

On the use of PCA for Diagnostics via Novelty Detection: interpretation, practical application notes and recommendation for use

Alessandro Paolo Daga¹, Alessandro Fasana¹, Luigi Garibaldi¹, and Stefano Marchesiello¹

¹ *Politecnico di Torino, Dipartimento di Ingegneria Meccanica e Aerospaziale*

C.so Duca degli Abruzzi, 24, 10129 Torino, Italy

alessandro.daga@polito.it

ABSTRACT

The Principal Component Analysis (PCA) is the simplest eigenvector-based multivariate data analysis tool and dates back to 1901 when Karl Pearson proposed it as a way for finding the best fitting d-1 hyperplane of a system of points in a d-dimensional (Euclidean) space.

Over the time, the PCA evolved in different fields with several different names and with different scopes, but, in its essence, it is always an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

Generalizing Pearson's purpose, the knowledge derived by such an analysis is mostly used to find a subspace which effectively and efficiently summarizes the original system of points by losing a minimum amount of information.

In the field of Diagnostics, the fundamental task of detecting damage is basically a binary classification problem which is in many cases tackled via Novelty Detection: an observation is classified as novel if it differs significantly from other observations. Novelty can, in principle, be assessed directly in the original space, but the effectiveness of the estimated novelty can be improved by taking advantage of the PCA.

In this work, the traditional PCA will be compared to a robust modification that is commonly used in the field of diagnostics to face the issue of confounding influences which could affect the novelty-damage correspondence. Comparisons will be made to shed light on the main misleading aspects of PCA, and finally, define a unique, theoretically justified procedure for Diagnostics via Novelty Detection.

1. INTRODUCTION

Novelty Detection (ND) is a well-known mechanism in data science. It is meant for the identification, from measured data, of patterns which are different from those characterizing the normal condition. Upon the removal of all the confounding influences (e.g., variations in the operational and environmental variables), ND is widely used in both Structural Health Monitoring and Condition Monitoring for diagnostics of both buildings and industrial machines [Worden, Manson, Fieller (2000), Yan, Kerschen, De Boe, Golinval (2005), Bellino, Fasana, Garibaldi, Marchesiello, (2010), Deraemaeker & Worden (2018), Daga, Fasana, Marchesiello, Garibaldi (2017 & 2019), Castellani, Garibaldi, Daga, Astolfi, Natili (2020)]. In particular, this work will deal with Vibration Monitoring, which exploits mechanical vibration for conveying diagnostic information out of the object of the analysis (e.g., a machine or a structure), and will be limited to Damage Detection, the first fundamental piece of diagnostics devoted to recognize anomalous conditions related to the presence of damage.

Damage Detection will be regarded as a Pattern Recognition problem, and we will deal with the important tasks of data pre-processing, features extraction and pattern processing via Principal Component Analysis and related algorithms for condition assessment and alarm triggering [Farrar & Doebling (1999)]. Statistical, Data-Driven, model will be used for "learning" information about the state of health of the machine. This is not a novel problem and is nowadays a common subject of Statistical and Machine Learning courses.

Science as revolving around measurements susceptible to comparison is a point of view shared by many scientists and philosophers of all times and places. And comparability of measurements requires some way of evaluating the

Alessandro Paolo Daga et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 United States License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

uncertainty in their values, leading to the nowadays called statistical reasoning [Stigler (1986)].

Historically, the turning point of statistical reasoning can be considered the introduction of Least Squares Regression. Even if the mathematical framework was already set by Laplace, Legendre and Gauss at the beginning of the 19th century [Stigler (1986), Nievergelt (2000)] and successfully applied in astronomy and geodesy, it is only after Galton and his pupil Pearson in the late 1800s that the regression takes its name and start to be used for applied statistics in biology and genetics (actually, with meaning of “regression toward the mean”, for justifying that “extreme” parents originate “less extreme” sons) [Stanton (2001)].

Least Squares literally means “minimization of the sum of squared deviations” and is at the present time a standard approach to approximate the solution of overdetermined systems meant for data fitting. Which means, to extract information from noisy measured data.

Nevertheless, this method shows a limit: it regresses an uncertain dependent variable over an independent variable assumed exact (i.e., in 2D, the deviations or residuals are computed as “vertical distances” by subtracting the measured dependent variable from its fitting curve for given values of the independent variable).

To overcome this issue, errors-in-variables models were introduced, leading to Orthogonal Regression, Deming regression and, more in general, Total Least Squares method [Van Huffel & Vandewalle (1991)]. As for Least Squares Regression, the mathematical framework was introduced way before its spread in statistics. The original model for Orthogonal Regression was proposed by Adcock in 1878 and was based on the idea of computing the residuals as the orthogonal distance measured points from the fitting curve. This was later generalized by Kummell in 1879 for a generic non-orthogonal distance. However, their ideas remained largely unnoticed for more than 50 years, until they were revived by Koopmans in 1937 and later propagated even more by Deming in 1943 [Cornbleet & Gochman (1979)].

The last historical piece of information regards Kendall in 1957, Hotelling in 1957 and Jeffers in 1967 who re-proposed the idea of using Principal Components in regression (i.e., Principal Component Regression) [Jolliffe (1982)]. But the Principal Components Analysis (PCA) was developed by Pearson (1901) with the precise scope of finding the line of best fit for a bidimensional point cloud, to give a mathematical form to the ideas of Galton (1886) who was solving such a problem by intuition, hand-drawing such fitting line on plots such as the one reported in Figure 1.

In practice, PCA is a search for a linear transform to bring data into a new coordinate system of uncorrelated variables: a “best fitting” line is generated to minimize the average squared distance from the points to the line; the next best-

fitting line can be similarly found from directions perpendicular to the first. By repeating this process, an orthogonal basis can be found, whose basis vectors can be referred to as Principal Components (PC). As a consequence, the PCs are ordered according to their explained variability, so that the first elements are often sufficient to picture the main information contained in the data, leading to a good reduced dimensionality approximation of the whole dataset (exactly as a regression line can be considered a good approximation for a 2D point-cloud of 2 variables showing a high linear correlation i.e., a Pearson correlation coefficient far from 0).

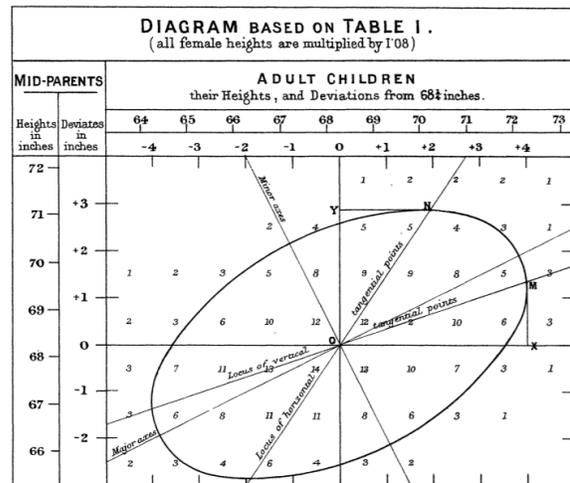


Figure 1. Galton (1886) handmade Orthogonal Regression

From a mathematical point of view, the PCA is the simplest of the true eigenvector-based multivariate analyses. In fact, the PCA linear transform, which corresponds to a rotation of the original coordinate system, can be found by eigen-decomposition of the covariance matrix of a dataset [Jolliffe (2002)] and corresponds to the matrix containing in its columns the eigenvectors of the covariance matrix ordered so that the corresponding eigenvalues are sorted in descending order. PCA is then an unsupervised learning algorithm relying entirely on the data themselves (i.e., data-driven).

Usually the original data are normalized before performing the PCA. The normalization consists of mean centering each variable. Some fields use in addition a normalization of each variable’s variance to 1 (i.e., a standardization of each variable to its z-scores). Pre-whitening (i.e., the operation of removing the correlation from data which then shows an Identity covariance matrix) is obviously not an option, but PCA can be used to find a whitening matrix for a dataset.

Starting from these considerations and adding the hypothesis of Normality, PCA can be used to give insight into the Mahalanobis Distance (MD), a multi-dimensional generalization of the 1-D distance from the mean,

adimensionalized and given in terms of number of standard deviations away from the mean. The MD, in fact, can be proved to “contain” PCA, so that an algorithm for Novelty Detection naturally arises from the unitless, scale-invariant and correlation-based results of MD.

The relationship of PCA and MD is well recognized by statisticians, but although a few authors have commented on it in the condition monitoring literature (e.g., Deraemaeker & Worden 2018), this has been rarely exploited.

This paper aims then to take advantage of such a relationship to derive an alternative formulation of the MD (similarly to that proposed by Brereton & Lloyd 2016 in the chemometrics field for improving Linear Discriminant Analysis) which opens to the evaluation of a novel Novelty Index which is possibly more robust to confounding influences. The methodology is described in next section and will be compared to the PCA-based algorithm of Yan et al. (2005), which can be considered as a benchmark for the condition monitoring community. At first, the algorithms will be tested on a synthetic dataset obtained from a 31-Degrees of Freedom (DOF) Bridge Model for Structural Health Monitoring similar to the model proposed in Yan et al. (2005). A simple assessment on the performance of both natural frequencies and time-signal statistics as features is also proposed to justify the use of common time-signal statistics for the diagnosis of high speed bearing from acquisitions taken on the test rig described in Daga et al. (2019).

2. METHODOLOGY

In this section, the main tools useful to develop the proposed methodology are explained in details, based on Jolliffe (2002).

2.1. Principal Component Analysis

Let us organize a d -dimensional dataset of n observations into a matrix $X_0 \in R^{d \times n}$. Let us preprocess this dataset by centering it (i.e., by removing the mean value of each variable, stored in the rows of the matrix) so as to produce the data-matrix X .

An unbiased estimator for the covariance of the dataset is:

$$S = 1/(n - 1) XX' \quad (1)$$

PCA corresponds to the solution of the eigenproblem

$$S V = V \Lambda \quad (2)$$

where V is the orthogonal matrix whose columns are the d eigenvectors v_j while Λ is the diagonal matrix of the d eigenvalues λ_j of the matrix S , sorted to have descending magnitude.

The matrix V can be used as a linear transform to decorrelate the dataset X , that is, to rotate the coordinate

system toward that identified by the eigenvectors of matrix S :

$$Z = V'X \quad (3)$$

If the eigenvectors in V are normalized to have unit length ($v_j'v_j = 1$), the transform is a pure rotation, and it can be proved that $\sigma_j^2 = var(z_j) = \lambda_j$. Namely, the diagonal Λ is the covariance matrix of Z . This corresponds to traditional PCA.

Adopting different normalizations is anyway possible. For example, V becomes the whitening matrix W if its eigenvectors are normalized so that $v_j'v_j = \lambda_j$. In this case $var(z_j) = 1$ so that the covariance matrix of Z_W is the identity matrix I . In this case, PCA is said to “whiten” the data, as it produces uncorrelated standard scores:

$$\begin{aligned} Z_W &= W'X = \Lambda^{-1/2}V'X \\ &= \Lambda^{-1/2}Z \end{aligned} \quad (4)$$

2.2. Principal Component Regression as PCA dimensionality reduction

Focusing on PCA Orthogonal Regression (PCA-OR), the direction given by the first eigenvalue is considered to correspond to the regression line, so that the regressed points $X_1 \in R^{d \times n}$ standing on this line can be simply found by projecting back the first PC scores alone. On the contrary, the remaining PC scores can be seen as the orthogonal residuals of the regression.

Considering that the scores for the j -th principal components can be found as:

$$\begin{aligned} z_j &= v_j'X \\ &= v_{j1}x_1 + v_{j2}x_2 + \dots + v_{jd}x_d \\ &= \sum_{k=1}^d v_{jk}x_k \end{aligned} \quad (5)$$

The regressed points are simply:

$$X_1 = v_1 z_1 = v_1 v_1'X \quad (6)$$

While the residuals corresponding to the scores of the $L = d - 1$ components other than the first can be found by removing the first eigenvector from matrix V to form a V_L such that:

$$Z_L = V_L'X \quad (7)$$

The PCA-OR is visualized in Figure 2.

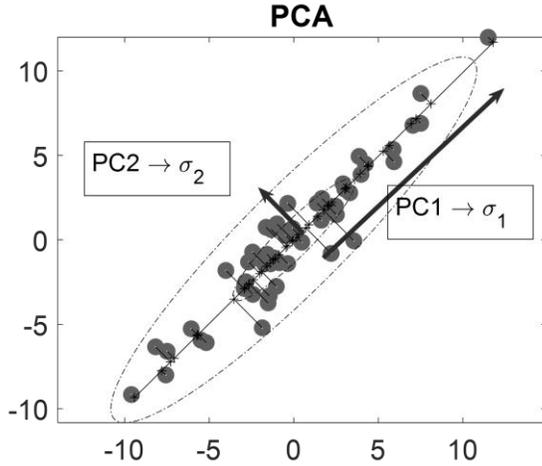


Figure 2. Principal Component Orthogonal Regression

2.3. PCA geometric interpretation

The geometric interpretation of PCA is related to the fact that an ellipsoid centred in the origin can be associated to any positive definite matrix such as the covariance S . Its equation can be proved to be:

$$X'S^{-1}X = 1 \quad (8)$$

The eigenvectors of S^{-1} define then the principal axes of the ellipsoid while the eigenvalues of S^{-1} are the reciprocals of the squares of the semi-axes length. This can be verified remembering that the eigenvectors of S^{-1} are the same as the eigenvectors of S and the eigenvalues of S^{-1} are the reciprocal of those of S . Indeed, using the inverse transformation $X = VZ$, one can get:

$$\begin{aligned} X'S^{-1}X &= Z'V'S^{-1}VZ \\ &= Z'\Lambda^{-1}Z = \sum_j z_j^2/\lambda_j = 1 \end{aligned} \quad (9)$$

which is the equation of an ellipsoid whose half principal axes are $\sqrt{\lambda_j} = \sigma_j$ long.

2.4. The Mahalanobis Distance and the proposed Reduced Mahalanobis Distance

The squared Mahalanobis Distance of a set of data from their centroid is given by:

$$MD^2 = X'S^{-1}X \quad (10)$$

It is interesting to note that the expression of the Probability Density Function of a generic Multivariate Normal Distribution is given by

$$f(X) = \frac{\exp\left(-\frac{1}{2}MD^2\right)}{\sqrt{(2\pi)^d|S|}} \quad (11)$$

From which it is clear that the MD is a multi-dimensional generalization of the 1-D distance from the mean given as number of standard deviations (i.e., the z-score).

This can be seen mathematically by using the PCA linear transform from X to Z :

$$\begin{aligned} MD^2 &= X'S^{-1}X = Z'V'S^{-1}VZ \\ &= Z'\Lambda^{-1}Z = \sum_j \frac{z_j^2}{\lambda_j} = Z'_W Z_W \\ &= \sum_j z_{Wj}^2 \end{aligned} \quad (12)$$

This proves that the squared Mahalanobis distance corresponds to the sum of squares of the whitened dataset. Hence, removing the first whitened component from the sum corresponds to merging PCA-OR and MD through PCA-whitening: this reduced MD (rMD), in fact, corresponds to the sum of the squared residuals from the orthogonal regression.

$$rMD^2 = \sum_{j=2}^d z_{Wj}^2 \quad (13)$$

It is also interesting to note that in the MD, as can be seen in eq.(12), the squared principal scores are normalized by the corresponding eigenvalues before the sum takes place.

By changing the initial value j , different rMDs can be produced.

2.5. The naturally arisen Novelty Detection and the confounders

The Mahalanobis Distance (MD) is widely used as a measure of novelty as it is unitless and scale invariant and takes into account the correlations of the multivariate data set. The idea to assess novelty is simple:

- during the training, the dataset X_{0h} corresponding to a healthy condition is used to compute the healthy mean vector μ_h and covariance matrix S_h
- the MD is used to produce normal Novelty Indices NI_h which will be used to find a suitable threshold Th limiting the number of False Alarms (i.e., limiting the number of NI_h exceeding such threshold)

$$NI_h = \sqrt{(X_{0h} - \mu_h)'S_h^{-1}(X_{0h} - \mu_h)} \quad (14)$$

- in operation, when a new dataset X_{0N} is tested for novelty, the derived eq.(14) will be used to obtain NI_N . These NIs will be compared to the previously defined threshold Th so that the acquisitions whose NI_N is exceeding the threshold will be considered an anomaly.

It is important to highlight that, if no confounding influence is assumed to affect the acquisition, damage will be the only

possible source of novelty left, so that the anomalous data-points can be considered symptoms of damage.

The MD NIs are also robust to strong quasi-linear confounding effects. In fact, if a strong confounder effect is present, this will be probably pictured by the first component. Nevertheless, the first PC scores will be normalized against a large variance before being summed up to form MD^2 . Hence, an automatic compensation is performed by MD NIs.

Anyway, if one is sure that the first PC is picturing only the confounder, the reduced MD can be used to compute NIs which are completely insensitive to such direction. This can be proved using a simple synthetic dataset obtained by drawing samples from known distributions, as in Figure 3.

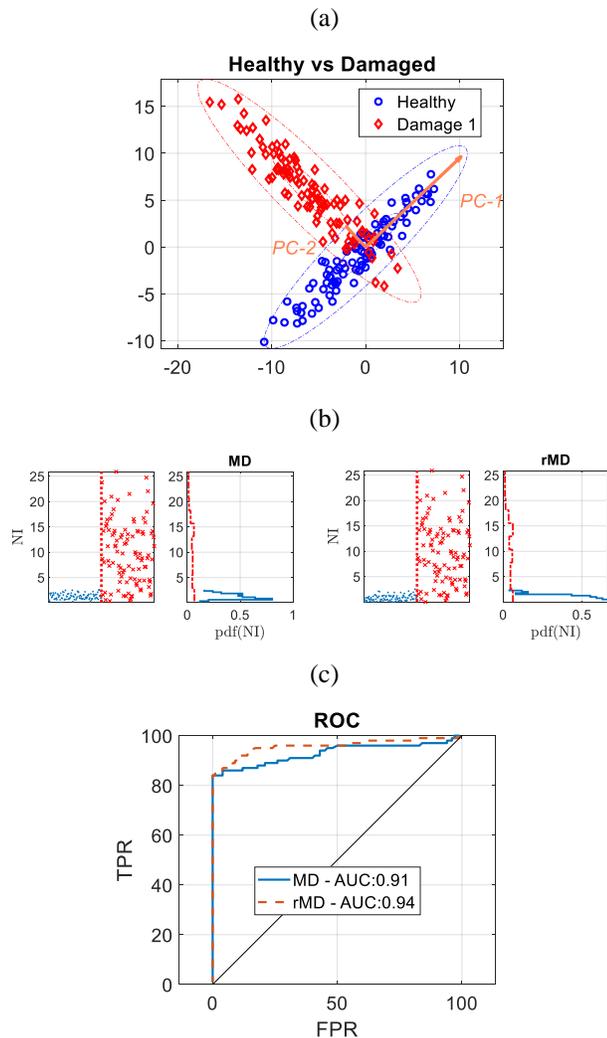


Figure 3. Synthetic example using bivariate normal distributions. (a) Draws from two distributions, the blue one is assumed healthy; (b) Mahalanobis NI vs Reduced Mahalanobis NI computed along the PC-2 direction only; (c) ROC with corresponding AUC (normalized to 1).

From Figure 3 it can be noticed that, if the damage develops along the second principal component, the reduced Mahalanobis distance dropping the information along PC-1 (a confounding factor) performs better than the traditional MD in terms of Receiver Operating Characteristic (ROC) of the corresponding Novelty Indices. This is better highlighted computing the Area Under the Curve (AUC) normalized to 1 (i.e., 0.94 against 0.91).

Obviously, in real-life applications the information of the direction of damage evolution in the feature space is unknown. In this case it is advisable to start with the full MD and then test the behavior of rMD by gradually removing components so as to tune the algorithms for the best performance.

2.6. Preprocessing

As already said in the introduction, data are commonly normalized for PCA. This is also the case of MD-NI and rMD-NI, even if one has to bear in mind that data centring will always occur on the same reference mean vector μ_h .

It is not common to standardize each variable before PCA. In any case, if the variables are in different units and can assume very different values, showing then very different variances, a standardization can be suitable, otherwise the first PC will practically correspond to the largest variance measurement, the second to the second largest variance measurement and so on. This corresponds to a PCA performed on the correlation matrix.

Obviously, it is meaningless to perform PCA on whitened data (i.e., with Identity covariance matrix). For such cases, the MD directly corresponds to the Euclidean distance of a point from the centroid (i.e., the sum of squared deviations from the mean over the different variables).

2.7. Considerations about the numerousness of the training set

As the proposed methodology is based on estimates of the mean and of the unbiased covariance matrix, the numerousness of the training set become relevant for ensuring a high enough statistical confidence of the estimates.

In order to assess the effect of the numerousness n of the training set for a dimensionality $d = 6$, which is of interest in this work, and $d = 12$, a Monte Carlo simulation is set up. After the consideration that the MD directly correspond to the Euclidean distance of a point from the mean if the original dataset shows identity covariance matrix, a draw from a multivariate Gaussian distribution (i.e., a normal distribution with zero mean and identity covariance matrix) of dimension d is repeated 1000 times for an increasing numerousness of samples. Taking the mean difference of MD and Euclidean distance, and the standard deviation of the difference, the graph in Figure 4 is produced.

As can be noticed, in both cases the knee of the graph is for a sample size $n = 100$, which proves to be enough for producing reliable MDs.

2.8. Benchmark method for condition monitoring

In the condition monitoring field, the paper of Yan et al. (2005) can be considered as a benchmark for PCA-based methods. The algorithm can be summarized with few relevant steps:

- Eigen-decomposing to find the transform matrix T : PCA is performed on the non-normalized dataset and the first m eigenvalues (i.e., column vectors) are used to generate the matrix T

$$X_0 X_0' = U \Sigma U' \quad (15)$$

$$T = [u_1 \dots u_m] \quad (16)$$

- Mapping, reducing the dimensionality and re-mapping to the original space for error computation

$$\widehat{X}_0 = T' T X_0 \quad (17)$$

$$E = X_0 - \widehat{X}_0 \quad (18)$$

- Computing the NIs in one of the two proposed ways, Euclidean or Mahalanobis based

$$NI_k^E = \|E_k\| \quad (19)$$

$$NI_k^M = \sqrt{E_k' S^{-1} E_k} \quad (20)$$

It is relevant to point out that in this methodology the original dataset is not normalized by purpose, so that the first PC always points to the centroid of the point-cloud. This ensures a sort of robustness, nevertheless, if the true PC-1 is not actually aligned to the expected, it can be detrimental. Secondly, for a better comparison with the proposed MD based algorithms, the Mahalanobis NI^M will be used in this work. Finally, it is common to set $m = 1$ considering that only one confounding factor (e.g., the temperature) is usually predominant. The authors, however, states that the selection of an appropriate dimension m is not so critical and when the number of environmental factors is not known a priori or is difficult to find by observing the eigenvalues, choosing a series of order m for verification may be considered.

In general, in order to select the number of PCs to get rid of, a clever way is to look for the knee in the eigenvalues spectrum [Cattel (1966)]. For the sake of code automation, both the selection of m for the benchmark method and of j for the proposed rMD method are, in first approximation, computed by counting the components explaining more than a percentage P of the overall variability (i.e., the components whose eigenvalues normalized over the sum of all the eigenvalues are higher than $P\%$ are automatically neglected, with P usually between 10 and 25%).

3. TEST ON DATA FROM A 31-DOF BRIDGE MODEL FOR STRUCTURAL HEALTH MONITORING WITH VARIABLE TEMPERATURE

In Yan et al. (2005) the here considered benchmark method is tested on the model of a 3-span bridge (Figure 5). In this work, a simplified 31-DOF scaled model of the same bridge is proposed. The same temperature gradients are applied to the structure (i.e., a linear gradient generated by letting the temperature of the left side – lowT – linearly varying from -15 to 15°C while the temperature of the right side is linearly changing from -15 to 45°C), and similar relations are used to model the effect of temperature on the elastic moduli of iron and concrete (and hence on the stiffnesses).

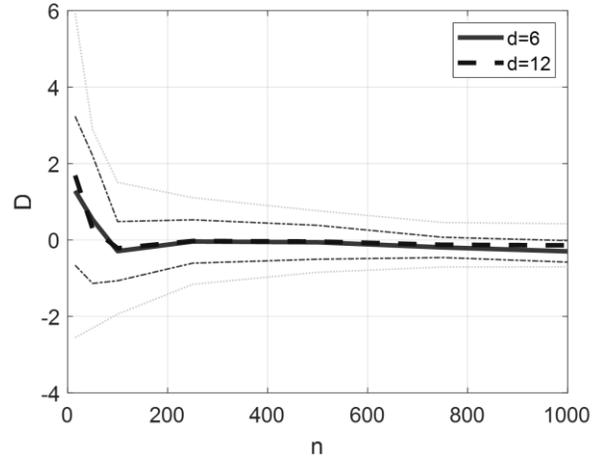


Figure 4. Averages of the difference D of the estimated Mahalanobis distances (continuous) and Euclidean distances (dashed) for n draws from a 6-dimensional multivariate Gaussian, considering 1000 Monte Carlo repetitions. $\pm\sigma$ confidence intervals of the estimated error are also given (d=12 – dotted and d=6 – dash-dotted).

The natural frequencies are computed for different temperatures conditions, and the parameters are tuned to produce results similar to those shown in Yan et al. (2005). In particular, the values of the first six natural frequencies are reported in Figure 7 for the healthy case and for a simulated damaged case, featuring stiffness reductions in the central region of each span.

Of particular interest is the plot of the natural frequencies normalized against the first frequency, which highlights a quasi-linear relationship among these features. Figure 6 visualizes the weak nonlinearity of the natural frequencies induced by the variation of the temperature.

In addition to Yan et al. (2005), this work was also interested into the comparison of the performance of the natural frequencies as features (i.e., frequency features) against other common time-series statistical features (i.e., time features). Hence, the response of a single DOF to a

random force applied in the middle of the bridge was also produced considering a proportional damping producing damping ratios bounded in the range 1-3,5%.

A Monte Carlo simulation was set up for aggregating information from 10 repetitions with a random forcing term, increasing the reliability of the features computed from the time response. Six common time-series statistical features were then extracted: the root mean square value, the skewness, the kurtosis, the crest factor, the maximum of the absolute value and the mean of the root of the absolute values. The resulting average values of the MC repetitions are reported in Figure 8.

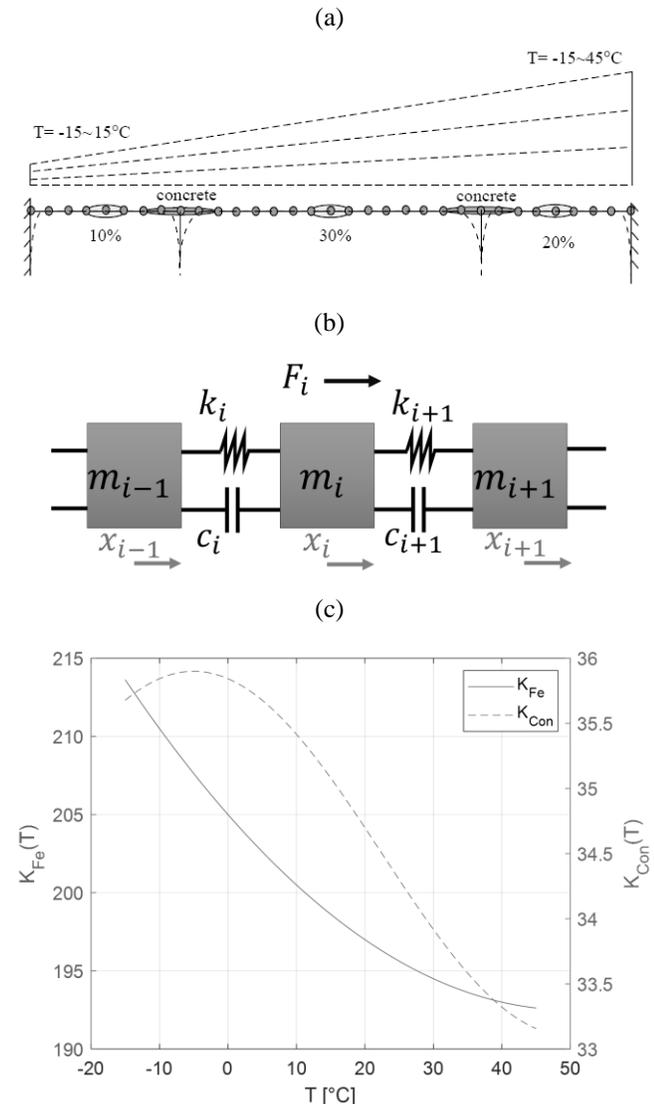


Figure 5. (a) Yan et al. (2005) 3 span bridge as a 31-DOF scaled model with a representation of the temperature gradient, the indication of the material and of the damage. (b) the m-c-k model for the i -th mass. (c) the stiffnesses as a function of temperature.

To faithfully reproduce the analysis in Yan et al. (2005), 5 particular temperature conditions will be considered, as reported in Table 1. Per each condition, the model estimated features will be repeated 100 times while being corrupted by random normal noise with standard deviation of 5% and 1% of the absolute value of the feature. Both the healthy set and the damaged set will be composed by 500 points at different temperature conditions, so that the temperature will play as a confounder for the diagnosis.

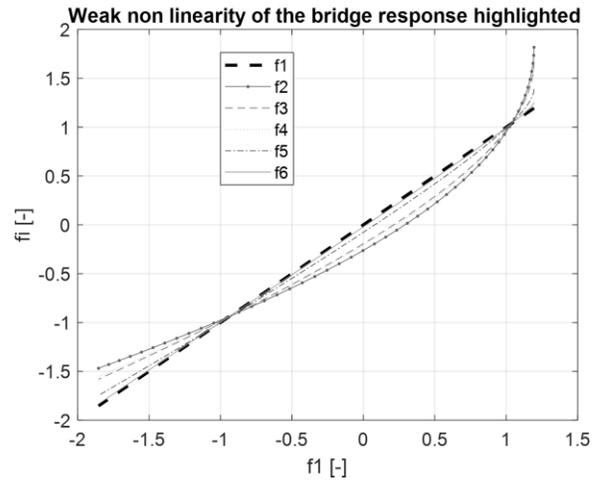


Figure 6. The natural frequencies normalized against the first frequency highlight a quasi-linear relationship

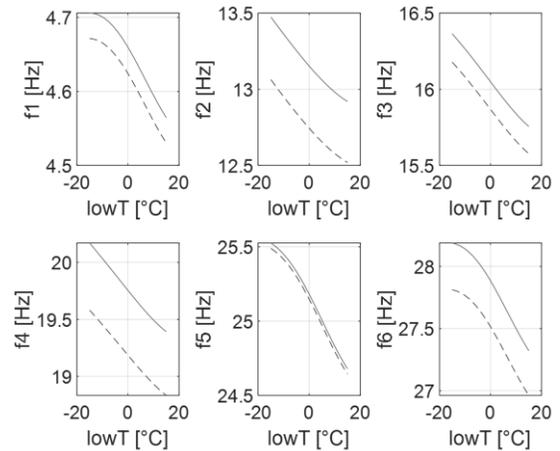


Figure 7. The first 6 natural frequencies as a function of the left side temperature ($lowT$) while the right side is changed from -15 to 45°C producing a linear temperature gradient on the bridge. Damaged features are drawn as dashed lines.

The results will be compared not only in terms of Novelty Indices, but also in terms of receiver operating characteristic (ROC) curves generated by increasing the level of a threshold and computing each time the true positive rate (TPR: the percentage of damaged points correctly identified over the total damaged) and the false positive rate (FPR: the

percentage of healthy points identified as damaged over the total healthy). The farther the ROC curve is from the 45° line, the better. In any case, the Area Under the ROC Curve (AUC) is added in the plots for a better comparison.

Table 1. Temperature conditions

#	T left (lowT) [°C]	T right [°C]
1	-15	-15
2	-7.5	0
3	0	15
4	7.5	30
5	15	45

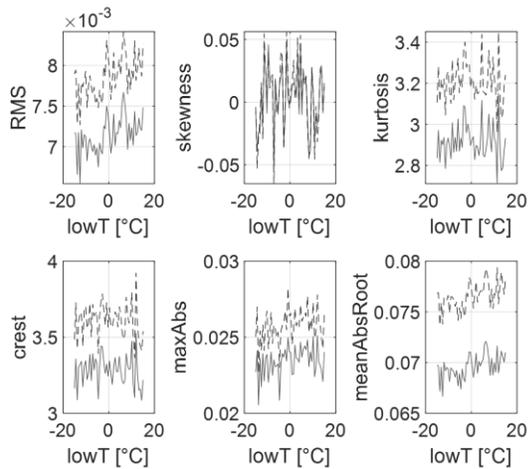


Figure 8. The six considered time-series statistical features. Damaged features are drawn as dashed lines.

The NIs are computed with the benchmark method Yan et al. (2005), here called B, with MD and with a reduced MD neglecting the first r components explaining more than the 25% of the overall variability (i.e., if their eigenvalues normalized over the sum of all the eigenvalues are higher than 0.25, the PCs are automatically neglected). In Figure 9 the results in terms of ROC curves are reported for the NIs resulting from the applied methods on the different features groups corrupted by 5% of noise. In particular, the first features group is composed by the 6 natural frequencies, the second by the 6 time-series features, and the third by all these 12 features together. From the pictures in Figure 9 several considerations can be drawn. As first, it can be recognized that, even if the amount of relative noise is the same for all the features, the frequency features prove to be much more affected by it, as probably the effect size of the damage is smaller than that in time-series features, which on the contrary proves to lead to quite satisfying ROC curves (i.e., the damage can be quite confidently detected). Furthermore, considering all the features together in a 12-dimensional problem, the damage detection slightly worsens with respect to Time Features alone, but not for rMD, whose AUC increases from 0.91 to 0.96.

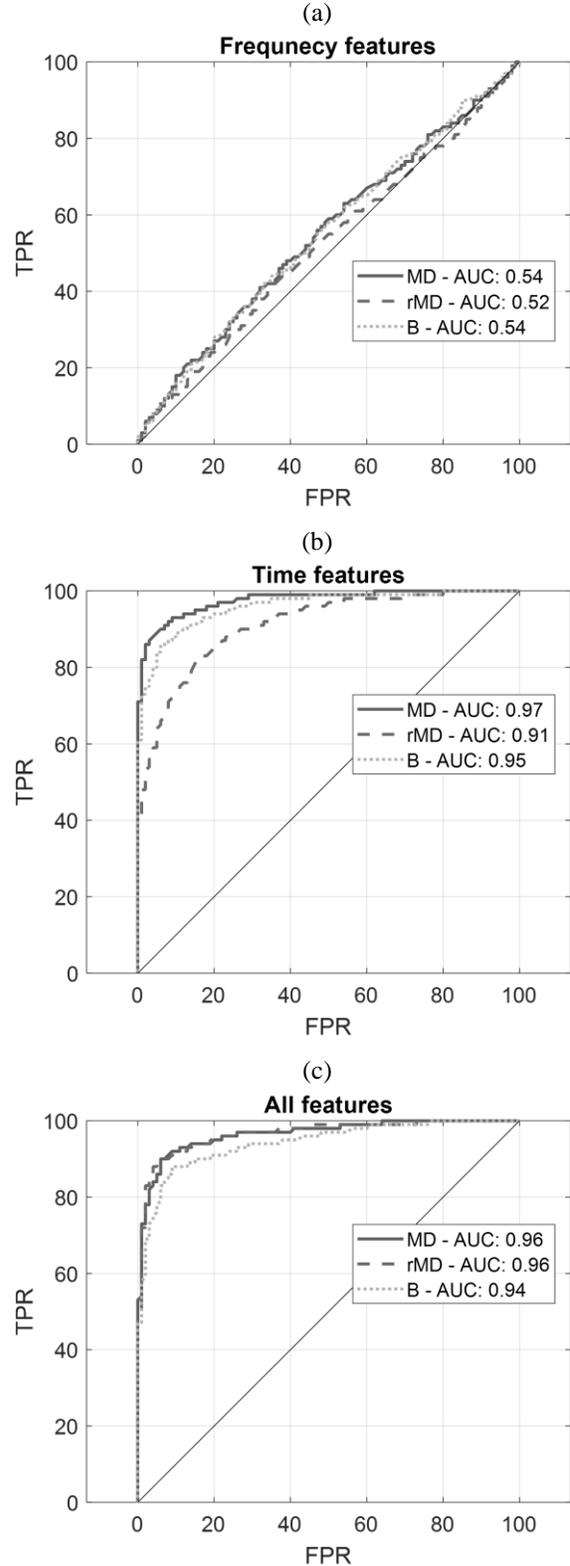


Figure 9. ROC curves for NIs computed with MD, reduced MD and Benchmark method – noise: 5%. a) frequency features, b) time-series features, c) all the 12 features together.

Then, focusing on the algorithms, it can be noticed that the proposed rMD algorithm is able to provide an improvement in the damage detection both for the frequency features and for all the features together.

For the time features alone unfortunately, the rMD worsens the damage detection of the very good MD-NIs.

In general, anyway, it can be said that the MD novelty indices prove to be superior to the benchmark method.

Similar considerations are still valid when 1% noise corrupts the features. The results reported in Figure 10 are again rewarding MD NIs, whose ROC curve are slightly better than the benchmark method and the rMD.

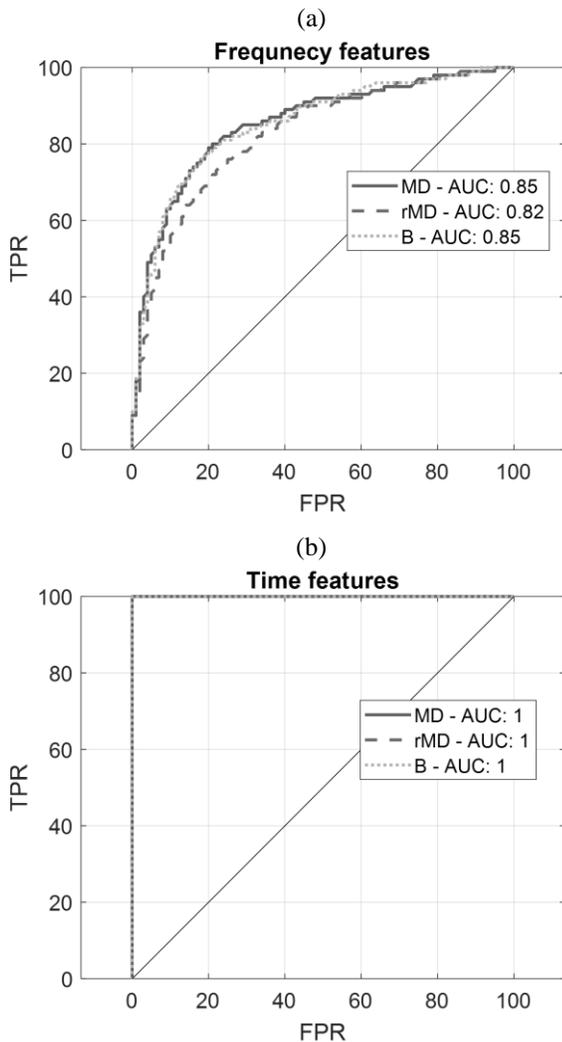


Figure 10. ROC curves for NIs computed with MD, reduced MD and Benchmark method – noise: 1%. a) frequency features, b) time-series features. Notice that the graph for all the 12 features together is omitted as equivalent to b).

In this case Figure 11 is reported to better understand the relationship of a certain ROC curve to the corresponding NIs. There, it easy to notice that time-series features NIs

shows a large separation of healthy and damaged condition. On the contrary, focusing on the frequency features NIs, their probability histograms result partially overlapped, indicating a lower effect size of the damage, leading to a more difficult diagnosis (reflected by the not so good ROC curve of Figure 10.a).

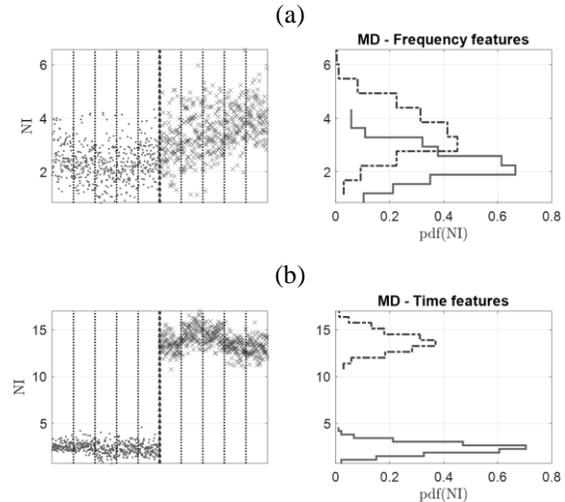


Figure 11. MD NIs visualized for frequency features (a) and for time-series features (b) in the case of 1% noise. The 5 temperature conditions are separated by dotted lines. Light gray dots refer to the healthy condition, dark gray crosses indicate the damaged condition. The corresponding probability histograms are also reported.

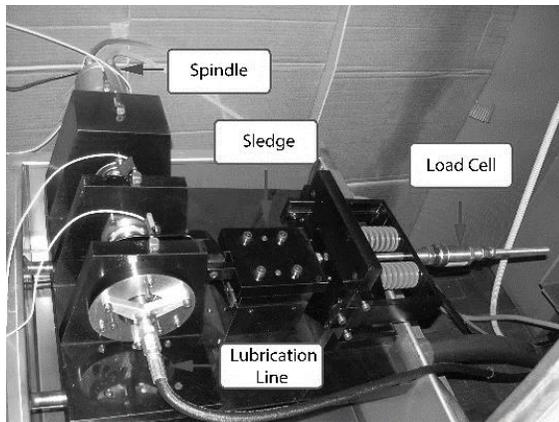
4. TEST ON EXPERIMENTAL DATA FROM POLI TO TEST RIG FOR HIGH SPEED BEARINGS VIBRATION MONITORING WITH VARIABLE SPEED

The dataset considered in this analysis comes from a test rig built by the Dynamic & Identification Research Group (DIRG), part of the Department of Mechanical and Aerospace Engineering of Politecnico di Torino, to test high-speed aeronautical bearings. The rig is fully described in Daga et al. (2019), but the main information is summarized hereinafter. The rig is made by a single direct-drive rotating shaft supported by two identical high-speed aeronautical roller bearings (B1 and B3 in Figure 12). B3 is known to be healthy while B1 is damaged on purpose with indentations of different size in different parts of the bearing (Rolling Element and Inner Ring) as described in Table 2. The third central bearing B2 is mounted on a sledge meant to load the shaft with a variable force, here left to 0 N, while the speed is reducing from 470 to 0 Hz (run-down acquisitions). Two tri-axial accelerometers located respectively on the B1 bearing support (accelerometer A1, as reported in Figure 12) and on the loading sledge (accelerometer A2). The acquisitions last for about $T = 50$ s at a sampling frequency $f_s = 102400$ Hz. In order to perform a significant analysis, the five selected features root mean

square, skewness, kurtosis, peak value and crest factor are extracted on one hundred independent chunks (about 0,5 s each) for each of the 6 channels in all the 7 health conditions (from 0A, healthy, to 6A). Finally, 100 observations in a 30-dimensional space (6 channels, 5 features) per each health condition are obtained. The dataset is visually summarized in Figure 13.

It must be said that frequency domain features other than the natural frequencies (which well applies to structural monitoring but not to industrial machines condition monitoring) could be found after a common procedure called Order Tracking [Fyfe & Munck (1997)], or using other spectrogram-based algorithms meant to remove the effect of the speed variation [Antoni et al. 2017]. Nevertheless, the additional computation of defect frequencies is usually not straightforward, not completely automated and requires additional algorithms which are out the scope of this analysis (e.g., envelope analysis to highlight the bearing damage frequencies – Randall & Antoni 2011).

(a)



(b)

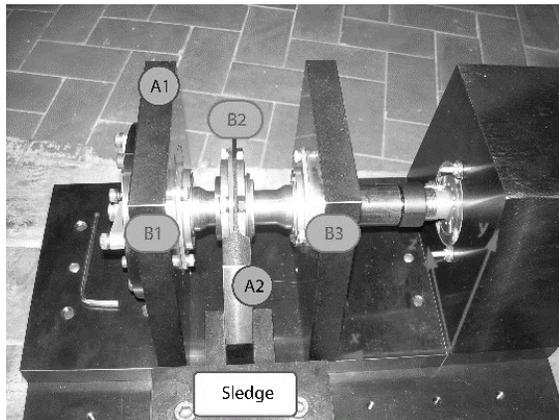


Figure 12. The experimental setup (a) and the detail of the triaxial accelerometers location (A1 and A2) and orientation (b)

Being the acquisitions made during the run-down of the machine, the speed naturally decreases from a maximum of 470 Hz to a still condition. This strongly affects all the features, but in particular RMS and peak value, which show important trends (Figure 13).

The speed is then a strong confounder which could potentially hide the diagnostic information.

It is also interesting to point out that the samples on which statistics are computed are 0.5s long. The rotation speed is decreasing from 470Hz to 0Hz, so that the first samples will be computed accounting for about 235 shaft rotations, while the last will capture less than few rotations. In order to ensure significance, then, the data was cut right before the complete stop. Nevertheless, it must be kept in mind that the features extracted from the last chunks will be less significant, but this does not affect much the analysis.

Table 2. Bearing B1 codification according to damage type (Inner Ring or Rolling Element) and size. The damage is obtained through a Rockwell tool producing a conical indentation of maximum diameter reported as characteristic size.

Code	0A	1A	2A	3A	4A	5A	6A
Damage type	none	Inner Ring	Inner Ring	Inner Ring	Rolling Element	Rolling Element	Rolling Element
Damage size [μm]	-	450	250	150	450	250	150

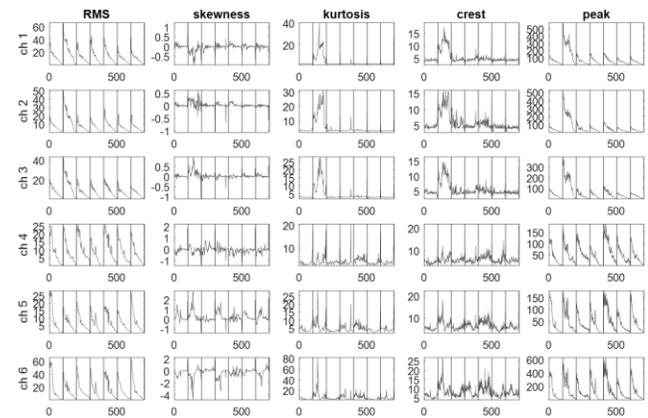


Figure 13. The considered dataset after features extraction for a load of 0 N while the speed is decreasing to a stop starting from 470 Hz. The black dotted lines divide the different damage conditions (0A to 6A). For each, 100 observations are plotted sequentially.

The described dataset was then treated with the previously introduced algorithms of MD and rMD and compared to the benchmark. In this case, being the dataset dimensionality $d = 30$, the first 20 PCs will be removed when using rMD.

The result in terms of NIs is reported in Figure 14. From the picture it is clear that the effect size of damage is quite large

and not affected too much by the decreasing speed, in fact the damaged NIs are quite far away from the healthy (notice the logarithmic scale).

It is interesting to notice that at the end of the run-downs for all the different health acquisition, the last sample tends to show an increase in the NIs which is due to the lower significance of such sample (i.e., few shaft rotation in the 0,5s chunk). In any case this effect is not particularly affecting the Novelty Detection.

Nevertheless, the separability is not perfect, as some damaged NIs are smaller than the maximum healthy NIs.

Finally, the ROC curve is reported for the three methods. From Figure 15 it can be noticed that the rMD is able to give a substantial improvement in the performance of the damage detection, pushing the ROC curve farther away from the 45° line (corresponding to a random damage detection).

For the sake of research completeness, Figure 16 reports the trend of the novelty detection performance in terms of AUC of the ROC curves obtained for a variable dimensionality reduction for both the rMD and the benchmark method. As it can be easily seen, rMD is able to effectively improve the performance, while the benchmark, in this experimental case, fails.

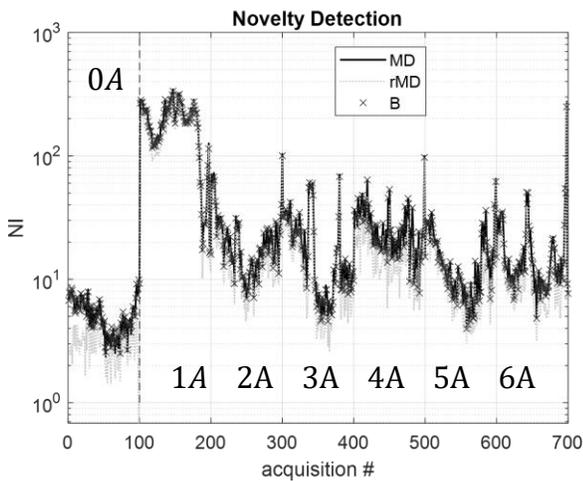


Figure 14. Novelty Detection of the experimental dataset with the different NIs. 0-100 samples are the healthy reference, 100-200 corresponds to 1A damage, and so on until 600-700 coming from 6A damage.

In this practical application then, as confirmed by Figure 16, it can be said that the MD novelty indices prove to be always superior to the benchmark method, while only rMD is able to really improve the damage detection.

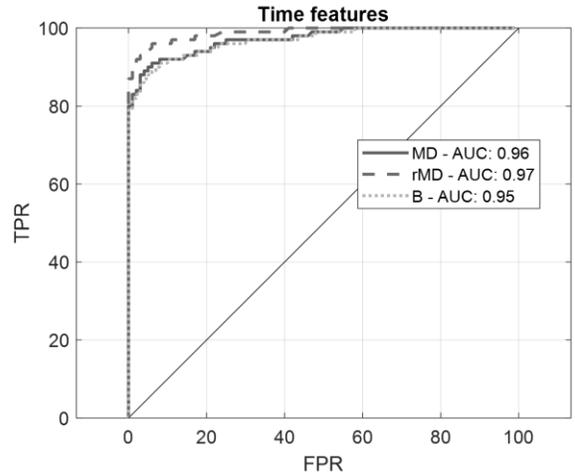


Figure 15. The ROC curves for NIs computed with MD, rMD (last 10 PCs) and Benchmark method on the experimental dataset.

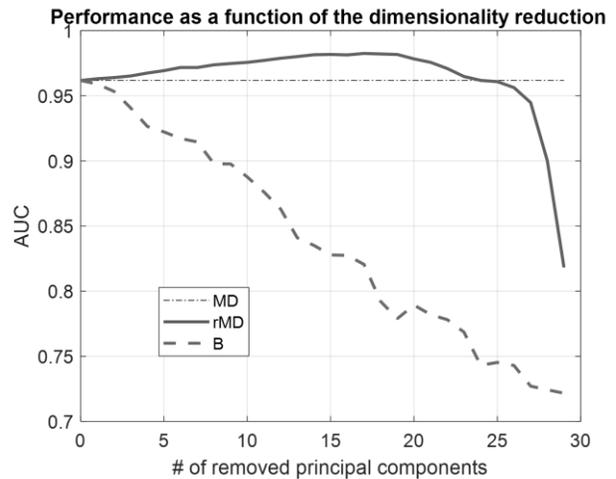


Figure 16. The AUC values of the ROC curves as a function of the different dimensionality reductions of both the rMD and the benchmark method on the experimental dataset.

5. CONCLUSION

Summarizing, in this paper PCA was analyzed in detail, so as to highlight both its essence and its mathematical formulation. PCA was then used to understand the Mahalanobis Distance, finding an alternative formulation which puts together PCA-Orthogonal Regression and Mahalanobis Distance. This naturally opens to the possibility of evaluating a reduced Mahalanobis Distance (reduced because only the information from last PCs is used to produce the rMD NI). This idea is in practice the same as that of the benchmark model: removing strong quasi-linear confounding influences by fitting a line (in case $m=1$) and considering as NI the distance of the points from such fit.

The proposed method was tested both on the 31-DOF scaled bridge model (a structural simulation similar to the assessment made in the paper describing the benchmark method [Yan et al. (2005)]), and on laboratory acquisitions from the PoliTO-DIRG test rig [Daga et al. (2019)] meant for studying high-speed aeronautical bearings.

The results proved that MD NIs are, in general, superior or comparable to the benchmark NIs. Furthermore, without dropping information, MD NIs can automatically compensate for the strong influences which are pictured in the first components (as these scores will be normalized on the variance along the component). Notice that, on the contrary, the benchmark method neglects the information contained in the direction of the fitted line ($m=1$).

Anyway, also the rMD NIs proved to be able, in some cases, to improve the diagnostic ability of the ND, in particular with the experimental test rig acquisitions. Obviously, as rMD also neglects some information, it is not wise to start directly with it, as the risk is to drop diagnostic information together with the quasi-linear confounding influences. It is then suggested to start with the regular MD-NIs and then to check whether rMD can lead to substantial improvements as for the experimental case here considered.

Finally, the bridge model was used to prove that time-series statistical features are potentially able to bring more diagnostic information than the natural frequencies. Obviously, in real-life structural cases, confounding influences other than the temperature alone can be present, and these can affect more the time features rather than the frequency features. Nevertheless, it can always be worth to give them a try given the low computational burden. On the other hand, condition monitoring of industrial machines is performed in relatively controlled environments, so that the confounding influences are typically related to variable operational conditions. In this case also, an analysis using common time features aggregated with multivariate statistics (e.g., with MD NIs) can deserve to be tested, as proved in this paper.

NOMENCLATURE

<i>DOF</i>	Degree of Freedom
<i>MD</i>	Mahalanobis Distance
<i>ND</i>	Novelty Detection
<i>NI</i>	Novelty Index
<i>OR</i>	Orthogonal Regression
<i>PCA</i>	Principal Components Analysis
<i>rMD</i>	reduced Mahalanobis Distance
<i>RMS</i>	Root Mean Square
<i>T</i>	temperature

REFERENCES

- Antoni J., et al., (2017). *Feedback on the Surveillance 8 challenge: Vibration based diagnosis of a Safran aircraft engine*. Mech. Syst. Signal Process. 97. DOI: 10.1016/j.ymssp.2017.01.037.
- Bellino, A., Fasana, A., Garibaldi, L., Marchesiello, S., (2010). *PCA-based detection of damage in time-varying systems*, Mechanical Systems and Signal Processing, Volume 24, Issue 7, Pages 2250-2260, DOI: 10.1016/j.ymssp.2010.04.009.
- Brereton, R.G., Lloyd G.R., (2016). *Re-evaluating the role of the Mahalanobis distance measure*. Journal of Chemometrics 30:134-143. DOI: 10.1002/cem.2779
- Castellani, F., Garibaldi, L., Daga, A.P., Astolfi, D., Natili, F. (2020). *Diagnosis of Faulty Wind Turbine Bearings Using Tower Vibration Measurements*. Energies. 13. 1474. DOI: 10.3390/en13061474.
- Cattell, R.B., (1966) *The Scree Test For The Number Of Factors*, Multivariate Behavioral Research, 1:2, 245-276, DOI: 10.1207/s15327906mbr0102_10
- Cornbleet, P.J., Gochman, N., (1979). *Incorrect Least-Squares Regression Coefficients*. Clin. Chem. 25 (3): 432-438. PMID 262186
- Daga, A.P., Fasana, A., Marchesiello, S., Garibaldi, L., (2017). *ANOVA and other statistical tools for bearing damage detection*, International Conference TSurveillance 9
- Daga, A.P., Fasana, A., Marchesiello, S., Garibaldi, L., (2019). *The Politecnico di Torino rolling bearing test rig: Description and analysis of open access data*. Mech. Syst. Signal Process. 2019, 120, 252-273.
- Daga, A.P., Fasana, A., Marchesiello, S., Garibaldi, L., (2019). *Machine Vibration Monitoring for Diagnostics through Hypothesis Testing*. Information, 10, 204.
- Daga, A.P., Fasana, A., Marchesiello, S., Garibaldi, L., (2019). *Confounding factors analysis and compensation for high-speed bearing diagnostics*, International Conference Survishno
- Deraemaeker, A., Worden, K., (2018). *A comparison of linear approaches to filter out environmental effects in structural health monitoring*, Mechanical Systems and Signal Processing, Volume 105. DOI: 10.1016/j.ymssp.2017.11.045.
- Farrar C.R., Doebling S.W., (1999). *Damage Detection and Evaluation II*. In: Silva J.M.M., Maia N.M.M. (eds) Modal Analysis and Testing. NATO Science Series (Series E: Applied Sciences), vol 363. Springer, Dordrecht. ISBN 978-0-7923-5894-7.
- Fyfe, K.R., Munck, E.D.S., (1997). *Analysis of Computed Order Tracking*, Mech. Syst. Signal Process. 11, DOI: 10.1006/mssp.1996.0056.
- Galton F., (1886). *Regression Towards Mediocrity in Hereditary Stature*, The Journal of the Anthropological Institute of Great Britain and Ireland, 15, pp. 246-263

- Jolliffe, I.T., (1982). *A Note on the Use of Principal Components in Regression*. Journal of the Royal Statistical Society: Series C (Applied Statistics) 31 (3). Wiley Online Library: 300–303.
- Jolliffe, I.T., (2002). *Principal Component Analysis, second edition*, New York: Springer-Verlag New York, Inc
- Nievergelt, Y., (2000). *A tutorial history of least squares with applications to astronomy and geodesy*, Journal of Computational and Applied Mathematics, Volume 121, Issues 1–2, Pages 37-72, DOI: 10.1016/S0377-0427(00)00343-5.
- Pearson, K., (1901). *On lines and planes of closest fit to systems of points in space*, Philosophical Magazine, Series 6, vol. 2, no. 11, pp. 559-572
- Randall, R.B., Antoni J., (2011). *Rolling Element Bearing Diagnostics - A Tutorial*. Mech. Syst. Signal Process. 25(2): 485-520, DOI: 10.1016/j.ymssp.2010.07.017
- Stanton, J.M., (2001). *Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors*, Journal of Statistics Education, 9:3, DOI: 10.1080/10691898.2001.11910537
- Stigler, S.M., (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge: Harvard. ISBN 0-674-40340-1
- Van Huffel S., Vandewalle J., (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*, SIAM, 1991, ISBN: 978-0-898712-75-9
- Worden, K. Manson, G., Fieller, N. R. J., (2000). *Damage detection using outlier analysis*, Journal of Sound and Vibration. DOI: 10.1006/jsvi.1999.2514
- Yan, A.M., Kerschen, G., De Boe, P., Golinval, J.C., (2005). *Structural damage diagnosis under varying environmental conditions – Part I: a linear analysis*, Mech. Syst. Signal Process. 19, 847–864, DOI: 10.1016/j.ymssp.2004.12.002.