Mechanical Systems and Signal Processing 119 (2019) 100–119

Contents lists available at ScienceDirect

Mechanical Systems and Signal Processing

journal homepage: www.elsevier.com/locate/ymssp

A Bayesian non-parametric clustering approach for semi-supervised Structural Health Monitoring

T.J. Rogers [a,*], K. Worden [a], R. Fuentes [a], N. Dervilis [a], U.T. Tygesen [b], E.J. Cross [a]

[a] Department of Mechanical Engineering, Dynamics Research Group, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK
[b] Ramboll Oil & Gas, Bavnehøjvej 5, DK-6700 Esbjerg, Denmark

ARTICLE INFO

Article history:
Received 17 April 2018
Received in revised form 28 August 2018
Accepted 5 September 2018

Keywords:
Structural health monitoring
Damage detection
Bayesian methods
Clustering
Semi-supervised learning

ABSTRACT

A key challenge in Structural Health Monitoring (SHM) is the lack of availability of data from a full range of changing operational and damage conditions, with which to train an identification/classification algorithm. This paper presents a framework based on Bayesian non-parametric clustering, in particular Dirichlet Process (DP) mixture models, for performing SHM tasks in a semi-supervised manner, including an online feature extraction method. Previously, methods applied for SHM of structures in operation, such as bridges, have required at least a year's worth of data before any inferences on performance or structural condition can be made. The method introduced here avoids the need for training data to be collected before inference can begin and increases in robustness as more data are added online. The method is demonstrated on two datasets; one from a laboratory test, the other from a full scale test on civil infrastructure. Results show very good classification accuracy and the ability to incorporate information online (e.g. regarding environmental changes).

1. Introduction

Structural Health Monitoring (SHM) [1,2] is an important area of research within engineering, seeking to detect and diagnose degradation in structures and systems before it can impede use or become a hazard. Given the maturity and availability of sensing hardware, a data-driven approach is commonly adopted. Here, statistical models can be used to detect similarity (or difference) between sets of data collected from a structure, which is, in turn, used to infer its health/condition. Data-driven approaches, if wanting to achieve more than novelty detection, require training data from multiple healthy and damage states which is a significant limitation.

In many cases, it will not be possible to acquire data covering all healthy conditions and damage scenarios, the main limitation being the cost of producing and subsequently damaging large valuable structures, e.g. within the aerospace industry or civil infrastructure. A particular challenge in civil infrastructure stems from the fact that structures are often unique. The existence of a number of different damage scenarios comes from the multiple mechanisms for damage that a structure might experience. For example, in an aerospace structure it would be desirable to detect degrading performance from fatigue damage accrual, but damage introduced by low-velocity impact is also of concern. In certain cases it will be unsafe to operate the structure with a given type of damage present, meaning that collection of data from this damage state prior to operation of the structure is not possible. Additionally, a structure will operate in a number of different operational and environmental

* Corresponding author.
E-mail address: trogers3@sheffield.ac.uk (T.J. Rogers).

https://doi.org/10.1016/j.ymssp.2018.09.013
0888-3270/

conditions, which result in significant changes to the measured dynamic behaviour. Continuing with the example of an aerospace structure, it is clear that there will be significant changes in behaviour between flight and taxiing. It is less obvious, however, that there may be other confounding influences such as crosswinds on landing or freezing temperatures which will affect the behaviour of the structure. One could go on attempting to imagine all the scenarios possible for changes in operating condition, but this is a fruitless exercise, as it quickly becomes apparent that collecting data from all these conditions is not feasible [1,3], not least because the operator normally has little or no control over these factors.

It is desirable, therefore, to consider methods which will allow the incorporation of operational data into the training of a given algorithm, which adapts, as time progresses. These methods are commonly referred to as *online learning* [4]. Rather than pure novelty detection from a known, healthy, baseline state, it would be beneficial to be able to first detect a new regime, then label it and be able to recognise that behaviour, should it occur, in the future. This is also sometimes called the *semi-supervised* learning approach [5], where new regimes are discovered in the data which are labelled in operation and incorporated into future analysis, this process of inspecting online leads to a partially labelled dataset.

This paper presents a process for using a Bayesian non-parametric clustering technique to learn clusters of data online without a training phase, or with restricted training data. Then, applying labels online to the clusters in a *semi-supervised* manner, the algorithm becomes an environmental/damage state classifier, reducing the occurrence of false positives as time progresses and the algorithm learns more normal states. The layout of the paper is as follows; Section 2 presents a short review of some key related work. Section 3 outlines the standard finite Gaussian Mixture Model, Section 4 the Bayesian formulation of the Dirichlet Process Gaussian Mixture Model. A method is proposed for application to SHM problems in Section 5 and this procedure is followed for two datasets in Section 6. Finally, a discussion of the method is presented, in light of the results, in Section 7.

2. Related work

Traditionally, approaches to SHM from a machine learning perspective have considered only *unsupervised* and *supervised learning* tasks[1] [1]. *Unsupervised learning* applications are dominated by two-class classification tasks based on outlier analysis [6,7]. A baseline healthy state is used to define a "normal" condition and then deviations from this can be detected in an online manner. The problem of *supervised learning* in SHM is usually concerned with regression or classification tasks which provide information regarding the type, location, or severity of damage in a structure [8].

Treatment of SHM as an *unsupervised learning* task has been mainly limited to an outlier detection problem, usually in a laboratory setting [9,10]. The challenge in this research has been in building algorithms that are robust to false alarm and environmental changes. A number of methods have been developed which handle this problem well [11,12]. However, a drawback to the most common approaches to dealing with confounding influences, is that they reduce SHM to a two-class problem, where distinction is only made between damaged and undamaged states. This fails to give additional information about the operating conditions of the structure, which would be useful for an operator to know, or, indeed, about any damage or performance anomalies that occur. To counteract this, a popular approach has been to consider clustering in an unsupervised manner. The most common approaches employ Gaussian Mixture Models [13–16], or other clustering techniques [17–20] in an offline manner. Tibaduiza et al. [21] present another unsupervised methodology based on self-organising maps of features from ultrasonic pitch-catch data.

The alternative to this, and preferred option when interested in additional information, is the *supervised learning* task. Here, a training dataset is formed which has information from all possible structure states; inference about the current state can now be made via pattern recognition or machine learning methods, where new observations/data are compared to the training set.

Although tools exist which perform very well in the supervised learning problem, a common stumbling block is the lack of availability of complete datasets for algorithm training. It is usually prohibitively expensive to acquire training data from all environmental conditions and damage states. For this reason, the development of algorithms which can be established/learn fully or partially online is of particular interest. Langone et al. [22] propose an adaptive learning algorithm based on a kernel PCA transformation, they demonstrate this by performing damage detection on a benchmark dataset – the Z24 bridge. The algorithm performs well on benchmark data but requires an initialisation and calibration phase before being fully operational; in this phase, the structure is assumed undamaged. The method also requires user input regarding thresholds and the expected number of clusters. Chen et al. [23] present a semi-supervised algorithm for damage detection based on a multi-resolution classification with adaptive graph filtering; the features are extracted by passing the input signals through a filter bank. A graph-filtering algorithm estimates the labels for unknown data given previously labelled features, and a regression step is able to compensate for missing data in the problem informed by the graph filter.

Finite Gaussian Mixture Models (GMM) in SHM have been used previously with promising results [24–26]; the strength of the GMM is in the ability of training data to shape clusters and form a probabilistic representation of the different possible states that the structure could be in, undamaged or damaged, with the possibility for multiple examples from each. The key

[1] For the purpose of this paper *unsupervised learning* is defined as a situation in which data is available without any labels or outputs. This can include the case where a dataset is collected from what is assumed to be a *normal* condition. The *supervised learning* task is treated as one where data is available with both the inputs and outputs (either continuous or labels) from which methods for classification or regression can be trained.

difficultly in implementing a finite GMM without a complete training set is the specification of the number of Gaussians in the mixture. The method proposed in this paper uses a Dirichlet Process (DP) clustering model to remove the need to pre-specify the number of clusters that are expected, while retaining a Bayesian formulation as opposed to methods such as affinity propagation [27].

DP models have been employed in a number of machine learning tasks including: Natural Language Processing [28] and topic modelling [29,30], where documents can be grouped according to thematic similarities. In image analysis, the model has been used to generate captions for images [31]; it has also found use in medical image analysis [32,33], for clustering regions of the brain from data collected by MRI or fMRI, and in genetic analysis [34,35]. In other medical applications, DP mixture models have been used for sorting neural spike data [36].

Previously, a DP mixture model has been shown to be effective in the feature selection step in SHM [37]. In that work, the outputs of the DP clustering model are used as features in a further analysis step – a particle-filter based damage progression model – where they are combined with a physical model. Only the number of clusters identified by the DP is used as a feature, which does not make full use of the Bayesian nature of the DP clustering method.

The approach adopted by the authors here makes use of the Bayesian properties of the DP to allow incorporation of prior knowledge and updates of belief given observed data. The aim is to avoid the need for a training dataset before the process begins, but retain flexibility to include any training data as a formal prior belief. In addition there is a reduction in the number of required user-tuned parameters in the model. In this way, a model is developed which can perform powerful *online learning* with minimal required *a priori* knowledge in terms of access to data or a physical model. The work in this paper aims to show how such a model can be implemented online for use in SHM. To achieve this, a novel feature selection approach is also explored, making use of Random Projection [38] of high dimensional frequency domain features.

3. Finite Gaussian Mixture Models

It is useful at this point to review the formulation of a standard Gaussian Mixture Model. Modelling data which is inherently non-Gaussian poses a challenge, as typically the inference becomes harder. It is possible to imagine that the data has been generated, not by some complex non-Gaussian process, but from a mixture of independent Gaussian distributions. In SHM, one could assume that during normal operation, features are clustered according to one Gaussian distribution; however, when damage occurs, the features are drawn from a separate Gaussian with different parameters. More Gaussians can be added to cover many different scenarios relating to changing operating conditions or different damage cases.

It is possible to construct a probabilistic model which describes this behaviour. First, one proposes a multinomial distribution $\boldsymbol{\pi}$[2], in which each element π_k is the probability that a data point comes from each class, $k = 1, \ldots, K$ for K classes, and $\sum_{k=1}^{K} \pi_k = 1$. In other words, $\boldsymbol{\pi}$ is merely the probability that the structure is in each state.

Each state of the structure is defined by its own Gaussian distribution which has a mean, $\boldsymbol{\mu}_k$, and covariance, Σ_k. This model is shown in Fig. 1 and it is possible to write it down as below:

$$\begin{aligned} \mathbf{x}_i \mid c_i &\sim \mathcal{N}\left(\boldsymbol{\mu}_{c_i}, \Sigma_{c_i}\right) \\ c_i &\sim \text{Mult}(\boldsymbol{\pi}) \end{aligned} \tag{1}$$

In order to use this model, the parameters must first be determined. The parameters of the model include: the number of clusters K; the mixing proportions $\boldsymbol{\pi}$; and the cluster parameters, $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_K\}$. This gives a total parameter vector, $\Theta = \{K, \boldsymbol{\pi}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1, \ldots, \Sigma_K\}$. Additionally, θ_k is defined as $\theta_k = \{\pi_k, \boldsymbol{\mu}_k, \Sigma_k\}$. Determining these parameters can be accomplished quite efficiently via Expectation Maximisation [4] for $\theta_{1:K}$ and for K either the Bayesian Information Criterion [39] or Akaike Information Criterion [40] can be used. This will give the maximum likelihood solution to the model given the currently observed data, however, a Bayesian solution to the problem has also been explored for SHM [41] or more generally, for the GMM, in [42].

4. Dirichlet Process Gaussian Mixture Models

The desirable modification to this hierarchical finite GMM is to make the inference over Θ Bayesian. This will give more robust estimates of the parameters, $\theta_{1:K}$ (i.e. the parameters in θ_k for all clusters $k = 1, \ldots, K$), and allow a probabilistic selection of K through use of the Dirichlet Process prior. The Bayesian approach allows incorporation of prior knowledge, such as the expected effects of damage, in a formal manner. Conversely, it also allows the data observed to shape the model belief.

Firstly, priors are placed over the cluster parameters, $\boldsymbol{\mu}_k$ and Σ_k. To help with inference over the model, these priors are chosen to be conjugate with the Gaussian distribution which is the likelihood, therefore, the prior over the means is a multivariate Gaussian and the prior over the covariances is an Inverse-Wishart ($\mathcal{IW}$). These prior distributions have their own hyperparameters associated with them which are, $\boldsymbol{\mu}_0, \kappa_0, \Sigma_0, \nu_0$. It is usual to combine these into a single prior distribution over the cluster parameters H.

[2] The convention adopted in this paper is to use **bold** lower case letters or symbols to represent vectors and UPPERCASE letters to represent matrices.

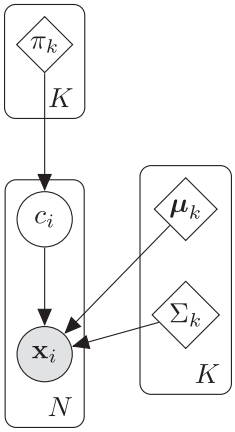

Fig. 1. Graphical model of a finite Gaussian Mixture Model with K components in the mixture.

$$\begin{aligned} H &= \mathcal{NIW}(\boldsymbol{\mu}_0, \Sigma_0, \kappa_0, \nu_0) \\ &= \mathcal{N}\left(\boldsymbol{\mu} \mid \boldsymbol{\mu}_0, \frac{\Sigma}{\kappa_0}\right) \mathcal{IW}(\Sigma \mid \Sigma_0, \nu_0) \end{aligned} \tag{2}$$

To perform Bayesian inference over the mixing proportions $\boldsymbol{\pi}$, as well as the cluster parameters, another prior must be specified. The sensible choice again is to choose the conjugate prior to the Multinomial distribution, which is a Dirichlet distribution governed by a strength parameter α, which is a single number when a symmetric Dirichlet distribution [43], is used, as in this case. Following [44], it is possible to take the limit of $K \to \infty$ and form an infinite Gaussian mixture model (IGMM) for which the generative model is shown in Eq. (3) and the graphical model is seen in Fig. 2.

$$\mathbf{x_i} \mid c_i \sim \mathcal{N}(\mathbf{x_i} \mid \boldsymbol{\mu}_{c_i}, \Sigma_{c_i}) \tag{3a}$$

$$\boldsymbol{\mu}_{c_i} \mid \Sigma_{c_i}, c_i \sim \mathcal{N}\left(\boldsymbol{\mu}_{c_i} \mid \boldsymbol{\mu}_0, \frac{\Sigma_{c_i}}{\kappa_0}\right) \tag{3b}$$

$$\Sigma_{c_i} \mid c_i \sim \mathcal{IW}(\Sigma_{c_i} \mid \Sigma_0, \nu_0) \tag{3c}$$

$$c_i \mid \boldsymbol{\pi} \sim \text{Mult}(\boldsymbol{\pi}) \tag{3d}$$

$$\boldsymbol{\pi} \sim \text{Dir}(\alpha) \tag{3e}$$

The strength of this formulation for a mixture model in the SHM context, is that only the hyperparameters need to be specified to use the model, there is no tuning of thresholds or calibration phase. Practically, this means that to implement the model, the operator does not need to specify a number of expected normal or damage conditions, which is difficult or impossible for a structure in operation. Nor does the user need to specify the changes that damage on the structure will

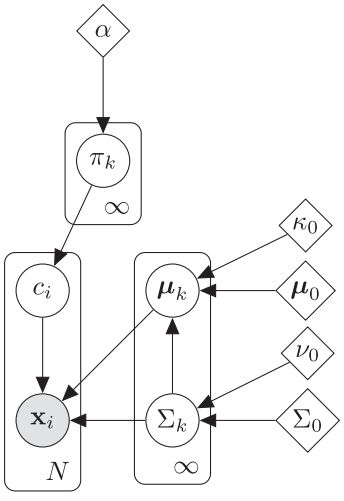

Fig. 2. Graphical model of the Infinite Gaussian Mixture Model.

introduce to the data (derived from the physical mechanism of damage or a large number of expensive tests); although in the presence of training data, this can be easily introduced by including clusters in the model where the prior parameters of those clusters are the posteriors of the parameters when the known data is added to the cluster.

A collapsed Gibbs sampler can be used to make efficient online inference over this model [45]. The collapsed Gibbs sampler refers to the process of analytically marginalising certain variables in the model. For the DPGMM, the cluster parameters, means and covariances, can be marginalised analytically removing the need to sample them. This gives a posterior distribution over each of the parameters against which new data can be assessed via the posterior predictive distribution, $p(\mathbf{x_i} \mid \mathcal{D}_{-i})$[3], the likelihood of point $\mathbf{x_i}$ given the rest of the observed data. Although potentially faster algorithms for variational inference in the Dirichlet Process mixture model exist [46,47], it is more practical to implement the Gibbs sampler when performing inference online. The nature of the Gibbs sampling solution is that each data point is assessed marginally in the sampler, this allows the addition of new points online rather than requiring batch updates.

For the case of a Gaussian base distribution, the Gibbs sampler proceeds as follows. The data are initially assigned to random clusters, then at each iteration one of the data points is chosen to be (re) assessed. This point is removed from its current cluster assignment, c_i, and the parameters of that cluster are updated. If that data point was the only point assigned to that cluster, it is destroyed and the total number of clusters, K, is updated. For each cluster, $k = 1, \ldots, K$, the prior likelihood that the point was drawn from that cluster k, is assessed. The prior is a Dirichlet Process prior, which for an existing cluster is equal to:

$$p(c_i = k \mid \mathbf{c}_{-i}, \alpha) = \frac{N_{-i,k}}{N + \alpha - 1} \tag{4}$$

It can be seen that the prior likelihood is governed by the hyperparameter, α, and the number of points currently assigned to that cluster, $N_{-i,k}$. The prior encourages clusters to grow, increasing α will make a higher number of clusters more likely. Since the information from the other data points should also be included in the clustering process, the likelihood term must be computed to get the posterior likelihood of the point belonging to each cluster, up to a constant. That is, compute: $p(c_i = k|\mathbf{x_i}, \mathbf{c}_{-i}, X_{-i,k}, \alpha, \boldsymbol{\beta}) \propto p(\mathbf{x_i} \mid X_{-i,k}, c_i = k, \boldsymbol{\beta}) p(c_i = k|\mathbf{c}_{-i}, \alpha)$ where $\boldsymbol{\beta} = \{\boldsymbol{\mu}_0, \Sigma_0, \kappa_0, \nu_0\}$, the prior constants of the base distribution. This is the posterior probability for the assignment of data point i to cluster k, given the data value, $\mathbf{x_i}$, the current cluster assignments, $\mathbf{c}_{-i}$, the data already assigned to that cluster, $X_{-i,k}$, and the hyperparameters α and $\boldsymbol{\beta}$.

The computation of the likelihood term, $p(\mathbf{x_i} X_{-i,k}, c_i = k, \boldsymbol{\beta})$, involves calculating the posterior predictive likelihood of that data point $\mathbf{x_i}$ being in cluster k. As data are added to each cluster the parameters of that cluster are updated via conjugate (closed form) updates to the Gaussian which defines it. The model requires a posterior distribution over the parameters of each Gaussian cluster: $\boldsymbol{\mu}_k$ the mean and Σ_k the variance. This leads to a prior over the cluster parameters[4],

$$\Sigma \sim \mathcal{IW}_{\nu_0}(\Sigma_0) \tag{5a}$$
$$\boldsymbol{\mu} \mid \Sigma \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \Sigma/\kappa_0\right) \tag{5b}$$
$$p(\boldsymbol{\mu}, \Sigma) \equiv \mathcal{NIW}(\boldsymbol{\mu}_0, \kappa_0, \Sigma_0, \nu_0) \tag{5c}$$
$$\propto |\Sigma|^{-((\nu_0+d)/2+1)} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Sigma_0^{-1}\Sigma^{-1}\right) - \frac{1}{2}\kappa_0(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)\right) \tag{5d}$$

The updates to the posterior parameters of the cluster are efficient since the priors have been chosen to be conjugate. The conjugate updates when n data points have been observed, are computed as shown,

$$\boldsymbol{\mu}_n = \frac{\kappa_0}{\kappa_0 + n}\boldsymbol{\mu}_0 + \frac{n}{\kappa_0 + n}\bar{x} \tag{6a}$$
$$\kappa_n = \kappa_0 + n \tag{6b}$$
$$\nu_n = \nu_0 + n \tag{6c}$$
$$\Sigma_n^{-1} = \Sigma_0^{-1} + S + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)^{\mathrm{T}} \tag{6d}$$

Here, S is defined as the sum of squares matrix around the sample mean, $\bar{x}$,

$$S = \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^{\mathrm{T}} \tag{7}$$

It can be shown that when updating a cluster by a single data point (adding or removing a single point), the updates can be carried out as Rank 1 updates to a Cholesky decomposition of the covariance matrix of the posterior, which significantly improves the speed of the computation [48,49]. The distribution of interest for calculating the likelihood term in the DP mixture model is sometimes referred to as the posterior predictive distribution $p(x \mid \mathcal{D}_k)$; the likelihood that a new point x was drawn from the posterior distribution of the currently observed data $\mathcal{D}_k$, in that cluster under the assumed prior. For the model being considered, this is given by a multivariate-t distribution with $\nu_n - d + 1$ degrees of freedom,

[3] The notation $-i$ is used to indicated all points except for point i

[4] Here $\mathrm{tr}(\cdot)$ indicates the trace operator

$$p(\mathbf{x} \mid \mathcal{D}_k) = t_{\nu_n - d + 1}\left(\boldsymbol{\mu}_n, \frac{\Sigma_n^{-1}(\kappa_n + 1)}{\kappa_n(\nu_n - d + 1)}\right)$$
$$= Z\left(1 + (\nu_n - d + 1)^{-1}(\mathbf{x} - \boldsymbol{\mu}_n)^{\mathrm{T}}\left(\frac{(\kappa_n(\nu_n - d + 1))}{\Sigma_n^{-1}(\kappa_n + 1)}\right)(\mathbf{x} - \boldsymbol{\mu}_n)\right)^{-(\nu_n - d + 1)/2} \tag{8a}$$

Where,

$$Z = \frac{\Gamma((\nu_n + 1)/2)}{\Gamma((\nu_n - d + 1)/2)(\nu_n - d + 1)^{d/2}\pi^{d/2}}\left|\frac{(\kappa_n(\nu_n - d + 1))}{\Sigma_n^{-1}(\kappa_n + 1)}\right|^{1/2} \tag{8b}$$

As the degrees of freedom of this distribution increases, it tends towards a Gaussian. Since the t distribution has similar shape to a Gaussian but with heavier tails, this has an interesting interpretation in the clustering model. When clusters have fewer points, a new point which is assessed in the tails of the distribution will have a higher likelihood than if a Gaussian were used. Practically, this will allow small clusters to still accept new points and reduce bias introduced from the small number of points defining the cluster.

Having computed the prior and likelihood for each of the existing clusters in the model, $k = 1, \ldots, K$, the prior and likelihood are calculated to account for the creation of a new cluster $k^\star$. The likelihood is calculated as in Eq. (8), where the parameters of the t distribution are equal to the prior parameters $\boldsymbol{\beta}$. The prior is calculated as,

$$p\left(c_i = k^\star \mid \mathbf{c}_{-i}, \alpha\right) = \frac{\alpha}{N + \alpha - 1} \tag{9}$$

Eqs. (4), (8), (9) allow the calculation of a value proportional to the posterior likelihood that the data point of interest $\mathbf{x}_i$, was a sample from any existing cluster or a new cluster. These likelihoods need to be scaled by the marginal likelihood, $\sum_{k=1}^{K+1}\tilde{p}(c_i = k \mid \mathbf{x}_i, \mathbf{c}_{-i}, X_{-i,k}, \alpha, \boldsymbol{\beta})$, where,

$$\tilde{p}(c_i = k \mid \mathbf{x}_i, \mathbf{c}_{-i}, X_{-i,k}, \alpha, \boldsymbol{\beta}) = p(\mathbf{x}_i \mid X_{-i,k}, c_i = k, \boldsymbol{\beta})p(c_i = k \mid \mathbf{c}_{-i}, \alpha) \tag{10}$$

Practically, this means summing $\tilde{p}(c_i = k \mid \mathbf{x}_i, \mathbf{c}_{-i}, X_{-i,k}, \alpha, \boldsymbol{\beta})$ for every existing cluster and the new cluster ($c_i = k^\star = K + 1$) and dividing each $\tilde{p}(c_i = k \mid \mathbf{x}_i, \mathbf{c}_{-i}, X_{-i,k}, \alpha, \boldsymbol{\beta})$ by this sum. This gives a multinomial distribution for the cluster label c_i of point i.

Sampling a cluster label c_i, from this distribution, the point is assigned to this cluster, either an existing cluster or a new cluster. If the point is added to an existing cluster then the parameters of that cluster are updated according to Eq. (6). If the point is assigned to a new cluster, that cluster is initialised from the $\mathcal{NIW}$ prior and the single point is added to it according to Eq. (6). The total number of clusters is also updated to reflect the increase, $K = K + 1$. Once these updates are made, another point is sampled and the process repeats itself.

Since the Gibbs sampler is a valid Markov Chain Monte Carlo (MCMC) method it is guaranteed that the normalised posterior distribution over the cluster labels will converge to the true posterior conditioned on α and $\boldsymbol{\beta}$ provided that the target distribution of the Markov Chain is that true posterior [43].

5. Online inference in the SHM context

Using a Gibbs sampling approach to assign cluster labels has a key advantage; each data point is assessed marginally. This means that new data points can be added into the data set and inference can proceed uninterrupted. Since the data in each cluster update the posterior parameters of that cluster, the cluster posterior distributions are refined by increasing the amount of data. The addition of data also allows for the creation of new clusters in a probabilistic manner without needing to pre-specify the total expected number of clusters or the parameters of those clusters, all without relying on heuristic measures. The learning of the number of clusters is a direct consequence of the Bayesian model form, it does not require expert knowledge or collection of a large training data set.

This behaviour can be exploited for use in an SHM context in three ways:

1. All data observed by a monitoring system refines the parameters of already known states, e.g. the normal condition, thus reducing false alarms.
2. When the behaviour of a structure changes, a new cluster is formed, triggering an alarm.
3. If, upon investigation, first of other available data (i.e. operational and environmental data) and if necessary of the structure itself, this alarm is not a result of damage, the cluster is given a label that allows classification of this separate undamaged state in the future.

This type of semi-supervised method allows the model to be continually updated so that all data collected are used to refine the model; this avoids the need to conduct many expensive long-term tests to acquire multiple normal state conditions and to observe the effects of all type of damage. It also allows all data collected by the monitoring system to be used as

additional information when making inference in the future. Therefore, the value of collecting data increases, as it is not only used for assessment of the structure, but also improves future operation of the SHM system.

Algorithm 1 A Gibbs Sampler for DP Clustering SHM Data with Forgetting and a Gaussian Base Distribution

function DP-FGS($\alpha, \boldsymbol{\mu_0}, \Sigma_0, \kappa_0, \nu_0, o_{max}$)
  $\boldsymbol{\beta} \leftarrow \{\boldsymbol{\mu}_0, \Sigma_0, \kappa_0, \nu_0\}$
  $N \leftarrow 0$ ▷The Number of Points Observed
  $C \leftarrow 0$ ▷Start with No Clusters
  for Each New Point Observed **do**
    $N \leftarrow N + 1$
    $o = \max(N - o_{max}, 0)$
    for randperm($i = o$ to N) **do** ▷Random Permutation of Last o Datapoints
      Remove point $\mathbf{x_i}$ from cluster c_i
      Update $\boldsymbol{\mu}_{\mathbf{c_i}}, \Sigma_{c_i}, K$
      for $k = 1$ to K **do**
        Calculate $p(c_i = k \mid \mathbf{c}_{-i}, X, \boldsymbol{\beta}, \alpha)$ ▷Predictive Posterior for Each Cluster
      end for
      Calculate $p\left(c_i = k^{\star} \mid \mathbf{c}_{-i}, X, \boldsymbol{\beta}, \alpha\right)$ ▷Predictive Posterior of a New Cluster for $\mathbf{x_i}$
      Sample new c_i from normalised $p(c_i \mid \mathbf{c}_{-i}, X, \alpha, \boldsymbol{\beta})$
      Add point $\mathbf{x_i}$ to cluster c_i
      Update $\boldsymbol{\mu}_{\mathbf{c_i}}, \Sigma_{c_i}, K$
    end for
  end for
end function

It is usual that an SHM system will be operational for an extended period of time, therefore, the size of the training dataset being considered in an *online learning* setting is constantly increasing. This introduces a challenge if the standard Gibbs sampling algorithm for inference in a DP mixture model were to be used. Since the Gibbs sampler would reassess all of the data points (calculating the posterior likelihood of each cluster label), at each iteration the algorithm would become progressively slower, to the point where it would not be feasible to continue. The proposed solution is to window the process so that only the previous o_{max} points that are added to the training set are considered in the Gibbs sampler. The value of this *forgetting factor* should be determined to be the maximum possible, given available computational power, since the early stopping of the Markov Chain may mean that the chain has not converged to the target distribution. It is worth considering that this is the case for all MCMC methods, whose convergence to the stationary distribution is guaranteed in the limit using the Strong Law of Large Numbers. The usual convergence checks for MCMC can be used, such as the $\widehat{R}$ statistic [43]; it is recommended, however, that the sampler is run for as many iterations as is computationally feasible. In an online setting, this is limited by the rate at which new data is being added to the process; the algorithm should be able to sample every point in the Gibbs sampler at least once between every new reading.

Pseudocode for the algorithm is shown in Algorithm 1, here it can be seen that only data points o_{max} samples back in time are reassessed. This introduces the additional hyperparameter to the model of how far back in time the sampler assesses. This parameter must be chosen *a priori* and is dependent on the system used with regard to the expected rate of change of behaviour, and computational requirements. Once datapoints will no longer be reassessed it is possible to discard them as the information can be contained in the cluster parameters thus leading to a more computationally and memory efficient implementation.

5.1. Hyperparameter selection

In many cases the choices of hyperparameters in the process, including $\boldsymbol{\mu}_0$, Σ_0, ν_0, κ_0, must be driven by prior knowledge of the system which can only come from an understanding of the structure as an engineering problem; additionally, the available computational resources will govern the range of feasible values.

If there is a case where no training data are available, it can pose problems in setting the hyperparameters for the clusters $\boldsymbol{\beta}$, and also the strength parameter α. In this case, pragmatism must take over. Normalisation of the data would allow the parameters in $\boldsymbol{\beta}$ to be set such that the prior cluster is a zero mean, unit variance Gaussian. It is clearly not possible to perform this normalisation in the absence of any training data. A sensible solution to this would be to implement a standard normalisation scheme, removing the mean and scaling by the standard deviation, where these quantities are calculated based on samples from a fixed period at the beginning of operation, either using the sample statistics, or by bootstrapping [50].

The choice of α poses a more difficult problem. This hyperparameter controls the likelihood that new clusters will be generated. It is not possible *a priori* to choose an optimal value for this parameter, since the spacing of the data in the feature

space is unknown. For many applications, there is a sensible range of values from which α can be set. Based on the authors' experience, it is recommended that α is set between one and 20 for most applications. Should problems be found with the process in operation, it is possible to repeat the analysis with a different value for α and if desired inference can be performed over α by placing a Gamma prior on the parameter [47].

5.2. A suggested decision making process

The algorithm returns more information than a usual novelty detection scheme due to its ability to cluster recurring feature sets into previously observed behaviour. Outlined here is one way in which this process could be used to aid decision making for SHM, as well as some of the considerations that should be made.

The simplest method to choose as the point at which an alarm is triggered is the creation of a new cluster, which in theory corresponds to the emergence of, as yet, unobserved behaviour. However, as the method progresses and clusters data online, for each assessment in the Gibbs sampler there is a non-zero probability that a new cluster will be created, although this probability can be very small. To protect against an unacceptable rate of false positives, a threshold can be introduced to ensure that alarms are not raised until a number of points are added to a new cluster. This threshold can be refined over the operation of the system as it does not affect the process of clustering the data itself. As a rule of thumb this can initially be set to be around five points for the critical mass in a cluster; this ensures that the process remains sensitive to changes in behaviour, but protects against small clusters being formed which don't correspond to actual structural changes, but are artefacts of the Gibbs sampler. The value of this threshold does not affect the progression of the algorithm and will likely be specific to individual use cases, its alteration online does not interrupt the algorithm.

A more robust system can be developed working on the assumption that damage causes ongoing changes in the behaviour of the system and that the structure cannot, of itself, return to an undamaged state. The affect of damage will not only cause a new cluster to be formed but points will continue to be added to this cluster as long as the structure is damaged. In view of this, it is possible to use the rate of growth of the clusters as indicators of the structures condition or operating behaviour. If a new cluster is created and grows at a significant rate (the extreme of which being all new points are added to it) this indicates a permanent shift in behaviour which could be associated with damage.

The problem remains of determining whether the change in behaviour is associated with damage to the structure or a change in operation. The primary method to separate damage from environmental variation is the choice of appropriate features to cluster [10]. Before the structure is inspected when an alarm is triggered, it is important to use all available data to assess reasons for changes in behaviour. The obvious suspects would be changes in environmental conditions: temperature, precipitation, etc. Other factors which will strongly influence the operational behaviour will include changes in use of the structure, such as change in loading or in the structural properties (e.g. changing topside mass on an offshore platform). It is worth considering at this point, the difference between observing a correlation with another measured variable related to the environmental conditions and establishing causation before deciding that the cause of a new cluster is benign. Discussion regarding this point can be found in [51]. Methods such as the Granger test [52] may help to provide insight as to whether a cause of the change in behaviour can be explained by other measured data.

6. Results

6.1. Three-storey building structure

The application of the DP mixture model is explored here using a benchmark dataset from a three-storey building structure (Fig. 3), produced by Los Alamos National Laboratory [53] for identification of damage under changing system behaviour. The experiment is a simplified three-storey building structure undergoing base excitation. Damage is simulated using a bumper attached between the second and third floors with the aim of representing a breathing crack in the structure. The structure is excited, nominally, along one axis only and mounted on linear bearings to minimise any torsional behaviour. The structure was tested in 17 states, aiming to represent a mixture of damaged and undamaged conditions, a summary of these states is shown in Table 1.

In the original report [53], a number of methods for feature extraction and classification are discussed and shown to be effective at detecting damage on this structure. Zhou et al. [54] show the use of an output-only approach to detect the introduction of damage. Figueiredo et al. [55] discusses the selection of AR model order in the context of damage detection on this structure, using the time series collected; whereas, Bandara et al. [56] show how frequency domain features (PCA projections of the FRF and coherence) can be used in the feature selection step and provide good classification results when used as inputs to a neural network.

The states where the changes made to the structure did not introduce nonlinearity (mass or stiffness changes) are considered to be environmental variation, and those which introduced nonlinearity (impacts of the bumper) are considered damage states. It can be seen that, in addition to the baseline condition, there are eight states representing environmental changes and eight representing damage.

50 measurements were made in every state, each comprising of a time series of 8192 data points, corresponding to 25.6 seconds of data. As is common, frequency domain features are extracted from the data. It is important here that the

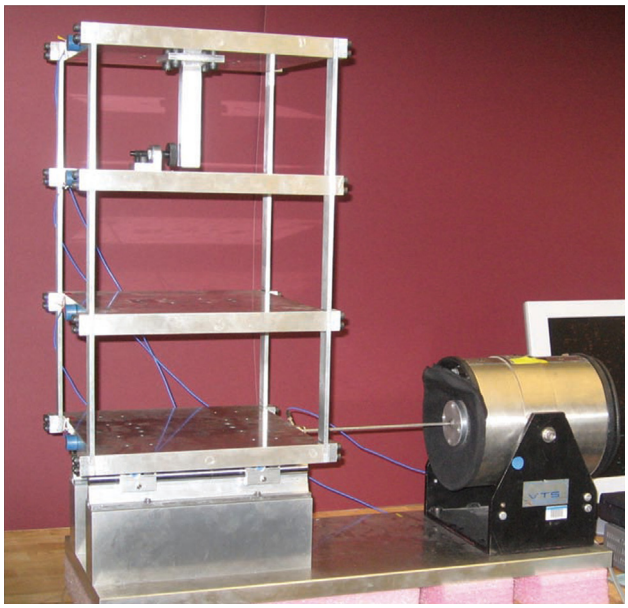

Fig. 3. Image showing setup of the three storey building structure, image reproduced from [53].

Table 1
Table reproduced from [53] showing 17 different states under which the structure was tested.

Label	State Condition	Description
State#1	Undamaged	Baseline condition
State#2	Undamaged	Added mass (1.2 kg) at the base
State#3	Undamaged	Added mass (1.2 kg) on the 1st floor
State#4	Undamaged	Stiffness reduction in column 1BD
State#5	Undamaged	Stiffness reduction in column 1AD and 1BD
State#6	Undamaged	Stiffness reduction in column 2BD
State#7	Undamaged	Stiffness reduction in column 2AD and 2BD
State#8	Undamaged	Stiffness reduction in column 3BD
State#9	Undamaged	Stiffness reduction in column 3AD and 3BD
State#10	Damaged	Gap (0.20 mm)
State#11	Damaged	Gap (0.15 mm)
State#12	Damaged	Gap (0.13 mm)
State#13	Damaged	Gap (0.10 mm)
State#14	Damaged	Gap (0.05 mm)
State#15	Damaged	Gap (0.20 mm) and mass (1.2 kg) at the base
State#16	Damaged	Gap (0.20 mm) and mass (1.2 kg) on the 1st floor
State#17	Damaged	Gap (0.10 mm) and mass (1.2 kg) on the 1st floor

clustering algorithm is also sensitive to features which can be extracted online. Although this limitation is minor, it does require some consideration when designing the identification algorithm.

Prior to the implementation of an SHM system, the use of such a system is justified, and the design of the system must be informed by operational evaluation [1]. This process considers the added benefit of investing in SHM; it also defines the parameters under which the system operates. These include considering the conditions in which the structure will operate, and the effect of this on any data acquisition scheme. A key step in SHM is feature selection; the challenge in this case is that many of the usual tools for feature selection are unavailable due to the lack of a training phase. It is necessary, therefore, to design the feature selection in such a way that it can: firstly, be computed online for all data that will be collected by the system; secondly, will give rise to features that are sensitive to changes in the structure that are of interest. In general this will be sensitivity to damage in the structure but not to environmental conditions. As is usual when dealing with measurements of acceleration of a dynamical system, data is first transformed into the frequency domain in batch. For vibration data, damage sensitive features are predominantly extracted in the frequency domain [57,58]. The additional benefit of using frequency domain features is that they can be invariant to the input to the system, e.g. the natural frequency (of a linear structure) is not affected by the forcing on the structure. This plays some role in the removal of environmental and operational changes.

Transformation of the blocks of 8192 time points into the frequency domain gives feature vectors which are 1024-dimensional real values in the Power Spectral Density (PSD), using Welch's method [59]. This high dimensionality is a significant hindrance to many algorithms, including the one presented in this paper. Not only does it add significant computational burden, in this case $\mathcal{O}\left(D^3\right)$, this complexity comes from the inversion of the covariance matrices which are size $D \times D$. But also many algorithms suffer from lack of sensitivity in high-dimensional spaces due to reliance on Euclidean distance metrics [60,61]. To avoid this it is possible to only consider other features which summarise the key properties of these high dimensional features, e.g. the natural frequencies and damping ratios of a system. However, a significant amount of information is lost when only these simple quantities are considered. It is desirable, therefore, to retain as much information as possible while also reducing the dimensionality of the feature space.

The usual manner to deal with this high dimensionality is to perform some type of dimensionality reduction such as Principal Component Analysis [4]. PCA, among other dimensionality reduction techniques, requires a representative training set of data which can be used to learn a linear projection onto a lower-dimensional space by accounting for maximum variance in each direction as the dimensionality increases. When designing an online SHM system, this does not represent a feasible approach since data are required to learn the optimal projection prior to any analysis using PCA, via the expectation maximisation method. The use of an online PCA projection also causes problems since the projection into the low dimensional space would be changing online; requiring the algorithm to fully recompute at each time step (running the Gibbs sampler multiple times to ensure convergence) which is not computationally feasible.

An alternative approach is to leverage a technique that has found widespread use in the compressive sensing community [62] — Random Projection (RP). The Johnson-Lindenstrauss theorem states that, when a set of high-dimensional data in Euclidean space is projected using a random matrix, the pairwise distances between the data are preserved with an error that can be quantified, allowing signals to be significantly compressed using RP [63,64]. By adopting a dimensionality reduction technique, which, rather than manually selecting features, does not require expert knowledge or a representative training set offers a number of advantages. The foremost of which (in this case) is the ability to begin operation of the SHM system immediately without a training phase and while preserving the pairwise distances between the full magnitude FRFs/coherences. In this way more information can be retained as opposed to the selection of some other low dimensional feature e.g. modal properties.

For this dataset, initially, the FRF and coherence at the top floor are considered; each of these is projected down onto ten dimensions using a random projection, where each element of the random matrix is an i.i.d. sample from the distribution $\mathcal{N}(0,1)$. These features are augmented with the area under the magnitude FRF at each floor including the base, giving a 24-dimensional feature vector. The addition of this feature is to capture the change in total energy being transferred to each floor as the structure state changes.

The algorithm was run with the parameters set as: $\alpha = 10, o_{max} = 200$, α is chosen by engineering judgement before looking at the data and o_{max} is limited by computation speed; therefore the Gibbs sampler reassessed only the previous 200 points, to save on computational burden. Fig. 4 shows the progression of the algorithm over time with each observation

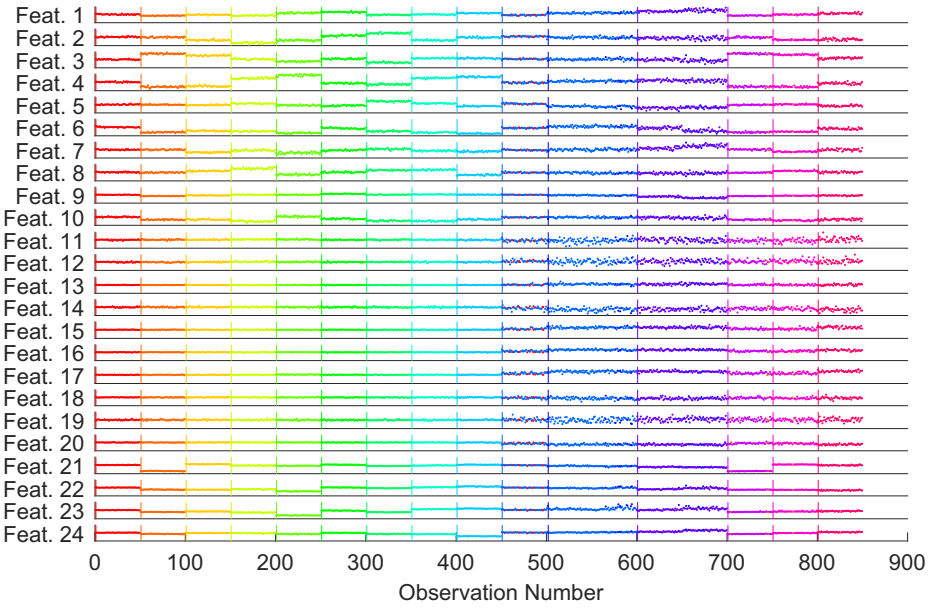

Fig. 4. Plot showing features used in clustering with colours indicating the clusters to which datapoints have been assigned for the online Dirichlet Process clustering with the full 24 dimensional feature space. The vertical lines of each colour indicate the initiation of that cluster. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

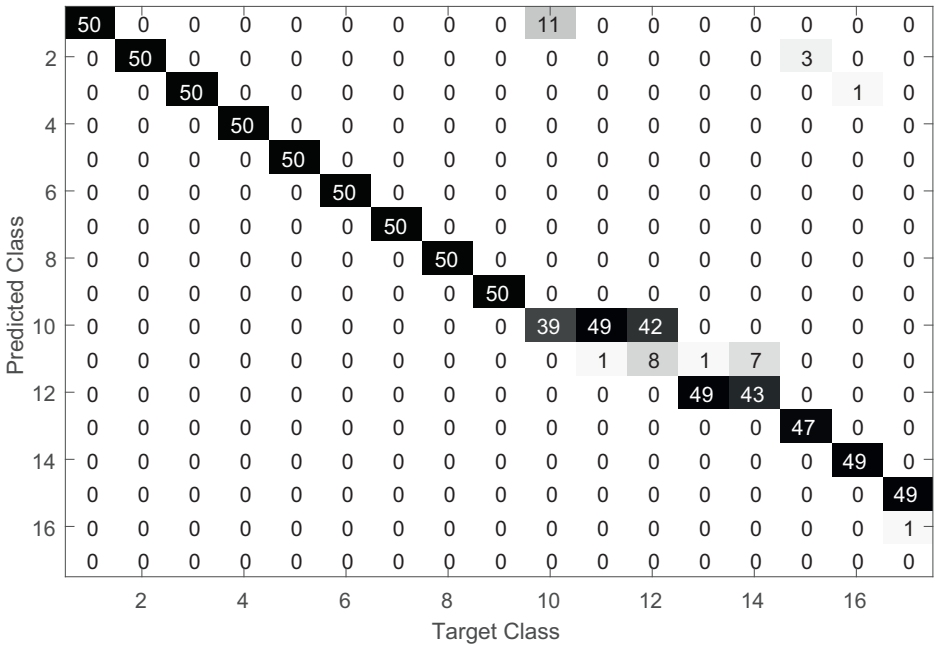

Fig. 5. Figure showing the confusion matrix for the implied states (cluster indices) from the online DP clustering when compared to the 17 known states for which the structure is tested using the full 24 dimensional feature space.

being a block of 8192 time series points from which the features are extracted. Vertical lines show the initiation of a new cluster at which point an intervention is triggered to label the newly-observed behaviour. By studying Fig. 4, one can see that 16 clusters have been detected. The damage introduced at observation 450, immediately triggers an intervention as a new cluster is formed.

Fig. 5 shows the confusion matix between the implicit states from the DP clustering and the known true states. For the initial nine states, baseline and eight environmental changes, there is perfect classification using the online DP clustering, the 9×9 matrix in the upper left is diagonal. This shows that while the algorithm would require further investigation when there is a change in environmental behaviour, the reappearance of these changes would then be correctly classified. For example, if there were seasonal changes in behaviour these would be classified correctly after the first appearance of the behaviour.

States 10 to 14 in the dataset correspond to increasing damage severity. It can be seen that for the smallest damage extent, despite triggering at the first damage observation, there is some confusion with the baseline state. Given these fifty observations it suggests that damage is occurring while the structure is operating under environmental conditions equivalent to State 1. As the severity of damage increases, the states are correctly classified into one of three clusters. State 15 corresponds to the lowest damage extent with the environmental change from state 2, which is classified well as a new damage case with only a small number of misclassifications into state 2. State 16 is equivalent to 15 except the environmental change is that seen in state 3, with similar results. State 17 corresponds to a larger damage extent with the environmental change from state 3. This is well classified as a new damage class.

It is useful however, to consider how varying the alpha parameter would affect the results shown for this case. For this reason the algorithm was additionally run with a number of different α values. If the system were running offline, inference could be performed over α to either select an optimal value or to learn the distribution in a Bayesian manner. Instead, the algorithm has been run with ten different fixed α values for a hundred different runs. Since the algorithm is stochastic, it is important to consider the distributions at different α values, not just a single result.

Fig. 6 shows the development of the false negative (FN) rate for increasing α. The boxplot shows the 25th and 75th percentiles as the top and bottom of each box, the sample median is shown by the red line. The "whiskers" show the interval of $\pm 2.7\sigma$ and outliers from this range are denoted by red crosses. The FN rate is defined here as the number of points in damage classes classified into an undamaged class. As seen in Figs. 4, 5, for the progressing damage scenarios in States 10 to 14, three clusters are created; distinction between these clusters is not included in the calculation of the FN or false positive (FP) rate. The FP rate was zero for all tests across all values of α, where an FP was defined as a point being classified into a cluster greater than 9 if it was in one of the first 9 states. For the results shown in Fig. 6, it can be seen that the FN rate is low across all levels of α. There is an increase in the FN rate as α tends to zero which is associated with data in the lowest damage extents, States 10 and 15, being misclassified as belonging to the healthy clusters associated with those environmental conditions. As the α value increases past 10, the FN rate has little variation with α, as the clusters are well separated; this stops the formation of more clusters. Fig. 7 shows this in a box plot where the distributions in the number of clusters are

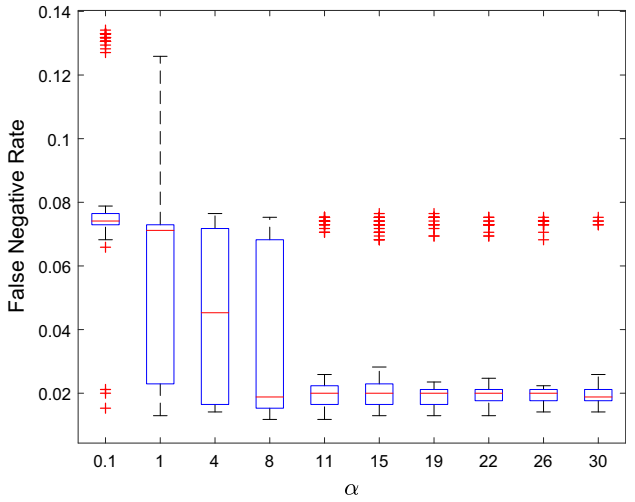

Fig. 6. Boxplot showing distribution of False Negative rates for 100 runs at the given levels of α.

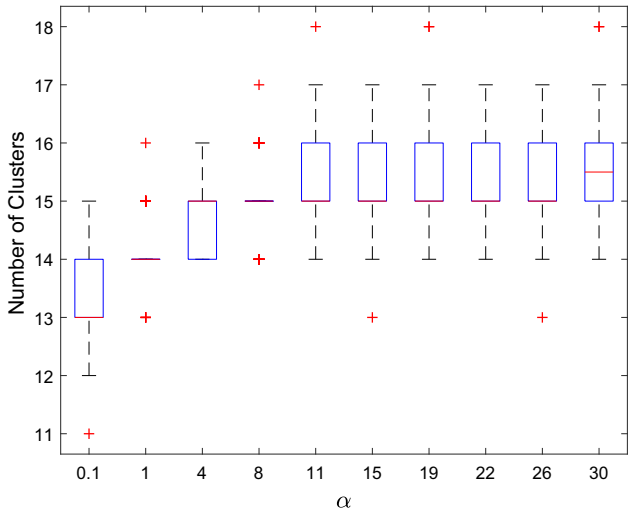

Fig. 7. Boxplot showing distribution of the number of clusters created for 100 runs at the given levels of α.

considered with varying α. The fact that the clusters are well separated, shows that in this range of α values, the number of clusters plateaus.

For the feature set shown in this experiment, which gives relatively good separation of clusters, the performance of the process is not significantly impacted by the choice of alpha within the range $\alpha \in [0.1, 30]$. This supports the *a priori* selection of $\alpha = 10$ as a starting point in engineering problems, where the data can be normalised to zero mean and unit variance and the parameters of the $\mathcal{NIW}$ prior are set $\boldsymbol{\mu}_0 = \mathbf{0}, \Sigma_0 = \mathbb{I}, \nu_0 = D, \kappa_0 = 1$ which corresponds to a unit Gaussian in D dimensions as a prior.

6.1.1. Feature selection to remove sensitivity to environmental changes

Should one wish to build a damage detection system that is insensitive to changes in the environmental conditions, it is possible to omit the features that are sensitive to this and perform the same inference procedure on a reduced feature set. The algorithm is re-run with a reduced feature set, where features are only sensitive to the damage condition not the environmental changes, with the same parameters as the previous analysis. This follows from the feature selection methodology shown in [10]; however, in the case of *online learning* these features must be chosen *a priori* based on engineering judgement.

Fig. 8 shows the confusion matrix when only 10 features which are damage-sensitive are used to perform the clustering. The algorithm is attempting to separate the damaged and undamaged classes from the dataset as defined in Table 1, where states one to nine are classified as undamaged and 10 to 17 as damaged. These ten features are chosen to be the randomly

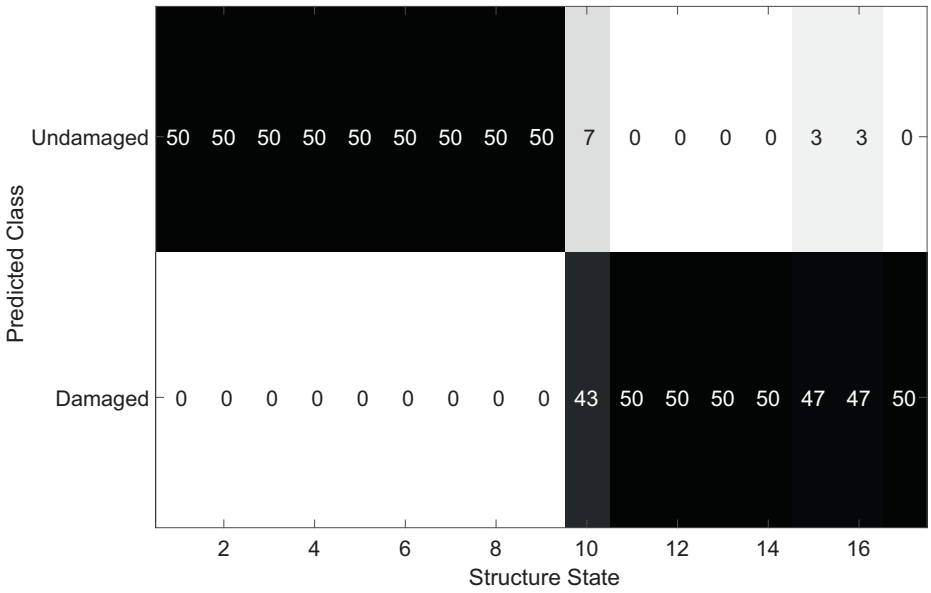

Fig. 8. Figure showing the confusion matrix for the implied states (cluster indices) from the online DP clustering with the reduced feature space (10 features). This compared to the 17 known states of the structure, states 1–9 correspond to undamaged behaviour and 10–17 correspond to damage, as is shown in Table 1.

projected coherence of the top floor. This feature selection does not require the damage state data. It is intuitive that, since the system is designed to detect a breathing crack in a structure that is approximately linear, the damage will increase any nonlinear behaviour which will cause significant change in the coherence but not in the FRF [65]. The coherence should also be broadly insensitive to the environmental changes that are expected to occur.

The DP clustering algorithm creates only two clusters in this case, without any tweaking of the hyperparameters. These two implicit states, upon inspection, correspond to the undamaged and damaged states. If only considering whether the system is labelled damaged or undamaged (Table 1), there are no false positives and 14 false negatives across the dataset of 850 observations, a FN rate of 0.017, defined as before. These results correspond to a sensitivity of 1 and specificity of 0.965 giving a total accuracy of 0.984. All of the false negatives occur at the lowest damage state (0.20 mm Gap), under differing environmental conditions. Despite this misclassification, the algorithm would raise a suitable alarm, even at the smallest damage extent, triggering an intervention.

The behaviour shown in the two cases above clearly demonstrates the ability of the algorithm to detect unknown states and to create new clusters to accommodate them. It also reveals that this does not remove the need for intelligent feature extraction based on sound engineering judgement. It is possible, given sufficient physical understanding of the structure, to imagine features *a priori* that will only be sensitive to changes of interest (e.g. insensitive to environmental conditions), and with the use of techniques such as RP to create feature spaces upon which the algorithm can operate. The choice of these features must be driven by engineering knowledge, in this case the assumption that a system whose behaviour is close to linear when undamaged will become more nonlinear with progressing damage but not with environmental changes [65].

6.1.2. Operating online without input information

In operation, an SHM system does not normally have access to measurement of the excitation source, as with a system tested under laboratory conditions. This is normally due to the difficulty in placing instrumentation in the load path of the structure, both practically and financially. In this case, features based on the FRF or coherence function become inaccessible due to their reliance on data regarding the forcing of the system.

It is desirable, therefore, to imagine a situation in which the proposed method would be applied on a dataset where this information is unavailable, the aim being to create a semi-supervised learning algorithm that is sensitive to damage on the structure. Again, using the intuition that the presence of damage on the structure will lead to increased nonlinearity in the structure [65], it is possible to determine a feature set that will be sensitive to damage; it is assumed here that the measurements of acceleration at all three floors are available, but not the forcing at the base.

In the same manner as before, the data arriving in windows of 8192 points can be converted into power spectra in the frequency domain with 1024 features. Operating directly on these power spectra will not yield a high sensitivity to damage and will be sensitive to changing environmental conditions. It is possible, therefore, to calculate the coherence between two of these output spectra rather than the traditional input–output coherence. This approach has been explored in [54], although here further signal processing is applied to create a damage-sensitive index based on the sum of the coherence

functions. This approach requires offline learning to set up a statistical control chart on this feature, a step that is not required in the current work.

This coherence between the two output spectra can be reduced in dimension in the same manner as previously, using RP, since using all spectral lines naïvely is not feasible computationally. The new algorithm here is tested using the projection of the coherence between the ground floor and the top floor (data channels 2 and 5) onto only three dimensions.

Figs. 9 and 10 show the progression of the algorithm in time, and the confusion matrix when this limited feature set is used. The performance is comparable to when the input information is also available to the algorithm. It is clear that this methodology is capable of detecting changes in behaviour associated with damage in an online semi-supervised manner (in the absence of training data at the start of the process), which is efficient in terms of memory and data storage requirements. The information returned from the method regarding the creation of new clusters is a simple trigger for intervention from engineers operating the system. In addition, as the system continues to collect data, its ability to correctly classify new data is enhanced as the cluster parameters are refined in a Bayesian manner.

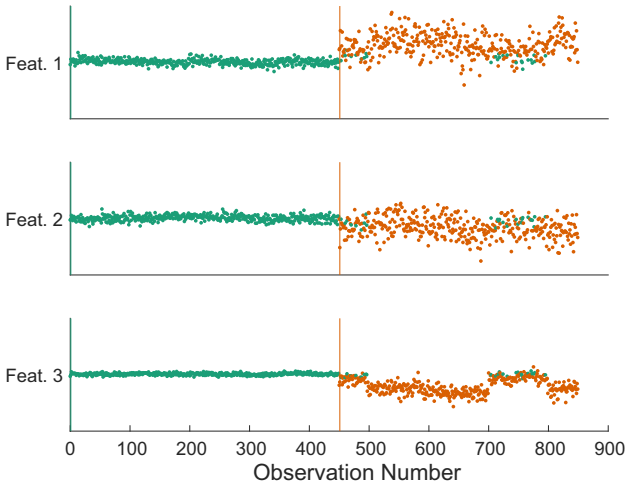

Fig. 9. Figure showing operation of the algorithm on the three dimension feature space created by randomly projecting the coherence between the ground floor and floor three.

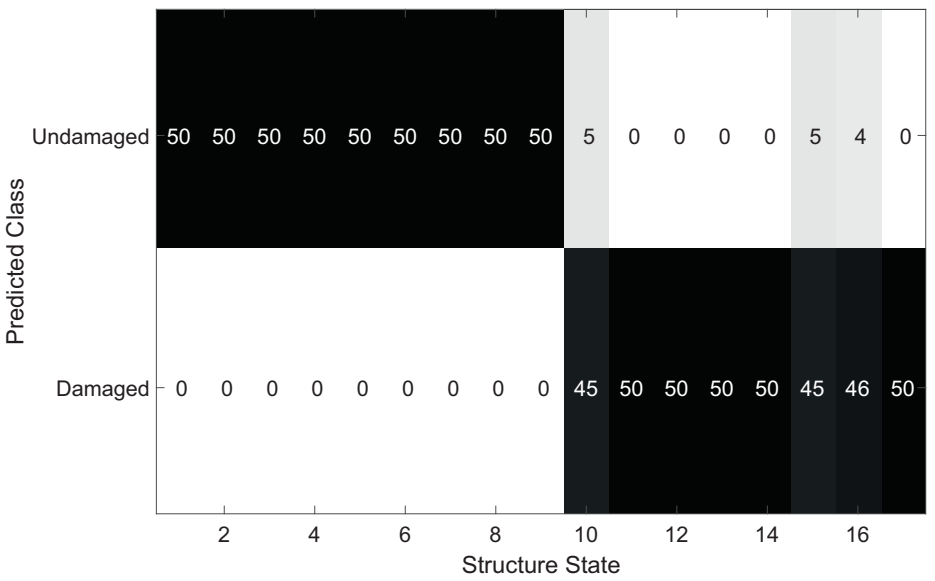

Fig. 10. Figure showing the confusion matrix for the implied states (cluster indices) from the online DP clustering with the randomly projected output coherence features. As in Fig. 8, the algorithm is attempting to classify between undamaged states (1–9) and damaged states (10–17) — see Table 1.

For this alternate feature set, the results shown in Figs. 9 and 10 are for an α value of 0.1. The results were run for a range of α values as before; however, for this experiment significant dependence on the choice of α was seen for the number of clusters created. This is shown in Fig. 11. To calculate the FN and FP, rates the problem was treated as a two class problem where, any points in cluster 1 were considered to be "normal" and points in any other cluster were considered "abnormal". This prescription leads to the true classes for each data point to be given by the State Condition column in Table 1. Box plots of the FN and FP rates for one hundred repeats at each α are shown in Figs. 12 and 13, both the FN and FP rates are very low for all values of α. The trend shown in Figs. 12 and 13 is a decrease in FN and increase in FP with increasing α. This is expected since the α parameter encodes the prior belief that data will be drawn from new clusters; intuitively, this states that with a given value of α a new cluster is just as likely as a cluster with α points in it already, see Eqs. (4) and (9).

Shown in Fig. 14 is the corresponding plot to Fig. 9 for a randomly chosen run of the algorithm with $\alpha = 30$; it can be seen that the FP rate is very low with only a single point misclassified, but as soon as the structure has damage introduced (point 450) multiple new clusters are created very quickly. To understand this behaviour, it is helpful to consider the pairwise correlation plots in Fig. 15. Since the data have been normalised online using the first fifty points, and the hyperparameters of the $\mathcal{NIW}$ prior are set to $\boldsymbol{\mu}_0 = \mathbf{0}, \Sigma_0 = \mathbb{I}, \nu_0 = D, \kappa_0 = 1$ as before; the increase in variance seen with the initiation of damage on the structure, and lack of separability of the clusters, leads to the creation of many new clusters (in this run 11 clusters in total). In other words, the prior encourages the process to make a mixture of unit variance Gaussian clusters, based on the normalised data. As damage progresses, the variance in the features increases despite the data being normalised to the lower variance portion of the signal. The process is, therefore, more likely to create a number of smaller clusters in the cloud of

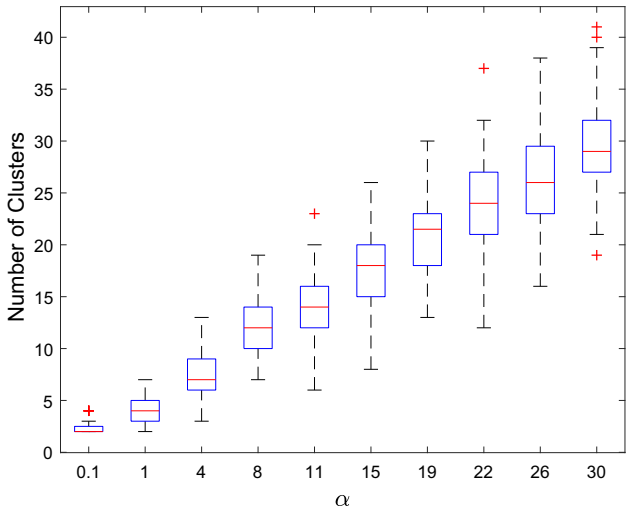

Fig. 11. Boxplot showing distribution of the number of clusters created for 100 runs at the given levels of α.

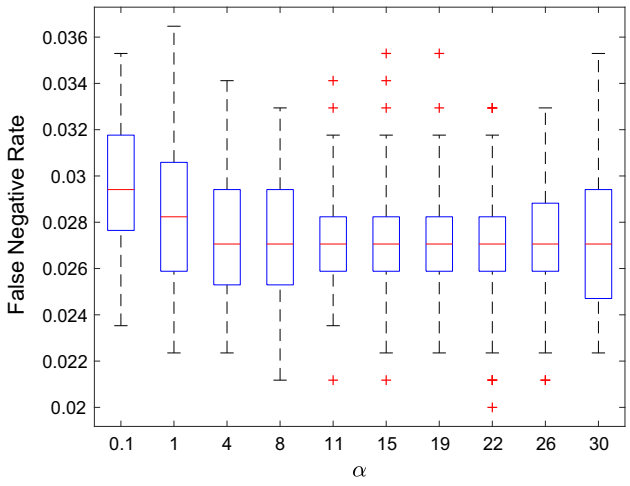

Fig. 12. Boxplot showing distribution of False Negative rates for 100 runs at the given levels of α when using the output only features.

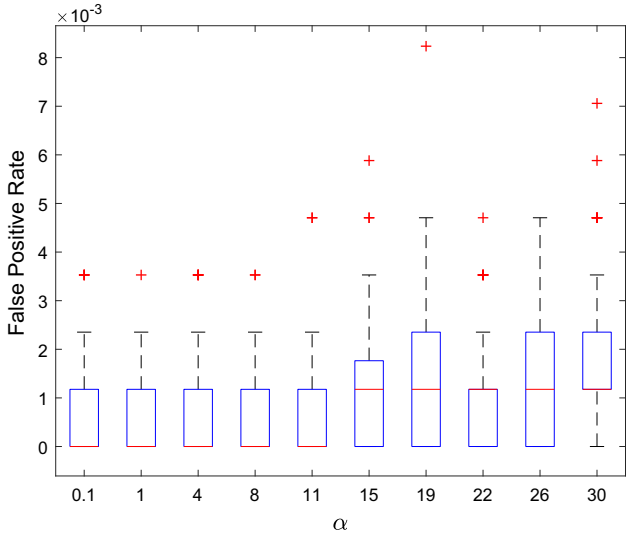

Fig. 13. Boxplot showing distribution of False Positive rates for 100 runs at the given levels of α when using the output only features.

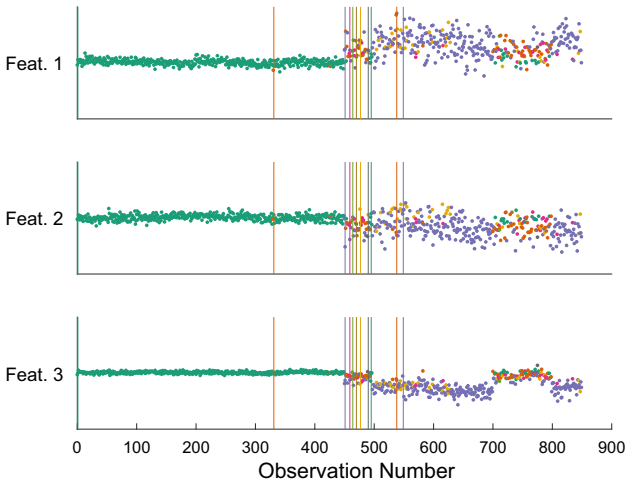

Fig. 14. Figure showing progression of the algorithm when $\alpha = 30$, vertical lines represent the initiation of a new cluster.

higher variance data instead of a single higher variance cluster. This effect is exacerbated by the higher α value, which favours the creation of more smaller clusters as the value increases.

The key question which must be asked is: in what way will this affect the operation of a system using this technique for SHM? The system remains resilient to false positives, and as discussed previously, techniques can be used to increase robustness to these. At the initiation of damage in the dataset a large number of new clusters are created which would lead to investigation, as discussed of other available environmental and operational data. The high number of alarms triggered would indicate a significant change in the structure, which in this case clearly corresponds to the damage being introduced.

6.2. Z24 bridge data

The now widely-known Z24 bridge dataset [66], has become a test-bed for many damage detection algorithms in SHM, particularly SHM of civil infrastructure. The dataset comprises of roughly one year of monitoring data from a bridge in Switzerland where damage was introduced deliberately toward the end of the monitoring programme. Researchers have most commonly used the first four natural frequencies of the bridge deck as damage-sensitive features; the difficulty in the dataset arises from the changes in environmental conditions which can confound damage detection algorithms. The most significant change is when a reduction in temperature is hypothesised to have caused stiffening of the deck asphalt leading to a rise in natural frequencies.

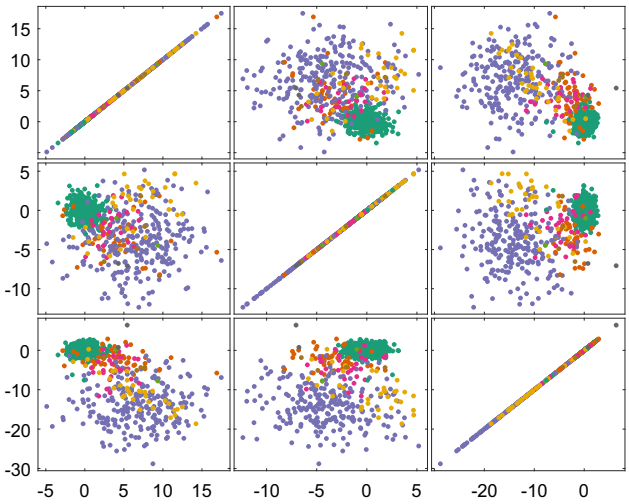

Fig. 15. Pairwise correlation plots for the process when $\alpha = 30$, where the colours shown correspond to those in Fig. 14.

This work makes no attempt to avoid these changes in behaviour due to environmental effects, instead it aims to demonstrate the ability of DP-based clustering to detect and subsequently classify different regimes of the structure. The data are tested with the parameters of the algorithm set as $o_{max} = 2000$ and $\alpha = 10$. Here, again, o_{max} is set on the basis of available computation time which is greater given the slower rate of arrival of the data points. α is set as before. Additionally to this, a threshold is introduced as discussed, to protect against false positives; this is required in this dataset due to the increased noise experienced in the full-scale test as opposed to the laboratory setting. The threshold was set at 50 data points; this was tuned on the basis of results from the initial section of the dataset, 500 data points. As previously mentioned, it may be possible to set a more robust trigger based on the rate of growth of the clusters, which may well constitute further work.

In the same manner as before, it is assumed that there is minimal training data available, only the first 500 data points. As the algorithm progresses, more clusters are created; this is shown in Fig. 16, a normal condition cluster (red) is quickly established. As the temperature cools three more cluster are created (orange, cyan and green) corresponding to the progression of freezing of the deck. Two other clusters are created, the dark blue one around time point 800 and the light blue one close to time point 1700. From inspection of the pairwise plots of each variable (Fig. 17) it appears that this light blue cluster corresponds to a shift and rotation in the normal condition. This could be caused by long term drift in the normal condition

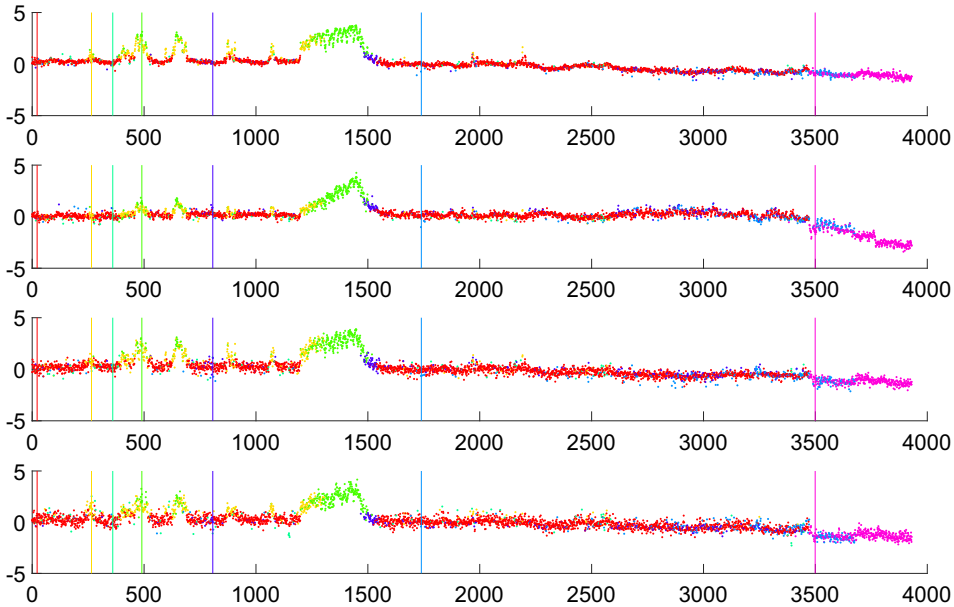

Fig. 16. Figure showing online DP clustering applied to the Z24 bridge data using the first four natural frequencies as the features.

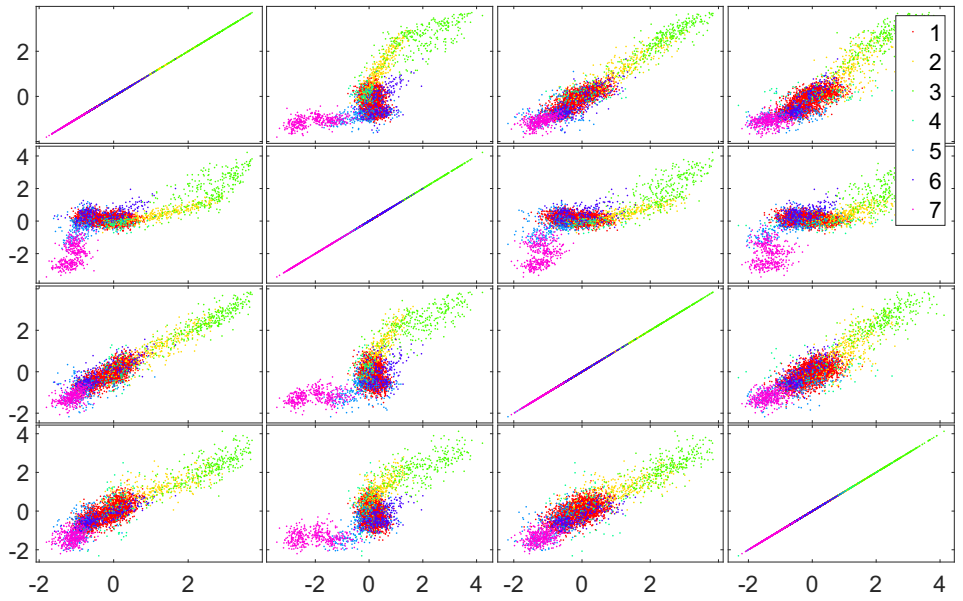

Fig. 17. The clusters found for the Z24 bridge by the online DP clustering are shown in the feature space.

which leads the distribution of points in this state to become non-Gaussian, possibly another affect of the varying ambient temperature, precipitating the creation of a second cluster to approximate the non-Gaussian distribution. Finally, the pink cluster is created only two data points after damage is introduced to the structure showing the method's ability, given the available feature set, to detect a change in behaviour corresponding to damage. In the Z24 dataset, there are two damage states induced, however, these are both classified into the same cluster when the DPGMM is run online. There are two reasons for this behaviour. The first is due to the lack of separation between the two damage state clusters in the feature space and the choice of α as shown on the three-storey bookshelf data. The second is that the data is normalised to the initial 500 points of data. The variance observed in this phase from which the hyperparameters of the cluster shape are set is greater that then separation between the two different damage clusters. This makes it difficult for the algorithm to create a new cluster for the second damage state due to the prior belief that has been encoded in $\boldsymbol{\beta} = \{\boldsymbol{\mu}_0, \Sigma_0, \kappa_0, \nu_0\}$.

It can be seen that once the algorithm has an explicit label assigned to the implicit label from the cluster assignment, subsequent data falling in that cluster can be correctly classified. The clusters relating to the different stiffness conditions of the deck, are able to classify these events from the second occurrence onwards, avoiding the need for unnecessary interventions, as would be the case with a simple novelty detection method. Fig. 17 shows the pairwise correlations of the first four natural frequencies of the Z24 bridge; it is in this feature space that the algorithm is operating. Here, it is clearer how the clustering algorithm is separating the feature space into a mixture of Gaussian distributions.

The results shown on the Z24 dataset, demonstrate the ability of the algorithm to deal with recurring environmental conditions while remaining sensitive to damage. It also makes clear that this approach to a damage identification algorithm will require more interventions/inspections shortly after the installation of the SHM system but with robustness increasing over time.

7. Discussion

The work presented in this paper introduces a methodology for incorporating a DP mixture model into an SHM system for online damage detection. The algorithm has been shown to perform very well on test data with multiple damaged and undamaged states. The method requires little user input and updates online with simple feedback to the user as to when intervention is required. Additionally, as clusters are assigned physically meaningful labels, additional information is available to the end user. It is believed that the method provides a promising approach for SHM when there is little or no availability of training data and inspections are possible to assign labels in a semi-supervised manner. There are a number of strengths to using this technique over a simple novelty detector or a non-probabilistic method such as affinity propagation [27]. The algorithm, unlike a basic novelty detector, can be run in a semi-supervised manner to assign labels to new behaviour states online. This additionally allows for multi-class classification as the algorithm progresses, allowing movement up the Rytter damage hierarchy [8], as more information is uncovered. The advantage of this over moving from a novelty detector to classifier online is that there is no further training phase required and the algorithm automatically incorporates both the classification and novelty detection. Unlike methods such as affinity propagation, the DP clustering algorithm has a

strong Bayesian foundation. By using a Bayesian technique, not only is a probability for every cluster provided, but there is a rigorous framework for the incorporation of prior knowledge. This would allow the use of an incompletely labelled training dataset to initiate the algorithm if certain states are known at the outset. This is achieved by assigning data points to clusters to update the cluster parameters, then excluding these points from the Gibbs sampling procedure to fix those clusters as explicit priors. If new points are added to these clusters, the parameters can continue to be updated via the conjugate update steps.

A modification to the normal DP algorithm has been proposed where the Gibbs sampler is truncated to consider only the previous $o_{\max}$ points in time. This allows the methodology to be applied online by stopping the computational complexity growing as more data are acquired. This limits the complexity at each step to be, naïvely, $\mathcal{O}\left(KD^3 o_{\max}\right)$ which is possible to compute online. However, it can be formulated such that the clusters undergo a rank one update to the covariance at each step which reduces the complexity to be $\mathcal{O}\left(KD^2 o_{\max}\right)$ on all but the first time step.

Another key advantage of the method is that, once the $\mathcal{NIW}$ hyperparameters have been set, there are only user-tunable hyperparameters, α and $o_{\max}$. If necessary, full Bayesian inference can be performed by placing a prior over the α parameter and performing inference, for example, via MCMC. The sensitivity of the process to this parameter has been discussed in terms of the affect on feature selection and normalisation. It has been shown, however, that problems which may occur from poor selection of this parameter are minimal, especially when clusters are well separated. Finally, it is possible to formulate the problem with a non-Gaussian base distribution, if the data are believed to be significantly non-Gaussian. It is worth considering whether this adds value to the inference procedure since computation time is severely increased in this case and many non-Gaussian datasets can be well represented by the Gaussian mixture model, especially when the number of mixtures does not need to be specified *a priori*.

It is noted that, although this highly flexible model has a benefit when data arrive online with an unknown number of states, there may be better tools to use in an offline state or if the problem is restricted to detection. It would be surprising if this semi-supervised method was able to compete with a fully supervised inference algorithm, since there is less information in the training phase. Although the method has been shown to work on a two-class novelty detection problem (Fig. 8) it is expected that other methods (e.g. robust outlier detection [11]) would perform better if data from the baseline were known. However, the performance shown in this paper is comparable with many offline supervised methods, particularly for the three-storey building structure [53,54,56].

For future work, it would be beneficial if the algorithm could be applied on a population level to allow labelling from one structure to inform inference on another, for example in the case of an offshore wind farm.

Acknowledgements

T. Rogers wishes to thank Ramboll for their support and the authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) through grant reference number EP/J016942/1.

References

[1] C.R. Farrar, K. Worden, Structural Health Monitoring: A Machine Learning Perspective, John Wiley & Sons, 2012.
[2] K. Worden, C.R. Farrar, G. Manson, G. Park, The fundamental axioms of structural health monitoring, Proc. R. Soc. London A 463 (2082) (2007) 1639–1664.
[3] H. Sohn, Effects of environmental and operational variability on structural health monitoring, Philos. Trans. R. Soc. London A 365 (1851) (2007) 539–560.
[4] D. Barber, Bayesian Reasoning and Machine Learning, Cambridge University Press, 2012.
[5] X. Zhu, Semi-supervised learning, Encyclopedia of machine learning, Springer, 2011, pp. 892–897.
[6] N. Dervilis, E. Cross, R. Barthorpe, K. Worden, Robust methods of inclusive outlier analysis for structural health monitoring, J. Sound Vib. 333 (20) (2014) 5181–5195.
[7] K. Worden, G. Manson, N.R.J. Fieller, Damage detection using outlier analysis, J. Sound Vib. 3 (2000) 647–667.
[8] A. Rytter, Vibrational Based Inspection of Civil Engineering Structures, Dept. of Building Technology and Structural Engineering, Aalborg University, 1993 (PhD thesis).
[9] K. Worden, G. Manson, D. Allman, Experimental validation of a structural health monitoring methodology: Part I. novelty detection on a laboratory structure, J. Sound Vib. 259 (2) (2003) 323–343.
[10] G. Manson, K. Worden, D. Allman, Experimental validation of a structural health monitoring methodology: Part II. novelty detection on a gnat aircraft, J. Sound Vib. 259 (2) (2003) 345–363.
[11] N. Dervilis, K. Worden, E.J. Cross, On robust regression analysis as a means of exploring environmental and operational conditions for SHM data, J. Sound Vib. 347 (2015) 279–296.
[12] E.J. Cross, K. Worden, Q. Chen, Cointegration: a novel approach for the removal of environmental trends in structural health monitoring data, Proc. R. Soc. London A 467 (2133) (2011) 2712–2732.
[13] R. Fuentes. On Bayesian Networks for Structural Health and Condition Monitoring, 2017.
[14] K.K. Nair, A.S. Kiremidjian, Time series based structural damage detection algorithm using gaussian mixtures modeling, J. Dyn. Syst. Meas. Control 129 (3) (2007) 285–293.
[15] E. Figueiredo, E. Cross, Linear approaches to modeling nonlinearities in long-term monitoring of bridges, J. Civil Struct. Health Monit. 3 (3) (2013) 187–194.
[16] J. Kullaa, Structural health monitoring under nonlinear environmental or operational influences, Shock Vibr. 2014 (2014).
[17] E. Figueiredo, G. Park, C.R. Farrar, K. Worden, J. Figueiras, Machine learning algorithms for damage detection under operational and environmental variability, Struct. Health Monit. 10 (6) (2011) 559–572.

[18] L. Yu, J.-H. Zhu, L.-L. Yu, Structural damage detection in a truss bridge model using fuzzy clustering and measured FRF data reduced by principal component projection, Adv. Struct. Eng. 16 (1) (Jan 2013) 207–217.
[19] A. Diez, N.L.D. Khoa, M. Makki Alamdari, Y. Wang, F. Chen, P. Runcie, A clustering approach for structural health monitoring on bridges, J. Civil Struct. Health Monit. 6 (3) (Jul 2016) 429–445.
[20] M.M. Alamdari, T. Rakotoarivelo, N.L.D. Khoa, A spectral-based clustering for structural health monitoring of the Sydney Harbour Bridge, Mech. Syst. Signal Proces. 87 (Mar 2017) 384–400.
[21] D.-A. Tibaduiza, M.-A. Torres-Arredondo, L. Mujica, J. Rodellar, C.-P. Fritzen, A study of two unsupervised data driven statistical methodologies for detecting and classifying damages in structural health monitoring, Mech. Syst. Signal Process. 41 (1–2) (Dec 2013) 467–484.
[22] R. Langone, E. Reynders, S. Mehrkanoon, J.A.K. Suykens, Automated structural health monitoring based on adaptive kernel spectral clustering, Mech. Syst. Signal Process. 90 (2017) 64–78.
[23] S. Chen, F. Cerda, P. Rizzo, J. Bielak, J.H. Garrett, J. Kovacevic, Semi-supervised multiresolution classification using adaptive graph filtering with application to indirect bridge structural health monitoring, IEEE Trans. Signal Process. 62 (2014) 2879–2893.
[24] K. Krishnan Nair, A.S. Kiremidjian, Time series based structural damage detection algorithm using Gaussian mixtures modelling, J. Dyn. Syst. Meas. Contr. 129 (3) (2007) 285.
[25] L. Qiu, S. Yuan, F.-K. Chang, Q. Bao, H. Mei, On-line updating Gaussian mixture model for aircraft wing spar damage evaluation under time-varying boundary condition, Smart Mater. Struct. 23 (12) (2014).
[26] L. Qiu, S. Yuan, H. Mei, F. Fang, An improved Gaussian mixture model for damage propagation monitoring of an aircraft wing spar under changing structural boundary conditions, Sensors (Basel, Switzerland) 16 (3) (2016) 291.
[27] B.J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (5814) (2007) 972–976.
[28] A. Vlachos, Z. Ghahramani, A. Korhonen. Dirichlet process mixture models for verb clustering, in: Proceedings of the ICML workshop on Prior Knowledge for Text and Language, 2008.
[29] D.M. Blei, Probabilistic topic models, Commun. ACM 55 (4) (2012) 77–84.
[30] C. Wang, J. Paisley, D. Blei, Online variational inference for the hierarchical Dirichlet process, Proc. Fourteenth Int. Conf. Artificial Intell. Stat. 15 (2011) 752–760.
[31] O. Yakhnenko, V. Honavar. Annotating images and image objects using a hierarchical Dirichlet process model, in: Proceedings of the 9th International Workshop on Multimedia Data Mining: held in conjunction with the ACM SIGKDD 2008, pp. 1–7, 2008.
[32] B. Thirion, A. Tucholka, M. Keller, P. Pinel, A. Roche, J.-F. Mangin, J.-B. Poline. High level group analysis of FMRI data based on Dirichlet process mixture models, in: Biennial International Conference on Information Processing in Medical Imaging, pp. 482–494, 2007.
[33] A.R.F. da Silva, A Dirichlet process mixture model for brain MRI tissue classification, Med. Image Anal. 11 (2) (2007) 169–182.
[34] N. Lartillot, H. Philippe, A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process, Mol. Biol. Evol. 21 (6) (2004) 1095–1109.
[35] N. Lartillot, N. Rodrigue, D. Stubbs, J. Richer, PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment, Syst. Biol. 62 (4) (2013) 611–615.
[36] F. Wood, M.J. Black, A nonparametric Bayesian alternative to spike sorting, J. Neurosci. Methods 173 (1) (2008) 1–12.
[37] D. Chakraborty, N. Kovvali, A. Papandreou-Suppappola, A. Chattopadhyay, An adaptive learning damage estimation method for structural health monitoring, J. Intell. Mater. Syst. Struct. 26 (2) (2015) 125–143.
[38] W.B. Johnson, J. Lindenstrauss, Extensions of lipschitz mappings into a hilbert space, Contemporary Math. 26 (189–206) (1984) 1.
[39] G. Schwarz et al, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464.
[40] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723.
[41] E. Figueiredo, L. Radu, K. Worden, C.R. Farrar, A bayesian approach based on a markov-chain monte carlo method for damage detection under unknown sources of variability, Eng. Struct. 80 (2014) 1–10.
[42] J. Diebolt, C.P. Robert, Estimation of finite mixture distributions through bayesian sampling, J. R. Stat. Soc.. Series B (Methodolog.) (1994) 363–375.
[43] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, D.B. Rubin, Bayesian Data Analysis, vol. 2, CRC Press Boca Raton, FL, 2014.
[44] C.E. Rasmussen. The infinite Gaussian mixture model. Advances in neural information processing systems, pages 554–560, 2000.
[45] R.M. Neal, Markov chain sampling methods for Dirichlet process mixture models, J. Comput. Graphical Stat. 9 (2) (2000) 249–265.
[46] D.M. Blei, M.I. Jordan. Variational methods for the Dirichlet process, in: Proceedings of the twenty-first international conference on Machine learning, p. 12, 2004.
[47] D.M. Blei, M.I. Jordan, Variational inference for Dirichlet process mixtures, Bayesian Anal. 1 (1) (2006) 121–143.
[48] M. Seeger. Low rank updates for the cholesky decomposition. Technical report, 2004.
[49] P.E. Gill, G.H. Golub, W. Murray, M.A. Saunders, Methods for modifying matrix factorizations, Math. Comput. 28 (126) (1974) 505–535.
[50] B. Efron, Bootstrap methods: another look at the jackknife, in: J.N. Kotz (Ed.), Breakthroughs in Statistics, Springer, 1992, pp. 569–593.
[51] K. Worden, E.J. Cross. On correlation and causality in structural dynamics, in: Proceedings of the 6th European Conference on Structural Control, Sheffield, 2016.
[52] C.W.J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica: J. Econ. Soc. (1969) 424–438.
[53] E. Figueiredo, G. Park, J. Figueiras, C. Farrar, K. Worden. Structural health monitoring algorithm comparisons using standard data sets. Technical report, Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 2009.
[54] Y.-L. Zhou, E. Figueiredo, N. Maia, R. Perera, Damage detection and quantification using transmissibility coherence analysis, Shock Vibr. 2015 (2015).
[55] E. Figueiredo, J. Figueiras, G. Park, C.R. Farrar, K. Worden, Influence of the autoregressive model order on damage detection, Comput.-Aided Civil Infrastructure Eng. 26 (3) (2011) 225–238.
[56] R.P. Bandara, T.H. Chan, D.P. Thambiratnam, Structural damage detection method using frequency response functions, Struct. Health Monit. 13 (4) (2014) 418–429.
[57] W. Fan, P. Qiao, Vibration-based damage identification methods: a review and comparative study, Struct. Health Monit. 10 (1) (2011) 83–111.
[58] C.R. Farrar, S.W. Doebling, D.A. Nix, Vibration–based structural damage identification, Philos. Trans. R. Soc. London A 359 (1778) (2001) 131–149.
[59] P. Welch, The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms, IEEE Trans. Audio Electroacoustics 15 (2) (1967) 70–73.
[60] C.C. Aggarwal, A. Hinneburg, D.A. Keim. On the surprising behavior of distance metrics in high-dimensional space, in: International conference on database theory, pp. 420–434, 2001.
[61] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is nearest neighbor meaningful?, International conference on database theory, Springer, 1999, pp 217–235.
[62] Y.C. Eldar, G. Kutyniok, Compressed Sensing: Theory and Applications, Cambridge University Press, 2012.
[63] S. Dasgupta, A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss, Random Struct. Algorithms 22 (1) (2003) 60–65.
[64] E.J. Candes, T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies?, IEEE Trans Inf. Theory 52 (12) (2006) 5406–5425.
[65] K. Worden, C.R. Farrar, J. Haywood, M. Todd, A review of nonlinear dynamics applications to structural health monitoring, Struct. Control Health Monit. 15 (4) (2008) 540–567.
[66] B. Peeters, G. De Roeck, One-year monitoring of the Z24-Bridge: environmental effects versus damage events, Earthquake Eng. Struct. Dyn. 30 (2) (2001) 149–171.