Random Matrix-Based Approach for Data Reduction and Sensor Selection with Application to Degradation Index Construction

1st Togo Jean Yves Kioye
Computer Science and Digital Society
Troyes University of Technology
Troyes, France
jean_yves.kioye@utt.fr

2nd Malika Kharouf
Computer Science and Digital Society
Troyes University of Technology
Troyes, France
malika.kharouf@utt.fr

3rd Khac Tuan Huynh
Computer Science and Digital Society
Troyes University of Technology
Troyes, France
tuan.huynh@utt.fr

***Abstract*—This paper relies on the random matrix theory to reduce data dimension and to identify useful data sources in the unsupervised context. A so-called random matrix based principal component analysis algorithm is thus developed and then applied to the well-known 2008 PHM dataset to build efficient but less costly degradation indices. A comparison of the degradation indices obtained with and without sensors selection confirms the performances of our proposed approach.**

***Index Terms*—Data reduction, sensors selection, random matrix theory, degradation index, 2008 PHM dataset.**

I. Introduction

Recent innovations in sensor technology have made condition monitoring simpler and more efficient. Indeed, human beings are no longer the only actor in the monitoring process, but are assisted by sensor networks that record the system condition over its whole life. Such recorded data are helpful for decision support that enables improvements in the system reliability, availability and safety, in environmental protection, as well as in maximizing companies profits. Nowadays, one can easily collect and store huge amounts of condition monitoring data of very different natures. This however leads to the increasing complexity for data analysis at high cost (because of the installation and maintenance of large sensor networks). In this context, how to reduce data without information loss and how to identify useful data sources are two big questions to be solved.

Data reduction consists in projecting data from a high dimensional space (with several sensors) into a lower dimensional space preserving as much as possible the structure of the data in the original space. Its principle is either to build new variables called synthetic or principal component by linear or non-linear combination of the set of sensors, or to make an optimal selection of relevant sensors that retain the majority of the information contained in the data. The processing of data in the reduced space allows to optimize the performances on different tasks in particular in classification and in regression because it allows to remove the redundant and not relevant information according to a criterion. In the literature, there exist several methods of dimension reduction by creating new synthetic variables. We can quote, for instance, the principal component analysis (PCA) [1], the factor analysis (FA) [2], the independent component analysis (ICA) [3]. To identify useful data sources, we are interested in the selection of corresponding sensors. An exhaustive literature research shows that such a selection can be done following two main approaches: (*i*) *filter* and (*ii*) *wrapper*. The first approach proposes to compute statistical measures of the variables, and then to filter out irrelevant variables. Very commonly, it does not consider the correlations among variables, and thus leads to low performances. Meanwhile, the second approach uses a well defined algorithm to find an optimal subset of variables. It allows high performances in a supervised context, but very costly in their implementation.

This paper also contributes a solution to the two above questions, but in an unsupervised context. An algorithm called *Random Matrix based Principal Component Analysis* (RM-PCA) is thus developed. It uses random matrix theory to determine eigenvalues and eigenvectors of a covariance matrix of the considered data. These eigenvalues and eigenvectors are the basis to reduce the data dimension and to select significant sensors of the network. To show its feasibility, we apply the RM-PCA algorithm to the well-known 2008 PHM dataset [4]. The aims is to construct from this dataset a less costly degradation index (i.e. an good degradation index built from a smaller number of sensors). Among numerous existing works on 2008 PHM dataset, we cite here two typical works of Wang *et al.* in [5] and Le Son *et al.* in [6]. Wang *et al.* proposed a method to select the important sensors by taking into account the trends and then using a significance test via a regression model. Based the sensors proposed by Wang *et al.*, Le Son *et al.* used the classical PCA to reduce the data dimension, and then constructed a degradation index as the euclidean distance between the obtained projections and the projections of the failure areas. We can remark that the selection approach proposed by Wang *et al.* is still heuristic, very costly in its implementation and requires the prior construction of a degradation index. It is therefore necessary to find an unsupervised method that allows to automatically select an optimal number of sensors that will allow to build less expensive degradation index with high performance. This is also the main contribution of our approach compared to most existing related works.

The remainder of this paper is structured as follows. In section (II), we first review results in random matrix theory needed for our approach. Then, we present our algorithms: RM-PCA and variables selection algorithms. In section (III), we present numerical results obtained by our algorithms on real data from aircraft engines.

II. Variables selection based on random matrix PCA

In this section we first present the needed mathematical tools from random matrix theory. Random matrix PCA (RM-PCA) and variables selection methods will then be developed.

Notations. The population covariance matrix, its eigenvalues and eigenvectors are notated as Σ, α_i and v_i respectively. The sample covariance matrix is given by $S_n = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T$, where y_i represent n copies of the observed vector $y \in \mathbb{R}^p$. Its eigenvalues and eigenvectors are given by $\hat{\lambda}_i$ and $\hat{u}_i$. The proposed estimator of Σ is notated by $\hat{\Sigma}$ with eigenvalues and eigenvectors are $\hat{\alpha}_i$ and $\hat{v}_i$ respectively.

A. Random matrix PCA framework

1) Marcenko-Pastur law [7]: In their seminal work [7], the authors prouve that the limit behaviour of the spectral density of the eigenvalues of S_n is intimately linked to the rate $\frac{p}{n}$ at large. More formally, assuming that the centred i.i.d. entries of y have a finite fourth moment and under the asymptotic regime $\frac{p}{n} \xrightarrow[n,p\to+\infty]{} c \in]0,+\infty]$, the spectral measure of S_n converges almost surely and in distribution to the Marcenko-Pastur [7] law where the density is given by:

$$f_{MP}(x) = \frac{\sqrt{b-x}\sqrt{x-a}}{2\pi c x} I_{[a,b]}(x) \tag{1}$$

with $a = \sigma^2(1-\sqrt{c})^2$ and $b = \sigma^2(1+\sqrt{c})^2$. The interval $[a,b]$ is called the bulk of Marcenko-Pastur law.

Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ be the ordered eigenvalues of S_n, Geman [8] showed that when $n, p \to +\infty$ with $p/n \to c \in (0,+\infty)$, the extreme eigenvalues verify:

$$\hat{\lambda}_1 \xrightarrow{p.s} b = \sigma^2(1+\sqrt{c})^2 \tag{2}$$

$$\hat{\lambda}_p \xrightarrow{p.s} a = \sigma^2(1-\sqrt{c})^2 \tag{3}$$

2) Spiked population model: In real data processing, one of the major difficulty is the presence of noise. This observation is at the origin of several methods allowing to clean up the eigenvalues of the sample covariance matrix S_n [9], [10]. One can mention heuristic methods like the elbow method and the Kaiser rule [11]. In random matrix theory, we consider the spiked population model [12], [13]. We consider the information-plus-noise model where the data matrix is corrupted by some additive noise. Let the observed vector $y \in \mathbb{R}^p$ be defined as:

$$y = s + e$$

Consider n independent and identically distributed copies of y and stocked in the $p \times n$ matrix $\mathbf{Y}$. Assuming independence between the information counterpart and the noise implies the following identity of corresponding covariance matrices.

$$\Sigma_y = \Sigma_s + \Sigma_e \tag{4}$$

it can also be written as follows

$$\Sigma_y = V \left(\begin{array}{ccc|c} \alpha_1 + \sigma^2 & & & \\ & \ddots & & 0_{r,p-r} \\ & & \alpha_r + \sigma^2 & \\ \hline & 0_{p-r,r} & & \sigma^2 I_{p-r} \end{array} \right) V^T ,$$

with $\Sigma_e = \sigma^2 I_p$, $\Sigma_s = diag(\alpha_1, \cdots, \alpha_r, 0_{p-r})$, where $V \in \mathbb{R}^{pp}$ is an orthogonal matrix and $I_{p-r} \in \mathbb{R}^{p-r}$ is a square identity matrix of size $(p-r) \times (p-r)$.

This model is also called spiked population model when the rank r of the information covariance matrix Σ_s is much smaller than the dimensions p and n. In the following, we give estimators of the covariance population eigenvalues, and estimators for the rank r and the noise unknown variance σ^2.

3) Population eigenvalues estimators.: Under the asymptotic regime $n, p \to +\infty$ with $p/n \to c \in (0,1)$, and under some additional statistical conditions, Baik and Silverstein [13] prove the following almost sure convergences of $\hat{\lambda}_i$ toward some functional of α_i

- For $1 \leq i \leq r$ we have

$$\hat{\lambda}_i \xrightarrow{a.s} \sigma^2 \phi(\alpha_i)$$

with

$$\phi(\alpha_i) = (\alpha_i + 1)(1 + \frac{c}{\alpha_i}) \tag{5}$$

- For $i > r$ we have

$$\hat{\lambda}_i \xrightarrow{a.s} \sigma^2(1+\sqrt{c})^2$$

We propose the following inverse formula to estimate the population covariance matrix Σ

$$\hat{\alpha}_i = \frac{1}{2\hat{\sigma}^2}(\kappa_i \pm \sqrt{|\kappa_i^2 - 4c\hat{\sigma}^2|}) \tag{6}$$

where $\kappa_i = \hat{\lambda}_i - \hat{\sigma}^2(c+1)$ and $\hat{\sigma}^2$ is an estimator of the unknown noise variance σ^2.

Estimation of the rank r and the variance σ^2. Estimating the number of the effective information variances, the so called spike eigenvalues, is a serious task. Passemier et al. [14] propose an interesting estimator for the number of spikes based on the successive differences of sample eigenvalues and using the spiked population model. However, this method is based on the assumption that sample eigenvalues are well separated and then becomes ineffective when a single sample eigenvalue is badly estimated.

We propose an estimator of the the rank r which is robust in the case of the presence of multiplicities of sample eigenvalue. Based on (2), the proposed estimator is given by:

$$r = \underset{i\in[|1,p|]}{\arg\min} \left(\hat{\lambda}_i > \sigma^2\left(1+\sqrt{p/n}\right)^2\right) + 1 \qquad (7)$$

In real case scenarios, the noise variance is unknown and then has to be estimated.The following algorithm summarise our imbricate method.

Algorithm 1 Estimation of the isolated variances number

```
1: Initialisation: $\hat{\sigma}^2 = \frac{1}{p}\sum_{i=1}^{p}\hat{\lambda}_i$
2: Compute $\hat{r}$, the estimator of $r$ using (7) by replacing $\sigma^2$ by $\hat{\sigma^2}$
3: Update $\hat{\sigma}^2 = \frac{1}{p-\hat{r}}\sum_{i=\hat{r}+1}^{p}\hat{\lambda}_i$
4: Return to 2. with the updated estimator of $\sigma^2$
5: Repeat 2-4 until convergence
6: Return $\hat{r} \leftarrow \hat{r}+1$.
```

4) Eigenvectors estimators: Based on the work of [15], we recall here the new estimator of Σ used for our RM-PCA algorithm. For the estimation of the new eigenvectors, classes of estimators called rotation invariant have been used. These estimators consider the hypothesis that there is no principal direction for the eigenvectors of Σ [16]. This assumption has two consequences. S_n has the same eigenvectors as Σ and the eigenvectors of S_n are not necessarily optimal estimators for Σ.

The work of [15] have allowed to obtain both an optimal and a rotation invariant estimator. This estimator is defined as follows.

Let $M(S_n)$ be the set of non-negative symmetric matrices having the same eigenvectors as the sample covariance matrix S_n. The optimal estimator of the population covariance matrix Σ verify :

$$\hat{\Sigma} = \underset{\Theta\in M(S_n)}{\arg\min} ||\Theta - \Sigma||^2 \qquad (8)$$

In [15], authors propose the following optimal estimator :

$$\hat{\Sigma} = \sum_{i=1}^{p}\xi_i\hat{u}_i\hat{u}_i^T \qquad (9)$$

with $\hat{u}_i$ the eigenvectors of S_n and $\xi_i =< \hat{u}_i, \Sigma\hat{u}_i >$, $i \in [|1,p|]$.

From a practical point of view, this estimator is useless since it depends on the unknown matrix Σ that we want to estimate. We should therefore look for a new writing of $\hat{\Sigma}$ which does not depend on Σ. For that we use the expression of ξ_i proposed in **[?]** :

$$\hat{\xi}_i = \frac{1}{n}\sum_{j=1}^{p}\frac{\hat{\alpha}_j^2}{(1-\frac{\hat{\alpha}_j}{\hat{\lambda}_i})^2} \qquad (10)$$

We propose to use (9) by injecting(10) to get an consistent estimator of the population covariance matrix. This estimator is used to propose our new approach: random matrix based principal component analysis (RM-PCA).

B. *Random matrix based PCA and variables selection algorithms*

1) RM-PCA algorithm: The spreading os sample eigenvalues phenomenon well highlighted by Johnsone [12] prove that sample covariance matrix S_n is no longer a consistent estimator for Σ under the modern regime, where both the number of parameters and the number of samples are large. The traditional PCA method reaches its limits. We propose here our RM-PCA method, where $\hat{\Sigma}$ is described in the previous paragraph. The corresponding algorithm is stated follows.

Algorithm 2 RM-PCA algorithm

```
Create $X$, the $p \times n$ data matrix
Compute the sample covariance matrix $S_n = \frac{1}{n}XX^T$ and the corresponding eigen-elements $\hat{\lambda}_i$ and $\hat{u}_i$
Compute $\hat{r}$, the number of the principal components following algorithm 1 in II-A3
Compute $\hat{\sigma}^2 \leftarrow \frac{\sum_{i=\hat{r}+1}^{p}\hat{\lambda}_i}{p-\hat{r}}$
Compute the population eigenvalues $\hat{\alpha}_i$ following equation (6) with the estimator $\hat{\sigma}^2$
Calculate $\widehat{\Sigma}$ following (9) with $\xi_i$ given by (10)
Extract the $\hat{r}$ largest eigenvalues and the corresponding eigenvectors, $\hat{\alpha}_1 \geq \cdots \geq \hat{\alpha}_{\hat{r}}$ and $\hat{v}_1, ......, \hat{v}_{\hat{r}}$ respectively
Projection into the subspace of dimension $\hat{r}$ :
    $\tilde{X} \leftarrow X^T\hat{V}_r,$
  where $\hat{V}_r = (\hat{v}_1, ......, \hat{v}_{\hat{r}})$.
```

2) Selection of variables: The selection of relevant variables is a problem in multidimensional statistics. It requires that the results obtained from the analysis of the extracted variables are better than or close to those obtained with all variables. It allows to avoid overlearning, redundancy of information and noise. There exist many approaches to variable selection based on principal components. With these methods the choice of important variables becomes complex if several components contain relevant information. An approach called Principal Feature Analysis (PFA) was introduced by [18]. It proposes to make a classification of the variables on the space of the principal components and then to select in each group a variable close to the center of gravity of the group. However, the number of groups to be defined during the classification is not exactly known, making it difficult to determine the number of variables to extract. An application of these methods are available in [19]. In this paper, we propose an extension of the PFA using the number of important variables r that we have determined through Algorithm 1 using random matrices. We perform a partitioning of the variables into r groups using the k-mean [20] algorithm, then in each partition we select the variable that best represents its group. We consider that the most representative variable of a group is the one with the greatest variability. This choice is based on the principe of principal components which seek to maximize the variance on the projected space. The step before the classification consists in calculating the coordinates of the variables. The procedure for obtaining the r relevant variables is summarized in algorithm 3 below:

Algorithm 3 Selection of variables

```
Spectral decomposition of $\hat{\Sigma}$ by extracting the eigenvalues $\hat{\alpha}_i$ and the eigenvectors $\hat{v}_i$.
$D \leftarrow diag(\sqrt{\hat{\alpha}_1} \cdots \sqrt{\hat{\alpha}_p})$ a diagonal matrix with the square roots of the eigenvalues.
$\hat{V} \leftarrow (\hat{v}_1, ......, \hat{v}_p)$ the eigenvector matrix.
    $$CV \leftarrow \hat{V}D,$$
    with $dim(\hat{V}) = p \times p$, $dim(D) = p \times p$, $dim(C_V) = p \times p$.
Classification of the variables of $C_V$ in $r$ clusters using K-means algorithm. We obtain $r$ groups $G_j$ (where $r$ is given by algorithm 1).
for $j <= r$ do
    $k \leftarrow \arg\max(var(x_1^j), var(x_2^j), ...., var(x_p^j))$
    $v_j \leftarrow x_k$
end for
We obtain $r$ variables $v_1$, $v_2$,.......,$v_r$.
```

III. Application to Aircraft Engine Data

To show the feasibility of the proposed method, we apply it to the 2008 PHM dataset. This dataset consists of multivariate time series that are collected from 218 identical and independent jet engines performed by NASA. It contains in total of 45918 observations and 26 variables that capture information on the operational profile, control tools and environmental conditions. At the beginning of the simulation (cycle 1) each engine starts its operation in a healthy state and then degrades until it reaches a failing state from which we decide not to operate the engine anymore (last cycle). Some descriptive tools for the data are available in [6], [**?**] and [4]. Figure 1 shows a turbojet engine as simulated by the jet engines model.

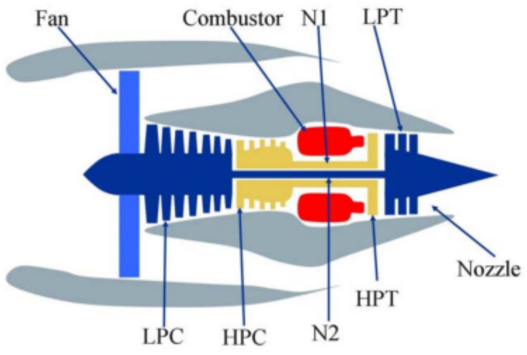

Fig. 1. Simplified schematic of the simulated aircraft engine

A. Primary Data Analysis

The data are made up of 26 different variables including

- the number of engines,
- the engine cycle,
- three operational parameters (altitude, speed, and angle),
- the measurements given by 21 sensors.

The measurements of the three operational parameters can be represented by six clusters in a 3-dimensional space as in Figure 2. The six observed clusters correspond to 6 operating conditions of the engines. Thus, it could be interesting to observe the data according to the operating conditions.

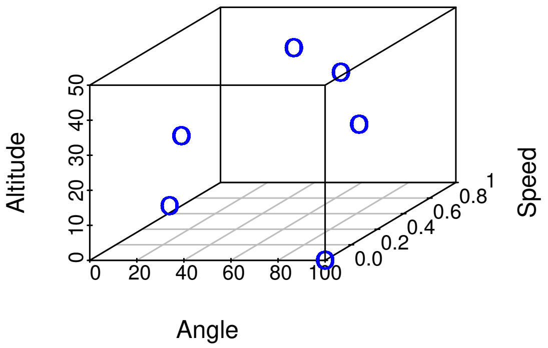

Fig. 2. Six modes of engine functioning

An initial trend analysis shows that it is necessary to construct an index representing the overall degradation of engines because the data from the sensors of each motor do not allow the observation of trends that are indicative of a convergence towards a faulty state. An illustration is presented in Figure 3 for the data of sensors 1 and 2 in engine 1. We

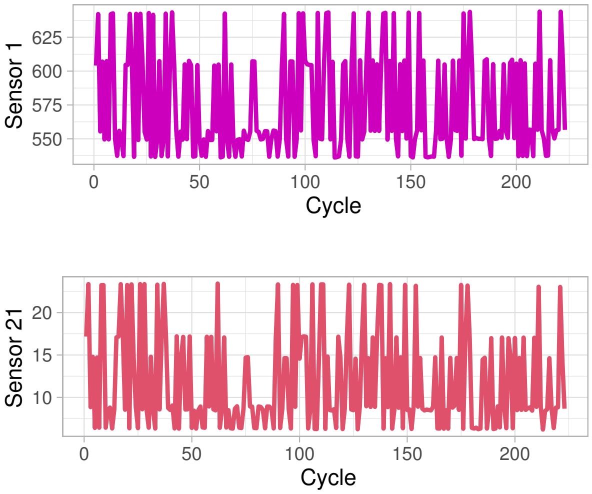

Fig. 3. A degradation indicator cannot be directly deduced from sensor data (Sensor 1 and 21 measurements for engine 1)

find the evolution curve of the signals from sensor 1and sensor 21 recorded during the operation of engine 1. These signals do not show any trend that could indicate a convergence of engine 1 towards a faulty state. All the sensors of the other motors show the same phenomenon.

By partitioning the sensor data by operational profile as in Figure 4, some trends can be observed.

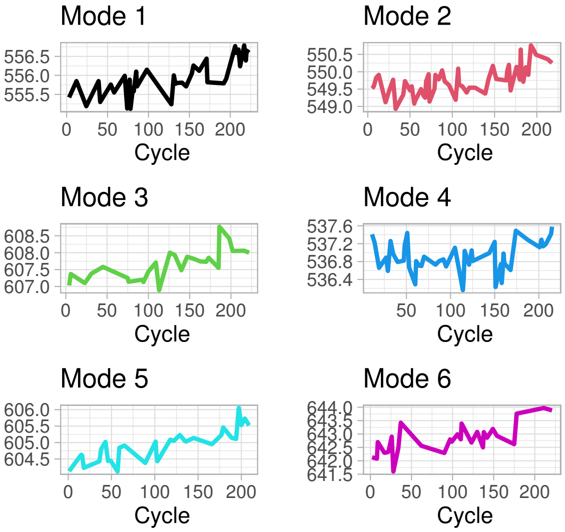

Fig. 4. The observation of data in the operating modes can reveal trends(engine 1 in 6 modes)

Indeed, Figure 4 shows the evolution curve of the signals from sensor 2 for motor 1 according to the six operating modes. These observations make it possible to highlight the trends. For some engines these trends are obvious but for others they are difficult to detect. Even if the trends are observable according to the operational mode, they do not allow to characterize the global degradation of the engines. So it is necessary to build a degradation indicator that could evaluate the global degradation of each engine. This indicator must be calculated taking into account the operational profile in which the engine is at each moment

B. Degradation Indicator Construction

The construction of our degradation indicator is based on the one proposed in [6] in which we replaced the classical PCA by RM-ACP. Such a process is schematized in Figure 5 and is described more in detail as follows.

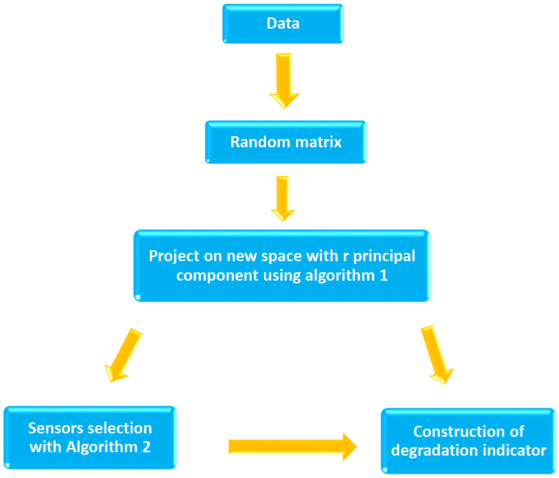
Data
Random matrix
Project on new space with r principal component using algorithm 1
Sensors selection with Algorithm 2
Construction of degradation indicator

Fig. 5. Procedure using data

- Let us denote by $X \in \mathbb{R}^{np}$ our centered data set. We will call by $X^d \in \mathbb{R}^{n_d p}$ a basis extracted from X corresponding to the data of the p sensors at the failure time (last cycle). X^d is the failure base. The failure of the engines occurred in one of the 6 modes of operation.
- We compute r on the set of data X according algorithm 1 in (II-A3).
- We extract from X^d the failures in the operational modes. Let $X^{di} \in \mathbb{R}^{n_{di}p}$ be the failures in mode i, $i \in [|1, 6|]$ and n_{di} the number of engines in failure in mode i.
- We apply RM-PCA algorithm to X^{di}.
- We project X^{di} in the new subspace of dimension r and obtain new coordinates in the subspace that we will note $di = (d^i_{1k}, ..., d^i_{rk}), k = 1, ..., n_{di}$ the projection of the failure instant of the engine j which is in the operational mode i at the time of its failure.
- We look for a unique center of failure for the engine having realized their failures in the i mode. We consider this center as the center of gravity of di it will be noted G^i. on the r dimensions the coordinates of $G^i = (G^i_1, ..., G^i_r)$
- The degradation indicator is considered as the distance between the projected data of each cycle t and the failure center of the mode i in which this cycle t is located. We will note by $(X^i_{j,1}(t), ..., X^i_{j,r}(t))$ the projection of sensors data on r dimension of engine j, at time t in the failure space of operational mode i (engine j is in mode i at time t). We obtain the degradation index $D^i_j(t)$ as

$$D^i_j(t) = \frac{\sqrt{(X^i_{j,1}(t) - G^i_1)^2 + \cdots + (X^i_{j,r}(t) - G^i_r)^2}}{Sd^i}, \tag{11}$$

where

$$Sd^i = \sqrt{\frac{1}{n_{di} - 1}\sum_{k=1}^{n_{di}}((d^i_{1k} - G^i_1)^2 + \cdots + (d^i_{rk} - G^i_r)^2)}.$$

Thus, for each cycle t, we look for its operating mode i, then we project the set of data from the sensors at cycle t onto the failure space of mode i and then we calculate the degradation index.

C. Degradation Indices Without Sensors Selection

Figures 6 and 7 show the degradation indices that we obtain with 19 sensors: 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 20, 21. In Figure 6, all the degradation indices of the 218 engines are represented. They indicate a decreasing trend that corresponds to the decrease in the health state of the engines as the number of cycles increases.

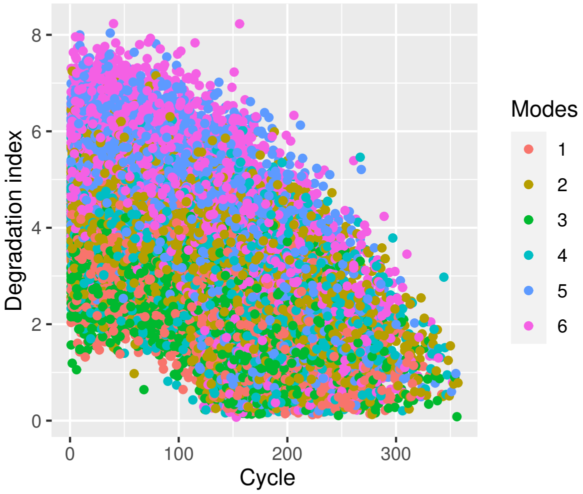

Fig. 6. Degradation indicator in all units with 19 sensors

Figure 7 presents a more detailed observation of the degradation indicator for engines 2, 100, 188, 218. The indicators at the time of failure converge to 0.

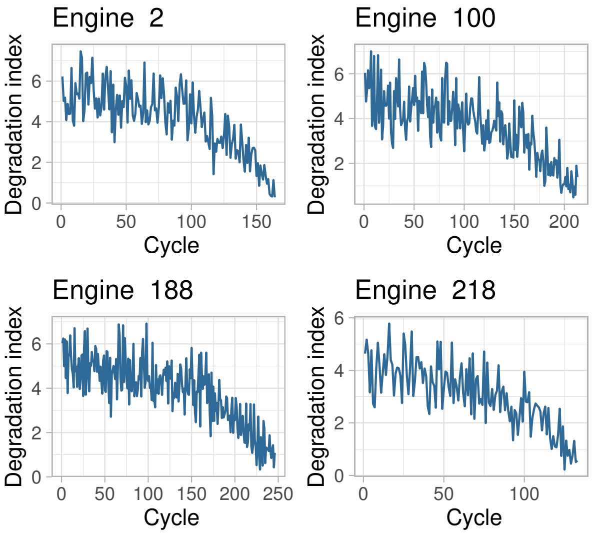

Fig. 7. Degradation indicator corresponding to component 2, 100, 188, 218 with 19 sensors

We can find that these indices allow us to specify the health state of the engines at each moment. The change in trend in the degradation indicator is a sign of fatigue or the presence of an anomaly in the operating state of the engines. Thus, as soon as a change in trend appears, companies can plan the maintenance of their systems.

D. Degradation Indices With Sensors selection

By applying Algorithm 2, we have selected 7 sensors among 19. The results of the classification on the two main dimensions are shown in Figure 8 and are reported in Table I. All the sensors that were alone in their groups were retained (i.e., sensors 4, 13, 3) for the groups with at least two sensors we retained the sensor with the greatest variance in its group (i.e., sensors 2, 7, 18, 20). Thus, the most important sensors are 2, 3, 4, 7, 13, 18, 20.

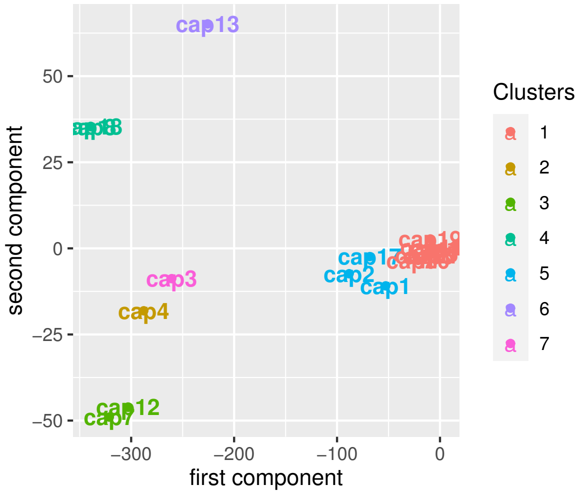

Fig. 8. Classification of sensors into 7 groups using algorithm 2

TABLE I
RESULTS OF SENSORS CLASSIFICATION

group number	Sensors number
Group 1	5, 6, 10, 11, 15, 16, 19, 20, 21
Group 2	4
Group 3	7, 12
Group 4	8, 18
Group 5	1,2, 17
Group 6	13
Group 7	3

The degradation indices that we obtain from the 7 selected sensors using our approach are presented in figures 9 and 10.

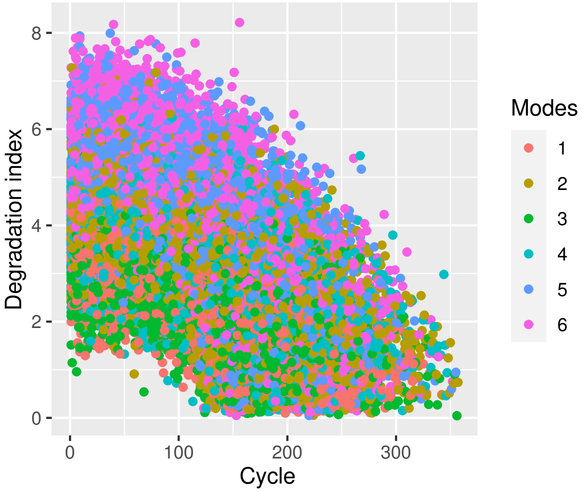

Fig. 9. Degradation indicator in all units with 7 sensors

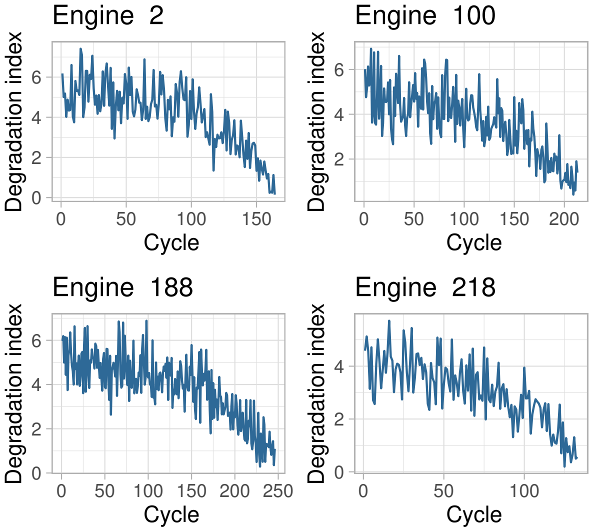

Fig. 10. degradation indicator corresponding to component 2, 100, 188, 218 with 7 sensors

We find that the selection of sensors gives the possibility to work with a reduced number of sensors and to have similar or very high performances than with all the sensors. The selection makes the construction of the degradation indices less expensive, because it does not require several sensors and the calculation time is less significant. The degradation indices obtained with sensors selection (see Figures 6 and 7) and without sensors selection (Figures 9 and 10) are very close.

IV. Conclusion

This work relies on the results of the random matrix theory to construct a new variance-covariance matrix before using them in dimension reduction via the RM-PCA algorithm and sensor selection. The application of our algorithms on the 2008 PHM chalenge data confirms the relevance of our proposed algorithms.

Our results on degradation indices can be improved by considering the case where the variance is colored. Future work on the use of our approach for the construction of the degradation index for the purpose of failure prognosis could allow us to pronounce the predictive power of our sensors.

References

[1] H. Hotelling (1933), "Analysis of a complex of statistical variables into principal components," Journal of Educational Psychology, 24 :417441.

[2] C. Spearman "General Intelligence, Objectively Determined and Measured," The American Journal of Psychology 15, no. 2 (1904): 20192. https://doi.org/10.2307/1412107.

[3] A. Hyvarinen, j. Karhunen, and E. Oja(2001) "Inde-pendent component analysis," John Wiley and Sons.

[4] A. Saxena, K. Goebel, D. Simon and N. Eklund, (2008, October). "Damage propagation modeling for aircraft engine run-to-failure simulation," In 2008 international conference on prognostics and health management (pp. 1-9). IEEE.

[5] T. Wang, J. Yu, D. Siegel and J. Lee, "A similarity-based prognostics approach for Remaining Useful Life estimation of engineered systems," 2008 International Conference on Prognostics and Health Management, 2008, pp. 1-6, doi: 10.1109/PHM.2008.4711421.

[6] K. Le Son, M. Fouladirad, A. Barros, E. Levrat and B. Iung, (2013), "Remaining useful life estimation based on stochastic deterioration models: A comparative study," Reliability Engineering System Safety, 112, 165-175.

[7] A. V. Marcenko, et L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," Mathematics of the USSR-Sbornik, 1967, vol. 1, no 4, p. 457.

[8] Geman, S. (1980), "A limit theorem for the norm of random matrices," The Annals of Probability, 8(2), 252-261.

[9] R. Cangelosi, A. Goriely"Component retention in principal component analysis with application to cDNA microarray data," Biology Direct 2, no. 2 (2007).

[10] I. T. Jolliffe "Principal components analysis," Springer Verlag, (1986).

[11] R. B. Cattell"The scree test for the number of factors," Multivariate Behavioral Research, 1:245276, (1966).

[12] M. I. Jonhstone, "On the distribution of the largest eigenvalue in principal components analysis," The Annals of statistics, 2001, vol. 29, no 2, p. 295-327.

[13] J. Baik and J. W. Silverstein,"Eigenvalues of large sample covariance matrices of spiked population models," Journal of multivariate analysis, 2006, vol. 97, no 6, p. 1382-1408.

[14] D. Passemier and J. Yao, "On determining the number of spikes in a high-dimensional spiked population model," Random Matrices: Theory and Applications, 2012, vol. 1, no 01, p. 1150002.

[15] J. Bun and A. Knowles, (2018) "An optimal rotational invariant estimator for general covariance matrices: The outliers," Preprint.

[16] J. Bun, "Application of Random Matrix Theory to High Dimensional Statistics," 2016. Thse de doctorat. Universit Paris-Saclay.

[17] R. Allez, J. Bun, Jol, and J. P. Bouchaud, "The eigenvectors of Gaussian matrices with an external source," arXiv preprint arXiv:1412.7108, 2014.

[18] Y. Lu, I. Cohen, X. S. Zhou and Q. Tian, (2007, September). "Feature selection using principal feature analysis," In Proceedings of the 15th ACM international conference on Multimedia (pp. 301-304).

[19] S. Parfait, (2010)."Classification de spectres et recherche de biomarqueurs en spectroscopie par rsonance magntique nuclaire du proton dans les tumeurs prostatiques," (Doctoral dissertation, Universit de Bourgogne).

[20] J. MacQueen, (1967). "Classification and analysis of multivariate observations," In 5th Berkeley Symp. Math. Statist. Probability (pp. 281-297).