i=1}^{n} \delta_{ji}^2}{n} - \left(\frac{\sum_{i=1}^{n} \delta_{ji}}{n}\right)^2$$

Considering (22) we obtain:

$$S(\Delta_j) = \left(\frac{\delta_{j1}}{n}\right)^2 \left[n \sum_{i=1}^{n} i^{2/D_{corr}(X_j)} - \left(\sum_{i=1}^{n} i^{1/D_{corr}(X_j)}\right)^2\right] \ . \tag{24}$$

Considering (16) and $D_{corr}(X_j) \in (1, \infty)$ from Lemma 1 it follows that:

$$S(\Delta_1) > S(\Delta_2) \ . \tag{25}$$

From Theorem 2 and the considerations of Section 3 it follows that the error of predictions made by a neural network will be higher for higher correlation dimension D_{corr} of data. This effect is independent of the embedding dimension d used to create data vectors.

In practical computations the correlation dimension of data is usually greater than 1. For reasonably long time series the condition (15) is in practice satisfied by more than 95% of distances between data points. Even though the condition (16) is not necessarily satisfied we show that in experiments fluctuations of the minimal distance between data points do not influence the behaviour of neither the $S(\Delta)$ nor the prediction MSE. Therefore, the applicability of Theorem 2 is in practical applications not much constrained by the assumptions made.

5 Experiments

In this section we present the results of experiments performed on meteorological data. Source data containing average monthly land-surface temperature series was obtained from [9]. These data sets consist of time series of temperatures recorded at some fixed points on Earth surface. Recording stations are spaced evenly every 0.5 degree latitude and longitude. Each time series contains temperature values from years 1930-2000. Sets of measurements from stations placed at the same latitude were used for testing.

Table 1 summarizes the data sets used in the experiments. Each time series $\{p_t\}$ contains $N = 852$ points. Values in each series were normalized to [0, 1].

Test time series were embedded in R^d with d being the number of input neurons using the sliding window technique. The correlation dimension D_{corr} of data calculated using the Grassberger-Procaccia algorithm [3] fell within the range [2.0734, 3.3532]. The variance of distances between data points was calculated and for the sets tested it fell within the range [0.1197, 1.5769].

To make sure that the effects described do not depend of the size of the

Set name	Latitude	Longitude	Number of time series
Beaufort Sea	69.75	120.25 to 139.75	40
Canada	52.25	100.25 to 119.75	40
Victoria Island	69.75	100.25 to 119.75	40

Table 1
Summary of data sets

network or the activation function tests were performed using all possible sets containing the parameters of the network summarized in Table 2. The number d of input neurons was chosen so that it covers the value $2D_{corr}+1$ implied by the Takens' embedding theorem [6]. Apart from that the value of $d = 16$ was also tested to provide an embedding space that has many more dimensions that the minimal number suggested by the embedding theorem. Apart from *logistic* and *linear* activation functions *softmax* and *tanh* were also tested but learning process convergence in case of these functions was very poor.

Parameter	Values
Number of input neurons (d)	6, 7, 8 and 16
Number of hidden neurons (h)	20, 32, 40, 100
Number of output neurons (k)	1
Neuron activation function	logistic, linear

Table 2
Summary of network parameters used in experiments

The learning process was performed using all but the last 150 vectors of each input set. Initial weights of each perceptron were drawn from a zero-mean unit variance gaussian. Then, weights optimization using a scaled conjugate gradient algorithm [5] was performed. Optimization was stopped when a change of all of the weigths in a single optimization step was smaller than 10^{-15}, but no later than after 1000 iterations. For data used in experiments no overfitting occurs so no early stopping method was employed.

After the weight optimization was complete the last 150 input vectors which had not been used for network training were forwarded and the prediction MSE was calculated. For each time series the whole process starting with random weights initialization was performed 10 times and the mean value of prediction MSE was recorded. For all tested time series the value of the mean

prediction error fell within the range [0.000069, 0.013415].

In Figures 2 and 3 the dependence between the correlation dimension of data D_{corr} and the variance of distance between data points $S(\Delta)$ is presented. Points on these plots mark values measured for each of the 40 time series in the data set. Solid line represents linear approximation obtained using least squares fitting.

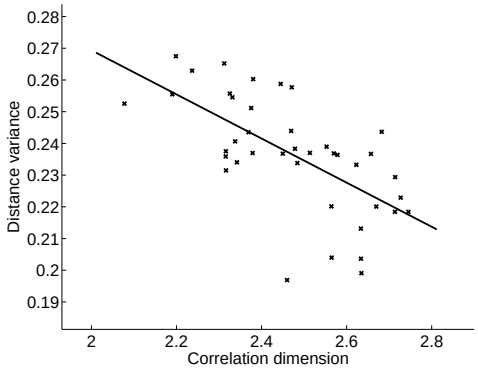

Fig. 2. Distance variance $S(\Delta)$ plotted against correlation dimension D_{corr} for Canada data set and embedding dimension $d = 8$

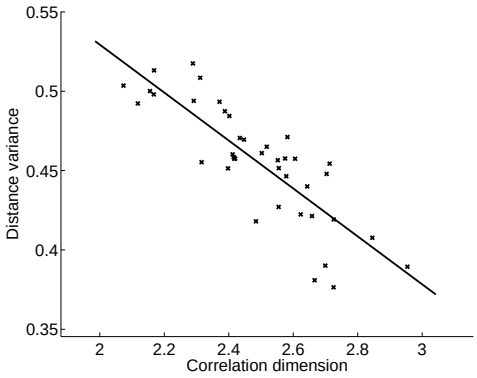

Fig. 3. Distance variance $S(\Delta)$ plotted against correlation dimension D_{corr} for Canada data set and embedding dimension $d = 16$

Values of slope of the approximation line for all data sets and all values of the embedding dimension are summarized in table 3.

Set name	$d = 6$	$d = 7$	$d = 8$	$d = 16$
Beaufort Sea	-0.016858	0.012158	-0.048584	-0.069922
Canada	-0.071730	-0.021860	-0.069565	-0.151065
Victoria Island	-0.084741	-0.056504	0.025281	-0.210228

Table 3
Slope of the approximation line for dependence between the variance of distance between data points $S(\Delta)$ and the correlation dimension of data D_{corr} for all datasets and all embedding dimensions

Only in two cases the slope of approximation line is positive. In all other cases the slope is negative indicating that the vairiance of distance between data points $S(\Delta)$ decreases with the correlation dimension D_{corr}.

In Figures 4 - 9 the dependence between prediction error and distance variance and dimensionality of data is presented. The figures present results for the Canada data set for three different sizes of the network. Also, two different

activation functions were used to obtain results presented in the figures. In all presented cases some outliers appear on the graphs but most of the sample follows the linear trend.

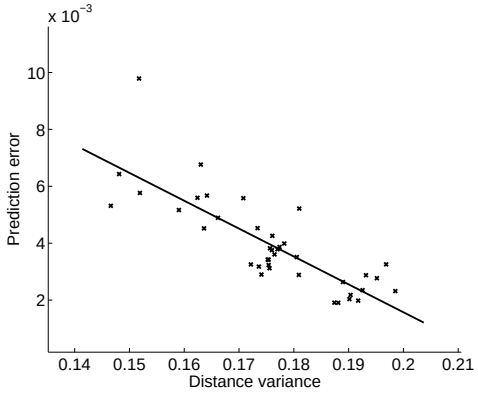

Fig. 4. Prediction error obtained using neural network with $h = 20$ hidden neurons and logistic activation function plotted against distance variance $S(\Delta)$ for Canada data set and embedding dimension $d = 6$

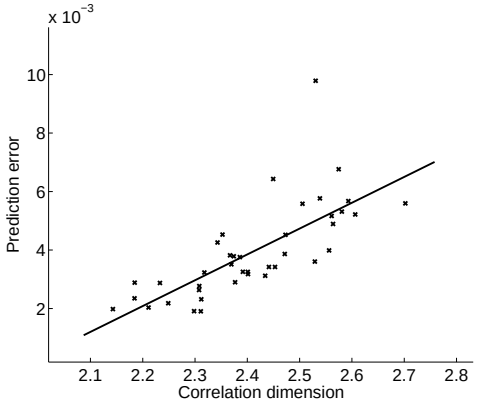

Fig. 5. Prediction error obtained using neural network with $h = 20$ hidden neurons and logistic activation function plotted against correlation dimension D_{corr} for Canada data set and embedding dimension $d = 6$

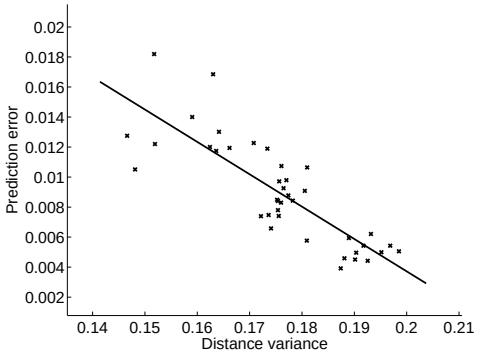

Fig. 6. Prediction error obtained using neural network with $h = 100$ hidden neurons and logistic activation function plotted against distance variance $S(\Delta)$ for Canada data set and embedding dimension $d = 6$

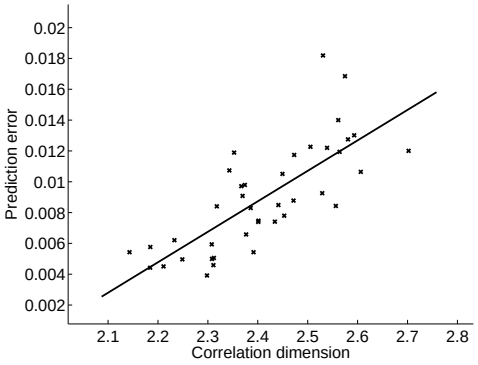

Fig. 7. Prediction error obtained using neural network with $h = 100$ hidden neurons and logistic activation function plotted against correlation dimension D_{corr} for Canada data set and embedding dimension $d = 6$

Results for all tested data sets and all network parameters are summarized in Tables 4 - 9. According to what has been shown in Section 4 when the variance of distances between data points $S(\Delta)$ increases the prediction error should decrease. This behaviour was in fact observed in the case of all experimental results. On the other hand as it was illustrated in Figures 2 and 3 the variance of distance between data points $S(\Delta)$ decreases with the correlation dimension

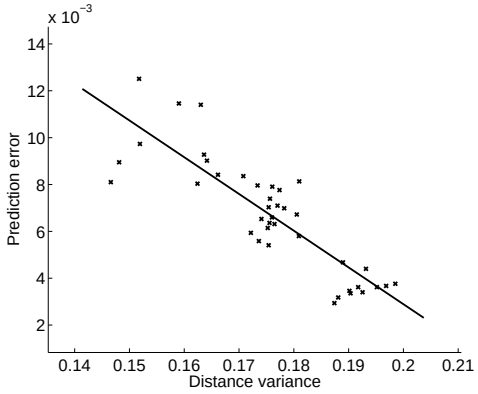

Fig. 8. Prediction error obtained using neural network with $h = 100$ hidden neurons and linear activation function plotted against distance variance $S(\Delta)$ for Canada data set and embedding dimension $d = 6$

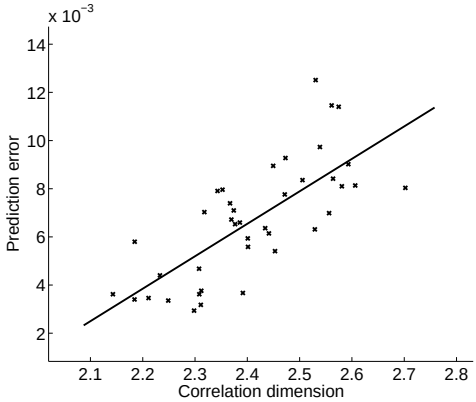

Fig. 9. Prediction error obtained using neural network with $h = 100$ hidden neurons and linear activation function plotted against correlation dimension D_{corr} for Canada data set and embedding dimension $d = 6$

of data D_{corr}. Therefore when the correlation dimension of data D_{corr} increases so does the prediction error.

The only cases when the prediction error decreases with the correlation dimension of data is when the variance of distance increases with data dimensionality. Such behaviour was observed for the Beaufort Sea data set for the embedding dimension $d = 7$ and for the Victoria Island data set for embedding dimension $d = 8$ and for the smallest neural network with $h = 20$ hidden neurons (see Tables 3, 5 and 9).

Network parameters	$d = 6$	$d = 7$	$d = 8$	$d = 16$
h = 20, logistic	-0.022685	-0.035179	-0.029861	-0.013479
h = 32, logistic	-0.030434	-0.025184	-0.032105	-0.024800
h = 40, logistic	-0.034045	-0.036522	-0.023291	-0.027377
h = 100, logistic	-0.097651	-0.110080	-0.109416	-0.011351
h = 100, linear	-0.010134	-0.014961	-0.025760	-0.018765

Table 4
Slope of the approximation lines for dependence between the prediction error and the variance of distance between data points $S(\Delta)$ for the Beaufort Sea dataset and all embedding dimensions

The experimental results presented in this section strongly support the hypothesis that the error obtained when using a multilayer perceptron to predict

Network parameters	$d = 6$	$d = 7$	$d = 8$	$d = 16$
h = 20, logistic	0.000807	-0.000554	0.001012	0.000962
h = 32, logistic	0.001124	-0.000143	0.001697	0.001912
h = 40, logistic	0.001190	-0.000412	0.000512	0.001867
h = 100, logistic	0.002038	-0.002169	0.004786	0.000652
h = 100, linear	0.000029	-0.000444	0.001753	0.001301

Table 5
Slope of the approximation lines for dependence between the prediction error and the correlation dimension of data D_{corr} for the Beaufort Sea dataset and all embedding dimensions

Network parameters	$d = 6$	$d = 7$	$d = 8$	$d = 16$
h = 20, logistic	-0.097986	-0.102614	-0.085330	-0.042275
h = 32, logistic	-0.123946	-0.099837	-0.115133	-0.049991
h = 40, logistic	-0.119613	-0.116101	-0.109517	-0.054916
h = 100, logistic	-0.215553	-0.116101	-0.181568	-0.018555
h = 100, linear	-0.156782	-0.146829	-0.151304	-0.068986

Table 6
Slope of the approximation lines for dependence between the prediction error and the variance of distance between data points $S(\Delta)$ for the Canada dataset and all embedding dimensions

Network parameters	$d = 6$	$d = 7$	$d = 8$	$d = 16$
h = 20, logistic	0.008829	0.002697	0.007537	0.007633
h = 32, logistic	0.011416	0.003034	0.008816	0.008963
h = 40, logistic	0.010714	0.002744	0.009556	0.009710
h = 100, logistic	0.019773	0.004464	0.016629	0.003091
h = 100, linear	0.013479	0.002453	0.013236	0.011302

Table 7
Slope of the approximation lines for dependence between the prediction error and the correlation dimension of data D_{corr} for the Canada dataset and all embedding dimensions

Network parameters	$d = 6$	$d = 7$	$d = 8$	$d = 16$
h = 20, logistic	-0.015829	-0.027500	-0.024749	-0.005833
h = 32, logistic	-0.035300	-0.036215	-0.031279	-0.011960
h = 40, logistic	-0.049242	-0.040283	-0.011250	-0.014746
h = 100, logistic	-0.094282	-0.071565	-0.069753	-0.011317
h = 100, linear	-0.106613	-0.067430	-0.060941	-0.006216

Table 8
Slope of the approximation lines for dependence between the prediction error and the variance of distance between data points $S(\Delta)$ for the Victoria Island dataset and all embedding dimensions

Network parameters	$d = 6$	$d = 7$	$d = 8$	$d = 16$
h = 20, logistic	0.001497	0.002501	-0.000045	0.002589
h = 32, logistic	0.002898	0.004230	0.002643	0.003090
h = 40, logistic	0.005606	0.005622	0.001429	0.004537
h = 100, logistic	0.009189	0.011919	0.010115	0.002878
h = 100, linear	0.009764	0.007608	0.004880	0.001909

Table 9
Slope of the approximation lines for dependence between the prediction error and the correlation dimension of data D_{corr} for the Victoria Island dataset and all embedding dimensions

values of a time series decreases with the variance of distance between data points. Implications of Theorem 2 are also confirmed by the results of the expetiments.

6 Conclusion

In this paper we have studied the influence of data dimensionality on the error obtained when predicting future values of some time series using a multilayer perceptron.

The two factors that we have taken into account are the variance of distances between data points $S(\Delta)$ and the correlation dimension D_{corr}. Both factors are defined on the embedding of the time series $\{p_t\}_{t\in\mathrm{N}}$ in some euclidean space R^d. We have theoretically shown that under certain conditions the variance

of distance between data points decreases with the correlation dimension of data. As it is hypothesized by some authors the decrease in the variance of distance between data points results in the increase of the prediction error.

Experimental results presented in this paper support these theoretical considerations. In most test cases the variance of distance between data points was observed to decrease with the correlation dimension of data. Consequently, for the same data sets the prediction error was observed to increase with the correlation dimension of data.

References

[1] Bishop Ch. M.: *Neural Networks for Pattern Recognition*. Oxford University Press (1995).

[2] Demartines, P.: *Analyse de donnees par reseaux de neurones auto-organises. Ph.D. dissertation (in French)*. Institut National Polytechnique de Grenoble - France (1994)

[3] Grassberger P., Procaccia I.: Characterization of strange attractors. Physical Review Letters, **Volume 50, Issue 5** (1983), 346–349

[4] Michalak K., Kwasnicka H.: *Correlation Dimension and the Quality of Forecasts Given by a Neural Network*. Lecture Notes in Computer Science, **Volume 3526** (2005), 332 – 341.

[5] Moller M. F.: *A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning*. Neural Networks, **Volume 6, Issue 4** (1993), 525 – 533.

[6] Takens F.: *Detecting Strange Attractors in Turbulence*. Proceedings of the Symposion on Dynamical Systems and Turbulence **(1983)**

[7] Tkacz G.: *Neural Network Forecasting of Canadian GDP Growth*. International Journal of Forecasting, **Volume 17** (2001), 57 – 69.

[8] Verleysen M., Francois D., Simon G., Wertz V.: *On the Effects of Dimensionality on Data Analysis with Neural Networks*. Lecture Notes in Computer Science, **Volume 2687** (2003), 105 – 112.

[9] Willmott, Matsuura and Collaborators' *Global Climate Resource Pages*. http://climate.geog.udel.edu/˜climate/