European Research Studies Journal
Volume XXV, Issue 2, 2022

pp. 587-596

Alternative Dimensional Reduction Methods on the Example of Data Preprocessing to Build a Ship Exhaust Model

Submitted 22/05/22, 1st revision 23/06/22, 2nd revision 15/07/22, accepted 30/07/22

Mariusz Dramski[1], Marcin Mąka[2]

Abstract:

Purpose: *Systems modeling is one of the basic research methods in scientific human activity. It should also be mentioned that the modeled systems are often multidimensional, which is an additional serious obstacle. Richard Bellman formulated the concept of the "dimensional curse" which says that as the dimensionality of a system increases, difficulties increase geometrically. In this article, the authors consider the problem of dimensionality reduction in order to build a model of the ship's exhaust emissions. It was observed that some data are incomplete or have little impact on the baseline variables.*
Design/methodology/approach: *Two methods of dimensionality reduction were applied (Pearson's linear correlation index and arc-angle index) and their suitability for this process was discussed.*
Findings: *Thanks to the methods used, it was possible to obtain information on the significance level of each of the model inputs. In addition, a lot depends on context, data availability, and much more. In any case, it is worth doing research in this direction.*
Practical implications: *We deal with modeling mainly in cases when we need to rely on measurement data, and the modeled system itself is unknown to us. The only thing the researcher has at his disposal is a certain set of measurement data, which very often lacks metadata. Nevertheless, even the basic information about input and output data allows you to create a model. However, the problem may be too much data. Although their storage itself is nothing difficult at present, the same sending them using means of communication (e.g. sending via the Internet) may already be troublesome.*
Originality value: *Typically, in the analysis of significance, methods that take into account the value of variance are used. Such a method is, for example, PCA (principal component analysis) (Sorzano 2014, Scholkopf 1997). The originality of the approach described in the article, however, consists in building a ranking of the significance of individual variables.*

Keywords: *Curse of dimensionality, exhaust emission, correlation, arc-angle index, dimensionality reduction, data dimensions, Pearson's correlation coefficient.*

JEL codes: *C20, C38, Q52.*

Paper type: *Research article.*

[1]*Maritime University of Szczecin, m.dramski@am.szczecin.pl;*
[2] *Maritime University of Szczecin, m.maka@am.szczecin.pl;*

1. Introduction

The examination of the significance of the input variables is aimed at determining whether the analyzed variable has an impact on the output variable. If it is absent (it is difficult to prove it on the basis of measurement data) or it is very low (easier to prove), then it may be considered whether or not to eliminate it from the system model building process. As a result, the dimensionality is reduced, which in turn may simplify the model. This problem is discussed, for example, in Piegat (2001) where the author even proposed a simple method of tracking the change in the value of the output variable from the value of the input variable. However, let us ask the question of how to define significance.

Let us assume that we are dealing with a multidimensional linear system in which the described variable has the form:

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n \tag{1}$$

In such a situation, the significance analysis is very simple. It is enough to read the values of the coefficients $a_1 \ldots a_n$. The higher the value of a given a_i, the higher the significance of the variable x_i will be. The a_0 coefficient can be ignored because there is no input variable next to it.

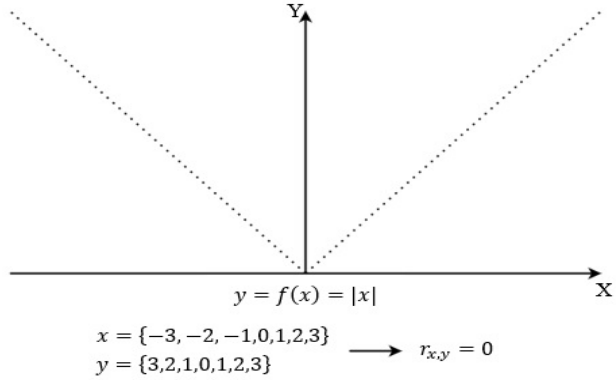
Of course, the described case is very simple. We rarely deal with a situation in which the modeled system is already described in an analytical form. In addition, there is always a risk that it is non-linear (as a rule, we do not have such information at the beginning). So, this simple method will not apply here.

2. Some Selected Methods of Dimensionality Reduction

The first method will be to use the Pearson's linear correlation coefficient (Pearson, 1895). It allows us to determine whether a given system is linear or not. With the values of this coefficient equal to 1 or -1 we deal with a linear system. Otherwise, it is a non-linear system, but in situations where the value of the correlation coefficient is close to 1, then such a system can be modeled using a linear model with little error. In Chimiak (2001) the author analyzed the usefulness of this coefficient in order to determine the significance of individual system inputs.

First of all, attention should be paid to the ease of use of the linear correlation coefficient. It is a tool known for many years and its interpretation is easy. In the same publication, the author presented a very simple experiment, which, however, shows some disadvantages of this approach. This is best illustrated in Figure 1.

Figure 1. *Linear correlation coefficient as an indicator of significance*

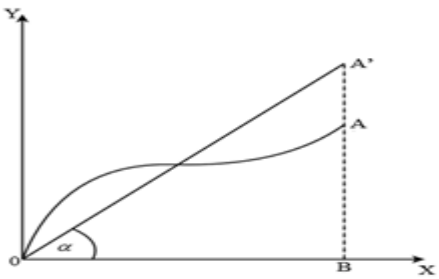


Source: *Chimiak (2001).*

We are dealing here with a simple linear function in the form f(*x*) = | *x* |. The dependence of the dependent variable on the explanatory variable is obvious. Nevertheless, with the given measurement data (described in the Figure), we obtained the value of the linear correlation coefficient $r_{x,y} = 0$. Of course, from the visual assessment of Figure 1, we can see that the explanatory variable is important. The more that it is the only variable describing the system. This value of the coefficient shows that the relationship is non-linear in this range of measurement data. Based on this experiment, according to Chimiak (2001), it can be concluded that the linear correlation coefficient may sometimes fail in the analysis of the significance of the input variables. However, in this article it was decided to use this method because such a situation as in Figure 1 is really rare and only happens in case of ideal experimental conditions (Chimiak-Opoka, 2001).

The second method is an arc-angle index proposed by Piegat (2001). To illustrate this method, let's assume that we have a given SISO (Single Input Single Output) system with a known analytical form and we can represent it using the OA curve as in Figure 2.

Figure 1. *Illustration of an arc-angle index*



Source: *Piegat (2001).*

At the outset, we will determine the length of the curve that represents our system. We express it in formula 2:

$$L = \int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx \tag{2}$$

Then, from the point 0, we draw a line 0A', the length of which is exactly the same as the length of the 0A curve. This line will form the angle α with the axis of the abscissa. The higher the value of this angle, the greater the significance of the given system input. Dividing $\frac{0B}{0A'}$, we get the cosine value of this angle. Taking into account the fact that $\cos 0 = 1$ and $\cos\frac{\pi}{2} = 0$, the formula for the significance index is as follows:

$$W(x) = 1 - \cos\alpha = 1 - \frac{\Delta x}{L} = 1 - \frac{b - a}{\int_a^b \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx} \tag{3}$$

Assuming that the data is preprocessed and normalized in the range (0 to 1), this formula will be simplified into the form:

$$W(x) = 1 - \cos\alpha = 1 - \frac{\Delta x}{L} = 1 - \frac{1}{\int_0^1 \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx} \tag{4}$$

Of course, the above formulas can be used when we are able to determine the dependence of the input x_I on the output y in the analytical form. Naturally, in the case of real data, where the goal is to build a model, this is not the case.

If we are dealing with real measurement data, then the length of the line 0A' is determined in a different way. The first stage (although not always necessary) is data normalization, preferably in the (0 to 1) range, especially when the domains of individual system inputs are very different from each other. Then we perform an analysis for each considered pair of inputs/outputs (you can also examine the relationship between the inputs or outputs themselves). We do this by dividing the measurement data set into n intervals, ensuring that each such interval has the same number of samples (as far as possible). The number of such intervals is usually determined experimentally.

A useful tool here can be the visualization of measurement data in the form of a graph. Too small number of ranges may make the estimation of significance too

inaccurate, while too large may lead to the fact that the final value of significance may be very much affected by e.g., a measurement error. Let us assume that we are dealing with a set of measurement data with the number of samples n = 100. Then the division into sub-intervals with the number of samples e.g., m = 2 or m = 50 is not correct.

For each resulting interval, we must then determine its center. We do it in a very simple way by calculating the arithmetic mean for the ordinate and abscissa variables. The points obtained in this way are then connected with a straight line. The first point is then connected with a straight line to the origin of the coordinate system, and the last point with the coordinates (1, y_n) where y_n is the y coordinate of the point being the center of the last interval.

As a result, we obtain a line whose beginning is determined by the beginning of the coordinate system and the end is determined by the point (1, y_n). We assume that the length of the broken line is equal to the length of the 0A' line (or the 0A curve) described above. Thus, the significance factor of the input x_i can be determined from the formula:

$$w(x_i) = 1 - \cos\alpha = 1 - \frac{1}{|0A'|} \tag{5}$$

3. Measurement Data

The input data (The Baltic Sea, the Pomeranian Bay area - about 25 000 records) has been received from two sources:

- automatic vessel identification system – AIS,
- LRS database containing technical data of ships (MARIT-IHS, 2021).

According to the SOLAS convention (SOLAS, 1974), AIS must be fitted to all seagoing ships above 300 GT gross tonnage and all passenger ships. The transmitted data streams, emitted at specified time intervals, contain data on the static and dynamic parameters of the ship, such as:

- identification numbers (MMSI, IMO),
- size (LOA, BOA, Draft),
- load capacity (DWT, GT),
- name,
- type,
- current position,
- course,
- speed.

For the output data, four quantitative parameters have been taken into account, taking into account emissions in four basic categories of exhaust components expressed in kilograms per hour [kg/h]:

- nitrogen oxides (NOx),
- sulfur oxides (SOx),
- carbon dioxide (CO2),
- particulate matter (PM).

The ranges of individual input and output variables are presented in the table:

Table 1. *Input and output data ranges*

L.p.	Name	Type	Minimum	Maximum	Unit
1	Type	In	n/a	n/a	n/a
2	DWT (deadweight tonnage)	In	86	37331	[t]
3	GT (gross tonnage)	In	123	22655	[t]
4	NT (net tonnage)	In	104	12073	[t]
5	Length	In	24	199.9	[m]
6	Width	In	7	23.78	[m]
7	Timestamp	In	-	-	[min]
8	Latitude	In	45.89699167	57.77976667	[deg]
9	Longitude	In	-7.780025	19.59539167	[deg]
10	SOG (speed over ground)	In	0.0	98.90000153	[kn]
11	COG (course over ground)	In	0.0	360.0	[deg]
12	NOx	Out	0.0	215.4630231	[kg/h]
13	SOx	Out	0.0	19.45352329	[kg/h]
14	PM	Out	0.0	3.584035714	[kg/h]
15	CO2	Out	0.0	10365.59668	[kg/h]

Source: *Own study.*

4. Experiments

So, let's analyze the significance of each input variable for each output variable. This was done using the two methods described above. Our system has 11 inputs and 4 outputs. In the case of a larger number of outputs, we can make a simple decomposition into several systems with one output. It does not matter much from the point of view of building the model. The problem, however, is the number of inputs. Assuming that we were dealing with binary information, then there may be as many as $2^{11} = 2048$ different unique input vectors at the input. Lowering the dimensionality by only one dimension causes the reduction of these vectors by 50%.

And we are talking only about binary values 0 and 1. In the case of real numbers (and there are such here) the number of unique vectors becomes almost impossible

to estimate. Thus, reducing the number of inputs will geometrically reduce the complexity of the model.

Table 2 shows the results of the linear correlation coefficients calculating the impact of each of the 11 inputs on the individual outputs. As is known, this coefficient takes values from -1 to 1. From the point of view of significance analysis, the sign "-" has no meaning. The absolute value of this coefficient may as well be taken as a result. Nevertheless, in Table 2 it was decided to leave negative values. Based on a short analysis of this table, we can conclude that the following parameters have the greatest impact on the values of individual output variables, GT, NT, Length, Width, Latitude, SOG and COG. The other inputs also have an impact, but possibly less.

As mentioned earlier, the linear correlation coefficient is not entirely suitable as a tool for significance analysis. Although we can reduce as many as 4 dimensions, only the quality of the model will answer the question of whether such a reduction was right. Usually, we construct the model in such a way that we first build a ranking of the input variables, taking into account their impact on individual outputs.

Then we can take two approaches, constructive and destructive. In the first case, we start with 1 or 2 of the most significant inputs and possibly add more dimensions later. The destructive approach is, of course, the opposite. The decision to add or subtract another dimension is made primarily after analyzing the credibility of the model. It should be remembered that the model error itself does not mean much because it may be acceptable for the training data, but very high for the test data (overfitting problem).

In the case of the linear correlation coefficient, it is also not possible to adopt a border value at which we can accept a given variable or reject it. Note that it is used to check whether the relationship is linear. The result may vary depending on the measurement data, even if you are considering the same system all the time.

Table 2. *Significance analysis using Pearson's correlation coefficient*

	NOx	**SOx**	**PM**	**CO2**
Type	-0.142	-0.085	-0.160	-0.148
DWT	0.301	0.088	0.298	0.290
GT	0.506	0.152	0.443	0.431
NT	0.436	0.135	0.398	0.387
Length	0.432	0.144	0.430	0.419
Width	0.546	0.166	0.555	0.543
Timestamp	-0.039	-0.073	-0.018	-0.016
Latitude	0.420	0.201	0.452	0.460
Longitude	-0.059	-0.030	-0.070	-0.071
SOG	0.616	0.294	0.624	0.641
COG	0.040	0.030	0.064	0.067

Source: *Own study.*

Thus, our system can take the form of a function:

$$f: \{GT, NT, Length, Width, Latitude, SOG, COG\} \rightarrow \{NOx, SOx, PM, CO2\} \quad (6)$$

We can treat the above equation as a starting point for building a model. It is up to the researcher whether he will adopt a constructive (recommended approach) or a destructive method. In the case of the constructive method, it may be possible to further reduce the dimensionality.

Now let's try to do exactly, the same experiment, but this time using the arc-angle index. The results are shown in Table 3. A split into 24 intervals was used. The normalization process was not performed as it is not necessary (described above). The results obtained turned out to be very interesting.

Table 3. *Significance analysis using the arc-angle index*

	NOx	**SOx**	**PM**	**CO2**
Type	0.649	0.043	0.015	0.991
DWT	0.000	0.000	0.000	0.026
GT	0.000	0.000	0.000	0.042
NT	0.000	0.000	0.000	0.155
Length	0.184	0.001	0.000	0.960
Width	0.782	0.055	0.005	0.995
Timestamp	0.007	0.000	0.000	0.626
Latitude	0.843	0.185	0.060	0.996
Longitude	0.752	0.086	0.032	0.995
SOG	0.628	0.011	0.005	0.991
COG	0.070	0.000	0.000	0.923

Source: *Own study.*

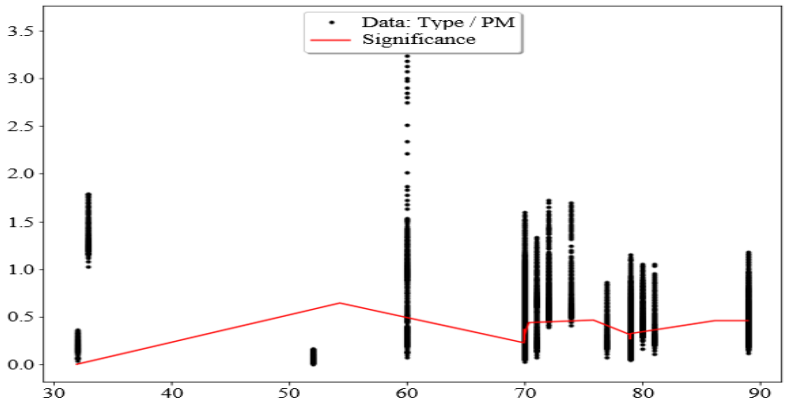
All results in Table 3 are rounded to three decimal places. In fact, none of the results had a value of 0. There is a very important thing to mention here. The arc-angle index does not uniquely evaluate the strength of the dependence of the output variable on the input variable. It also depends on many factors, such as the distribution of measurement data.

When using this method, it should also be remembered that it does not determine the linearity of a given relationship. The conclusion is as follows, the arc-angle index value itself is not significant. It is important that from the experiment we can build a ranking of the input fields that can be taken into account to build the model. Figure 3 shows an example of how the arc-angle index is computed. The greater the length of the broken line (marked in red), the higher the value of this index.

Let us assume, therefore, that we select for the model those input variables which for the NOx, SOx and PM outputs showed values greater than 0. All values of the significance of the input variables, in the context of CO2, showed high values. The very selection of the variables used to build the model does not have to be fully

compliant with the ranking. Moreover, consider that the output space is multidimensional.

Figure 3. *An example of calculating the arc-angle index*



Source: *Own study.*

For one dimension, you can just consider taking only values, e.g., greater than 0.500. It is also worth making a simple, subjective analysis of variables when the relationship seems obvious to us. Thus, we can conclude that the key variables for pollutant emissions are: Type, Length, Width, Latitude, Longitude and SOG. Therefore, we limited the number of dimensions of the input space from 11 to 6. In this way, we can obtain the function described by the equation:

$$f:\{Type, Length, Width, Latitude, Longitude, SOG\} \rightarrow \{NOx, SOx, PM, CO2\} \quad (7)$$

The question that should be asked is whether a further reduction is possible. We can clearly see that the values of the output variables can depend on the Length and Width inputs. It is customary to determine, first of all, the length of the ship. Therefore, it may turn out that to build the model it is enough to take into account only one of these variables. Thus, we only consider 5 input variables. A similar question can be asked for the variables Latitude and Longitude. Do geographic coordinates affect exhaust emissions? As we consider the limited data from a small area of the Baltic Sea, we can assume that their actual impact is negligible.

5. Conclusions

Richard Bellman's term "dimensional curse" (Bellman, 1957) perfectly describes the problems in modeling large-dimensional systems. This article describes two approaches to this problem, using Pearson's linear correlation coefficient and the arc-angle index. Both methods allowed the reduction of a certain number of input variables. Unfortunately, both of them also have some drawbacks. The definition of the linear correlation coefficient is unambiguous. Every researcher knows exactly what it is for. Nevertheless, we are constantly looking for new applications also for

known research methods. Arc-angle index, on the other hand, is a method that was designed specifically for significance analysis. Naturally, it will also be sensitive to the distribution of measurement data, the number of samples, etc. So how to judge if the method used is effective?

The best method of assessing whether the number of inputs has been reduced correctly is, of course, to evaluate the model. Earlier it was mentioned that a correct model should be characterized by the minimum error and the highest possible reliability. While model underfitting is very easy to detect, we sometimes have a problem with overfitting. Therefore, one should remember to divide the measurement data set into training data and test data in order to be able, for example, to use cross-validation techniques. The construction of the model is also a very interesting material for analysis.

However, due to the volume of the article, the authors decided to present this topic in a separate paper. It should also be mentioned that principal component analysis (PCA) is also used to reduce dimensionality. It was described, inter alia, in Krzanowski, (2005). However, this article focuses on the two methods previously mentioned.

References:

Biagi, L., Brovelli, M., Zamboni, G.A. 2011. Dtm Multi-Resolution Compressed Model For Efficient Data Storage and Network Transfer. Int. Arch. Photogram. Remote Sens. Spatial Inf. Sci., Xxxviii-4/W25, 7-13.

Bellman, R.E. 1957. Dynamic Programming. Princeton University Press P. Ix.

Chimiak, J. 2001. Analysis of Inputs Significance in Miso Systems. Part I. Comparative Analysis of Correlation Coefficient and New Arc-Angle Index. Annuals of Appliable Informatics. Technical University of Szczecin, 8, 37-62 (in Polish).

Chimiak-Opoka, J. 2001. New Method For Variables Selection in Miso Systems - Testing in Ideal Experimental Environment. 6th Session on Computer Science. Technical University of Szczecin, Faculty of Computer Science and Information Systems, 425-430 (in Polish).

Krzanowski, W. 2000. Principles of Multivariate Analysis: A User's Perspective. Oxford University Press.

Marit-His. 2021. Maritime.Ihs.Com.

Piegat, A. 2001. Fuzzy Modeling and Control. Springer-Verlag, Berlin/Heidelberg, Germany.

Pearson, K. 1895. Notes on regression and inheritance in the case of two parents. Proceedings of the Royal Society of London, 58, 240-242.

SOLAS. 1974. Documents Issued at the 1974 Conference www.imo.org.

Sorzano C.O.S., et al. 2014. A survey of dimensionality reduction techniques. Machine Learning (cs.LG); Quantitative Methods (q-bio.QM).

Scholkopf, B., et al. 1997. Kernel Principal Component Analysis Proc. of ICANN, 583-589.