Proceedings of the Federated Conference on
Computer Science and Information Systems pp. 27–34

ISBN 978-83-60810-51-4

Curvilinear component analysis in search of dimensionality and structure of some gearbox data

Anna Bartkowiak
University of Wrocław, Institute of Computer Science
and Wrocław High School of Applied Informatics
Wrocław, Poland
Email: aba@ii.uni.wroc.pl

Radosław Zimroz
Wrocław University of Technology
Vibro-Acoustic Science Laboratory
Wrocław, Poland
Email: Radoslaw.Zimroz@pwr.wroc.pl

***Abstract*—Our aim is to explore some gearbox vibration data characterized by 15 power spectra amplitudes. The data were gathered for two gearboxes: one being in a good state (set B) and the the other in a bad state (set A). In turn, each of the sets was gathered when the machine was operated under small/no load, and under full load. This gives 4 data sets to compare. We are concerned with two topics: 1. Is it possible to compare in a simple way the structure of the four obtained subsets? 2. Could the number of variables be reduced without losing essential information content of the data? To answer both these questions, we use a visual tool, the CCA (Curvilinear Component Analysis) method proposed by Demartines and Herault. Using this tool, we are able to answer positively the above two questions. We use the CCA algorithm in special layout, applying DxDy plots, with emphasis to estimating the overall intrinsic dimensionality of the four groups of data and capture visually the differences in their structure.**

I. Introduction, The aim of our research

IN OUR research we use vibration data gathered in the Vibro-Acoustic Science Laboratory, Wroclaw University of Technology. The goal is to build a mathematical model permitting on the base of the recorded data to make a monitoring of the state of the gearbox implanted into the machine. To do this, we need to know the structure of the data. There was some research how to do it, mainly when using linear methods [1], [17], [18]. By inspecting the data by pairwise scatterplots, we come to the conclusion that the dependency among the considered variables shows some tendency for nonlinearity [19]. To elucidate that observation, we turned to non-linear methods. There are many of them: see [16] for a short review of some of them. After considering some other nonlinear methods, we concentrated on the method named CCA (Curvilinear Component Analysis), developed by Demartines and Herault (D&H) [6], [9]. The method works on inter-point distances using specific cost function giving favor to inter-point proximities in the output space (small distances are dilated). When working locally (that is, with small neighborhoods), CCA permits to unfold the non-linear structures of the data yielding as output some flat manifolds of lower dimension. It allows also to make projections of the original data to the obtained manifold of lower dimension k. As shown by the authors, the method can be applied to data with quite a serious deviation from linearity, like folded submanifolds. The horseshoe or U-shaped distributions provide simple illustration of linear manifolds embedded in 3D space. In case of dimension k equal to 2 or 3 it is possible to visualize the data in the 2D or the 3D scatter plot.

We have chosen this method for several favorable properties:

α. It is able do deal with folded data and unfold in.
β. It is able to project data points to lower-dimensional flattened manifolds.
γ. It is generative, that is the mapping can be done only for representatives of the data set; after that is is possible to map the remaining data to the reduced space by a simple location algorithm.

Let us notice also, that the CCA method, as a distance based method, is not influenced by assumptions on the normality (Gauss) of distribution of the data; as a matter of fact, this method does not need assumptions on distribution of the data.

We consider data recorded by [1]. They recorded data for condition monitoring of two gearboxes: one being in a good and the other in a bad state. The obtained data are named B ('good' gearbox) and A ('bad' gearbox) appropriately. An earlier application of CCA [4] carried out using the data for the good state gearbox found that the set is composed of two different subsets of different dimensionality, thus can not be modelled by one common Gaussian distribution [4]. Now we continue these investigations on augmented data by taking into account also the bad state gearbox, i.e. by considering both data sets. Each of them was split into two subsets: NLOAD and LOADD (see Section 2 for more details).

Our aim is to obtain answers to two questions:

1) Is it possible to compare in a simple way the structure of the four obtained subsets?
2) Could the number of original variables ($d = 15$) be reduced to carry out further analysis with smaller number of variables?

We will show that the answer for these two questions is positive.

The paper is scheduled as follows: The above text constitutes Section 1 being introduction presenting the problems we want to solve. Section 2 describes shortly the data and their dimensionality evaluation yielded by classic PCA performed

* This paper was in part financially supported by Polish State Committee for Scientific research 2010-2013 as research project no. N504 147838.

both for groups B ('good') and A ('bad') and their splits into NLOAD and LOADD subgroups. Section 3 introduces the CCA method and shows DxDy plots (a modified version of the originally proposed in [6], [9] dydx plots). The possibility of estimating the intrinsic dimensionality of the data is emphasized. In Section 5 some summary of the results is presented.

II. Data used for the analysis and their intrinsic dimensionality estimated by PCA

There may be an ambiguity when talking about dimensionality of data. Generally, dimensionality in a common sense means the number of variables, that is, the number of columns in a data matrix (we denote it using the symbol d). However, it may happen that these variables are linearly dependent, which means that some of them may be expressed as linear function of the remaining ones. Thus, an essential question is: how many original (observed) variables are necessary to reproduce the entire set of data? In such a case we ask for true or intrinsic dimensionality of the data.

A. Data recorded from two gearboxes

We use part of the data gathered and analyzed by Bartelmus and Zimroz [1]. The data are given in the form of a matrix $\mathbf{X}$ of size $n \times d$, with n denoting the number of rows and d the number of columns of the data matrix $\mathbf{X}$. We have two such data matrices: one, set B, of size $n = 951 \times 15$ representing the machine in good state, and the other, set A, of size $n = 1232 \times 15$, representing the damaged machine. The 15 variables denote amplitudes of power spectra obtained from the Matlab PSD function. Each row of the matrices represents one data vector (instance, segment), it contains numerical values $d = 15$ variables, named $pp1, \ldots, pp15$ and used for further analysis; at the same time it may be viewed as a d-dimensional data point located in the d-dimensional Euclidean space R^d.

It was stated in [1], [3] that the analyzed data sets (B, good, and A, bad) may be split into 2 types of data vectors: those corresponding to time instances when the machine has worked under small or no load, and those working under normal load condition. These two types of points constitute two subgroups of the data referred up from now as the NLOAD and LOADD subgroups.

B. A preliminary dimensionality analysis by PCA

PCA (Principal Component Analysis) is one of the most frequently used methods for reduction of multivariate data [7], [12]. Using eigenvectors of the covariance matrix (or correlation matrix) of a data matrix $\mathbf{X}$ of size $n \times d$ one projects the data vectors (called also data points) to a lower dimensional subspace of dimension, say k. This yields k new variables, called principal components and denoted as $PC_1, ..., PC_k$. The derived PCs have a number of favorable properties [7], [12]. How to find the proper dimension k? This is usually done by inspecting the eigenvalues computed from the correlation matrix of the data. In [17], [18] it was concluded that taking the correlation matrices for the performed analysis, the proper dimension for the analyzed data is $k = 2$ or $k = 3$.

In Figure 1 we show the respective eigenvalues calculated for the sets B and A and their subsets NLOAD$_B$, LOADD$_B$, NLOAD$_A$ and LOADD$_A$ appropriately.

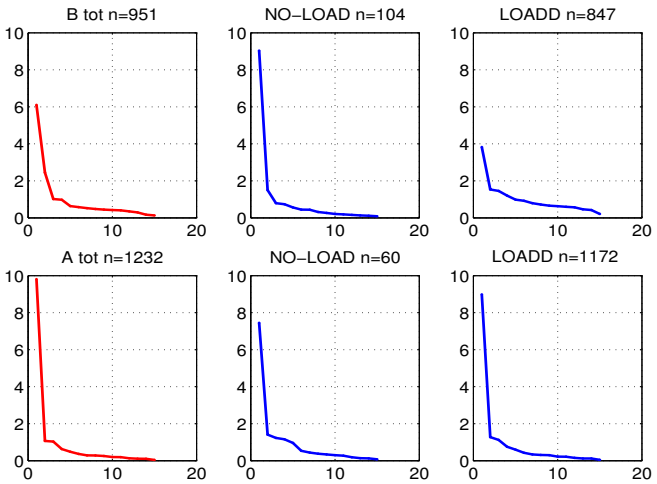


Fig. 1. Eigenvalues for the entire set B (top) and set A (bottom), with their subsets subsets NLOAD$_B$, LOADD$_B$, NLOAD$_A$ and LOADD$_A$ appropriately.

The results were obtained from correlation matrices of the respective groups, thus the sum of the diagonal elements is equal to 15, and this is the sum of the all 15 eigenvalues of each correlation matrix, and also the amount of variance (*inertia*) contained in the data. Looking at the eigenvalues displayed in Figure 1 one may notice a peculiar configuration. All the profiles are roughly similar. In every panel we observe one big dominant eigenvalue, and next a slow decay of the others. For example, for **set A** there is one big dominant eigenvalue explaining an 0.65, 0.50, and 0.60 part of the total variance in sets A, NLOAD$_A$ and LOADD$_A$. The contributions of the first 10 eigenvalues are:

λ	1	2	3	4	5	6	7	8	9	10
A	9.8	1.1	1.0	0.6	0.5	0.4	0.3	0.3	0.3	0.2
NLD	7.5	1.4	1.2	1.2	1.0	0.5	0.4	0.4	0.3	0.3
LDD	9.0	1.3	1.1	0.8	0.6	0.4	0.3	0.3	0.3	0.2

According to Kaiser's rule [10] the eigenvalues smaller than 1.0 may be considered as 'noise'. However, according to the principles used in factor analysis, Figure 1 may be considered as a *scree* graph, and to indicate the noise, we look for that dimension k_0, starting from which the decay of the eigenvalues exhibits a linear pattern. In our case this may be the value: $k_0 = 6$ or $k_0 = 7$. How good are the individual variables pp1, ..., pp15 reproduced by subsequent principal components? The reproduction formula reads [10]:

$$\mathbf{R} = \sum_{k=1}^{d} \lambda_k \mathbf{a}_k \mathbf{a}_k^T, \qquad (1)$$

with $\mathbf{R}$ denoting the analyzed correlation matrix, λ_k is the

kth eigenvalue of $\mathbf{R}$ (ordered in descending order), and $\mathbf{a}_k$ the corresponding eigenvector. Details of representation of the diagonal of $\mathbf{R}$ by $k = 2, 3, 6 and 10$ principal components are shown in Table I ($\mathbf{R}$ denotes here correlation matrix calculated from set A).

TABLE I
REPRODUCTION OF THE DIAGONAL OF CORRELATION MATRIX OF SET A BY $k = 2, 3, 6$ AND 10 PRINCIPAL COMPONENTS

var	k=2	k=3	k=6	k=10
1	0.90	0.90	0.92	0.94
2	0.64	0.90	0.96	1.00
3	0.54	0.75	0.88	0.99
4	0.77	0.82	0.86	0.93
5	0.90	0.92	0.93	0.96
6	0.41	0.54	0.98	1.00
7	0.88	0.88	0.89	0.92
8	0.83	0.83	0.85	0.94
9	0.71	0.75	0.91	0.98
10	0.79	0.82	0.85	0.93
11	0.70	0.74	0.83	1.00
12	0.78	0.85	0.85	0.92
13	0.81	0.82	0.89	0.92
14	0.62	0.77	0.82	0.99
15	0.60	0.61	0.97	1.00

Looking at the data shown in Table I one may state that generally, for $k = 3$ the reproduction ratio of the amplitudes of individual variables is equal or above 75%, except three variables: pp6 (only 54%), pp11 (only 74%) and pp15 (only 61%].

For $k = 6$ the reproduction ratio is equal or over 85%, except pp11 (only 83%) and pp14 (only 82%), which is not bad, especially when taking into account that the reproduction ratio for 6 other variables amounts over 90%. Possibly, some noise is here reproduced too.

Generally, PCA is not providing a clear indication what is the intrinsic dimension of the data.

III. CURVILINEAR COMPONENT ANALYSIS (CCA), BASIC CONCEPTS

The curvilinear analysis (CCA) was introduced by Demartines and Herault (D&H) [5], [6], see also [9], [14] for further applications of the method. The basic assumption underlying this method is that the multivariate data with d variables are located truly in a manifold of lower dimension, say p ($p < d$), moreover this subspace is somehow folded, which makes the relations between the observed variables – when viewed in R^p – to appear as non-linear ones. The problem to solve is formulated as follows:

We have N data vectors $\mathbf{x}_1, ..., \mathbf{x}_N$, each data vector, viewed as a data point, is located in d-dimensional data space R^d, that is $\mathbf{x}_i \in R^d$, $i = 1, \ldots, N$. This space is called the input space. Each data point has d components constituting observed values of the variables $X_1, \ldots, X_d$. The main idea is to find a mapping of the given N points $\{\mathbf{x}_i\}$, $i = 1, \ldots, N$, to a lower dimension subspace R^k ($k < d$) called the output space $\mathcal{Y}$. The obtained projections in R^k will be denoted as $\mathbf{y}_i$, $i = 1, \ldots, N$.

How to find a proper mapping? This can be done in many ways. The authors of CCA have chosen a distance-based method: For every pair of points $(\mathbf{x}_i, \mathbf{x}_j,\ i \neq j)$ belonging to the input space $\mathcal{X}$ take the inter-point distance X_{ij} in the input space and – basing on some criterion E expressing 'error' or 'cost'– find the corresponding points $(\mathbf{y}_i, \mathbf{y}_j,\ i \neq j)$ yielding in the output space $\mathcal{Y}$ the corresponding inter-point distance Y_{ij}. The distance between two points (i, j) may be defined in many ways, the simplest and most popular is the Euclidean distance. To find the proper mapping one needs to solve an optimization problem: namely to find values of the $\mathbf{y}_i$-s which minimize the assumed error function E. D&H [6], [5] have considered the following error function:

$$E = \frac{1}{2} \sum_i \sum_{j \neq i} (X_{ij} - Y_{ij})^2 F(Y_{ij}, \lambda_y)\,. \tag{2}$$

The function $F(Y_{ij}, \lambda_y)$ is chosen as a bounded and monotonically decreasing function, in order to favor local topology conservation (as happens in SOMs). In particular, $F(.)$ may be defined as $F(Y_{ij}) = 1$ for $Y_{ij} < \lambda$ and $F(Y_{ij}) = 0$, otherwise. This means that in every step, for given i, the algorithm updates only those values of $\mathbf{y}_j$ which are near to $\mathbf{y}_i$. It is also possible to define $F(Y_{ij}, \lambda_y)$ in a more soft way, e.g. by a power function, descending with the no. of the iteration.

D&H have called initially their proposed method as VQP neural network (Vector Quantization and Projection). The method was intended to provide an alternative to Kohonen SOM and not to be confined to a fixed *a priori* grid of neurons (in SOMs typically rectangular or hexagonal). Thus, the intention of D&H was to elaborate a topologically correct mapping resembling the true shape of the data cloud at hand ("instead of performing a vector quantization under the constraint of a predefined neighborhood between neurons, quantization and mapping functions are separately performed by two layers of connection" of the network [6]).

D&H's algorithm was designed to carry out a non linear dimension reduction and unfold high dimensional data structure towards its mean manifold. $F_\lambda(Y_{ij})$, for fixed λ, allows first for a global and then to a local unfolding. Two data transformations processes are identified: unfolding and projections. These two processes are visualized by so called *dydx* plots [6], [9], [8], which visualize the changes.

D&H [5], [6] proposed to visualize the changes in the Y_{ij}s as compared to the original input distances X_{ij}s in the form of a scatterplot called by them *dydx* plot. Let k denote the actual dimension of output space, to which the mapping is performed. Let dy denote any distance Y_{ij} evaluated in the actual output space $\mathcal{Y}$, of dimension k. Let dx denote any distance X_{ij} in the input space. Then dydx plot is the scatterplot composed of pairs (dy,dx) with dy taken as the 'independent' (abscissa) and dx as the 'dependent' (ordinate) variable. The authors have attached some definite meaning to the outlook of a dydx plot constructed for strongly nonlinear and folded data (see [5], [6], [8] for further details).

The most interesting aspect of the proposed CCA algorithm is to make a projection to the 2D (or 3D) Euclidean space,

which permitted – for $k = 2$ or $k = 3$ – to obtain a 2D (or 3D) representation of the data with approximate vision of their shape; a representation not confined to a fixed *a priori* grid of neurons. The problem of sufficiency (quality of the obtained representativeness) was not risen.

The concept of investigating the in intrinsic dimensionality appears in [9], p.632 and is described in the following words – referred in the following by us as **Herault's intrinsic dimension paradigm**:

When searching for the (unknown) intrinsic dimension of the input data, we choose the output dimension by dichotomy: if the distribution lies on the first diagonal, we can lower the output dimension, and if the distribution becomes thicker, the output dimension is too small.

IV. Analysis of the gearbox data

A. Our methodology

• Introducing DxDy plots

Our main goal to obtain an estimate of the intrinsic dimensionality of the data. D&H proposed to use for this the dydx plots (see [6], [9], [8] or Section 3 above). However, after several trials, we come to the conclusion, that more natural and interesting for us is to apply reverse plots, which we will call DxDy plots – speaking in full words: Distance-x Distance-y plots. They are constructed by taking Dx (equal to any X_{ij}) as the 'x' variate, and look how much this value was distorted by CCA projection yielding the correspondent Dy (equal to any Y_{ij}) in the output space $\mathcal{Y}$ and put on the DxDy plot as the 'y' coordinate. To our opinion, the introduced by us DxDy plot is appropriate, when the data are neither strongly curvilinear nor much folded, which is the case for our data (for which we found that apparent non linearity was caused by the non-homogeneity of the data, that is, of the sets B and A; after splitting the data into homogeneous subgroups, NLOAD and LOADD, the non linearity has disappeared).

In the following – when considering our two problems formulated in the introduction, we will consider only the DxDy plots, just introduced above.

• Design of our experiment

What concerns the practical aspect of applying the CCA algorithm, it needs evaluation of a distance matrix of size $N \times N$, and the size of this matrix may be very large. For example, for data with size $N \approx 1000$ the number of pairs is enormous (equal $1000 \times 1000 = 1,000,000$) and the calculations would be very long; moreover, such way on analysis would yield only one answer, without any repetition. To operate on repeated samples in a balanced design seemed to us to be a better way yielding more informative results.

Generally, both sets B and A are large: the cardinalities of the sets are:

$|B| = 951$, $|A| = 1232$.

After splitting the sets into the NLOAD and LOADD counterparts, the respective cardinalities are:

for set B: $|NLOAD_B| = 104$, $|LOADD_B| = 847$,

for set A: $|NLOAD_A| = 60$, $|LOADD_A| = 1172$.

One may notice that the cardinalities in the NLOAD and LOADD subsets are very unbalanced. We decided to perform the CCA in samples established in the following way:

For the set B the sample size for the NLOAD$_B$ subset equals 104, thus we taking the entire subgroup. However we repeat the analysis 3 times with different number of epochs $ep = 20, 50, 200$.

In the LOADD$_B$ subset we draw 3 samples, also of size 104, and repeat for each sample the CCA mapping, with fixed number of epochs $ep = 30$.

For the set A the layout is similar as for set B, except that the sample size is now equal to 60.

The calculations of CCA were performed using the MATLAB function CCA [13]. Each sample was standardized to have means equal 0 and variances equal 1. Principal components were used as starting point for the CCA function.

All calculations were performed twice: firstly using all 15 variables, and next using only variables 1–8.

• Graphs for inspecting

As explained above, we had four subsets to analyze: NLOAD$_B$, LOADD$_B$, LOADD$_A$. For each of them we have constructed and displayed 15 (=5*3) DxDy plots. They all were put together into a matrix with 5 rows and 3 columns constituting one panel of display. The rows of the matrix correspond to mappings to $k = 3, 4, 5, 6$ dimensional subspaces; the columns contain repetitions for the same k, however operated on new samples.

The DxDy plots for all the combinations described above are displayed in Figures 2 – 5 exhibiting

Figure 2: subset NLOAD$_B$, $d = 15$ (top) $d = 8$ (bottom)

Figure 3: subset LOADD$_B$, $d = 15$ (top) $d = 8$ (bottom)

Figure 4: subset NLOAD$_A$, $d = 15$ (top) $d = 8$ (bottom)

Figure 5: subset LOADD$_A$, $d = 15$ (top) $d = 8$ (bottom)

Each figure has two panels: the upper one corresponds to calculations with $d = 15$ and the bottom one - with $d = 8$ variables.

B. Results for sets B and A

We have noticed that the DxDy plots, depicted for our data, have an interesting property: the points (Dx,Dy) are spread around the diagonal $y = x$. With increasing k, the mass around the diagonal becomes thinner and tinner and finally, for some k_0 all data points are located exactly on the diagonal. This means that the input space $\mathcal{X}$ and the output space $\mathcal{Y}$ contain the same information about the spread of the data points. The value k_0 constitutes the upper bound for the intrinsic dimension of the elaborated data set. Considering k larger as k_0 does not change the display: when taking $k > k_0$ we see again and again the same diagonal with all the data points crowded over it. Referring to the Herault's intrinsic dimension paradigm, the rule for the upper bound of dimensionality of a data set might be formulated as follows: Take that k, for which all the points (Dx,Dx) are covered by the main diagonal of the respective DxDy plot.

Now let's look at the plots displaying the DxDy plots obtained from the CCA for the four investigated subsets.

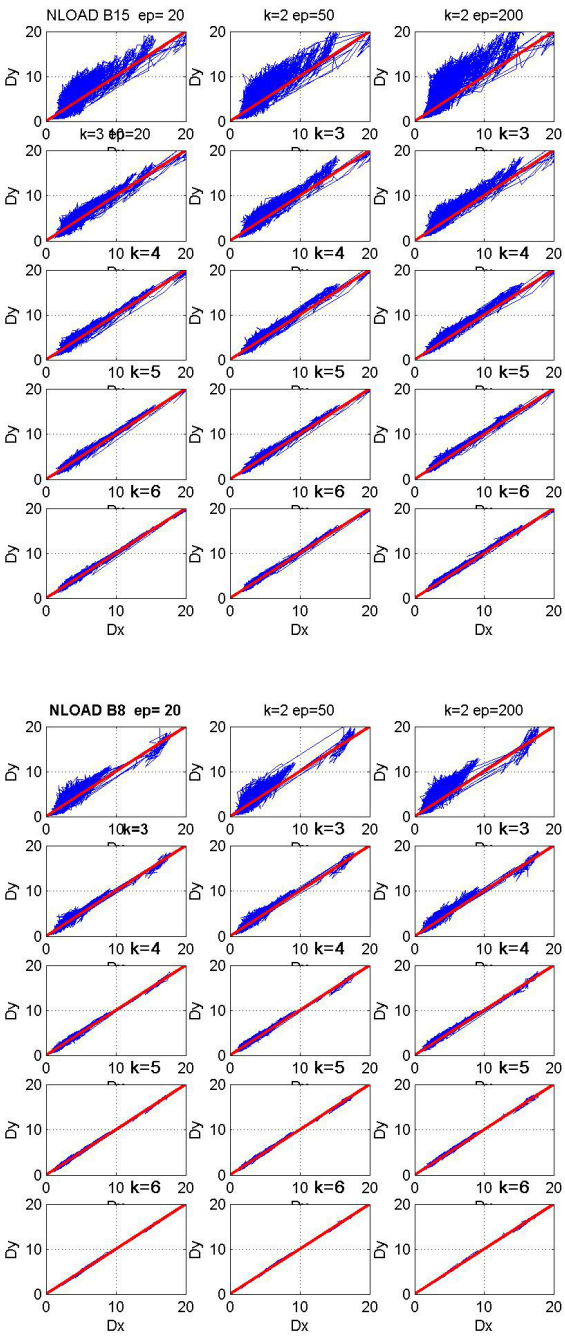


Fig. 2. Set B. DxDy plots for data subset NLOAD for $d = 15$ (top) and $d = 8$ (bottom). Each DxDy plot represents a sample of size $n = 104$ from subgroup NLOAD. Three columns correspond to three different numbers of epochs: 20, 50, 200. Rows: Results assuming output dimension $k = 2, 3, 4, 5, 6$.

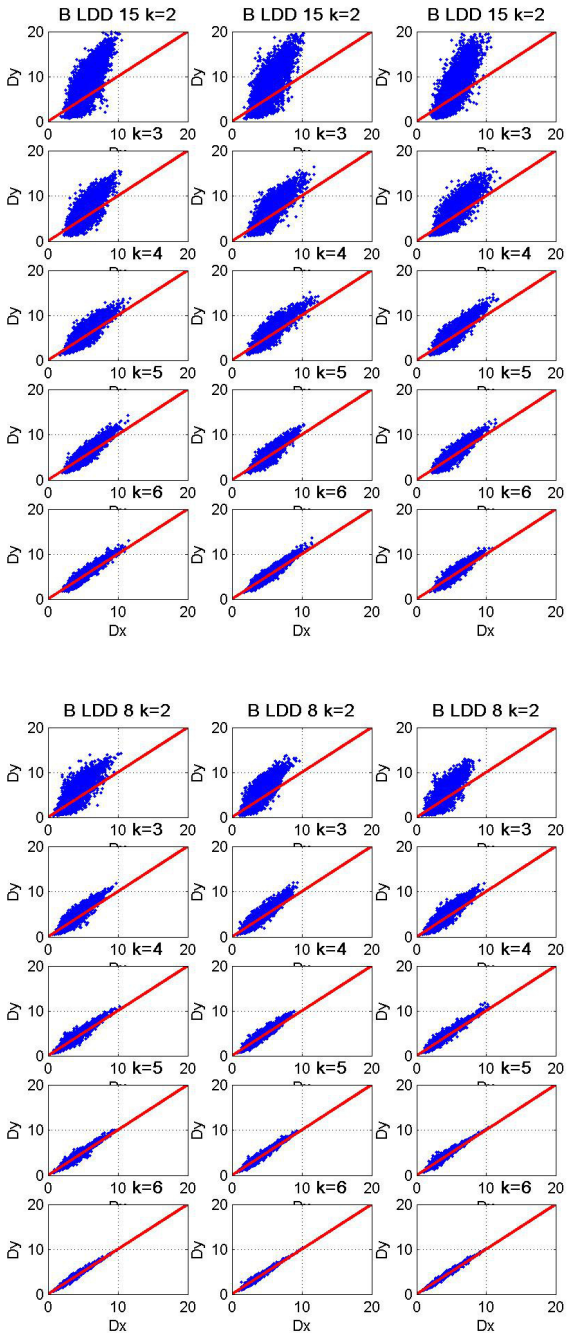
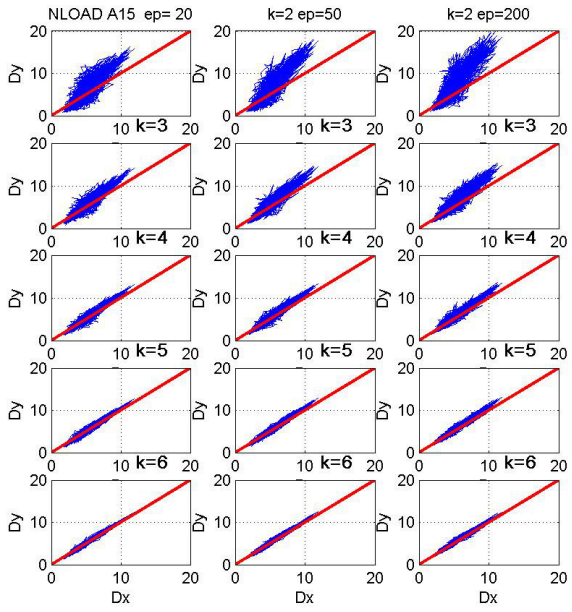


Fig. 3. Set B. DxDy plots for data subset LOADD for $d = 15$ (top) and $d = 8$ (bottom). Three columns correspond to three samples different in each row, containing $n = 104$ data vectors each. Rows: Results assuming output dimension $k = 2, 3, 4, 5, 6$.



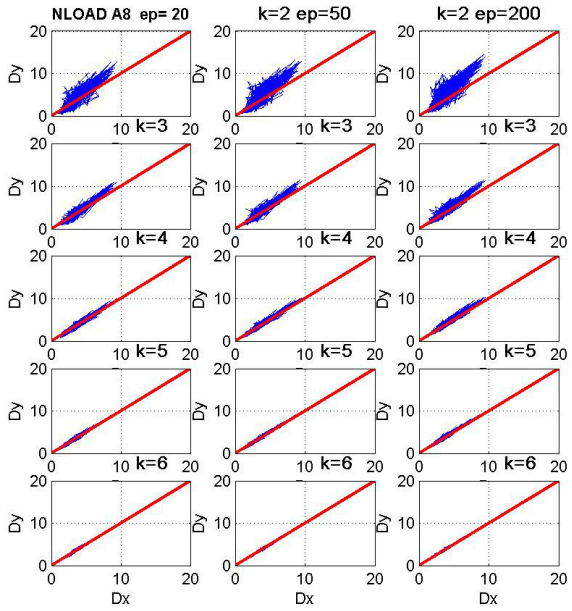
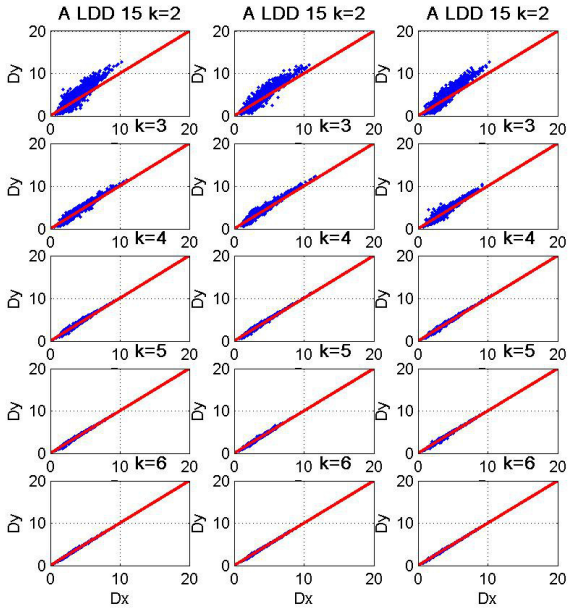


Fig. 4. Set A. DxDy plots for data subset NLOAD for $d = 15$ (top) and $d = 8$ (bottom). Samples of size $n = 60$ from subgroup NLOAD. Three columns correspond to three numbers of epochs: 20, 50, 200. Rows: Results assuming output dimension $k = 2, 3, 4, 5, 6$.



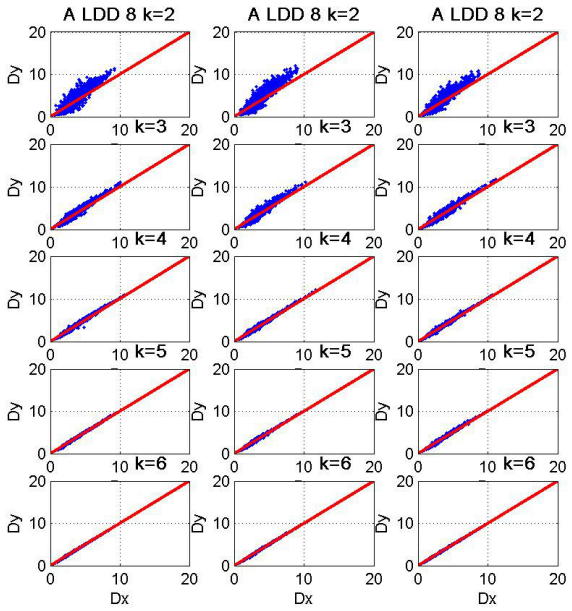


Fig. 5. Set A. DxDy plots for data subset LOADDD for d=15 (top) and d=8 (bottom). Samples of size n=60 from subgroup LOADDD. Three columns correspond to three different samples. Rows: Results assuming output dimension $k = 2, 3, 4, 5, 6$.

Looking at the plots 2 – 5, we are able to make a lot of interesting comparisons concerned the structure appearing in the pairs of subsets NLOAD$_B$, LOADD$_B$, NLOAD$_A$, LOADD$_A$, inspected horizontally and vertically.

Generally, one may notice that, with growing k denoting the dimension of the output space $\mathcal{Y}$, the spread of the (Dx,Dy) points about the main diagonal is becoming tinier and tinier; for $k = 6$ the (Dx,Dy) points coincide practically with the diagonal, that is with the straight line $y{\equiv}x$ of the display. This happens in all the eight displayed panels. Therefore we may infer that the output space for $k = 6$ and the distances there-in are practically the same as the inter-point distances in the input space $\mathcal{X}$ where the observed data are located.

How to fix the dimension k_0, that is, the value, which could be accepted as intrinsic dimension of the data? This should be the value, when the spread of the (Dx,Dy) points about the diagonal line $y{\equiv}x$ of the display is practically none. What does it mean? The Herault's paradigm is not precise in this aspect and does not give any strict mathematical criterion, when this can be assumed. We think, this could be formulated in terms of residual variance. We do not do it here and our following analysis is based on visual inspection of the graphs only.

Now let's try to establish the intrinsic dimension of the four analyzed subsets.

• Displays in Fig. 2 and Fig. 4 for subsets NLOAD$_B$ and NLOAD$_A$. One may state that the displays look rather similar. From visual inspection we judge that k_0 in upper panels (all variables) amounts k_0=4, and in bottom panels k_0=3. It appears that when taking 8 variables only, we have smaller spread of the (Dx,Dy) points in the bottom panel, which is reasonable, because we have here less data (columns 8-15 of the data were not accounted for).

• Displays in Fig. 3 and Fig. 5 for subsets LOADD$_B$ and LOADD$_A$. One may state that the displays have some common patterns, however their intrinsic dimensionality looks to be different. Generally, the spread of the (Dx,Dy) points around the corresponding diagonal is in Fig. 3 larger as in Fig. 5.

Considering upper panels, from all pp1–pp15. From visual inspection of the plots for set B we judge that k_0 is higher than 6 (case not shown in the upper panel of Fig. 3), and for set A it is only it is only k_0=4.

Considering bottom panels, from pp1–pp8 only. From visual inspection of the plots for set B we judge that k_0 is higher than 6 (case not shown in the upper panel of Fig. 3), and for set A it is only k_0=4.

Such results are reasonable, because it is known [17], [18] that the interdependence structure of the 15 variables pp1–pp15 differs considerably in sets B and A. Let us say also that the above statements (about the intrinsic dimensionality of the subsets) were obtained from inspecting 3 independent runs of the algorithm CCA carried out with independent samples. This strengthens the credibility of the obtained results.

The above comments give answer to our **first question** formulated in the Introduction. Indeed, we got a tool which in a simple explorative manner shows differences in intrinsic dimensionality among sets of data. In turn, differences in intrinsic dimensionality mean difference in interdependence structure among the variables characterizing the data.

To get answer to the **second question** (reduction of the number of variables), one should inspect the panels displayed in Figs 2–5 vertically. Without hesitation one may state that in Figs 1, 4 and 5 the top and bottom panels are very similar, which means that taking all 15 variables, or only 8 first variables, one obtains the same displays.

Summarizing the results: Looking at the plots 2-5 we find there answers to the questions formulated in the Introduction. It may be stated that

1.The subsets NLOAD and LOADD have different structure in set B and set A.

2. The difference in the structure may be noticed both when considering $d = 15$ and $d = 8$ variables.

3. It makes sense to reduce the number of variables taken for analysis. We got a clear indication that the number of variables may be reduced; in particular $d = 8$ (with variables pp1–pp8) may be as good as $d = 15$ with variables (pp1–pp15). Let's say, that the subset pp1–pp8 was our first choice, other subsets might be even better – this needs further exploration.

V. Discussion and closing remarks

We have investigated the algorithm CCA (Curvilinear component analysis) [5] as used for estimation of the upper bound of the dimensionality of a multivariate data set. This was in context of the NLOAD and LOADD subgroups obtained for the set B and set A of the gearbox data. After depicting so called DxDy plots it is possible to get a clear idea whether two groups of data are similar in their structure and have similar intrinsic dimensionality. There is a lot of statements that can be deduced from comparisons of scatterplots shown in Figures 2 – 5. For lack of space we have not discussed them here.

Our most important results are:

1. We were able to find an upper bound on the dimensionality of the analyzed data set.

2. It is reasonable to reduce - the so far considered 15 variables – to 8 variables only (meaning the first 8 out of 15 variables).

We did not try and do not have our opinion whether the CCA mapping (that is, the lower dimensionality coordinates yielded by this technique) is of great use in discriminant analysis. Some people have expected this and were disappointed [11], [15]. In our our opinion, for making discriminant analysis, one needs methods dedicated specifically to discriminant analysis. The CCA method is a general purpose method, and is not oriented towards discrimination and classification of data.

The novelty in our paper is:

(i) The CCA method was for the first time applied to an extended data set ('good' and 'bad' data, more than 2000 time segments) for gearbox condition monitoring.

(ii) We got a better (more convincing than for PCA) estimation of the intrinsic dimension of the analyzed data using CCA.

(iii) We have arrived to our conclusions in a non parametric way without making any assumptions about probability distributions of the data, in particular about normality (Gauss distribution).

There are some open problems:

1) Criterion for intrinsic dimensionality. The Herault's intrinsic dimension paradigm should be formalized in a strict way.
2) Reduction of the number of variables. We have carried out the analysis using 15 variables named pp1-pp15. We have shown that 8 variables make sense. A search for the best subset should be performed, taking into account other criteria. The goal could be to build a non linear discriminant function for recognizing items from the 'good' and the 'bad' state of the machine.

VI. Acknowledgements

The authors thank to two anonymous referees for their constructive comments.

References

[1] W. Bartelmus and R. Zimroz, "A new feature for monitoring the condition of gearboxes in nonstationary operating systems", *Mechanical Systems and Signal Processing* vol. 23 no. 5, 2009, pp. 1528–1534.

[2] A. Bartkowiak and R. Zimroz, "Outliers analysis and one class classification approach for planetary gearbox diagnosis", *Journal of Physics: Conference Series* 305 (1) , art. no. 012031 DOI: 10.1088/1742-6596/305/1/012031

[3] A. Bartkowiak and R. Zimroz, "Data dimension reduction and visualization with application to multi dimensional gearbox diagnostics data: comparison of several methods". *Diffusion and Defect Data Pt.B: Solid State Phenomena* 180, 2012, pp. 177-184 DOI: 10.4028/www.scientific.net/SSP.180.177

[4] A. Bartkowiak and R. Zimroz, "Curvilinear dimensionality reduction of data for gearbox condition monitoring". At *18th Int. Multiconference Advanced Computer Systes ACS* 30 May - 1 June 2012, Miedzyzdroje. Manuscript, 2012.

[5] P. Demartines and J. Herault, "Curvilinear component analysis: A self-organizing neural network for non-linear mapping of data sets". *IEEE Trans. on Neural Networks*, vol. 8 no. 1, 1997, pp. 148–154.

[6] P. Demartines, "Analyse de données par réseaux de neurones auto-organisés". *PhD thesis*. Institut National Polytechnique 1994.

[7] Q. He, R. Yan, F. Kong and R. Du, "Machine condition monitoring using Principal Component representations", *Mech. Systems and Signal Processing* vol. 23, no. 2, 2009, pp. 446-466.

[8] J. Herault, A. Guerin-Dugue, P. Villemain: "Searching for the embedded manifolds in high dimensional data, problems and unsolved questions". *ESANN'2002 proceedings, European Symposium on Artificial Neural networks*, Bruges (Belgium), d-side publi, pp. 173–184.

[9] J. Herault, C. Jausions-Picaud and A. Guerin-Dugue: "Curvilinear component analysis for high-dimensional data representation: I. Theoretical aspects and practical use in the presence of noise". In J. Mira and J.V. Sanchez (Eds), *Proceedings of IWANN'99*, vol. II, Springer, Alicante (Spain) June 1999, pp. 625-634.

[10] I.T. Jolliffe, *Principal Component Analysis*, 2nd Edition, Springer, New York, 2002.

[11] S. Masiello, A.M. Esposito, et al., "Application of Self Organied maps and Curvilinear Component Analysis to the discrimination of the Vesuvius Seismic Signals". WSOM Paris 2005, pp. 387–395.

[12] I. Trendafilova, M. Cartmell and W. Ostachowicz, "Vibration based damage detection in an aircraft wing scaled model using principal component analysis and pattern recognition". *J. of Sound and Vibration* vol. 313, nos. 3-5, 2008, pp. 560-566.

[13] J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas, *SOM Toolbox for Matlab 5*. Som Toolbox Team, Helsinki University of Technology, Finland, Libella Oy, Espoo 2000, 1–54. {Vesanto} http://www.cis.hut.fi/projects/.

[14] M.Voiry, K. Madani, V. Amarger and J. Bernier, "Data dimensionality reduction for neural based classification of optical surfaces defects". *Internaional Scientific Journal of Computing (Computing)* vol. 8, issue 1, 2009, pp. 32-42.

[15] M. Voiry, K. Madani, V. Amarger and J. Bernier, "Optical devices diagnosis by neural classifier exploiting invariant data representation and dimensionality reduction ability". In: F. Sandoval et al. (Eds): *IWANN 2007, LNCS 4507*, pp. 1098-1105, 2007. ©Springer-Verlag Berlin Heidelberg 2007.

[16] H. Yin, "Nonlinear principal manifolds – adaptive hybrid learning approaches". In: E. Corchado, A. Abraham, and W. Pedrycz (Eds.): *HAIS 2008, LNAI 5271*, 2008, 15-29, Springer.

[17] R. Zimroz and A. Bartkowiak, "Investigation on spectral structure of gearbox vibration signals by principal component analysis for condition monitoring purposes". *Journal of Physics Conference Series* 305 (1), 2011, art. no. 012075 (2011)

[18] R. Zimroz and A. Bartkowiak, "Two simple multivariate procedures for monitoring planetary gearboxes in non-stationary operating conditions". *Mech. Syst. Signal Process*. 2012, http://dx.doi.org/10.1016/j.ymssp.2012.03.022

[19] R. Zimroz and A. Bartkowiak, "Multidimensional data analysis for condition monitoring: features selection and data classification". *The Ninth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies Technologies, CM2012\MFPT2012*. BINDT, 11-14 June, London. Electronic Proceedings, 2012, art no. 402, pp. 1-12.