

The Politecnico di Torino rolling bearing test rig: Description and analysis of open access data

*Original*

The Politecnico di Torino rolling bearing test rig: Description and analysis of open access data / Daga, Alessandro Paolo; Fasana, Alessandro; Marchesiello, Stefano; Garibaldi, Luigi. - In: MECHANICAL SYSTEMS AND SIGNAL PROCESSING. - ISSN 0888-3270. - STAMPA. - 120:(2019), pp. 252-273. [10.1016/j.ymssp.2018.10.010]

*Availability:*

This version is available at: 11583/2716028 since: 2018-10-31T10:52:16Z

*Publisher:*

Elsevier

*Published*

DOI:10.1016/j.ymssp.2018.10.010

*Terms of use:*

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

*Publisher copyright*

Elsevier postprint/Author's Accepted Manuscript

© 2019. This manuscript version is made available under the CC-BY-NC-ND 4.0 license  
<http://creativecommons.org/licenses/by-nc-nd/4.0/>. The final authenticated version is available online at:  
<http://dx.doi.org/10.1016/j.ymssp.2018.10.010>

(Article begins on next page)



# POLITECNICO DI TORINO

## THE POLITECNICO DI TORINO ROLLING BEARING TEST RIG: DESCRIPTION AND ANALYSIS OF OPEN ACCESS DATA

ALESSANDRO PAOLO DAGA, ALESSANDRO FASANA, STEFANO MARCHESIELLO, LUIGI GARIBALDI

DYNAMIC AND IDENTIFICATION RESEARCH GROUP - DIRG, DIPARTIMENTO DI INGEGNERIA

MECCANICA E AEROSPAZIALE – DIMEAS

<https://doi.org/10.1016/j.ymssp.2018.10.010>

CITE AS:

A.P. Daga, A. Fasana, S. Marchesiello, L. Garibaldi, *The Politecnico di Torino rolling bearing test rig: Description and analysis of open access data*, *Mechanical Systems and Signal Processing* 120 (2019) 252–273. doi:10.1016/j.ymssp.2018.10.010.

THE POLITECNICO DI TORINO ROLLING BEARING TEST RIG: DESCRIPTION AND ANALYSIS  
OF OPEN ACCESS DATA

Politecnico di Torino

Dipartimento di Ingegneria Meccanica e Aerospaziale – DIMEAS

Corso Duca degli Abruzzi, 24, I-10129 Torino, Italy

Tel.: +39 011 0903397; fax: +39 011 0906999.

E-mail address: [alessandro.fasana@polito.it](mailto:alessandro.fasana@polito.it), [alessandro.daga@polito.it](mailto:alessandro.daga@polito.it)

**Keywords:**

Bearing test rig; open access data; damage detection; damage evolution; big data dimensionality; statistical analysis; ANOVA; Fisher's LDA; PCA; Mahalanobis Distance; Outliers Analysis.

**Abstract**

Nowadays, machines-diagnostics via vibration monitoring is rising an always growing interest thanks to the huge and accurate amount of health information which could be extracted by the raw data coming from accelerometers. Damage severity, type and location of a fault are the kind of information which are buried in the time records.

The scope of this paper is double: first, to present the huge amount of data which have been acquired on the rolling bearing test rig of the Dynamic and Identification Research Group (DIRG), in the Department of Mechanical and Aerospace Engineering at Politecnico di Torino and to share them with the scientific community; secondly, to present a statistical approach analysis and its performances as example of a simple technique to be fruitfully adopted for comparison. To this goal, a detailed presentation of the test rig is given, which comprehends different working conditions up to 30000 rpm, damage types and levels, various sensors positions and directions as well as an endurance test. The related time records can be downloaded from [ftp://ftp.polito.it/people/DIRG\\_BearingData/](ftp://ftp.polito.it/people/DIRG_BearingData/).

Afterward, tried-and-tested statistical tools are exploited to learn the information about bearing damages from this massive amount of data. This "data mining" will be performed using inferential statistical techniques as analysis of variance (ANOVA), applied on usual statistical features, which characterize of the signal. A linear discriminant analysis (LDA) in the configuration proposed by Fisher will be also used to see if the data were classifiable in a multidimensional space with this basic algorithm. Finally, an Outlier Analysis based on Mahalanobis distance will be formulated, so as to distinguish a damage condition from the healthy state (training data), compensating when possible for environmental (temperature) and operational (speed and load) variations.

## 1. Introduction

This paper introduces the data collected over a rig set up at DIRG Lab in the Department of Mechanical and Aerospace Engineering at Politecnico di Torino, specifically conceived to test high speed aeronautical bearings, whose accelerometric acquisitions at variable rotational speed, radial load and damage level, together with temperature measurements, are being made available as open access data. This test rig will be described in detail in the next section of the text. This novel, open access, dataset could be in fact used as a benchmark to assess and compare in a proper way the performance of any proposed diagnostic algorithm regarding rolling bearings [15]. The presence of hidden diagnostic information was also tested in this paper through statistical investigation.

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data, and can rely on tried-and-tested techniques, widely applied in the scientific fields as well as in many everyday life devices. Inferring properties about a population, trying to extract information from a large, representative, data set, is a task that can be easily tackled by Inferential Statistics, which provides reliable tools to derive estimates, test hypotheses and perform clustering or classification of data points into groups. This can be applied to the field of condition monitoring to infer the presence of damages in rotating machinery, namely diagnosing malfunctions and allowing an effective preventive maintenance regime, fostering the reliability and the safety while resulting less expensive in terms of both repair cost and down-time. Focusing on vibration monitoring, many high-level techniques are available to reveal information about the exact position of a fault, the amount of damage and even the expected time to failure. These procedures, although effective, usually need to be tailored for each machine, relying on human contribution and interpretation of the results. Furthermore, they tend to be computationally expensive and of difficult real-time implementation. In this respect, lower level algorithms, based on statistical models, are a fast alternative, able to disclose a fault presence and to automatically trigger an alarm in case of deviation from the normal (heathy) condition. Therefore, a data mining is performed on the raw data to learn health information, switching to a supervised pattern recognition problem, based on labelled samples of extracted, meaningful features.

In particular, a preliminary statistical analysis is set up in this work. A Principal Component Analysis is first used for dimensionality reduction to a 2D representation. Then a univariate ANalysis Of VAriance (ANOVA) is conducted, together with the corresponding multi-comparison Fisher's test, trying to seek for the most informative features and sensors positions. Multivariate tools are used as well, so as to "fuse" all the information from the different channels and features. Fisher's Linear Discriminant Analysis (LDA) is tested to understand how the data are spread in the

multidimensional space and if it is possible to perform a satisfactory classification with this well-known algorithm.

Then, limiting the training phase to the healthy labelled measures, a Mahalanobis distance based Outlier Analysis is also proposed to detect novelty. When a new data is too far from the healthy distribution with respect to a threshold, a damage is likely to have occurred, then an alarm should be triggered. The number of false and missed alarms is used to evaluate the performances of the Outlier Analysis applied to damage detection.

## **2. Description of the experimental setup**

The parts of this section are devoted to the description of the test rig, the sensors and the acquisition system. The aim is to give a precise and comprehensive explanation of both the test rig and of the measurements, performed in various conditions and with different damaged bearings.

### **2.1 The test rig**

The test rig is depicted in Figure 1a and basically consists in a high-speed spindle, driving the rotation of a shaft.

The bearings of the spindle, whose characteristics are unknown, are grease lubricated and their temperature is limited by a liquid (glycol/water) refrigeration circuit. The body of the spindle is fixed to a single and extremely rigid support which rests on a very massive steel base plate. The speed of the spindle is set through the control panel of an inverter but can not be actively controlled: not only the spindle has no keyphasor transducer or tachometer to detect its actual speed but also there is no feedback to the controller of the inverter. As a direct consequence, the actual speed of the shaft is always lower than the ideal one and the difference increases with the applied load.

The same plate carries a couple of supports (positions B1 and B3) for the outer rings of two identical roller bearings. The inner rings of these bearings are connected to a very short and thick hollow shaft, specifically designed for speeds up to 35000 rpm (Figure 1c).

The shaft was originally part of a complete gearbox and carried a spur gear, which actually drove the rotation. The spur gear, because of the applied torque, generated a contact force with radial and tangential directions which insisted on the couple of roller bearings. The same radial force exerted by the spur gear is replaced, in the laboratory test bed, by the load applied by a third and larger roller bearing (B2, Figure 1b/1c). The outer ring of bearing B2 is linked to a precision sledge, whose motion is orthogonal to the shaft: when the sledge is pulled through the rotation of a nut, two parallel springs (green in Figure 1a) are compressed and produce the required load. A static load cell allows for measuring the resulting force, whose direction is purely radial. With this particular design, the work required by the spindle is limited to overcome the dissipated energy (no external

appliance is connected to the shaft), but the bearings are correctly loaded.

The main geometrical properties of the three bearings, specifically manufactured for this high speed aeronautical applications, are listed in Table 1. The lubrication of bearings B1, B2 and B3 is guaranteed by the injection of oil into the hollow shaft; small radial pipes then bring the centrifuged oil (the steady state operative speed is at least 6000 rpm) to the inner rings and consequently to the rollers.

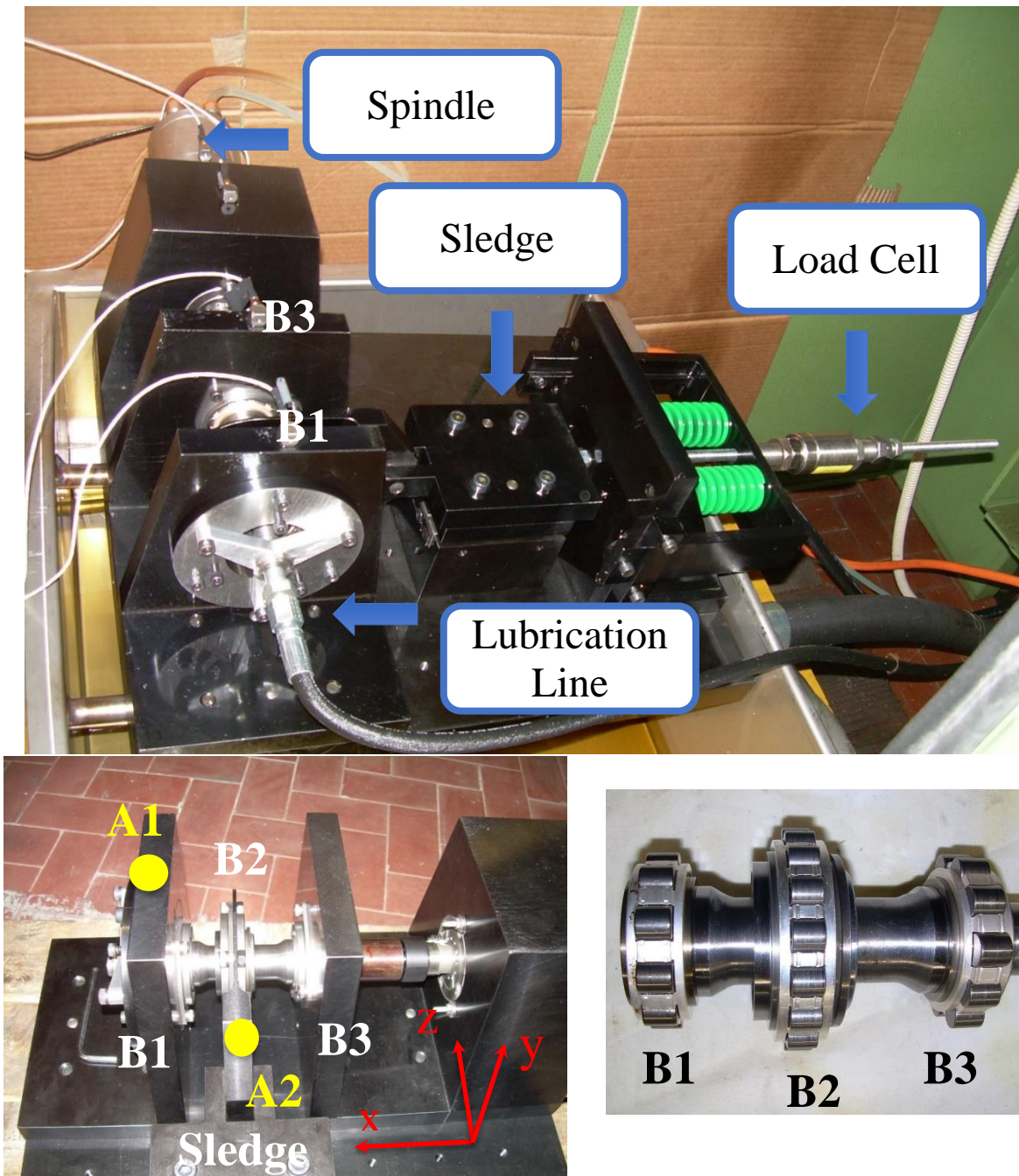


Figure 1: The test rig a) general view of the test rig; b) positions of the two accelerometers and the reference system; c) the shaft with its three roller bearings



Table 1: main properties of the roller bearings

	Pitch diameter	Rollers diameter	Contact angle	Rolling elements
	D (mm)	d (mm)	$\phi$ (°)	Z
<b>B1 &amp; B3</b>	40.5	9.0	0	10
<b>B2</b>	54.0	8.0	0	16

## 2.2 Sensors

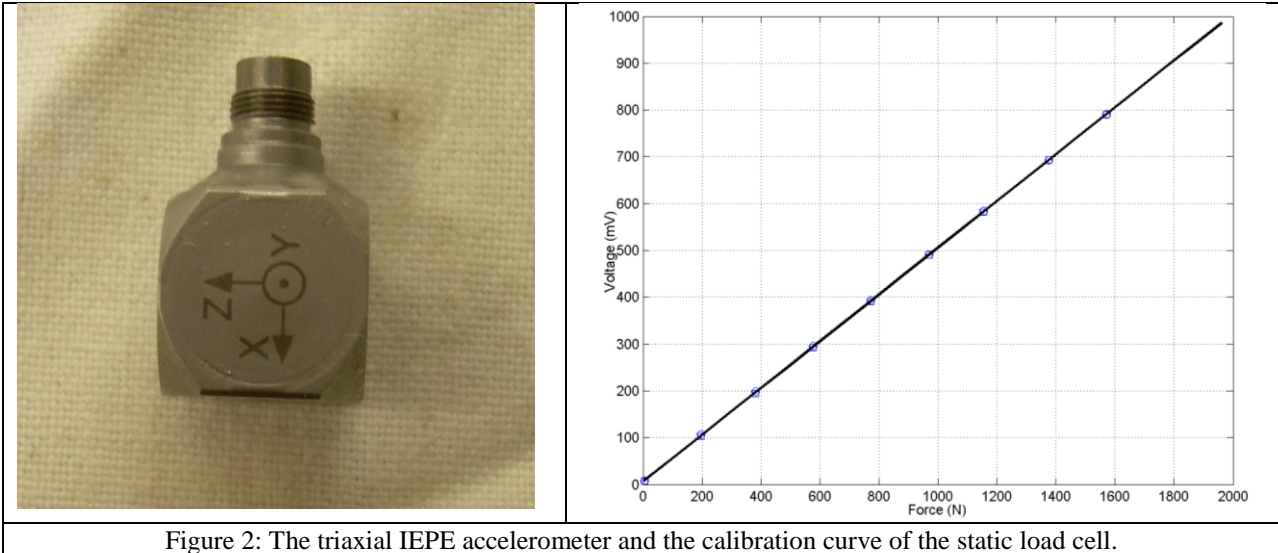


Figure 2: The triaxial IEPE accelerometer and the calibration curve of the static load cell.

The test rig has been variously sensed during its working life, but data collection has always been mainly focused on the accelerations of the supports of the bearings and of the electric spindle. The files described in Section 3 contain the accelerations of the two most significant points of the structure, A1 and A2 in Fig. 1b, located respectively on the support of the damaged bearing under test B1 and the support of the larger bearing dedicated to the application of the external load B2.

The accelerometers are of the triaxial IEPE type, with frequency range 1-12000 Hz (amplitude  $\pm 5\%$ , phase  $\pm 10^\circ$ ), nominal resonant frequency 55 kHz and nominal sensitivity 1 mV/ms<sup>-2</sup>.

The radial force on the central bearing has been measured by a static load cell whose sensitivity – 0.499 mV/N, Figure 2b – has been measured by repeated load cycles, showing an almost null hysteresis cycle.

A K-type thermocouple and dedicated digital thermometer with 0.1 °C resolution has finally been placed as near as possible to the external ring of the damaged bearing. The temperature has been manually recorded during the “Endurance” test (section 3.2) and passed from about 20°C, in the morning and before the start of each test, to about 60-70°C at the end of the day.

## 2.3 Acquisition

The digital data acquisition has been achieved by an OR38 signal analyser, produced by OROS, whose accuracy on the input channels is: phase  $\pm 0.02^\circ$ , amplitude  $\pm 0.02$  dB, frequency  $\pm 0.005\%$ .

The analogue-to-digital transformation is performed by a 24 bits delta-sigma converter and is synchronous on all channels (no multiplexing); the range of every channel was independently set between the minimum ( $\pm 17$ mV) and maximum ( $\pm 40$ V) allowable limits so as to avoid saturation of the channels but also to reach the optimal amplitude resolution.

The specific sensitivities of the accelerometers have been set into the OR38 so that the files listed in Appendix A contain the acceleration time histories in  $\text{m/s}^2$ . The six recorded channels correspond to the outputs of the couple of accelerometers placed in points A1 and A2 (Figure 1b); Table 2 specifies the direction of the acceleration for each channel.

Table 2: direction of the measured accelerations (see also Fig. 2b)

	<b>Channel 1</b>	<b>Channel 2</b>	<b>Channel 3</b>	<b>Channel 4</b>	<b>Channel 5</b>	<b>Channel 6</b>
<b>Direction</b>	Axial, x	Radial, y	Radial, z	Axial, x	Radial, y	Radial, z

Every data set has the same duration  $T$  and the same sampling frequency  $f_s$ . The sampling frequency must not be divided by the number of channels so that each time domain acceleration record consists of  $T \times f_s$  samples.

The values of  $T$  and  $f_s$  are given in Section 3 and are different for the two investigations performed on the bearings.

## 3. Description of the tests

Two different experimental sessions are reported in this paper. In the first the accelerations are relative to bearings with different damages, running at different speeds and under different loads; the second reports the behaviour of a single damaged bearing undergoing a long (about 330 hours) test at constant speed and load.

### 3.1 Variable speed and load

The bearing in position B1 (Fig. 1a) has been designed to quite simply be removed from its support, so to have the chance to check the response of the system when bearings with different types and dimensions of damage are mounted.

Table 3 gives the names of the damaged items, with 0A indicating the undamaged case. Localised faults on the elements have been produced by using a Rockwell tool, causing a conical indentation on the inner ring or on a single roller (1A to 6A); the size of the resulting circular area is given by



its approximate measured diameter (150, 250, 450  $\mu\text{m}$ ). An example of indentation (case 4A) is presented in Figure 3a.

Table 3: List of the defects of the various bearings mounted in position B1.

Name	Defect	Dimension ( $\mu\text{m}$ )
0A	NO DEFECT	----
1A	Diameter of an indentation on the inner ring	450
2A	Diameter of an indentation on the inner ring	250
3A	Diameter of an indentation on the inner ring	150
4A	Diameter of an indentation on a roller	450
5A	Diameter of an indentation on a roller	250
6A	Diameter of an indentation on a roller	150

Every bearing, starting from 0A to 6A, has undergone the same testing procedure:

- a brief run at the minimum speed (100 Hz) and no load, so to check the correct mounting;
- application of the static load: at first 1000 N, then 1400 N and finally 1800 N;
- increment of the speed of the shaft from 0 Hz to 500 Hz with steps equal to 100Hz;
- measurement of the accelerations as soon as a steady speed of the shaft was reached.

The whole procedure took about 30 minutes and, because of the limited power of the inverter, the higher rotation speeds could not be reached with the higher loading conditions. Table 4 gives the list of the speed-load combinations.

Table 4: list of the tested load and speed cases.

Nominal load (N)	Nominal speed (Hz)				
	100	200	300	400	500
0	100	200	300	400	500
1000	100	200	300	400	500
1400	100	200	300	400	----
1800	100	200	300	----	----

The time histories of the six channels (Table 2) have been collected with sampling frequency  $f_s=51200$  Hz for a duration of  $T=10$ s. Data are recorded in files whose names have the following format: CnA\_fff\_vvv\_m.mat.

C: root of the file name, common to all files;

n: integer value from 0 to 6, indicating the kind of the defect, e.g. 0A, 1A, ..., 6A (Table 3);

fff: integer value from 100 to 500, indicating the nominal speed of the shaft (Hz);

vvv: integer value corresponding to the voltage of the load cell (mV), indicating the applied load;

m: integer value, indicating if the measurement has been repeated (m=2) or not (m=1);  
.mat; Matlab® file extension.

For example, C4A\_100\_702\_1.mat contains the first registration of acceleration signals produced by bearing 4A (defect on a roller), at the nominal rotational speed of 100 Hz (i.e. 6000 rpm), under a load of about 1407 N which corresponds to an output of the load cell of 702 mV (sensitivity 0.499 mV/N).

Each file contains a matrix with the same name of the file (apart from the .m extension) with 512000 rows (time samples) and 6 columns (one for each channel).

The list of all filenames is given in Appendix A, Table A.1.

### **3.2 Endurance**

An endurance test has been performed on bearing 4A to monitor the evolution of the defect, initially a conical indentation with a maximum diameter of about 450µm (Figure 3.a). It has to be stressed that before starting this experiment, the bearing had already been installed on its actual complete gearbox, running at various load and speed conditions, not documented here because of a non-disclosure agreement. The condition of the roller at the end of this first investigation (about 19 hours) is reported in Figure 3.b; it can be noticed that the border of the indentation is not as regular as before and that a small amount of material has been removed from the roller surface. The evolution of the defect has been inspected also after 70, 144 and 332 hours of the endurance test, and is highlighted by the pictures in Figure 3. Unfortunately, the photos are not so uniform because they have been taken by different persons on different microscopes, but the trend is clear: with time passing larger parts of material are removed and, at the same time, the small “crest” around the area of the indentation is flattened.

It is worth noticing that the lubrication oil has been changed before starting the endurance test. The new oil has almost the same viscosity (15 cSt at 40 °C, ASTM D 445) as the previous one but it not specifically designed for high speed applications. For example, its resistance at generating foam does not meet the required aeronautical standard. Contrary to the original oil, it has the enormous advantage of not being neurotoxic, which is a compulsory requirement for our laboratory experiments.

Every measurement of the endurance assessment has been performed at the same conditions:

nominal speed: 300 Hz (i.e. 18000 rpm)

nominal load: 1800 N.

The time histories of two triaxial accelerometers have been regularly recorded and each registration

lasts for 8s with a sampling frequency of 102.4 ksps.

An indication of the temperature was given by a thermocouple fixed as near as possible to the outer ring of the bearing. Every day, at the beginning of the test and immediately before switching the spindle on, the temperature of the test bed and the bearing was about 18-20 °C (room temperature); the heating of the bearing, because of the thermal inertia of the massive structure and the lubricating oil (about 15 litres), lasted for some hours. An almost steady temperature could be reached only after roughly three hours so that the temperature associated to the first file of each working day is 5-6 °C lower than the final value, as reported in Table A.2 in appendix A.

All file names are in the form: E4A\_nnn.mat

E4A\_: root of the file name, common to all the files;

nnn: integer value from 007 to 226, indicating the number of the acquisition.

For example, E4A\_083.mat contains the acceleration signals recorded almost at the end of the day, after 7.5 hours of running time.

Each file contains a matrix with the same name of the file (apart from the .m extension) with 819200 rows (time samples) and 6 columns (one for each channel).

Table A.2 in Appendix A gives the list of the files, 66 in total, with the corresponding temperatures and the acquisition hour, given as the time passed since switching the spindle in the morning.

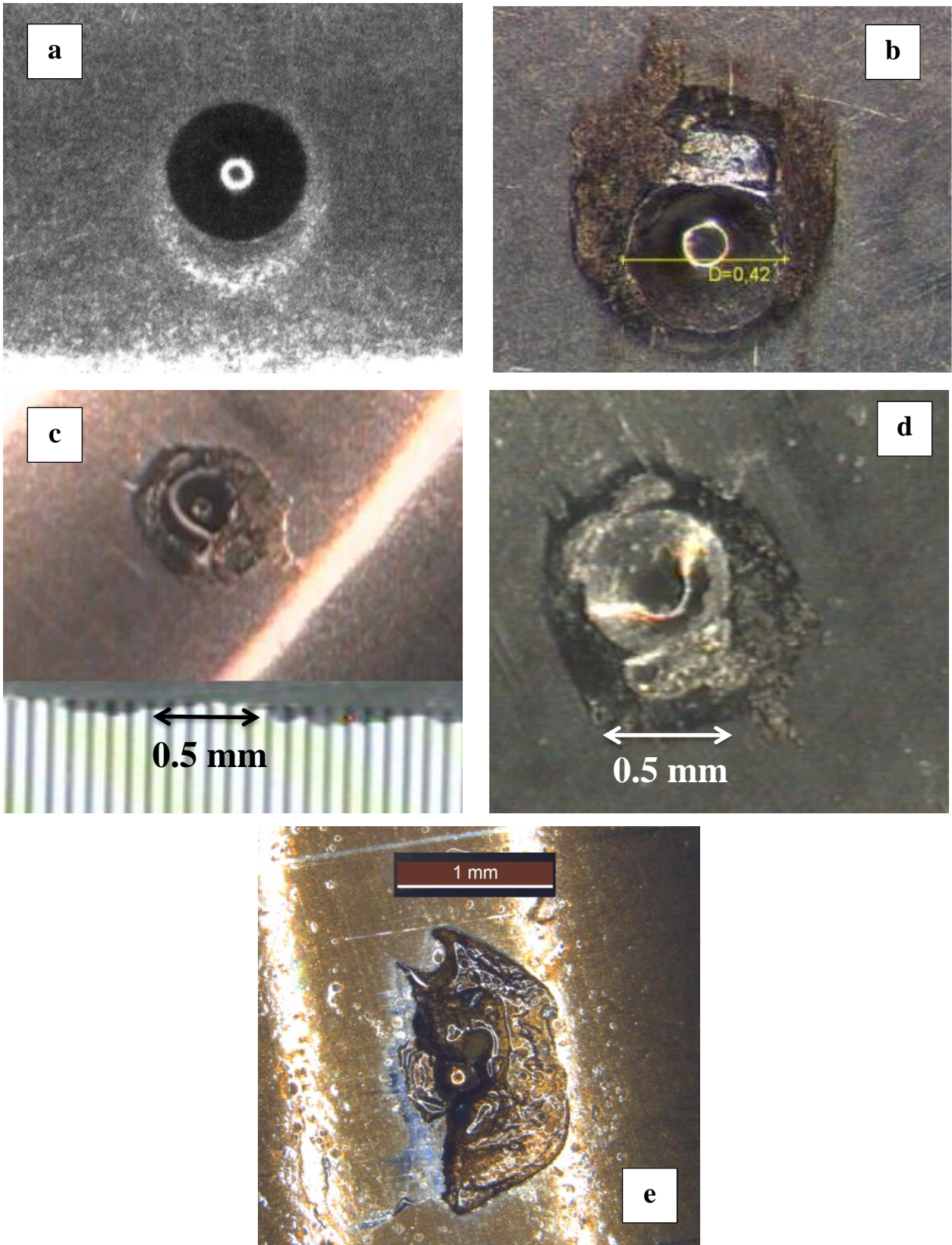


Figure 3: The progression of the indentation of bearing 4A; a) initial defect; b) after 19 hours at various load and speed conditions; c) after 70 hours; d) after 124 hours; e) after 232 hours – c), d) and e) at 300 rpm and 1800 N

#### 4. Procedure

The analysis of the huge dataset available, starts with a fundamental pattern recognition operation: the features selection and extraction. This is essential to enhance characteristics not evident otherwise.

The procedure pursues with a Principal Component Analysis, to visualize the high dimensional dataset on a 2D approximated plot.

The Analysis of Variance (ANOVA) together with Fisher's LSD post hoc test is then used to assess whether some feature, obtained from a particular sensor position and direction, is more informative than the others. This is an exploratory tool to explain univariate observations, so it does not account for correlation. A multivariate analysis is then preferable. The founding consideration of ANOVA could be extended to obtain a simple multivariate classification tool, the Fisher's Linear Discriminant Analysis, which is much more suitable for diagnostic purposes than a MANOVA. The procedure concludes with an Outlier Analysis, meant to detect the damaged acquisitions which will be ideally far from the nominal healthy condition used to train the algorithm.

The scheme in Figure 4 summarizes the proposed sequence.

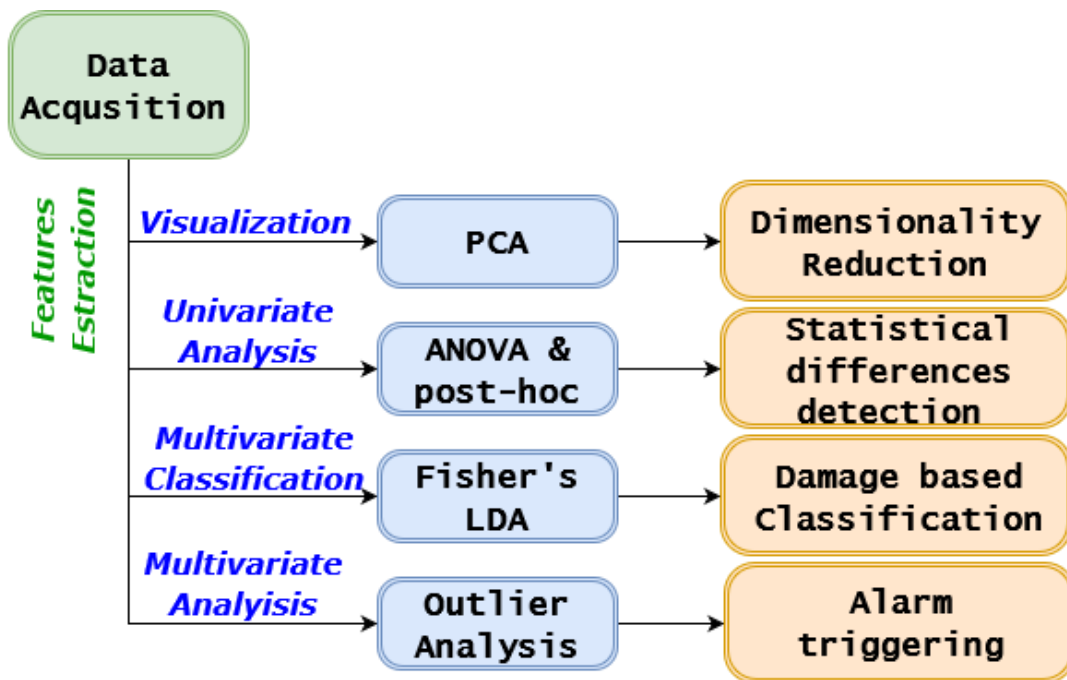


Figure 4: Procedural scheme

##### 4.1 Features extraction [10]

As introduced before, the available data-set of raw acceleration may contain a lot of diagnostics information but those are unfortunately “hidden”. In order to explore the data and foster the learning, some derived quantities called features are usually extracted, as they summarize the signal, allowing for a simplified and better interpretation. In general, many different features are

commonly used in diagnostics, but these can be classified in two main categories: time-domain and frequency-domain features. The frequency-domain features are known to be more stable with respect to changes in the overall machine configuration (e.g. speed, load, ...), but have the drawback of being less general, requiring a deeper prior knowledge of the machine under analysis (geometry of the bearings, dimensions, effective running speed...).

In this paper, the most common time-series features have been selected, coping with the need of speed and automation of the analysis.

*Root mean square, skewness, kurtosis, peak value* and *crest factor* (peak/RMS) have been then computed on 0,1s chunks of the original 10s acquisitions generating 100 data points for each of the 17 acquisitions described in section 3.1 (see Table 4 and Table 5). This has been chosen in accordance to the considerations about the curse of dimensionality reported in **Appendix C**.

Table 5: Labels associated to the pairs frequency-load

Label:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
<b>f</b> [ $10^3$ Hz]	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	5	5
<b>F</b> [kN]	0	1	1.4	1.8	0	1	1.4	1.8	0	1	1.4	1.8	0	1	1.4	0	1

The data-set will be then composed by 7 differently damaged conditions, from 0A (healthy), to 6A, containing 1700 measurements in a 30-dimensional space (6 channels, 5 features).

#### 4.2 Principal Component Analysis (PCA) [11,12]

The PCA is an orthogonal space transformation which aims to get a set of uncorrelated variables starting from a correlated dataset. The initial space is then rotated so that the axes will match the directions in which the maximum variance is obtained, namely the principal components, which are uncorrelated linear combinations of the original variables.

This rotation will then bring the first principal component to explain the largest variance, while each succeeding component will show the highest possible variance under the constraint of orthogonality with the preceding ones. An example is given in Figure 5 for a bi-normal distribution. This can be efficiently accomplished via eigenvalue decomposition (or alternatively singular value decomposition) of the covariance matrix of the data, after mean centering. Since PCA is also sensitive to the scaling of the original variables, a standardization of the raw data set is advisable, so that the PCA will be performed on Z-scores ( $z = (x - \mu)/\sigma$ ). Alternatively, the data variance-covariance matrix could be substituted by the correlation matrix to solve scale issues.

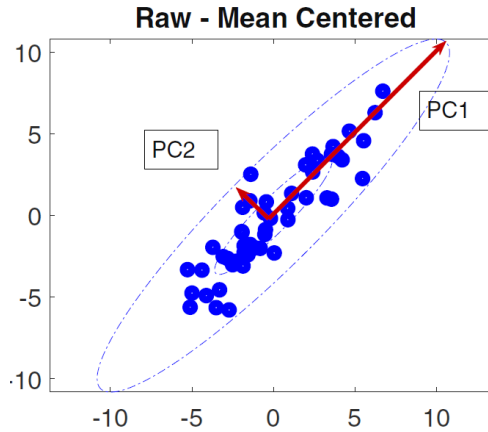


Figure 5: PCA disclosing principal components for a bi-normal distribution.

In this work, the PCA will be used to reduce the original 30-dimensional space to the bi-dimensional one defined by the first two principal components. After neglecting the low-variance components, an approximated visualization of the data is possible. Although useful for a preliminary visual investigation, this does not really help diagnostics as the health information is likely to be neglected, requiring a more challenging detection.

### 4.3 ANOVA [1,2]

Analysis of variance (ANOVA) is a statistical inference method used to analyse univariate data acquired from different groups, deducing properties of their underlying distributions. It basically tests the omnibus (variance based) null hypothesis  $H_0$  that no significant difference is detectable among the mean of the different groups so that the measured phenomena (groups) results non-related. Hence, this may be used as a diagnostic tool to assess whether the damaged distribution, in this case 1700 points for each of the 6 groups, differs from the healthy one per each channel and feature combination (30 cases).

ANOVA is a powerful exploratory tool, but in case of  $H_0$  rejection, it is not able to explain which population differs the most and from which one of the others. Multiple two-groups tests (ANOVA basically simplifies to a Student's t-test for two classes) can be used in this case, but this increases the chance of incorrect rejection of a true null hypothesis (statistical type I error), so the usual post hoc analysis relies on multi-comparison tests such as the Fisher's Least Significant Difference (LSD).

From a mathematical point of view, ANOVA assumes a linear model, so that an observation of the  $j^{th}$  group corresponds to a random draw from a normal distribution centered around the group mean  $\mu_j$  and with a standard deviation  $\sigma$  equal for all the groups, i.e.

$y_{ij} = \mu_j + \varepsilon_{ij}$	$\varepsilon_{ij} \sim N_j(0, \sigma_j)$	$\sigma_j = \sigma \forall j = 1: G$
-------------------------------------	--	--------------------------------------



This involves then three hypotheses for the group distributions:

- Normality
- Homoscedasticity (homogeneity of variances)
- Independence of observations

If these are verified, the overall variance  $\sigma_t^2$  can be divided into a *within groups variance*  $\sigma_{wg}^2$ , roughly an average of the  $\sigma_j^2$  weighted according to the numerosness  $n_j$  of each group, a so called “pooled estimate” of the variance, and a *between groups variance*  $\sigma_{bg}^2$ , the squared deviation of the groups means  $\mu_j$  from the overall mean  $\bar{y}$ .

$$\sigma_t^2 = \sigma_{wg}^2 + \sigma_{bg}^2$$

In order to test the hypotheses, some tried-and-tested techniques are available, such as the *Q-Q plots and Jarque-Bera or Lilliefors tests* for Normality and the *Fisher-Snedecor F-test, Bartlett's test* for Homoscedasticity.

However, ANOVA is generally considered robust to violations of these assumptions, in particular for the simplest case in which all the groups under analysis show equal numerosness,  $n_j = n \forall j = 1:G$ .

When  $\sigma_{bg}^2$  and  $\sigma_{wg}^2$  are statistically equivalent, their ratio is distributed according to the Fisher-Snedecor's  $F_{(G-1, N-G)}$ , so that the statistical summary F (the ratio) can be compared to the theoretical distribution (*Fisher's F test*). The result will be read in terms of p-values, namely the probability of F to be the same as or more extreme than the actual observed ratio. In general, when this probability is lower than 5%, the H0 hypothesis is rejected at a confidence  $1 - \alpha = 95\%$ , so the F ratio is usually directly compared to the so called critical value  $F_{(G-1, N-G)}^{5\%}$  (see Table 6 and Figure 6).

Table 6: Relevant ANOVA expressions in the special case  $n_j = n$  in each group

$\sigma_{bg}^2 = \sum_{j=1}^G \frac{n_j}{N} (\bar{y} - \mu_j)^2$	$\sigma_{bg}^2 \sim \chi_{(G-1)}^2$	$F = \frac{\sigma_{bg}^2 / G - 1}{\sigma_{wg}^2 / N - G} \sim F_{(G-1, N-G)}$	<b>H0 accepted (same mean) if:</b>  $F \leq F_{(G-1, N-G)}^\alpha$ (critical value)
$\sigma_{wg}^2 = \frac{1}{N} \sum_{j=1}^G \sum_{i=1}^{n_j} (y_{ij} - \mu_j)^2$	$\sigma_{wg}^2 \sim \chi_{(N-G)}^2$		
<i>N = nG: overall numerosness, G: number of groups.</i>			

It is worth to remember that the ratio F can be used as a measure of separation among two

distributions; in fact the farther two groups are, the bigger will be the between-groups variance, leading to larger values of F, which corresponds to lower p-values. This will be used in Fisher's LDA classifier.

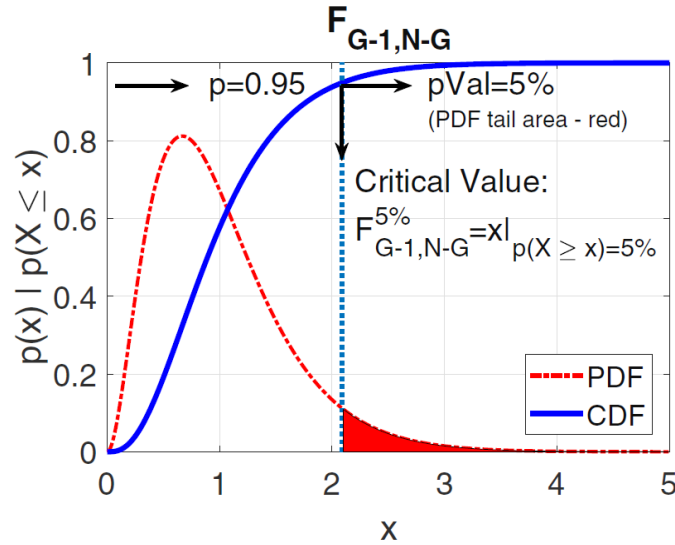


Figure 6: Fisher-Snedecor's  $F_{(6,11893)}$  critical value ( $\approx 2.1$ ) for a p-value of 5% - one sided test

#### 4.4 Fisher's LSD multi-comparison [13,14]

As stated in section 4.3, ANOVA is unsuitable to compare one single distribution with respect to the ones of the other group, but it can be anyway used as a starting point for further post hoc analysis, namely multi-comparison tests. Remembering that ANOVA simplifies to a Student's t test if just 2 groups are considered (as, in terms of distributions,  $F_{(1, N-2)} \equiv t^2$ ), Fisher's Least Significant Difference (LSD) generalizes an individual t-test referencing it to the pooled estimate of the standard deviation for all groups, reducing the type I error (incorrect  $H_0$  rejection).

Therefore, a limit range of  $\pm LSD/2$  may be built around the mean of each group, so that intersecting ranges will be considered not significantly distant, as it would be hard to recognize them with enough confidence. Table 7 summarizes all the steps from Student's t test to Fisher's LSD formulation.

Table 7: LSD formulation starting from Student's t test, for equal  $n_j = n$  in each group

Student's t test:	Fisher's generalisation:	Fisher's LSD:
$t = \frac{\mu_i - \mu_j}{\sqrt{\frac{\sigma_i^2 + \sigma_j^2}{n}}} \sim t_{(2n-2)}$	$t = \frac{\mu_i - \mu_j}{\sqrt{\frac{\sigma_{wg}^2}{n}}} \sim t_{(N-G)}$	$LSD =  \mu_i - \mu_j  \leq t_{(N-G)}^{\alpha/2} \sqrt{\frac{\sigma_{wg}^2}{n}}$
$H_0: \text{same mean if } t \leq t_{(2n-2)}^{\alpha/2}$		

#### 4.5 Multivariate Analysis: Fisher's LDA [8]

Since many sensors and features are available, it would be wise to “fuse” the information contained in the original data set trying to improve the results. A multivariate analysis of variance (MANOVA) could be used, but this shares the same limitations as ANOVA, as the resulting p-values will be of difficult interpretation. Luckily enough, simply extending the considerations introduced for ANOVA, Fisher's Linear Discriminant Analysis (LDA) may be deduced, enabling a rough but simple, classification. Being a supervised algorithm, a classifier function will be built on the basis of labelled training samples; this will be later tested on a validation set, so as to assess the quality of the discrimination.

From the mathematical point of view, LDA seeks for a direction  $w$  in which the distance between the projected class means is maximum. This distance will be normalized by the measure of the within-class scatter along  $w$ . The measure of separation is then given by the ratio  $\sigma_{bg}^2/\sigma_{wg}^2$ , which can be generalized for a multivariate space.

For the sake of simplicity, the formulas are proposed considering only 2 groups, as graphically reported in Figure 7.

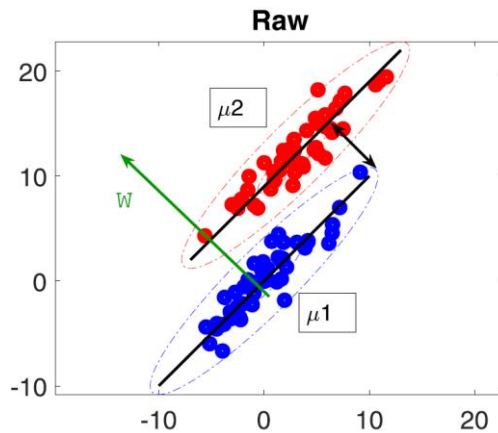


Figure 7: LDA idea - find the direction on which the projected samples will show the higher separation  
The separation as a function of the direction  $w$  is

$$J(w) = \frac{w' S_{bg} w}{w' S_{wg} w}$$

where  $S_{bg}$  and  $S_{wg}$  are the estimates of the Between Groups covariance matrix and of the Within Groups covariance matrix.

$$S_{bg} = (\bar{\mu}_2 - \bar{\mu}_1)(\bar{\mu}_2 - \bar{\mu}_1)'$$

$$S_{wg} = \sum_{k \in C_1} (\bar{y}^k - \bar{\mu}_1)(\bar{y}^k - \bar{\mu}_1)' + \sum_{k \in C_2} (\bar{y}^k - \bar{\mu}_2)(\bar{y}^k - \bar{\mu}_2)'$$

Hence, from the maximization of the separation  $J(w)$  it is possible to find the optimal direction for classification:

$$\arg \max_w J(w) \rightarrow w \propto S_{wg}^{-1}(\bar{\mu}_2 - \bar{\mu}_1)$$

Extending this to multiple classes, it can be found that, when  $w$  is an eigenvector of  $S_{wg}^{-1}S_{bg}$ , the separation will be equal to the corresponding eigenvalue. With reference to section 4.2 (PCA), it is enough to focus on the principal components of the  $S_{wg}^{-1}S_{bg}$  matrix.

$$w \propto PC \text{ of } S_{wg}^{-1}S_{bg}$$

It is important to remember that, as for the ANOVA, an assumption of homogeneity of the variance-covariance matrices of the classes should be verified, using for example a *Box's M test*.

LDA is very interesting from a theoretical point of view as it opens to the multivariate classification, but it works well only for the rare case of linearly separable classes.

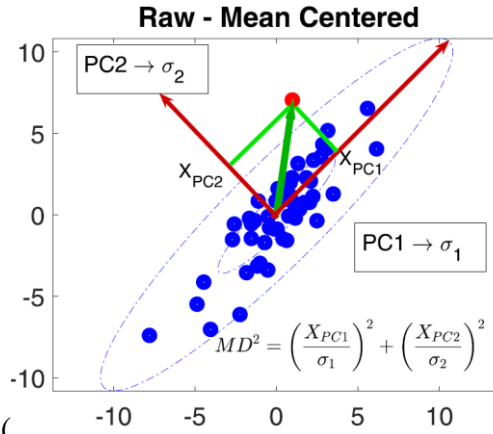
Modifications of LDA as the Quadratic Discriminant Analysis (QDA) solve this issue and are also more robust to violations of the assumptions, but will not be tested in this work. Indeed, the purpose of the paper is to diagnose a damaged condition, not to optimize a classifier. In this regard, an Outlier Analysis is preferable. As a matter of fact, the same performance may be reached (QDA is based on squared distances) with the advantage of having a shorter training, only focused on the healthy class.

#### 4.6 Outlier Analysis

In statistics, a discordant measure is defined *outlier* when it is inconsistent with the others, and can be ascribed to a different physical mechanism. The judgement on discordancy is based on a measure of the distance of the sample from the underlying distribution of the dataset. This distance usually takes the name of Novelty Index (NI), and corresponds to a 1-D space reduction.

A single threshold will be finally defined so as to discern possible outliers.

The simplest measure of discordancy is the Eulerian distance from the mean, but this does not account for the correlation of the data (different variances on the principal components), so an alternative formulation is usually preferred. In this regard, the Mahalanobis Distance (MD) is the optimal candidate for assessing the novelty, as it corresponds to a Eulerian distance on the



standardized principal space ( ). It is then based on  $S$ , the estimated variance-covariance matrix of the data, so as to get a unitless and scale invariant measure of distance:

$$MD(X) = \sqrt{(X - \mu)^T S^{-1} (X - \mu)} = NI(X)$$

Such Novelty Indices will be then compared to some objective criterion to judge if a datum is far enough from the healthy distribution used for training to be considered *novel*.

It is wise to remember that novelty is not necessarily ascribable to the presence of a damage; indeed, if not accounted in the training, a variation in the operational and environmental conditions could also induce the detection of outliers. Further considerations are added in the final part of this section.

Anyway, the outlier detection reduces to a selection of a suitable threshold. This operation is not trivial, as in general the training distribution may be far from normality. Luckily, some probability techniques are robust to non-normal data.

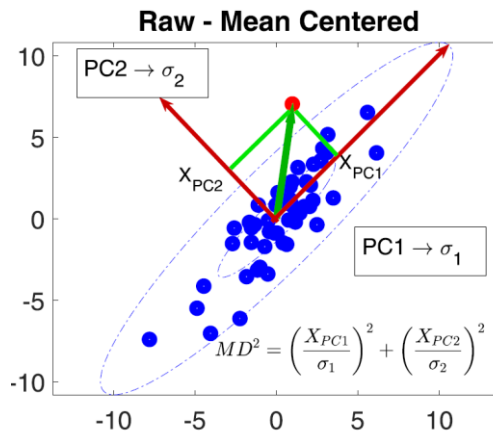


Figure 8: Mahalanobis Distance geometric interpretation, see Appendix B

Let us consider the well-known Chebyshev's inequality for example. The statement according to which "nearly all values are close to the mean — no more than  $1/k^2$  of the distribution's values can be more than  $k$  standard deviations away from the mean" holds for a wide class of 1-D probability

distributions. This hypothesis usually overestimates the probability of the extremes, considering fatter distribution’s tails (the tail of a normal distribution decays more rapidly than that), but a first guess value for the threshold can be defined.

To improve the thresholding operation, Worden [6] proposes to repeat several Monte Carlo simulations (MC) of a  $p$ -dimensional Gaussian distribution, drawing  $n$  observations in  $p$  variables and computing the NIs. Extracting the maximum NI for every MC repetition and focusing on the maxima CDF, a threshold could be found, taking for example the 99th percentile of such a distribution (Figure 9).

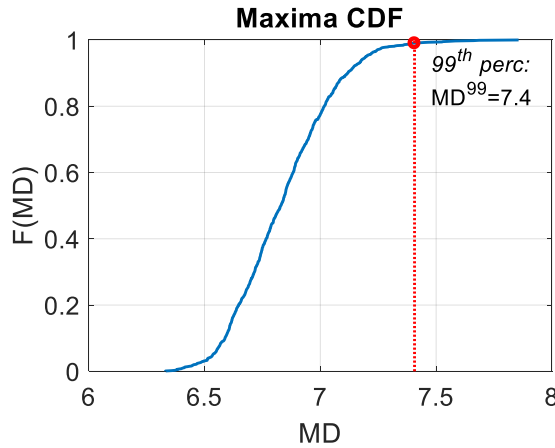


Figure 9: Maxima CDF for 1000 MC repetition of  $n = 100$  samples in a dimension  $d = 30$ .

It is relevant to point out that the use of MD based NIs brings the advantage of allowing some compensation for variations of the operational and environmental conditions under linearity (or quasi linearity) assumption, because of the intrinsic training data standardization on its own PCs’ standard deviations (see Appendix B).

## 5. Analysis and Results

As introduced in the previous section, the preliminary step is the features extraction and the re-organization of the huge original dataset (Section 4.1). Seven matrices, one per each condition (0A to 6A), have been organized in 1700 rows, made by 100 samples for every speed and load combination (17 conditions, ordered as defined in Table 5), and 30 columns (5 features extracted for each of the 6 available channels). This set corresponds then to 7 clouds of 1700 points in a 30-dimensional space.

The first step of the proposed procedure is a visual exploration of these clouds, allowing a qualitative, preliminary analysis. The PCA (Section 4.2) has been then used to summarize the data matrices in the 2D plane generated by the first two principal components of the healthy (0A) set, as reported in Figure 10.

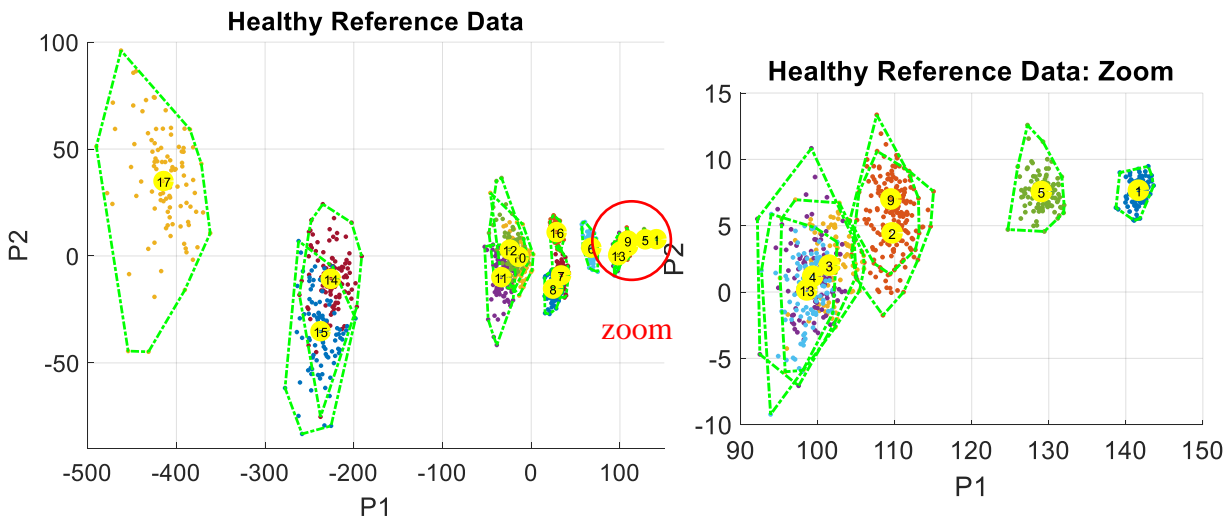


Figure 10: Healthy data for the 17 speed and load combinations (Table 5) \_ 68% and 99% (1 and 3 sigma equivalent) ellipses compared to the convex hull.

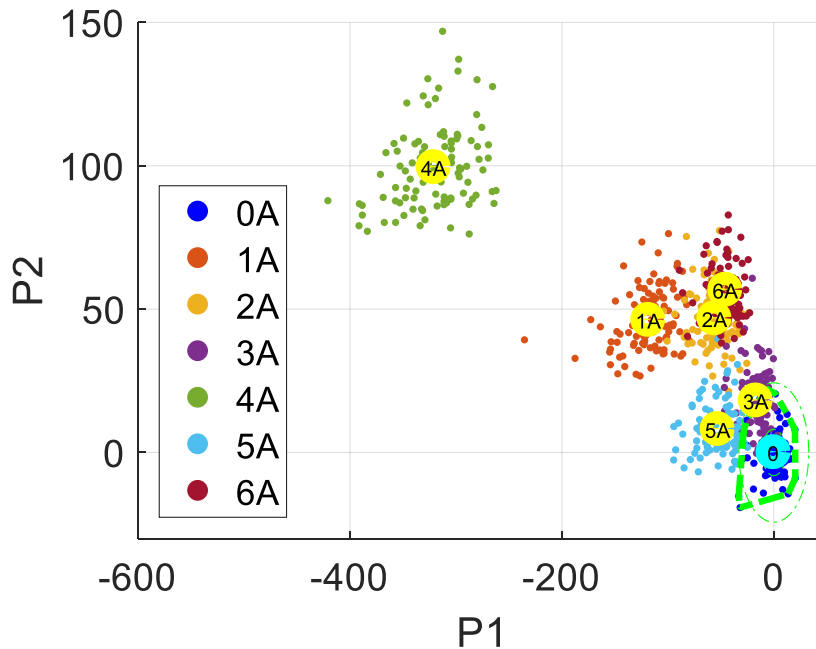


Figure 11: Healthy data compared to damaged acquisitions, centered on the same reference – work condition 12.

From the picture it is easy to notice that the effect of speed is much more relevant than the load. Indeed, after neglecting the zero-load condition (1,5,9,13 labels), which is anyway not very common, the data could be clustered in equal speed subgroups, almost regardless from the load (2-3-4, 6-7-8, 10-11-12 and 14-15 clusters).

In Figure 11 on the contrary, the acquisition 12 (300 Hz, 1800 N) is considered, at all the different damage levels, after removing the mean value of the healthy set (0A). In this case the diagnostic information is very effectively pictured, as the most damaged conditions (1A and 4A) result to be the furthest from the healthy cluster. This condition will be enhanced by the following outlier



analysis.

It is useful to keep in mind that this dimensional reduction is just a simplified projection which is neglecting a lot of information, but it is able to underline the strong effect of the working conditions on data distribution. Because of this consideration, the outlier analysis will begin by comparing the damaged conditions to the healthy one at the same operational conditions.

As a second step, the univariate ANOVA (Section 4.3) is performed for each channel (sensor position and direction) and feature to assess the diagnostics-ability of each combination.

Although the assumptions of normality and homoscedasticity were not completely met, after the consideration about the robustness of ANOVA (see section 4.3), the results of this analysis were deemed to be accepted. In all the 30 tests, indeed, the ANOVA p-values resulted almost negligible, meaning that the omnibus null hypothesis  $H_0$  (no significant difference detectable among the groups) is rejected at a very high confidence.

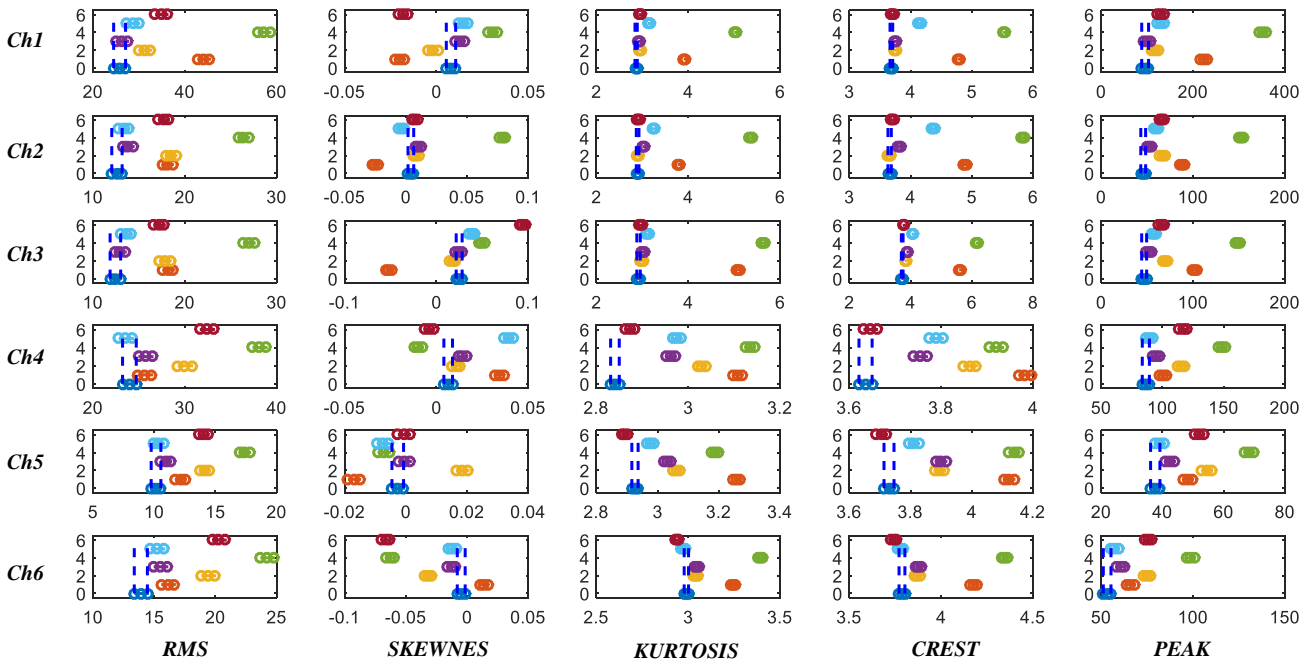


Figure 12: ANOVA post-hoc, Multicomparison result. For different channels and features, all the 6 damage conditions are compared to the healthy reference (0A) through LSD limits.

A focus on the more informative multi-comparison post-hoc test (Section 4.4) was then the natural continuation, so as to get a comparison of the 6 damage conditions with the healthy reference through LSD limits. Looking at Figure 12, it's easy to notice that *kurtosis* and *crest* are the best features, able to discriminate the most damaged 1A and 4A conditions in all the channels. They also seem to be quite consistent with the damage. Anyway, the damage is clearly producing a detectable effect on all the considered features, as already proved, at a very high confidence, by ANOVA

(negligible p-values for all the feature-channel combinations).

The point is that, in particular for the less damaged conditions, all the features are adding some information, so that it would be wiser not to discard any of the 30 dimensions.

Then, a multivariate classification has been conducted in the third step of this analysis. This allows to assess if it's possible to discern the different damage conditions using the information from the whole set of 6 channels and 5 features altogether. It is worth to remember that such a 7 class classification is not common in diagnostics, in which the damage detection phase is usually separated from the damage type and severity identification phase. Every other classifier, considering only one damage per time, would produce a better classification; QDA and kNN were also tested, but their application is beyond the scope of this paper. Nevertheless, this is proposed to test if a statistical difference among the groups can be recognized also in the multivariate space of the features (all together).

The LDA classifier (Section 4.5) has been trained on the first 60 samples for each of the 17 operational conditions, while the remaining set has been used as a validation set. Two confusion matrices have been produced so as to assess the classifier performance in terms of percentages of the classified samples (approximated) against the true, expected class (to be read by rows). Table 8 shows that the linear discriminant analysis, despite correctly recognizing 1A and 4A in most of cases, shows some troubles with the other damages and in particular with the undamaged condition which is correctly recognized only in half of the cases.

Table 8: LDA confusion matrices in rounded percentages by rows.

LDA		Training (in sample)							Validation (out of sample)						
		Target Class							Target Class						
		0A	1A	2A	3A	4A	5A	6A	0A	1A	2A	3A	4A	5A	6A
Output Class	0A	51	0	10	12	0	9	15	48	0	9	15	0	12	14
	1A	12	73	0	4	1	6	0	12	73	1	4	1	6	0
	2A	14	0	59	13	0	4	7	16	0	52	17	0	5	7
	3A	25	0	11	43	0	11	7	22	0	12	44	0	13	7
	4A	1	4	0	0	81	10	0	1	4	0	1	80	11	0
	5A	17	1	6	5	0	63	5	15	1	8	8	0	61	4
	6A	13	0	8	12	0	7	57	14	0	8	12	0	7	57

Unfortunately, in many real-life industrial applications, it's not advisable to let machines run in a damaged configuration, so the corresponding acquisition will be hardly available. In this case an

Outlier Analysis trained only on a healthy dataset will be preferred.

The discordant measures will be efficiently underlined by computing the Mahalanobis distance based Novelty Indices and comparing them to the MC evaluated threshold (Section 4.6).

It is wise to remember that novelty does not imply the presence of a damage. Indeed, possible variations of operational and environmental conditions could lead to similar deviations. This is underlined by the analysis on the entire data set reported in Figure 13. In the picture, the effect of the strong variations of the speed and load is already evident in the healthy (training) case. Although all the conditions are used in the training, the too large range of variation causes a high rate of False Alarms (FA). Furthermore, only the biggest damages can be recognized effectively, while many Missed Alarms (MA) are present in all the other conditions.

An independent analysis for each of the 17 measured operational conditions has been then performed, reporting the results in terms of FA and MA in

Table 9 and graphically in Figure 14 for just a couple of interesting conditions (work conditions 3 and 12). In this case the results are really improved, as the alarm rates are very good for almost all the conditions, while just at low speed, for conditions 3 and 4 it seems difficult to diagnose the 3A damage.

Indeed, a consistency between NIs and damage is evidenced, as they change almost monotonically with the defect severity. This could be then used to extract further information about the size of the damage. Additionally, in most of cases (as in Figure 14 - condition 12) all the damaged conditions show a wide distance from the healthy state, so that the threshold may be increased to reduce the FA rate without worsening the MA rate.

The last part of the analysis is meant to show the ability of MD OA to compensate for linear or quasi linear variations of the operational conditions, in accordance to what introduced in the previous section. A test at a constant speed of 300 Hz while load changes from 0 to its maximum value is then performed, agglomerating acquisitions 9 to 12. Indeed, by training the algorithm on a variable load, healthy set, allows to filter out quite well the effect of load variation, as it is easy to notice in Figure 15.

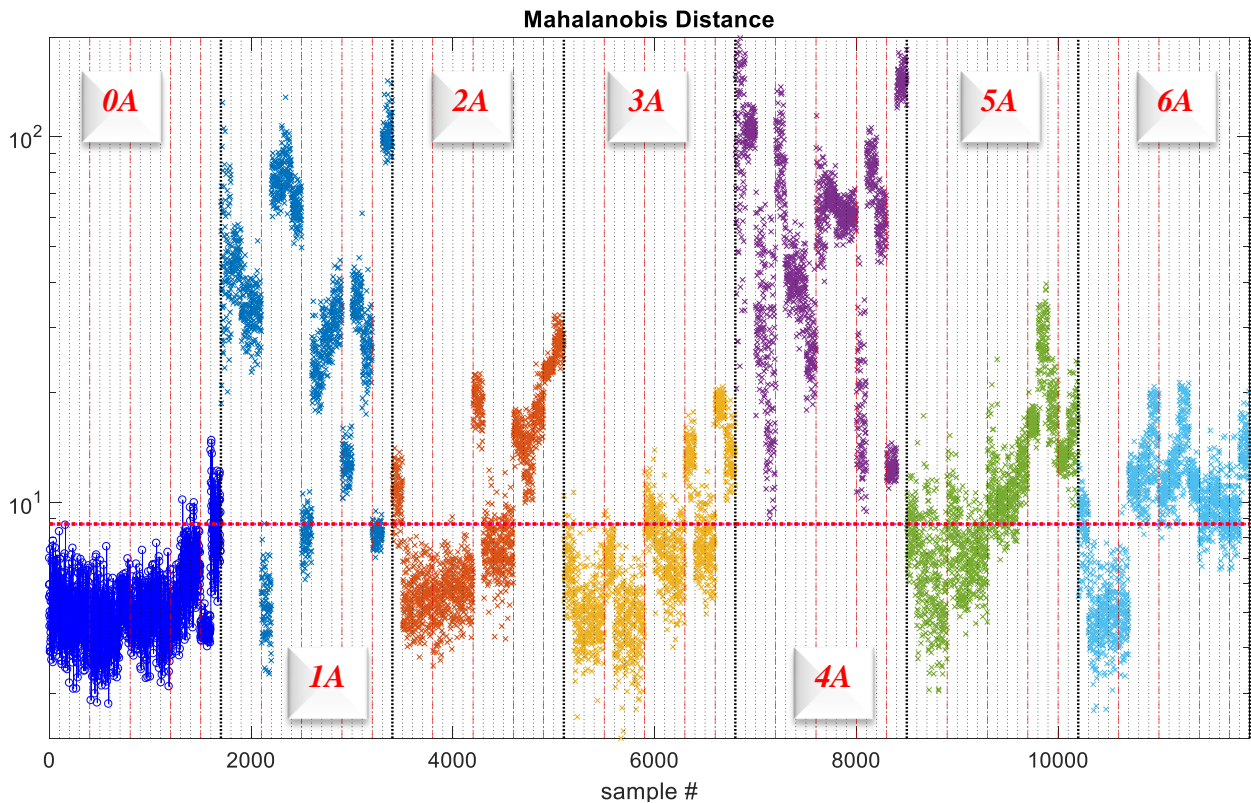
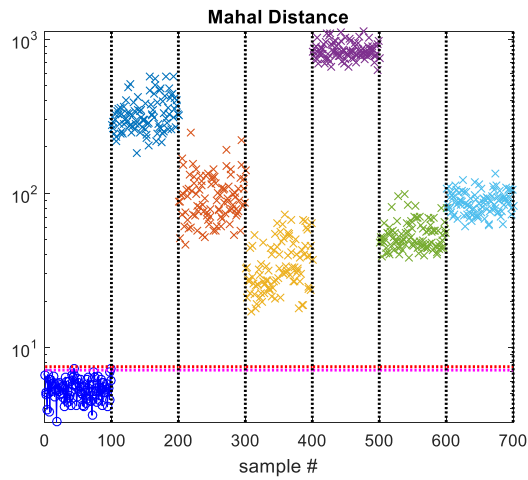


Figure 13: Mahalanobis distance referred to the entire available dataset, without discerning the different operational conditions; in red the threshold: the 99<sup>th</sup> percentile of the maxima distribution from 1000 Monte Carlo repetitions.

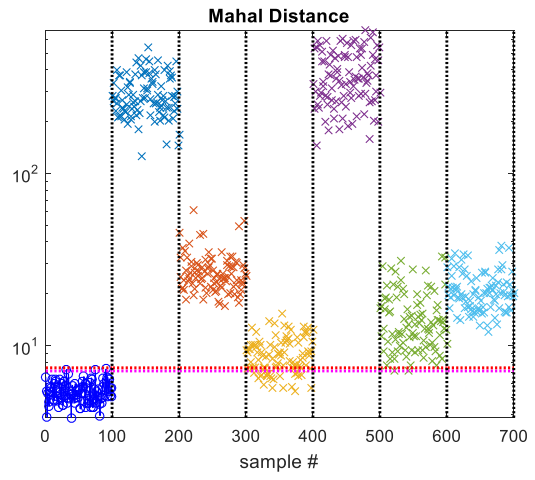
Table 9: False and Missed Alarms in % for the 17 operational conditions, considered independently and compared to their own reference healthy acquisitions (see Figure 14); the 99<sup>th</sup> percentile of the maxima distribution from 1000 Monte Carlo repetitions, used as a threshold, is reported as well.

	FA		MA						MC 99% threshold
	0A	1A	2A	3A	4A	5A	6A		
1	2	0	0	0	0	0	0	7,42	
2	2	0	0	0	0	0	0	7,37	
3	3	0	0	21	0	2	0	7,38	
4	4	0	0	7	0	0	0	7,46	
5	3	0	0	0	0	0	0	7,42	
6	1	0	0	0	0	0	0	7,43	
7	1	0	0	0	0	0	0	7,37	
8	3	0	0	0	0	0	0	7,41	
9	3	0	0	0	0	0	0	7,44	
10	1	0	0	0	0	0	0	7,37	
11	4	0	0	0	0	0	0	7,41	
12	2	0	0	0	0	0	0	7,39	
13	4	0	0	0	0	0	0	7,43	
14	2	0	0	0	0	0	0	7,49	
15	4	0	0	0	0	0	0	7,42	
16	1	0	0	0	0	0	0	7,40	
17	0	0	0	0	0	0	0	7,38	

average threshold: 7,41

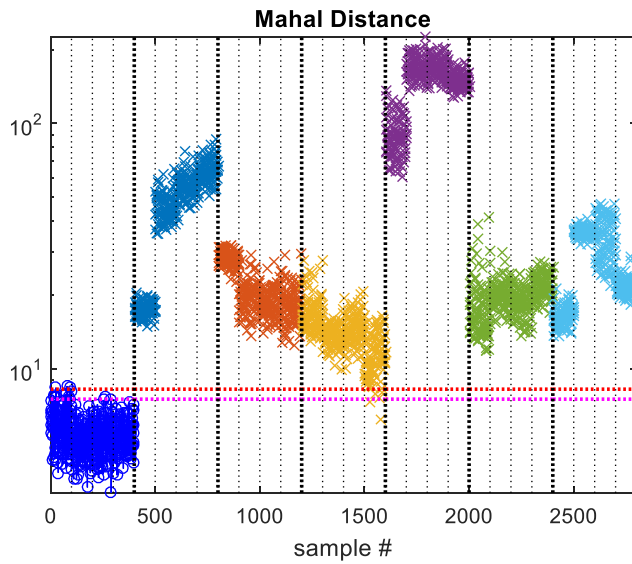


Work condition 12



Work condition 3

Figure 14: Mahalanobis distance for certain operational conditions.



Alarms:	
0A	3 %
1A	0 %
2A	0 %
3A	0,5 %
4A	0 %
5A	0 %
6A	0 %
MC 99% threshold: 8,2	

Figure 15: Mahalanobis distance for operational conditions 9-12 (constant speed 300 Hz).

## 5.1 Endurance

The main aims of this part of the test was to verify how fast the progression of damage is, in standard operative conditions (load and speed) for this type of roller bearing, and to highlight the influence of the lubricant oil. Indeed, the bearing was not brought to failure and the physical variations of the indentation, in both shape and extension (Figure 3), are quite limited. The identification of damage evolution, given such small variation, then results in a challenging task.

It is interesting to propose the same multivariate statistical analysis on the endurance data, to see if it's able to extract information about the 4A damage evolution with prognostics purposes.

A projection on the first principal components is reported in Figure 16. The column on the left is referred to the healthy configuration, on which a first PCA is performed. With this space transformation, the all endurance data result in a single cloud very far from the original 0A and even further from 4A (both taken in the same speed and load conditions of 300 Hz and 1800 N). Conducting a PCA on the first endurance dataset (at a life of 19h) on the contrary, as in the right column of the picture, three different clusters can be highlighted, corresponding to the groups of acquisitions 19-70h (End1), 70-124h (End2), 124-223h (End3). This is due to the mounting and dismounting of the bearing at the scope of inspecting the damage and producing the pictures in Figure 3 c (t=70h) and d (t=124h) respectively.

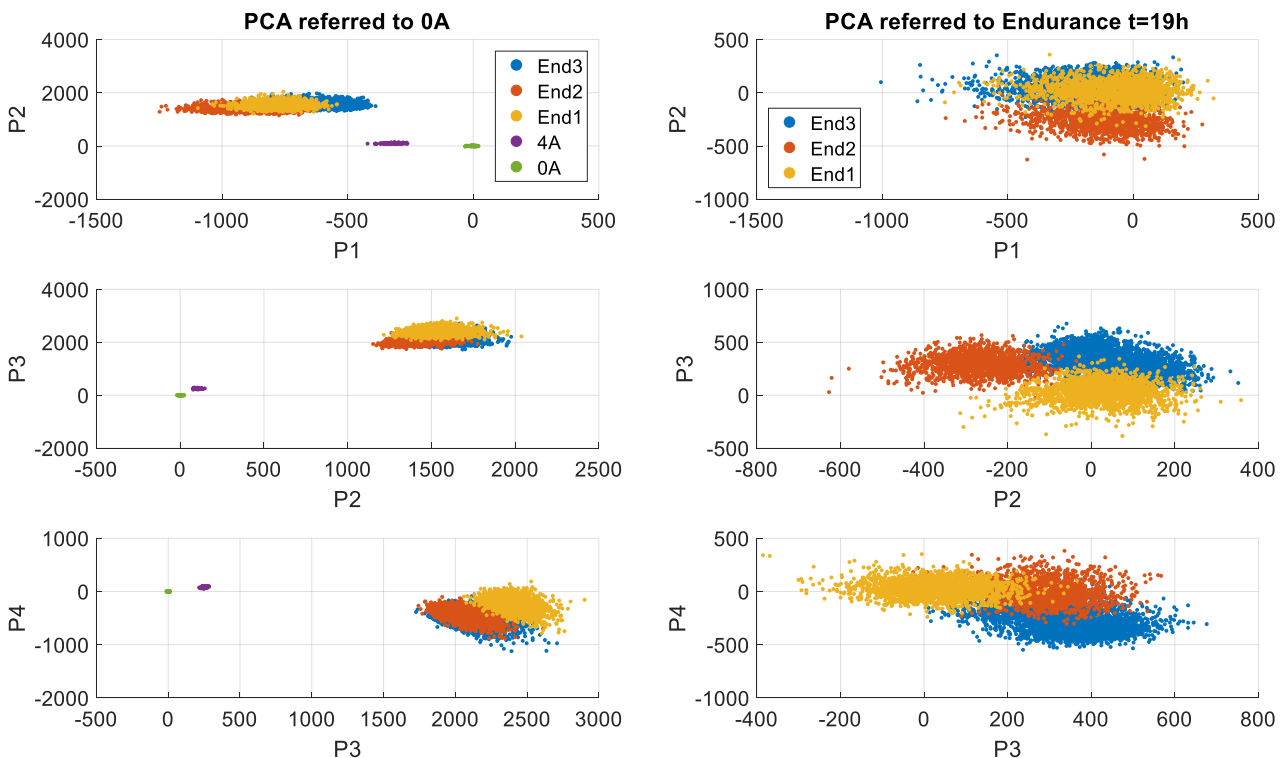


Figure 16: On the left PCA referred to the healthy dataset; on the right PCA based on the first endurance set available, at a lifetime of 19h. The first four principal components are considered to produce the 3 different plots.

Similar conclusion may be deduced using the MA based novelty indices, which can be easily computed similarly to the previous analysis, but which are giving a more complete picture of the condition as no information is neglected.

In this case, focusing on Figure 17, it is possible to understand that, despite the effect of the mounting and dismounting, which in such a small test-rig has a great impact, a deviation of the bearing condition is present. In particular, in the first hours, NIs are quickly increasing but after about 20 hrs (File 038) they stabilize to a lower value.

This seems to be coherent with the practical knowledge according to which the defect widens in time while flattening, so that the detection becomes harder.

Noticeably enough, not only the undamaged state 0A but also the damaged condition 4A are very far from the results of the endurance test. Since no other change in other physical phenomena was recorded between the two sets of acquisitions, this great difference is completely due to the variation of the lubricating oil of the bearings, as mentioned in section 3.2.

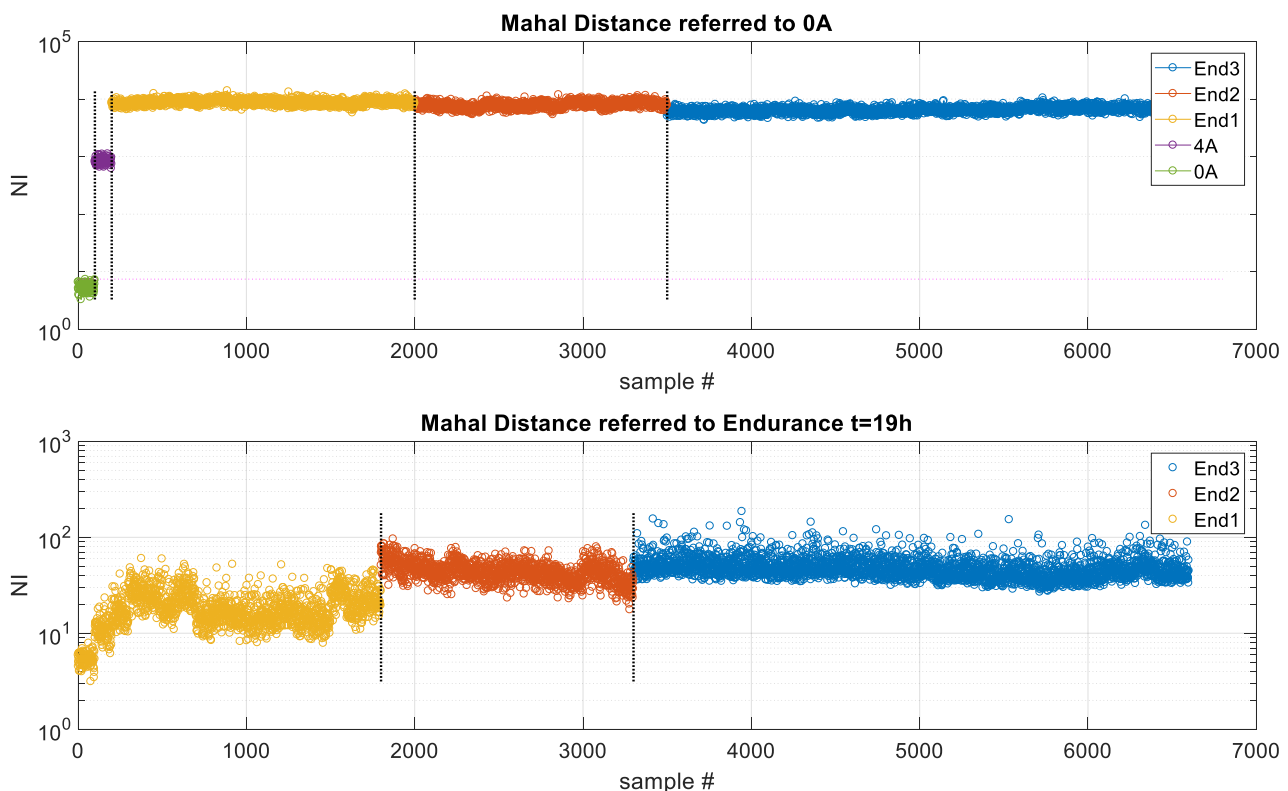


Figure 17: Above, MD referred to the healthy dataset. Below, MD based on the first endurance set available, at a lifetime of 19h. Lifetimes of 70h and 124h (dismounting) are highlighted.

Finally, the damage evolution is probably too weak to be detected using the simple features proposed, which results much more sensitive to the complete disassembly of the bearing and to the lubricant oil variation. This will be an additional challenge to whoever will be interested in analysing the data.



## 6. Conclusions

This paper has been conceived to reach a double scope: to present in details the rig used for testing bearing damage detection algorithms and to share a coherent selection the complete dataset, and to show that also a pure time domain statistical approach can achieve a good level of accuracy to the detection of a damage presence. The latter is also proposed as comparison reference for those wishing to test their own algorithms or novel approaches.

Hence, the first part of the paper is devoted to the description of the rolling bearing testing rig developed at the DIRG lab of Politecnico di Torino, where a high-speed spindle drives a shaft supported by a couple of identical bearings. Different damage conditions have been imposed on one of the bearings, and accelerations have been recorded in different positions and directions, under various load and speed conditions. The evaluation of the local damage on a roller has also been evaluated by monitoring the bearing under the same speed and load conditions for about 230 hours.

A selection of the acquired records can be downloaded from [ftp://ftp.polito.it/people/DIRG\\_BearingData/](ftp://ftp.polito.it/people/DIRG_BearingData/).

In the second part of this paper, several different supervised and unsupervised techniques have been tested to compare their performances in extracting information from the big dataset available, also proposing a procedure of analysis.

As preliminary step, the PCA has been used to visualize the entire dataset in a bi-dimensional plot. Then, 30 univariate ANOVA with the consequent LSD multi-comparison test have been carried out, highlighting that all the channel and all the features are able to give different pieces of information for the different damages, so it is wiser to mix all the available intelligence using multivariate tools rather than neglecting contributions.

The LDA was then tested, pointing out that different speed, load and damage conditions were recognizable even if not in a really effective manner, as they were not linearly separable.

The MD OA has been finally set up to detect deviations from the healthy condition taken as reference.

Different operational conditions have been separated by using different training and comparing the relative damaged sets. The results are more than satisfactory in terms of False and Missed Alarms rate, also considering that the damage severity can be deduced as well.

This holds in stationary conditions, but we could take advantage of the intrinsic ability of MD OA of compensating for simple linear or quasi linear hidden effects, by proposing an analysis on more conditions grouped together.

Unfortunately, when the entire 17 operational conditions are considered altogether, the method

could well recognize only 4A and 1A damages (the more severe), reducing its reliability. But limiting to groups of acquisitions at constant speed, the compensation for a variable load can be performed in a very effective way.

These results are very interesting also in terms of quickness, simplicity and full independence from human interaction, making the MD OA suitable for real time implementation.

## References

- [1] R.A. FISHER, *The Use of Multiple Measurements in Taxonomic Problems*, Ann. Eugen. 7 (1936) 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.
- [2] C.G. Gonzalez, S. da Silva, M.J. Brennan, V. Lopes Junior, *Structural damage detection in an aeronautical panel using analysis of variance*, Mech. Syst. Signal Process. 52–53 (2015) 206–216. doi:10.1016/j.ymsp.2014.04.015.
- [3] A.M. Yan, G. Kerschen, P. De Boe, J.C. Golinval, *Structural damage diagnosis under varying environmental conditions - Part I: A linear analysis*, Mech. Syst. Signal Process. 19 (2005) 847–864. doi:10.1016/j.ymsp.2004.12.002.
- [4] A.M. Yan, G. Kerschen, P. De Boe, J.C. Golinval, *Structural damage diagnosis under varying environmental conditions - Part II: Local PCA for non-linear cases*, Mech. Syst. Signal Process. 19 (2005) 865–880. doi:10.1016/j.ymsp.2004.12.003.
- [5] A. Bellino, A. Fasana, L. Garibaldi, S. Marchesiello, *PCA-based detection of damage in time-varying systems*, Mech. Syst. Signal Process. 24 (2010) 2250–2260. doi:10.1016/j.ymsp.2010.04.009.
- [6] K. Worden, G. Manson, N.R.J. Fieller, *Damage Detection Using Outlier Analysis*, J. Sound Vib. 229 (2000) 647–667. doi:10.1006/jsvi.1999.2514.
- [7] K.I. Penny, *Appropriate Critical Values When Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance*, Appl. Stat. 45 (1996) 73–81. doi:10.2307/2986224.
- [8] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, *The Mahalanobis distance*, Chemometrics and Intelligent Laboratory Systems, 50 (1) (2000), pp. 1-18
- [9] K. Worden, *Structural fault detection using a novelty measure*, Journal of Sound and Vibration, 201 (1) (1997), pp. 85-101
- [10] A. Widodo, E. Y. Kim, J. Son, B. Yang, A. Tan, D. Gu, B. Choi, J. Mathew, *Fault diagnosis of low speed bearing based on relevance vector machine and support vector machine*, Expert Systems with Applications, 2009, Vol 36, ISSN 0957-4174

- [11] W. Martinez, A. Martinez, J. Solka, *Exploratory Data Analysis with MATLAB®*, CRC Press (2008). doi:10.1201/9780203483374.
- [12] I.T. Jolliffe, *Principal Component Analysis*, Second Edition, Springer. 30 (2002) 487. doi:10.2307/1270093.
- [13] Yosef Hochberg, Ajit C. Tamhane, *Multiple Comparison Procedures*, August 2011, ISBN: 978-0-470-56833-0
- [14] J. W. Tukey, *Exploratory Data Analysis*, Addison-Wesley (1977), ISBN 0-201-07616-0
- [15] W.A. Smith, R.B. Randall, *Rolling Element Bearing Diagnostics Using the Case Western Reserve University Data: A Benchmark Study*, MSSP. Vol 64–65 (2015) Pag. 100–131. doi:10.1016/j.ymsp.2015.04.021.

## APPENDIX A

This appendix lists the names of the files that can be downloaded from [ftp://ftp.polito.it/people/DIRG\\_BearingData/](ftp://ftp.polito.it/people/DIRG_BearingData/).

Table A.1: file names for the eight bearings with different damages, from 0A to 6A

1	C0A_100_000_1.mat	C1A_100_000_2.mat	C2A_100_000_1.mat	C3A_100_000_1.mat
2	C0A_100_505_1.mat	C1A_100_502_2.mat	C2A_100_506_1.mat	C3A_100_505_1.mat
3	C0A_100_706_1.mat	C1A_100_702_2.mat	C2A_100_701_1.mat	C3A_100_699_1.mat
4	C0A_100_899_1.mat	C1A_100_898_2.mat	C2A_100_901_1.mat	C3A_100_906_1.mat
5	C0A_200_505_1.mat	C1A_200_502_2.mat	C2A_200_506_1.mat	C3A_200_505_1.mat
6	C0A_200_706_1.mat	C1A_200_702_2.mat	C2A_200_701_1.mat	C3A_200_699_1.mat
7	C0A_200_899_1.mat	C1A_200_898_2.mat	C2A_200_901_1.mat	C3A_200_906_1.mat
8	C0A_300_505_1.mat	C1A_300_502_2.mat	C2A_300_506_1.mat	C3A_300_505_1.mat
9	C0A_300_706_1.mat	C1A_300_702_2.mat	C2A_300_701_1.mat	C3A_300_699_1.mat
10	C0A_300_899_1.mat	C1A_300_898_2.mat	C2A_300_901_1.mat	C3A_300_906_1.mat
11	C0A_400_505_1.mat	C1A_400_502_2.mat	C2A_400_506_1.mat	C3A_400_505_1.mat
12	C0A_400_706_1.mat	C1A_400_702_2.mat	C2A_400_701_1.mat	C3A_400_699_1.mat
13	C0A_500_505_1.mat	C1A_500_502_2.mat	C2A_500_506_1.mat	C3A_500_505_1.mat
1	C4A_100_000_1.mat	C5A_100_000_1.mat	C6A_100_000_1.mat	
2	C4A_100_496_1.mat	C5A_100_498_1.mat	C6A_100_500_1.mat	
3	C4A_100_702_1.mat	C5A_100_700_1.mat	C6A_100_705_1.mat	
4	C4A_100_895_1.mat	C5A_100_900_1.mat	C6A_100_909_1.mat	
5	C4A_200_496_1.mat	C5A_200_498_1.mat	C6A_200_500_1.mat	
6	C4A_200_702_1.mat	C5A_200_700_1.mat	C6A_200_705_1.mat	
7	C4A_200_895_1.mat	C5A_200_900_1.mat	C6A_200_909_1.mat	
8	C4A_300_496_1.mat	C5A_300_498_1.mat	C6A_300_500_1.mat	
9	C4A_300_702_1.mat	C5A_300_700_1.mat	C6A_300_705_1.mat	
10	C4A_300_895_1.mat	C5A_300_900_1.mat	C6A_300_909_1.mat	
11	C4A_400_496_1.mat	C5A_400_498_1.mat	C6A_400_500_1.mat	
12	C4A_400_702_1.mat	C5A_400_700_1.mat	C6A_400_705_1.mat	
13	C4A_500_496_1.mat	C5A_500_498_1.mat	C6A_500_500_1.mat	

Table A.2: file names of the endurance test, with temperature on the bearing and acquisition time, given as an increment from the daily beginning of the test

### Current bearing life: 19 hours

<i>File number</i>	<i>Temperature (°C)</i>	<i>Time from the morning start (h)</i>
007	67.9	1.5
013	74.4	4.5
018 *	74.5	<b>7.0</b>
021	60.6	1.5
027	68.2	4.5
033	69.0	7.5
038	59.1	1.5
044	65.7	4.5
050	66.9	7.5

057	59.2	1.5
063	64.9	4.5
068 *	65.8	<b>7.0</b>
071	59.0	1.5
077	63.9	4.5
083	64.7	7.5
087	59.0	1.5
093	64.2	4.5
099	65.2	7.5

**Current bearing life: 70 hours**

**The bearing is dismounted. Picture in Figure 3.c**

**Current bearing life: 73 hours**

File number	Temperature (°C)	Time from the morning start (h)
101	57.6	1.5
107	64.6	4.5
113	65.1	7.5
118	58.4	1.5
124	63.7	4.5
130	64.3	7.5
134	58.5	1.5
140	62.5	4.5
146	62.8	7.5
149	57.4	1.5
155	62.0	4.5
161	62.7	7.5
163	57.1	1.5
169	61.6	4.5
175	62.4	7.5

**Current bearing life: 124 hours**

**The bearing is dismounted. Picture in Figure 3.d**

**Current bearing life: 127.5 hours**

File number	Temperature (°C)	Time from the morning start (h)
183	55.4	1.5
186	60.8	4.5
189	61.2	7.5
191	57.2	1.5
194	60.9	4.5
197	61.5	7.5
199	58.4	1.5
202	62.0	4.5
205	62.8	7.5
207	57.1	1.5
210	60.4	4.5

213	60.6	7.5
216	56.0	1.5
219	59.9	4.5
222	60.9	7.5
225	56.5	1.5
228	60.2	4.5
230 **	60.7	<b>6.5</b>
231	56.1	1.5
234	60.2	4.5
236 **	60.6	<b>6.5</b>
237	56.1	1.5
240	59.6	4.5
243	60.2	7.5
245	53.4	1.5
248	58.1	4.5
251	58.7	7.5
253	57.3	1.5
256	59.8	4.5
259	59.9	7.5
260	57.1	1.5
263	60.5	4.5
266	60.9	7.5

**Current bearing life: 223 hours**

**The bearing undergoes various undocumented tests with light load.**

**Current bearing life: 232 hours**

**The bearing is dismantled. Picture in Figure 3.e**

---

## **APPENDIX B – The Mahalanobis Distance: Geometric interpretation [3,4,5,8,9]**

The Mahalanobis Distance (MD) is a unitless and scale invariant measure of distance of a point from a distribution. It is based on  $S$ , the variance-covariance matrix estimated from the data. If the number of samples  $n$  is high enough with respect to the space dimension  $d$  (Appendix C), the reported formulation is a meaningful measure of distance.

$$MD(X) = \sqrt{(X - \mu)^T S^{-1} (X - \mu)}$$

This simple formula can be decomposed into 3 main steps, which should follow a centering of the data on the centroid of the reference distribution, which corresponds the healthy condition used in the training phase of this diagnostic-purpose paper.

The first step consists then into a rotation of the feature space to the principal components space which basically comes from a Principal Component Analysis (namely the eigenvectors of  $S$ ).

Later, by rescaling these new variables on their own standard deviation (the square root of the corresponding eigenvector), a new standardized space can be found, in which the new variables will

be an uncorrelated set of linear combinations of the original features.

Table 10: Relevant expressions for standardized PCA

<i>Covariance Unbiased Estimator</i>	$S = \frac{1}{n-1} X'X$
<i>Eigenproblem</i>	$Sa_j = l_j I a_j$
<i>Normalized Eigenvectors</i>	$w_j = \frac{a_j}{\sqrt{l_j}}$
<i>Transform to standardized PCs space</i> ( <i>Z</i> : principal scores)	$Z = XW$

The last step corresponds to an additional non-linear transform of such a space, simply performed by squaring its components. This is the only difference comparing MD approach to PCA.

The squared Mahalanobis distance is then the distance from the origin in this squared, standardized, centered plane (basically the sum of those squared components) and corresponds to a Eulerian distance on the standardized principal space.

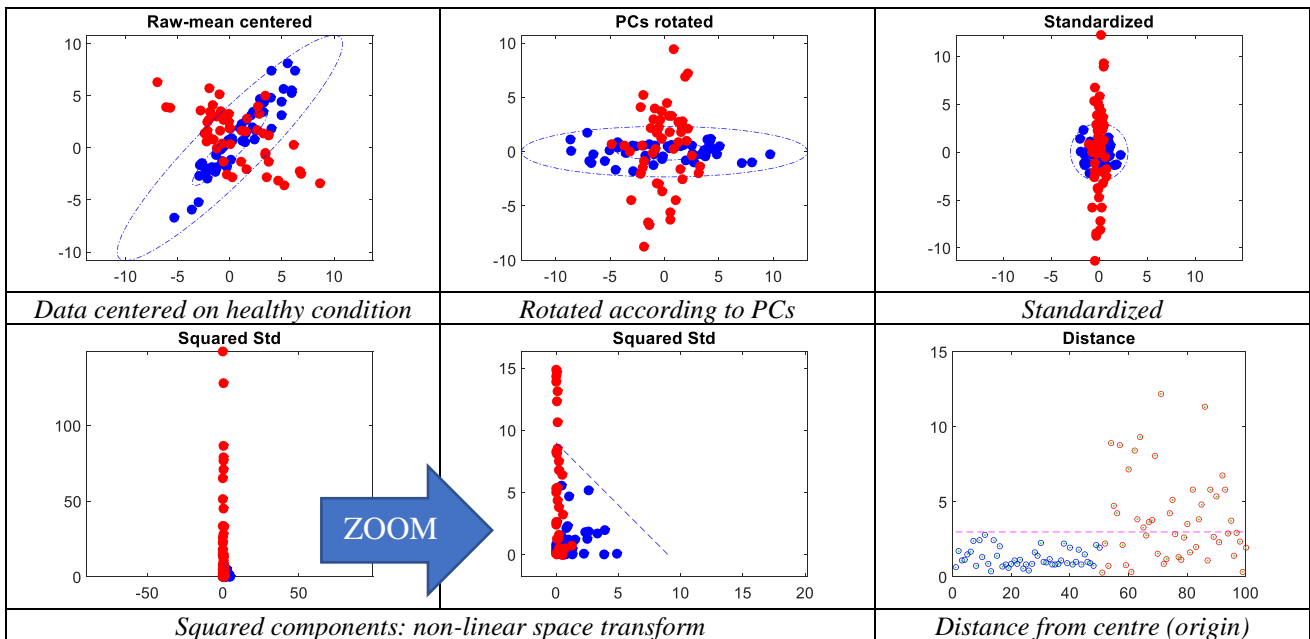


Figure 18: Mahalanobis equivalent procedure on a 2D simplified plane, for 2 simulated normal classes (blue: healthy, red: damaged).

*Do note that the the Centering and Standardization of the space is unique and based on the healthy reference. All the damaged acquisitions will be projected in this plane.*

### APPENDIX C – The Curse of Dimensionality: sample numerousness effect

In general, the main problem of statistical based algorithms is the reliability of the estimated parameters. Many of the proposed methods rely on the unbiased covariance estimator

$S = \frac{1}{n-1} X'X$ . In this case the estimation error will be higher if the number of samples is not high



enough according to the dimension of the considered feature space, as a side effect of “the curse of dimensionality”. A method to quantify this error is here proposed via Monte Carlo repetitions.

An insight of the problem of interest in this paper ( $n = 100, d = 30$ ) is then obtained repeating 1000 Monte Carlo simulations with  $n = 100$  samples drawn from a standard normal of increasing dimensionality,  $d$ .

Theoretically, being the exact covariance matrix equivalent to the identity, the Mahalanobis distance should always be equivalent to the Eulerian distance. Unfortunately, the MD is based on the estimate  $S$  of the covariance matrix, and this will introduce some differences among MD and ED, which will be only due to the estimation error. It is sufficient then to compare the results of the two procedures to get an idea of the accuracy of the covariance matrix estimate.

In order to simplify the comparison, the analysis is done focusing on the maximum distances in the sets, averaged on 1000 repetitions, also computing  $\pm\sigma$  limits. It is wise to remember that the ED trend could also be deduced by Extreme Value Theory as the  $1 - 1/n$  critical value of the a  $\chi^2_{(d)}$  distribution, as the ED (and MD too) are square roots of sum of squares [7].

It can be noticed that the Maxima EDs increase with dimension  $d$ , as an expected effect of the “curse of dimensionality”; a threshold should then follow this behavior so as not to affect the outlier detection.

From Figure 19, it can be seen that, for  $n = 100$ , the two distances start diverging for space dimensions bigger than 10. Indeed, when the numerosness  $n$  is too low, the samples are not “filling” the entire space, and the estimate of  $S$  will be poorer, compromising the accuracy of the MD.

Focusing on  $d = 30$  (of interest in this paper), it can be noticed that the MD and ED confidence intervals are still intersecting, pointing out that the MD computation is anyway reliable from a statistical point of view.

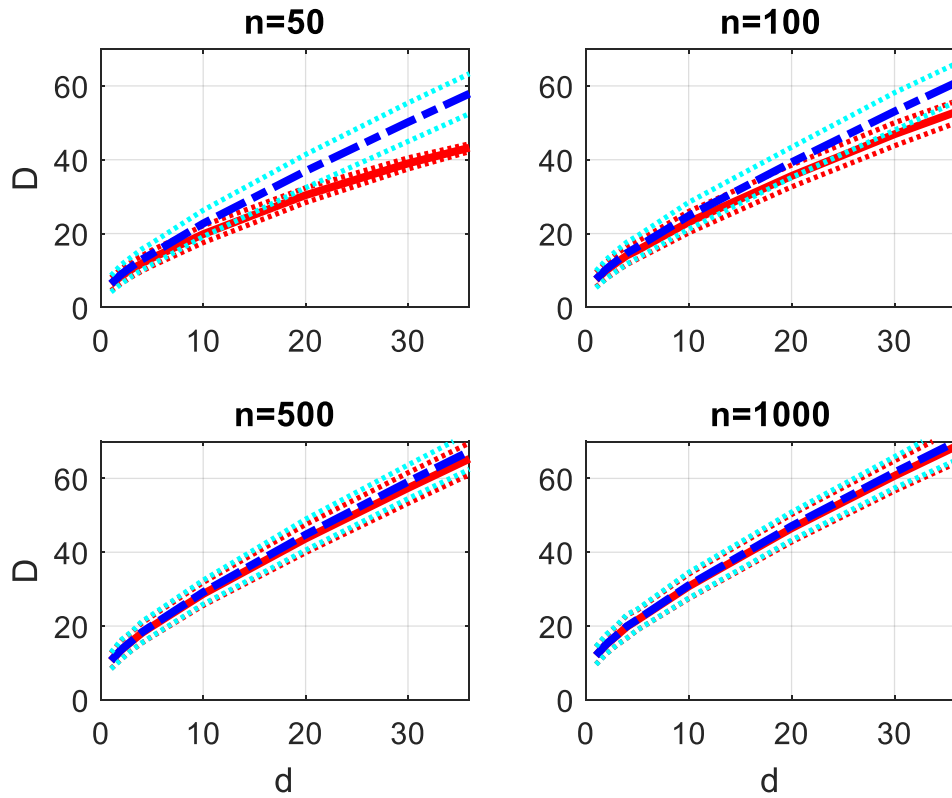


Figure 19: Averages of the Mahalanobis distances (red) and Eulerian distances (blue) for the Maxima with respect of the space dimension  $d$ , in 4 sets with growing number of samples  $n$ , considering 1000 Monte Carlo repetitions;  $\pm\sigma$  confidence intervals of the estimations are also given in cyan (Eulerian) and magenta (Mahalanobis).