© Universiti Tun Hussein Onn Malaysia Publisher's Office

IJIE

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

Data Clustering based on Gaussian Mixture Model and Expectation- Maximization Algorithm for Data-driven Structural Health Monitoring System

Sharafiz A. Rahim[1*], Graeme Manson[2], M. A. Aziz[1]

[1]Mechanical Engineering Department,
Universiti Putra Malaysia, Serdang, 43400, MALAYSIA

[2]Mechanical Engineering department,
University of Sheffield, Sheffield, S1 3JD, UNITED KINGDOM

*Corresponding Author

DOI: https://doi.org/10.30880/ijie.2021.13.07.020
Received 16 August 2021; Accepted 2 September 2021; Available online 30 September 2021

Abstract: Data groups generated by a system often inherit dynamics characteristics unique in data distribution parameters. A degradation in structural health can affect the dynamic behavior hence the probability distribution parameters. Based on the probabilistic and Expectation-Maximization (EM) algorithm, Gaussian Mixture Model (GMM), one can cluster data groups that may overlap with different data groups based on different orientations and shapes. This article explores GMM probabilistic model applied on vibration data set generated by aircraft wing box structure for Structural Health Monitoring (SHM) application. In the data processing stage, the high dimensional data is transformed into lower dimensions using Kernel Principal Component Analysis (KPCA). KPCA transforms the continuous signal into discrete data, allowing the ellipsoids' fitting (clusters) on the data spread. Based on the baseline data set (undamaged structural condition) and several components (loading class and damage class), the fitting is performed using GMM driven by EM. This paper shows that GMM-EM based data clustering model is an effective clustering probability model in fitting the data density in the presence of operational variations. It highlights clustering of reduced vibration data using KPCA in the interest of SHM based on the baseline's initial parameters.

Keywords: Gaussian mixture model, expectation maximization, structural health monitoring, kernel principal component analysis

1. Introduction

The motivation for choosing GMM in the current study is to establish a predictive model for distinguishing damage feature (DF) from loading feature (LF) using a clustering technique based on the Gaussian distribution model. The basic idea is that generally, data points are assumed to adopt a normal distribution, and each of the density models is associated with different data labels or categories. Clustering using GMM-EM based on the normal data distribution is an appropriate tool for many real physical systems. It gives an advantage by incorporating ellipsoids enclosing the data points based on the mean and the covariance (the shape of the ellipsoid) naturally, depending on the data behavior produced by the system. This data distribution depends on the mean and covariance as well the mixing proportions in the probability framework (Figure 1).

*Corresponding author: sharafiz@upm.edu.my

2021 UTHM Publisher. All rights reserved.

The rationale is, when the system is excited by a random signal, the data generated is likely to inherit a Gaussian distribution. In addition to the assumption, the more data samples acquired or measured from the system, the behavior should resemble a normal distribution as stated by the theory of Central Limit [1], [2] and [3]. From an SHM perspective, the health state of a system can be compared with a reference data set given by an undamaged state using statistical pattern recognition. Any pattern that deviated from the reference set can be reckoned that it is likely to have a defect or anomalies in the structure with the exception that the operational and environmental variations have been countered [1], [2], [4] and [5].

In a vibration-based damage detection (VBDD) approach, the effects of the damage or loading variations on the data produced in the vibration test can be significant and misled about the true health state of the structure. This is because both effects provide changes to the vibration data. The driving or fundamental frequencies can be altered when there is damage (due to a reduction in material stiffness) inflicted on the structure or when operational and environmental changes present [4], [7], [8] and [20].

Through the GMM-EM algorithm, by incorporating clusters (ellipsoids), the severity of damage can be monitored by looking at how far the clusters translate or separate from the baseline set (undamaged condition). For vibration signals under the presence of damage and loading variations, the data from different categories/labels can overlap and intersect with each other. Concerning this requirement and properties of GMM calculated based on maximum likelihood make GMM a more flexible clustering technique compared to the linear k-means clustering [4], [10] and [13]. Due to the high dimensionality of the measured vibration data, a dimensional reduction technique is required. In this study, the KPCA is utilized to transform the data signal into a discrete group of data points, allowing GMM clustering to be incorporated to predict the likelihood of the data points generated by GMMs specified by the mean and covariance. Section 1 in the Introduction presents the overview of the work from the method applied using GMM and EM algorithms. A brief description of the experiment is provided here. Section 2 highlights the methodologies used in this work. The first part describes the experiment performed in this work, while the second part details the GMM, and EM algorithm applied to the data set. Section 3 presents a discussion about the results. Finally, a conclusion of this work is summarized in Section 4.

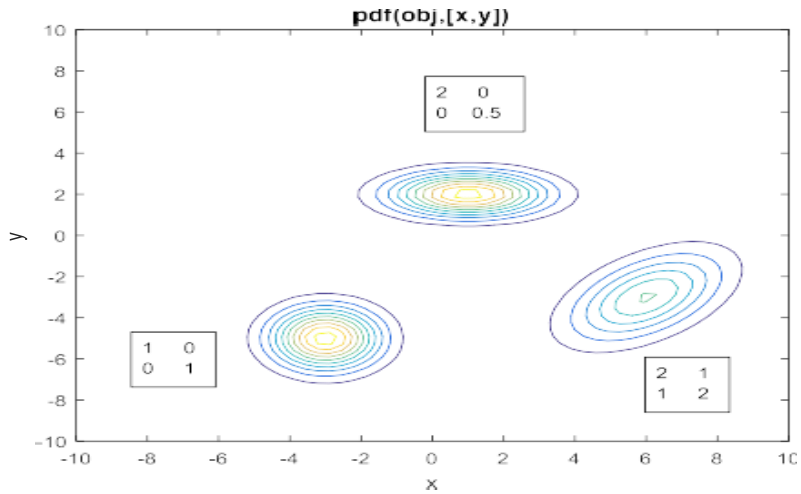

Fig. 1 - Clusters are fitted into data points based on covariance of different orientations to describe the size and orientation of each cluster

2. Methodology

2.1 An Experiment of a Wing Box with Attached Liquid-Tanks

The purpose of this section is to introduce the experimental configurations and loading variations that are accounted for in the work objectives. The unique of this work is by the operational load variables (by refilling and emptying) and the way the experimental work is carried out in cyclic and follow the systematic loading (increasing) and unloading (decreasing) of the liquid mass into/ from the added tank (as shown in Fig. 2 (a)).

2.1.1 Test Structure

The structure used is a stiffened aluminium panel to represent an aircraft wing box and it is a similar structure used in [1] as shown in Figure 4. The top sheet of the wing box is a 750 X 500 X 3 mm aluminium sheet. The structure is stiffened by two ribs of the length of C-channel riveted to the shorter edges and two stiffening stiffeners composed of angle sections, which are bolted along the length of the sheet. Free-free boundary conditions are approximated by suspending the wing box from a substantial frame using springs and nylon lines of heavy-duty type attached at the corners of the top sheet. The wing box structure is a weight of around 6.464 kg.

2.1.2 Damage Initiation

A sequential increase of damage severity is introduced into an inboard stiffener of the wing box by 16%, 33%, 50%, and 66%. These cuts correspond to the depth of 4 mm, 8 mm, 12 mm, and 16 mm respectively. The saw-cuts are introduced directly into one of the stiffeners by using a hacksaw without removing the stiffener from the structure to avoid any influence of boundary conditions on the vibration response as highlighted in the literature (as shown in Fig. 2 (b)) [20].

2.1.3 Data Acquisition

The acquisition system used in the test is a DIFA SCADAS III of 16-channel and high-speed data acquisition system, controlled by LMS software running on a Dell desktop PC. The measurements are recorded using a frequency range of 0-2048 HZ with a resolution of 0.25 Hz. The wing box is excited with a white Gaussian signal through an LDS shaker powered by an amplifier of a similar brand. The response is measured using PCB piezoelectric accelerometers mounted vertically on top of the wing box. The excitation signal is measured by a standard PCB force transducer.

The base measurements used in the test are FRFs acquired using sensor 1 which is located near the horizontal edge of the plate (Figure 3). Before locating the best place to attach the sensors, an impact test is done to detect the location where significant energy amplitude present using a hover method. In this test, the structure is excited using a random signal. The signal is applied with the Hanning window to improve the signal continuity and the measurements are performed with 8-averages. The measurements are taken from seven accelerometers mounted below the water tank and the wing plate and the signal that shows the highest sensitivity to DF and LF is considered (Fig. 4).

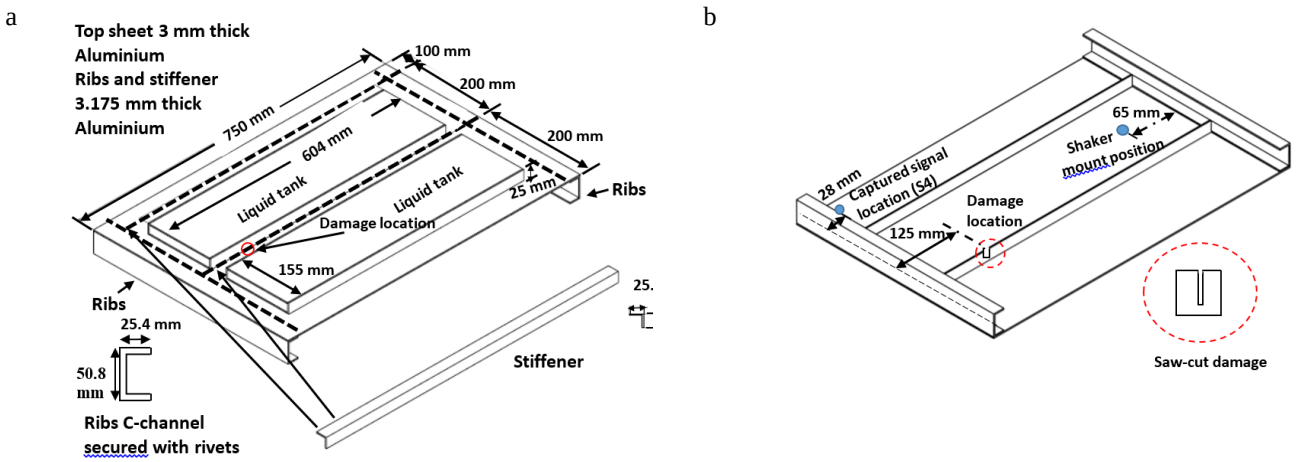

Fig. 2 - (a) The aircraft wing box with a refillable and dischargeable liquid tank; (b) second picture

2.2 Feature Selection

Damage sensitive and loading sensitive features are selected based on the moving of the frequency peaks because of the loading and damage effects. In vibration based SHM, any changes to material stiffness (due to damage) or changes in structural mass will shift the frequency peaks. This indication is used in determining the DS and LS features on the data signal. The selected DS and LS feature comprised of the spectral lines from 350 to 450 is established as the data feature to apply with GMM algorithm. This selected feature of frequency spectral lines consists of five different mass loadings with each mass loading groups encompassed of four different damage severities and one undamaged (normal) condition as illustrated in Fig. 5.

2.3 GMM-EM Algorithms

GMM algorithm is formulated by choosing enough number of Gaussian components C, means μ, covariance Σ and mixing proportions π_C to describe the Gaussian distributions for the data set. It uses an Expectation and Maximization algorithm (EM) to fit data points associated to the parameters of the GMM model, it consists of the means, covariance and mixing proportions of the Gaussians components as stated in equation (1). The EM is evaluated based on the log likelihood where it is checked for convergence as displayed in equation (1). Based on the same initial values and computed previously, the EM algorithm will then update the parameter values, evaluate the maximum log likelihood of the posterior probability, and repeat the steps until the criteria for convergence is met (follow equation 1).

$$\ln p(\mathrm{X}|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\left\{\sum_{c=1}^{C}\pi_c\mathcal{N}(\mathrm{x}_n|\mu_c,\Sigma_c)\right\} \quad (1)$$

Where the Gaussian distribution is based on the multivariate form,

$$\mathcal{N}(\mathrm{x}|\mu,\Sigma)=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\mathrm{x}-\mu)\Sigma^{-1}(\mathrm{x}-\mu)^{\mathrm{T}}\right\} \quad (2)$$

where μ is the *D*-dimensional mean vector, Σ is the *D* X *D* covariance matrix and $|\Sigma|$ denotes the determinant of Σ.

Derive equation (2)) with respect to the means μ_c of the Gaussians components to zero, the maximum of the log likelihood function is:

$$0=-\sum_{n=1}^{N}\overbrace{\frac{\pi_c\mathcal{N}(\mathrm{x}_n|\mu_c,\Sigma_c)}{\sum_j\pi_j\mathcal{N}(\mathrm{x}_n|\mu_j,\Sigma_j)}}^{r(z_{nc})}\Sigma_c(\mathrm{x}_n-\mu_c) \quad (3)$$

Note that the posterior probabilities or the responsibilities appear on the right-hand side of the derivatives of the log likelihood function in equation (3).

Multiplying by Σ_c^{-1} and rearrange the equation (3) gives

$$\mu_c=\frac{1}{N_c}\sum_{n=1}^{N}r(z_{nc})\mathrm{x}_n \quad (4)$$

The mean for the c_{th} Gaussian component μc is computed by taking a weighted mean of all data points in the data set whose weighting factor for data point x_n is given by the posterior probability $r(z_{nc})$ in which the component c is responsible for generating the data point x_n. Defining Nc as total number of data points effectively assigned to cluster c. The sum of this soft membership or fractional weight assigned to cluster c is described as

$$N_c=\sum_{n=1}^{N}r(z_{nc}) \quad (5)$$

The covariance matrix in the frame of maximum likelihood solution is given in terms of the weighted responsibility for the component *c* that generates the data point can be stated as

$$\Sigma_c=\frac{1}{N_c}\sum_{n=1}^{N}r(z_{nc})(\mathrm{x}_n-\mu_c)^T(\mathrm{x}_n-\mu_c) \quad (6)$$

It is the equivalent form with the corresponding result for a single Gaussian with each data point weighted by the corresponding posterior probability $r(z_{nc})$ and multiplied by the inverse N_c.

π_c is the mixing coefficient for the c^{th} component given by the average responsibility which the component takes for explaining the data points. In other words, it is the total responsibility allocated to cluster *c* normalized by the total number of data *N*, given as

$$\pi_c=\frac{N_c}{N} \quad (7)$$

The first step is the E-step that choose the parameter means μ and covariance $\sum$ and mixing coefficients as fixed values. The probability of x, p(x) assigned to component c with a weighted Gaussian π_c is normalised by the total values of c. Here the responsibility of data point x belongs component c is reinstated as:

$$r_{ic} = \frac{\pi_c \mathcal{N}(\mathrm{x}_i | \mu_c, \Sigma_c)}{\sum_{c'=1}^{C} \pi_{c'} \mathcal{N}(\mathrm{x}_i | \mu_{c'}, \Sigma_{c'})} \quad (8)$$

Practically, r(z_{nc}) is a number of data points by number of clusters that sums to one over the index c. If x is very likely to be under the Gaussian component c, it will get high responsibility value r(z_{nc}). The denominator just makes the sum of r(z_{nc}) equal to one. The second step in EM is the maximization step. It starts with the probability assignment r(z_{nc}) and update the clusters' parameters μ_c, Σ_c and π_c. The parameters are weighted by r(z_{nc}) so that if x_n is a strong member of cluster c, this weight will be nearly one, but if x_n is not very well explained by cluster c then it will not influence the average very much. Using the same initial values of the GMM parameters belonging to the undamaged condition, EM algorithm will then update the parameter values, evaluate the maximum log likelihood of the posterior probability, and repeat the steps until the criteria for convergence is achieved (following equation 1).

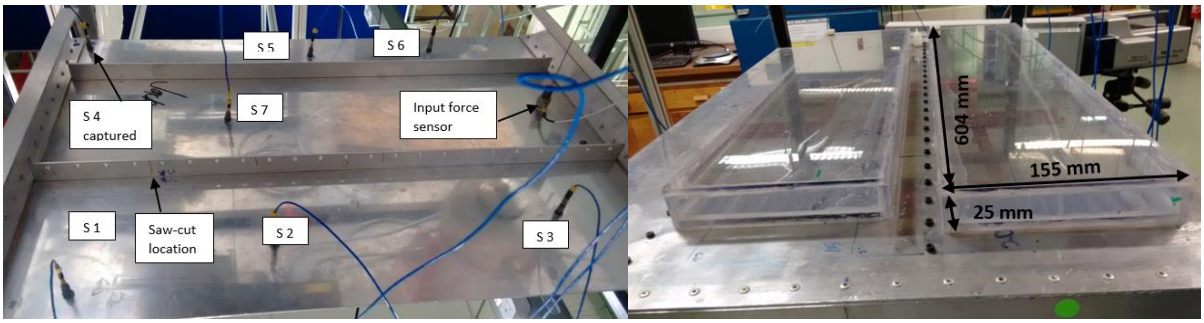

Fig. 3 - Beneath the wing box where piezoelectric sensors and input sensor are located (left). On top of the plate where the water tanks are attached to (right)

3. Results and Discussion

Incorporating the GMM-EM algorithm on data obtained from KPCA (Fig. 5), 25 new ellipsoids are established, comprised of 5 loading classes (empty, quarter-full, half-full, three-quarter full, and full load). Within those loading classes, another 5 different damage levels (UD, D1, D2, D3, and D4) are predicted using the clustering algorithm. The results from Fig. 6 show the fitting of ellipsoids on whole data using the GMM-EM algorithm based on the initial conditions specified by the UD (baseline) class consists of complete all five loading classes. The data signal obtained in Frequency Response Function (FRF) in Fig. 4 is first transformed into discrete variables using KPCA to allow the clustering process. The GMM algorithm fits the ellipsoidal on each Gaussian component produced from different structural health and loading conditions very well (Fig. 6). All clusters show consistent translation, moving away from the undamaged (UD) state as the damage level increases. It reveals accurate prediction of damage groups that use only the baseline data set as their initial conditions. The result is encouraging, considering that all the means and covariance of the initial values of the test data are based on the undamaged condition encompassing all loading conditions. The center point (0,0) lies right at the center of the projection space (Fig. 6)

A zoom on the quarter-full load class shows the data belonged to UD, D1, and D2 lying close to each other (Fig. 7). Applying the GMM-EM algorithm can predict the data points that are most likely to belong to their actual label. However, for data points that fall apart from the high-density group of undamaged and small damage, they are most likely to be clustered within the probability distribution of higher damage group (Fig. 8). Results in Fig. 8 and Fig. 9 reveal clearer clusters fitting based on the parameters of the baseline set. On Half-full load (Fig. 9), the clusters are predicted accurately, especially for the data points UD and D1 class which the data points from both classes heavily overlap. This feature is the key advantage of the GMM that allows clustering of overlap data groups. Practically, this model can establish a predictive model by observing the degree of cluster separation from the baseline (UD) set. It implies that as the clusters of damage states move away from the baseline clusters, the damage is more severe and may require greater attention from the maintenance personnel. All the ellipsoids revealed an excellent correlation to damage severities that describe an increase in damage severity as the clusters separate more from the baseline data set. No other parameters are required except the total number of components (data

groups) including both damage and loading classes which need to be specified. The number of components of loading classes can be easily identified using KPCA as they are clearly separated after the transformation (Fig. 5).

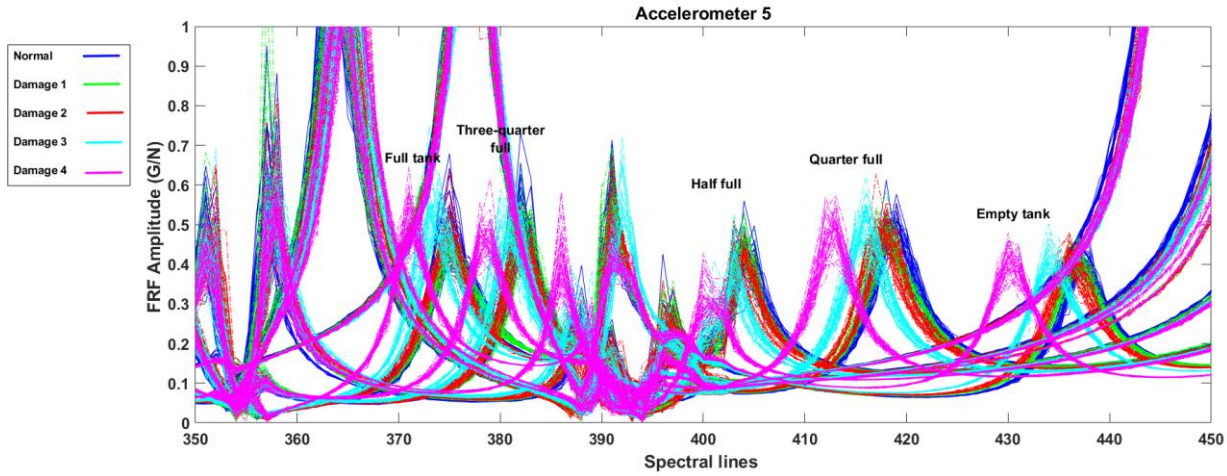

Fig. 4 - The data signal in the lower range obtained from the accelerometer 5, which is chosen as the input feature for the GMM

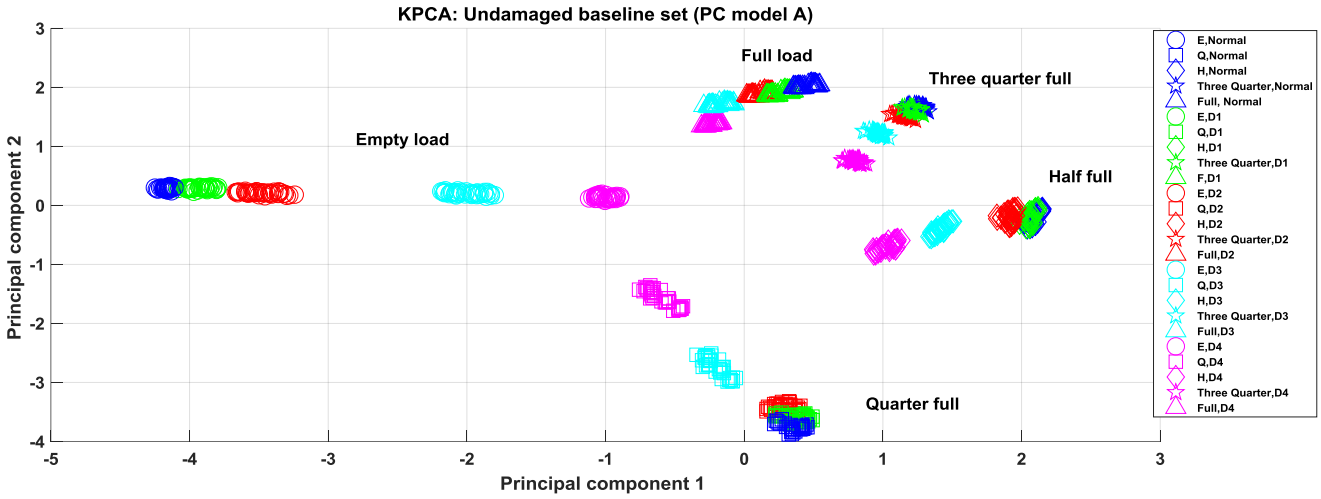

Fig. 5 - Transforming the data into discrete variable using KPCA before applying GMM-EM algorithm

4. Conclusion

The GMM clustering takes the initial values (mean and covariance) from the baseline model. The advantage of such a model is that it only needs to establish the parameters from the undamaged state compared to a more costly damaged state. This study demonstrates that GMM based on the maximum likelihood and EM function can correctly predict data groups belonging to various damage groups. Based on the means and covariance of the baseline set, the algorithm forms clusters (ellipsoidal shapes) on each of the data groups that are likely to fall into each GMM distribution. KPCA based data transformation is useful before applying the GMM-EM method. KPCA transforms the data variables into discrete data groups that allow clustering to be performed on the discrete data groups. The challenging part is that when the data points of one group intercept with the other group. Defining the GMM parameters (mean, covariance, and mixing coefficients) based on the baseline data set can define a cluster that overlaps with the undamaged group (baseline). The study demonstrates the incorporation of KPCA with the GMM-EM algorithm to identify various damage severities groups within changing loading mass through clustering. This method has a potential application in identifying different data groups based on Gaussian distribution and EM for the interest of SHM. Based on healthy baseline data groups to specify the initial parameters, clustering on different damage groups can be effectively performed throughout the loading classes.

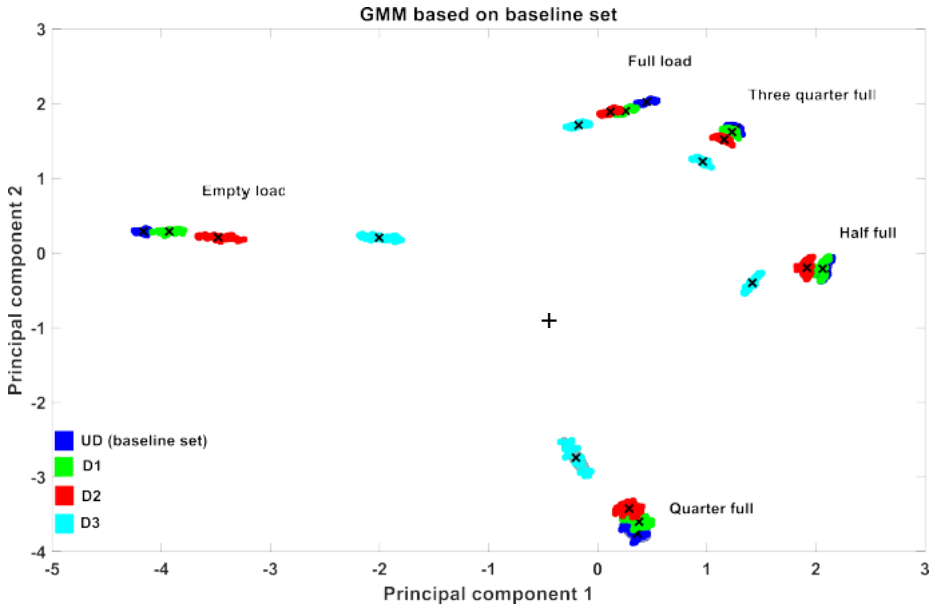

Fig. 6 - Data clustering based on the initial means and covariance of the baseline data group (undamaged condition)

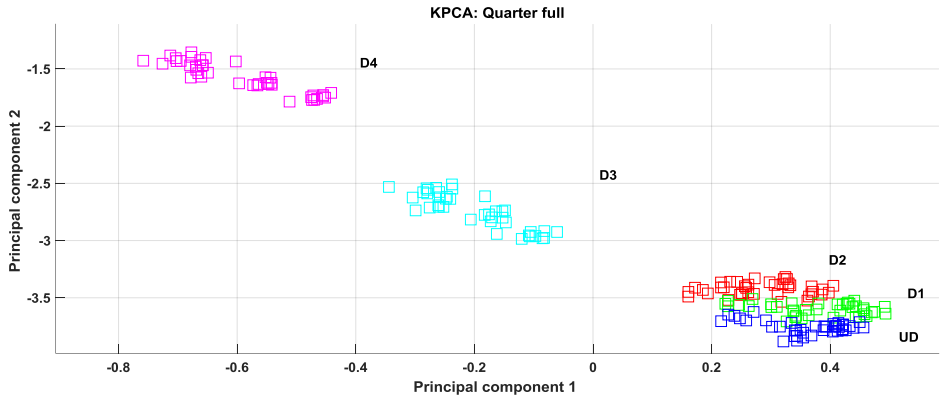

Fig. 7 - A zoom in on a quarter full load in Figure 7, highlights the data points on Quarter-full load (before clustering)

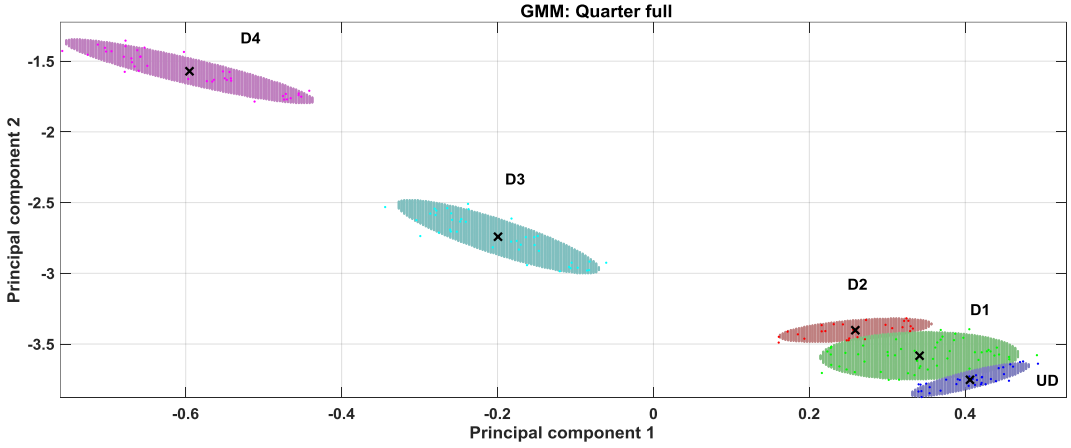

Fig. 8 - Zooming on data clustering on a quarter full load in Figure 7, illustrates the clustering on Quarter-full load data points

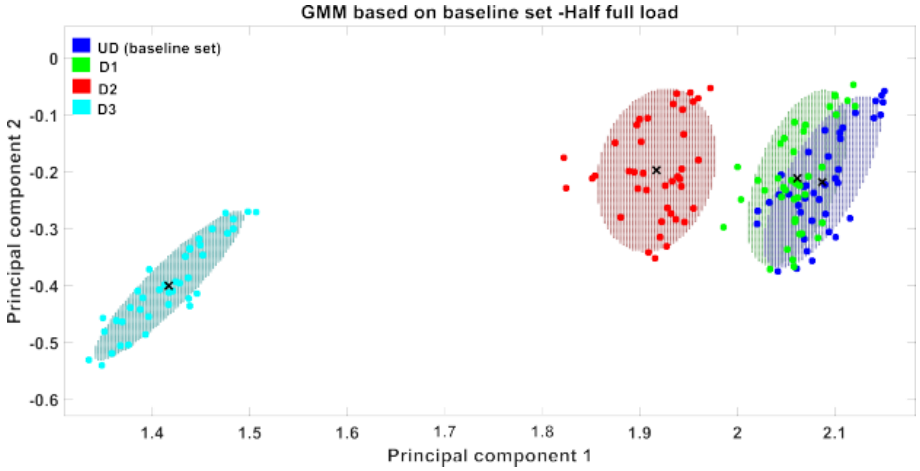

Fig. 9 - Zooming on a half full load in Figure 7, shows the Gaussian density model nicely fitted into each ellipsoid

Acknowledgements

This work is part of the author's PhD research work at the University of Sheffield from 2014 until 2018. He would like to thank his supervisor Dr Graeme Manson for his intellectual guidance put into this work. He also like to thank Prof Keith Worden from the University of Sheffield for sharing invaluable insights on this topic and his novel idea for this topic. This work has been supported by Ministry of Education Malaysia and Universiti Putra Malaysia where the author currently attached to.

References

[1] Bull, L.A., Worden, K., Fuentes, R., Manson, G., Cross, E.J., Dervilis, N. (2019). Outlier ensembles: A robust method for damage detection and unsupervised feature extraction from high-dimensional data. Journal of Sound and Vibration, 453, 126 –150

[2] Cross, E.J., Manson, G., Worden, K., Pierce, S.G. (2012). Features for damage detection with insensitivity to environmental and operational variations. Proceeding Royal Society Mathematics Physics Engineering Science, 468, 4098–4122

[3] Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer

[4] Farrar, C.R., Worden, K. (2013). Structural Health Monitoring: A machine learning perspective. John Wiley & Sons Ltd

[5] Rahim, S. (2018). Investigating the Effect of Variable Mass Loading in Structural Health Monitoring from a Machine Learning Perspective. PhD thesis, University of Sheffield

[6] Wah, W. S. L., Chen, Y., Roberts, G, W., Elamin, A. (2017). Damage Detection of Structures Subject to Nonlinear Effects of Changing Environmental Conditions. Procedia Engineering, 188, 248–255

[7] Ghazi, R. M., Marzouk, Y. M., Büyüköztürk, O. (2018). Conditional classifiers and boosted conditional Gaussian mixture model for novelty detection. Pattern Recognition, 81, 601–614

[8] Li, K., Ma, Z., Robinson, D., Ma, J. (2018). Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering. Applied Energy, 231, 331-342

[9] Qiu, L., Yuan, S., Chang, F., Bao, Q., Mei, H. (2014). On-line updating Gaussian mixture model for aircraft wing spar damage evaluation under time-varying boundary condition. Smart Mater Structure, 23

[10] Patel, E., Kushwaha, D.S. (2020). Clustering Cloud Workloads: K-Means vs Gaussian. Procedia Computer Science, 171, 158–167

[11] Qiu, L., Fang, F., Yuan, S. (2019). Improved density peak clustering-based adaptive Gaussian mixture model for damage monitoring in aircraft structures under time-varying conditions. Mechanical Systems and Signal Processing, 126, 281–304

[12] Sobkowicz, J.H., Zimroz, R., Pitera, M., Wyłomańska, A. (2020). Informative frequency band selection in the presence of non-Gaussian noise – a novel approach based on the conditional variance statistic with application to bearing fault diagnosis. Mechanical Systems and Signal Processing, 145, 126

[13] Valencia, L.D.A., Fassois, S.D. (2017). Gaussian Mixture Random Coefficient model based framework for SHM in structures with time–dependent dynamics under uncertainty. Mechanical Systems and Signal Processing, 97, 59–83

[14] Silva, M., Santos, A., Santos, R., Figueiredo, E., Sales, C., Costa, J.C.W.A. (2017). Agglomerative concentric hypersphere clustering applied to structural damage detection. Mechanical Systems and Signal Processing, 92, 196–212

[15] Wah, W.S.L, Owen, J.S., Chen, Y.T., Elamin, A., Roberts, G.W. (2019). Removal of masking effect for damage detection of structures. Engineering Structures, 183, 646–661
[16] Chen, H., Ma, H., Chu, X., Xue, D. (2020). Anomaly detection and critical attributes identification for products with multiple operating conditions based on isolation forest. Advanced Engineering Informatics, 46, 101139
[17] Yu, J. (2012). A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. Chemical Engineering Science, 68, 506–519
[18] Li, H., Hansman, R.J, Palacios, R., Welsch, R. (2016). Anomaly detection via a Gaussian Mixture Model for flight operation and safety monitoring. Transportation Research Part C: Emerging Technologies, 64, 45-57
[19] Fuentes, R., Joyce, R.S.D, Marshall, M.B., Wheals, J., Cross, E.J. (2020). Detection of sub-surface damage in wind turbine bearings using acoustic emissions and probabilistic modelling. Renewable Energy, 147, 776–797
[20] Avci, O., Abdeljaber, O., Kiranyaz, S., Hussein, M., Gabbouj, M., Inman, D.J. (2021) A review of vibration-based damage detection in civil structures: From traditional methods to Machine Learning and Deep Learning applications. Mechanical Systems and Signal Processing, 147, 107077
[21] Manson,G. (2002). Identifying damage sensitive, environment insensitive features for damage detection. Proceeding Third Conference Identification Engineering System. University of Wales, Swansea