



HAL
open science

Clustering acoustic emission data streams with sequentially appearing clusters using mixture models

Emmanuel Ramasso, Thierry Denoeux, Gael Chevallier

► To cite this version:

Emmanuel Ramasso, Thierry Denoeux, Gael Chevallier. Clustering acoustic emission data streams with sequentially appearing clusters using mixture models. *Mechanical Systems and Signal Processing*, 2022, 181, pp.109504 (23). 10.1016/j.ymssp.2022.109504 . hal-03772732

HAL Id: hal-03772732

<https://hal.science/hal-03772732>

Submitted on 8 Sep 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Clustering acoustic emission data streams with sequentially appearing clusters using mixture models

Emmanuel Ramasso

Institut FEMTO-ST (UMR CNRS 6174), Université Bourgogne Franche-Comté, Département Mécanique Appliquée, Besançon, France

Thierry Denœux

*Université de technologie de Compiègne, CNRS, Heudiasyc, Compiègne, France
Institut universitaire de France, Paris, France*

Gaël Chevallier

Institut FEMTO-ST (UMR CNRS 6174), Université Bourgogne Franche-Comté, Département Mécanique Appliquée, Besançon, France

Abstract

The interpretation of unlabeled acoustic emission (AE) data classically relies on general-purpose clustering methods. While several criteria have been used in the past to select the hyperparameters of those algorithms, few studies have paid attention to the development of dedicated objective functions in clustering methods able to cope with the specificities of AE data. We investigate how to explicitly represent clusters onsets in mixture models in general, and in Gaussian Mixture Models (GMM) in particular. We propose the first clustering method able to provide, through parameters estimated by an expectation-maximization procedure, information about when clusters occur (onsets), how they grow (kinetics) and their level of activation through time. This new objective function accommodates continuous timestamps of AE signals and, thus, their order of occurrence. The method, called GMMSEQ, is experimentally validated to characterize the loosening phenomenon in bolted structure under vibrations. A comparison with four standard clustering methods on raw streaming data from five experimental campaigns shows that GMMSEQ not only provides useful qualitative information about the timeline of clusters, but also shows better performance in terms of cluster characterization.

Keywords: Acoustic emission, clustering, onsets, continuous timestamps, loosening of bolted joints.

Email addresses: emmanuel.ramasso@femto-st.fr (Emmanuel Ramasso), thierry.denoeux@utc.fr (Thierry Denœux), gael.chevallier@univ-fcomte.fr (Gaël Chevallier)

URL: <https://github.com/emmanuelramasso> (Emmanuel Ramasso), <https://www.hds.utc.fr/~tdenoeux/> (Thierry Denœux)

1. Introduction

ASTM standard E1316 [1, 2] defines Acoustic Emission (AE) as the detection of the subnanometric displacements of the surface of a material induced by the propagation of an elastic wave generated by a sudden and permanent change in the material integrity. This capability makes the AE technique particularly relevant to gain insights into the behavior of a material, a structure or an equipment under usage [3, 4, 5, 6, 7, 8] and accounts for its wide use in applications related to material testing, Structural Health Monitoring (SHM) and process monitoring and control.

Original AE data take the form of a *data stream* recorded by sensors attached onto a structure (Figure 1). The sensors, converting the subnanometric displacements into voltage signals, have to be read continuously in order to catch all events originating from the material. The data stream is then segmented using a wave-picking algorithm with the aim to detect damage-related (non-noise) AE signals [9, 10, 11, 12, 13, 14]. In feature-based interpretation of AE signals, a *feature extraction* step is performed in which AE signals are represented in a common feature space. The set of feature vectors represents an AE data set for a given experiment and is generally stored in an $N \times d$ feature matrix

$$\mathbf{X} = [\mathbf{x}_1^\top, \dots, \mathbf{x}_i^\top, \dots, \mathbf{x}_N^\top] \quad (1)$$

where $\mathbf{x}_i^\top \in \mathbb{R}^d$ is the transposed feature vector computed from the i -th AE signal. The timestamps of AE signals are their instants of occurrence and are denoted as t_i with

$$0 = t_0 < t_1 < \dots < t_i < \dots < t_N = T, \quad (2)$$

where T is the data stream duration. AE data have several special characteristics in terms of data processing [15]; in particular timestamps are continuous and *unequally-spaced in time*, i.e., we generally have

$$\frac{t_i - t_{i-1}}{t_j - t_{j-1}} \neq 1. \quad (3)$$

For a given AE signal, the d features generally belong to a standard list of AE features, some of which are listed, for example, in [16]. In the sequel, we suppose that the features have been extracted using the algorithm introduced in [13], which implements in MATLAB[®] common features available in the Mistras AEWIn[®] software. Unsupervised feature selection is not tackled in this paper. The reader interested in this topics can refer, for instance, to the aforementioned references and [17, 18, 19, 20].

Since a huge amount¹ of AE signals can be detected in a data stream, it is difficult, except for specific configurations, to know the ground truth for a sufficient large amount of AE signals. Lack of knowledge about the source of AE signals prevents us from using supervised learning methods for AE data interpretation, or even for anomaly detection when “normal” condition data are available [22, 23]. *Unsupervised learning*, particularly based on clustering, is generally used to extract information from AE data. This is the main scope of the present work.

Clustering methods have been applied for decades to interpret AE data [24, 25, 26]. The most commonly used methods are the K-means [27, 28], the fuzzy C-means (FCM) [29, 30], the Gustafson-Kessel (GK) algorithm [31, 21] and Gaussian Mixture Models (GMM) [32, 33]. A clustering method computes membership degrees of feature vectors to clusters; a hard partition is then

¹The number of AE signals depends on the materials and the type of loading. For example, hundreds of thousands of signals were detected for composite materials during quasi-static tests reported in [21].

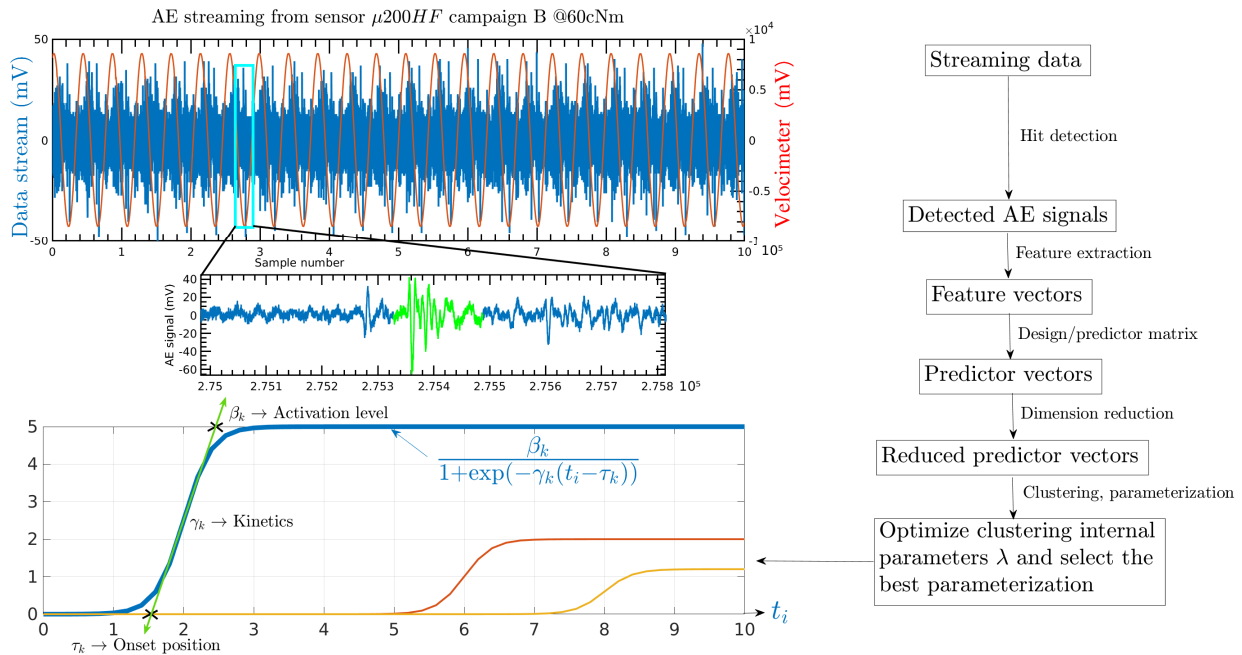


Figure 1: Proposed methodology and illustration on a real data stream extracted from one of the campaigns.

obtained by assigning each vector to its maximum-membership cluster. Based on this partition, data interpretation is generally necessary to determine the correspondence between clusters and damages. From a statistical point of view, the criteria used in classical clustering methods rely on the assumption that AE signals are *independent and identically distributed* (iid). Therefore, the partition, after applying one of those methods, does not depend on the ordering. From a physical point of view, this seems counterintuitive in most of AE-related applications since the progression of a damage type is known to depend on the preceding damage states and damage accumulation [34, 35]. To cope with this problem, we need a clustering method taking the time distribution of AE signals into account. Note that “time” can be replaced by any monotonically increasing measure such as cumulative loading or cycles.

Clustering methods dedicated to time-series have been developed in the past [36, 37] but only a few of them are able to manage continuous and uneven timestamps as shown in the recent review [38]. One of the first attempts to modify standard approaches (such as K-means, FCM or GMM) so as to accommodate temporal data with such timestamps was presented in [39]. The authors modeled the time series as piece-wise linear functions and proposed a distance measure between slopes. Using this distance, they derived a modified version of the FCM method.

Clustering methods describe a data set through a set of parameters such as, for example, the cluster centers in the K-means algorithm or the means, covariance matrices and proportions in Gaussian Mixture Models. Parameters are identified from data by iteratively optimizing an objective function that is an explicit function of these parameters. Additionally, clustering methods also depend on hyperparameters such as, e.g. the number of clusters, which cannot be optimized in this way. For hyperparameters, we need to perform a grid search by varying them and evaluate their impact on the clustering result. We also need a selection criterion (for determining, e.g., the optimal number of clusters). For each set of hyperparameter values, the objective function is first

optimized to estimate the parameters of the clustering method. Then, the best configuration is selected. The learning process may, thus, be time-consuming.

50 In AE, hyperparameters are determined, in the majority of cases, by computing a criterion such as Davies-Bouldin or Silhouette, which focus on the shape of clusters. As a consequence, to be selected, hyperparameters must lead to compact and well-separated clusters in the feature space [33, 40, 41, 42, 43]. This is the definition of natural clusters often used in AE, originally based on [44]. An alternative is to find hyperparameters that lead to clusters characterized by onsets which
55 are well distributed in time or load. This approach was developed recently because some authors found limitations in shape-based criteria for interpreting AE data [45, 21, 46, 47, 48, 49, 15, 50]. However, to the best of our knowledge, *the onset times of clusters have not been considered in an objective function so far. Taking into account onset times for clustering AE data is the main objective of the present work. The proposed clustering method, called GMMSEQ, treats onsets as*
60 *parameters that can be optimized together with the other cluster parameters directly from data, which was not possible before.*

GMMSEQ relies on a modification of the original GMM to account for the fact that AE signals are indexed by continuous timestamps. More specifically, the proportions in the mixture are assumed to vary in time according to a model of evolution based on sigmoid functions (Figure 1).
65 Each sigmoid function allows us to represent:

- The level of activation of a given damage related to the cumulative number of signals generated by this damage;
- The growth rate of the damage driven by the slope of the sigmoid function at the origin and related to the kinetics of the damage;
- 70 • The instant of the damage onset.

Therefore, this new clustering method makes it possible to identify when a damage first occurs (onset), how it grows (kinetics) and how it accumulates (cluster progression). The method makes a step beyond the standard approach to AE analysis by characterizing damage progression through three parameters estimated from the data.

75 The ability to represent onset times, kinetics and activation level makes this approach relevant for applications in which the chronology, sequence or timeline are of key importance. For example, Sawan et al. argued for *an AE analysis approach that seeks to separate observations into the greatest number of clusters with distinct evolution behavior*. For that, they used a GMM in its original form and represented their results by means of a cumulated number of hits per cluster.
80 However, the analysis is subjective without a proper identification of parameters related to the evolution, which is a common problem in AE analysis based on damage progression. In a previous paper [21], a criterion was proposed to identify clusters with different proportions and the authors experimentally observed that the timeline suggested by the clusters (obtained by different methods) was improved compared to standard shape-based criteria.

85 Another advantage of GMMSEQ is that the compactness and separability of clusters in the feature space is still of primary importance because it is managed explicitly using a mixture model with the possibility to adapt the distribution to the data. While being developed to take AE data characteristics into account, this new clustering method can be applied to other temporal data for which onsets, growths and cumulative progression of clusters are relevant to the analysis. The
90 proposed optimization procedure assumes that all data are available at once (offline analysis).

The model and the estimation algorithm are presented in Section 2. The method is then illustrated in Section 3 using simulated and real data sets. Data sets and codes are shared on Dataverse [51] and Github².

2. Gaussian Mixture Model with sequentially appearing clusters (GMMSEQ)

95 The GMMSEQ method introduced in this paper is based on a GMM with time-varying proportions. The model is described in Section 2.1, and parameter estimation is addressed in Section 2.2.

2.1. Model

Gaussian mixture models. In a mixture model, the data are supposed to follow a probability distribution defined as a weighted sum of K distributions:

$$p(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k g(\mathbf{x}_i; \boldsymbol{\theta}), \quad (4)$$

where π_k denotes the proportion of component k and g can be, for example, a Gaussian, Gamma or Student-t probability density function (pdf); the vector of all parameters is represented by $\boldsymbol{\theta}$. In AE data clustering, GMM's [32] have been widely used [33, 52, 53, 54, 55, 23] and are considered in the following developments. Each component in the mixture (4) is then a Gaussian pdf:

$$\phi(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right), \quad (5)$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are, respectively, the mean and covariance matrix of component k .

After observing a realization $\mathbf{x}_1, \dots, \mathbf{x}_N$ from an iid sample, the likelihood function is

$$L(\boldsymbol{\theta}; \mathbf{x}_1, \dots, \mathbf{x}_N) = \prod_{i=1}^N \sum_{k=1}^K \pi_k \phi(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (6)$$

100 where $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K, \pi_1, \dots, \pi_{K-1})$. The maximum likelihood estimates (MLE's) cannot be computed in closed form and are usually computed numerically using the Expectation-Maximization (EM) algorithm [56].

New model. We propose to modify (6) in order to incorporate a time-dependency of the data through the proportions:

$$p(\mathbf{x}_1, \dots, \mathbf{x}_N; \boldsymbol{\theta}) = \prod_{i=1}^N \sum_{k=1}^K \pi_{ik} \phi(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (7)$$

where the bold subscript in π_{ik} emphasizes the difference with (6), i.e., the proportion of each cluster k are now dependent on the timestamps t_i through additional variables α_{ik} :

$$\pi_{ik} = \frac{\alpha_{ik}}{\sum_{\ell=1}^K \alpha_{i\ell}}, \quad k = 1, \dots, K, \quad (8)$$

²The project is publicly available at https://github.com/emmanuelramasso/MIXMOD_SEQUENTIAL.

where $\alpha_{i1} = 1$ for $i = 1, \dots, N$ and

$$\alpha_{ik} = \frac{\beta_k}{1 + \exp[-\gamma_k(t_i - \tau_k)]}, \quad k = 2, \dots, K, \quad i = 1, \dots, N. \quad (9)$$

Parameters τ_k , β_k and γ_k in the logistic (sigmoid) activation functions (9) must satisfy the following constraints:

$$0 \leq \tau_k \leq T, \quad \beta_k \geq 0, \quad \gamma_k \geq 0 \quad (10)$$

for $k = 2, \dots, K$. As illustrated in Figure 1 (bottom-left), the degree of activation α_{ik} of the k -th cluster depends on the real timestamps of AE signals through a sigmoid function delayed by τ_k , with upper limit β_k and slope γ_k . The delay τ_k represents the onset time of cluster k . The proportions π_{ik} in (8) are equal to the normalized activation degrees. An example of how proportions can vary in time is shown in Figure 5b.

The apparently simple modification of GMM brought about by making the proportions in (6) dependent on time allows us to tackle the problem mentioned in Section 2, concerning the inability of standard approaches to manage continuous and irregularly spaced-in-time timestamps. By associating a sigmoid function to each cluster, GMMSEQ is, to our knowledge, the first clustering method dedicated to AE data able to estimate, directly from data, parameters related onsets, growth and kinetics of clusters. Its performance will be studied in Section 3. Parameter estimation in this new model requires a specific, and more complex EM algorithm, described in the next section.

2.2. Parameter estimation

Maximum-likelihood parameter estimation in GMM's is usually carried out using the EM algorithm [57], an approach that will also be used here. However, there is no closed-form expression to update parameters τ_k , β_k and γ_k in the M-step, which makes it necessary to use a gradient algorithm. The algorithm is described in detail below.

Objective function. The first step is to write down the complete-data log-likelihood function for our model:

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^N \sum_{k=1}^K y_{ik} \log \pi_{ik} + y_{ik} \log \phi(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (11)$$

where $\boldsymbol{\theta} = (\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K, \{\tau_k, \beta_k, \gamma_k\}_{k=2}^K)$ is the parameter vector, and the y_{ik} 's are binary cluster-membership indicator variables such that $y_{ik} = 1$ if observation i belongs to k , and $y_{ik} = 0$ otherwise. Here, variables y_{ik} are missing. At each iteration q of the EM algorithm, we thus replace ℓ_c by its conditional expectation given the observed data, which yields the so-called *auxiliary function* Q [56]:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = \underbrace{\sum_{i=1}^N \sum_{k=1}^K y_{ik}^{(q)} \log \pi_{ik}}_{Q_1} + \underbrace{\sum_{i=1}^N \sum_{k=1}^K y_{ik}^{(q)} \log \phi(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{Q_2}, \quad (12)$$

with $y_{ik}^{(q)} = \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[Y_{ik} | \mathbf{x}_i]$.

We can observe that the term Q_2 on right-hand side of (12) is identical to that of the auxiliary function for a standard GMM (6) with fixed proportions, for which parameter updates that maximizes $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ are known (and recalled below). The difference between our EM procedure and

125 the usual one for GMM's thus resides in the maximization of the first term Q_1 with respect to the parameters defining the proportions π_{ik} . This procedure is detailed below.

E-step. In the E-step, we compute the conditional expectations $y_{ik}^{(q)}$ from the current parameter values as [32]:

$$y_{ik}^{(q)} = \frac{\phi(\mathbf{x}_i; \boldsymbol{\mu}_k^{(q)}, \boldsymbol{\Sigma}_k^{(q)})\pi_{ik}^{(q)}}{\sum_{l=1}^K \phi(\mathbf{x}_i; \boldsymbol{\mu}_l^{(q)}, \boldsymbol{\Sigma}_l^{(q)})\pi_{il}^{(q)}} \quad (13)$$

with ϕ given by (5).

M-step for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. In the M-step, parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are first updated by maximizing Q_2 . The update equations are [32]:

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{1}{N_k} \sum_i y_{ik}^{(q)} \mathbf{x}_i, \quad (14)$$

with $N_k = \sum_i y_{ik}^{(q)}$, and

$$\boldsymbol{\Sigma}_k^{(q+1)} = \frac{1}{N_k} \sum_i y_{ik}^{(q)} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})^\top. \quad (15)$$

M-step for τ_k , β_k and γ_k . Since no explicit update equations for parameters τ_k , β_k and γ_k can be obtained, they need to be updated by an iterative optimization procedure. To enforce the constraints (10), we first introduce the following auxiliary variables:

$$\tau_k = \frac{T}{1 + \exp(-\xi_k)}, \quad \beta_k = b_k^2, \quad \gamma_k = g_k^2. \quad (16)$$

The calculation of the derivatives of Q_1 with respect to ξ_k , b_k and g_k is detailed in Appendix A. Using the gradient, we can then use any unconstrained nonlinear optimization procedure. In the experiments reported in Section 3, we used a trust region algorithm implemented in the MATLAB 2020b Optimization toolbox. Several schemes were implemented and tested, included a Generalized EM [58] and various optimization algorithms. The trust region method globally provided the best results.

2.2.1. Regularisation of the τ_k 's

In the considered application related to acoustic emission data clustering, physical knowledge can be available suggesting when, in the timeline of a test, some damages must have occurred. This information can be provided, for example, as prior values τ_k^{prior} for τ_k (for some or all k depending on the application). In this case, the auxiliary function (12) can be replaced by a regularised version

$$Q_r(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) - \lambda \|\boldsymbol{\tau} - \boldsymbol{\tau}^{\text{prior}}\|_2^2, \quad (17)$$

135 where λ is a regularization coefficient, $\|\cdot\|$ is the L_2 norm and $\boldsymbol{\tau}$ is the vector of τ_k 's for which a prior value is available. In that case, the updating is modified for $\boldsymbol{\tau}$ according to Equation A.8.

3. Experiments

The GMMSEQ method is first illustrated on a toy data set in Section 3.1. It is then applied to real experimental data from a mechanical system in Section 3.2.

140 *3.1. Simulated data set*

Model for data generation. A simulated data set was generated from the following model with $K = 4$ clusters:

$$\boldsymbol{\mu}_1 = [1 \ 1], \quad \boldsymbol{\mu}_2 = [2 \ 3], \quad \boldsymbol{\mu}_3 = [3 \ 5], \quad \boldsymbol{\mu}_4 = [5 \ 6],$$

and

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.2 \end{bmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.3 \end{bmatrix}, \quad \boldsymbol{\Sigma}_4 = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}.$$

The numbers of observations in the four clusters were set to [1000, 1000, 3000, 1000] and the time-stamps were generated randomly from a uniform distribution

$$t_i = t_{i-1} + \mathcal{U}_{[0,1]},$$

starting from $t_1 = 0$. The value of T is thus $\max_i t_i$ and the length of the data is equal to 6000. The parameters defining the time-varying proportions were set as follows:

$$\boldsymbol{\beta} = [2.72 \ 10.1 \ 30.2], \quad \boldsymbol{\gamma} = [0.009 \ 0.015 \ 0.012],$$

and the τ_k were initialized to the time stamps t_i with $i \in \{488, 1990, 2472\}$.

Because the proportions in this model vary with time, the mixture density also depends on time. Contours of the mixture density are depicted in Figure 2 at four different time steps showing the gradual emergence of the four clusters. Figure 3 shows a contour plot of the likelihood function (assuming the correct number of clusters) as a function of β_4 and γ_4 , with the other parameters fixed at their maximum likelihood estimates.

Model selection. In practice, we need a criterion allowing us to select the number of clusters automatically. For mixture models, common choices are the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Integrated Completed Likelihood (ICL) [59, 60]. We used these three criteria to select the best GMMSEQ model for the simulated data. For each run, the model was initialized by:

1. A standard GMM (with a Matlab 2020b implementation provided in the Statistics and Machine Learning Toolbox);
2. The K-means algorithm (provided in the same toolbox as for GMM);
3. A segmentation of the data into K blocks of the same size and a Gaussian distribution fitted to each block.

For the three types of initialization, 10 runs were performed for a number of clusters varying from 2 to 10. After convergence, for each number of clusters, the best run was selected according to the value of likelihood computed by (7). The values of AIC, BIC and ICL were finally computed for each model. The criteria are plotted against the number K of clusters in Figure 4. AIC shows an evolution presenting an “elbow” from which the number of clusters can be chosen, while BIC and ICL have a minimum for the correct number of clusters ($K = 4$). For ICL, depending on the runs, the evolution can show several local minima as in the figure or a single one, but the global minimum is always located at the correct value. For different runs, BIC and AIC showed a consistent elbow-shaped behavior with a minimum that can be more or less pronounced for the correct number of clusters. This study with simulated data suggests that the three criteria have the ability to provide the correct number of clusters.

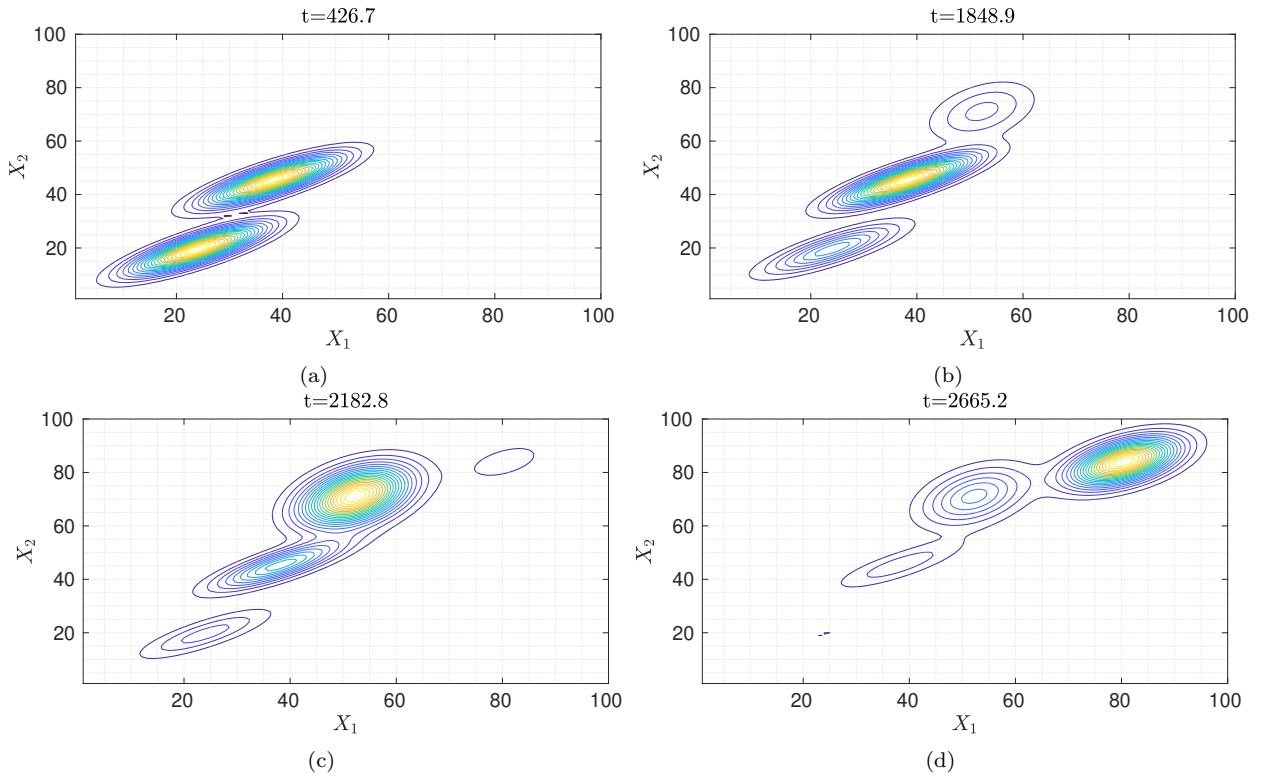


Figure 2: Contours of the mixture density at four successive time steps. The corresponding proportions π_{ik} correspond to the locations of the square markers in Figure 5.

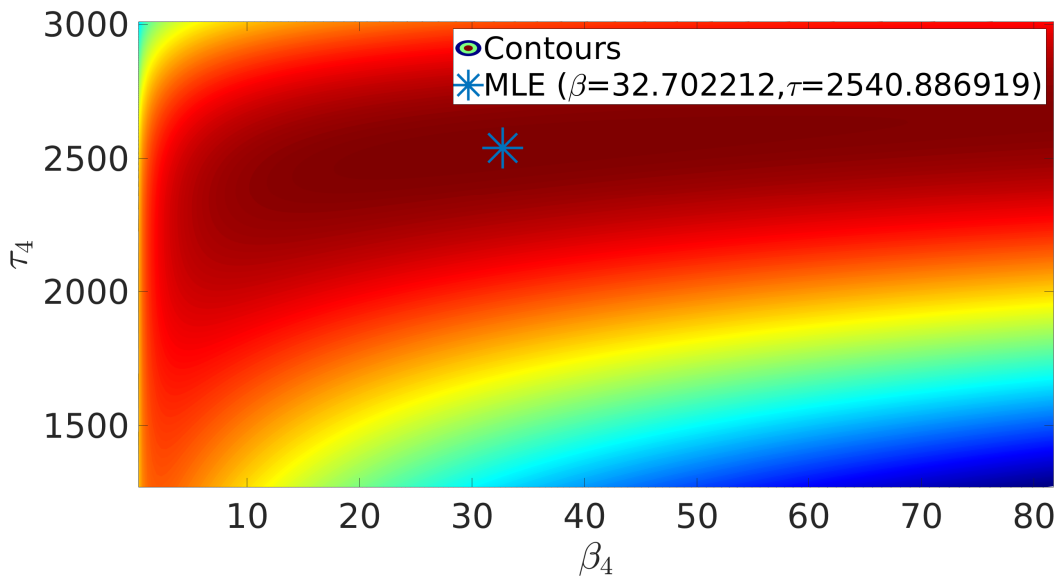


Figure 3: Contours of the log-likelihood in the plane (β_4, τ_4) , the other parameters being fixed to their maximum likelihood estimates.

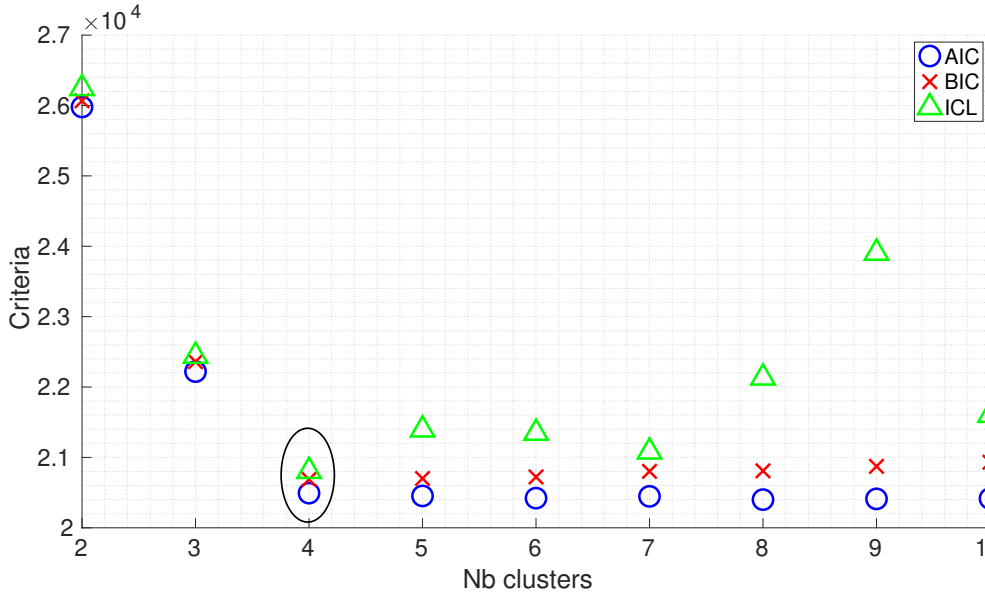


Figure 4: AIC, BIC and ICL of GMMSEQ models with different numbers of clusters for the simulated data.

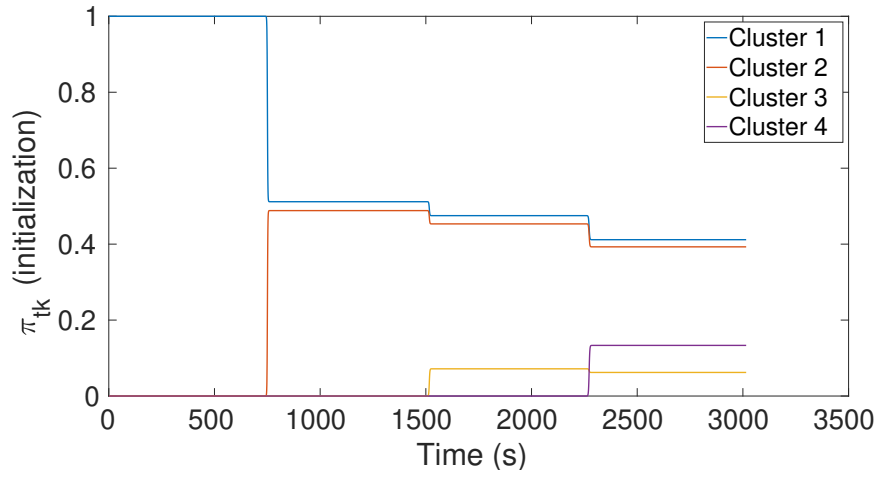
Onset estimation. In the introduction, we motivated this work by the need to process data with continuous timestamps and we proposed a model including an estimation of the onsets of clusters. The time-varying proportions (π_{ik}) are represented in Figure 5 using the true values of parameters, the estimated ones, and the values used in the initialization of GMMSEQ. We can observe that the estimated proportions (β), kinetics parameter (γ) and onsets (τ) are very close to the true ones, validating the identification procedure in GMMSEQ.

3.2. Real data

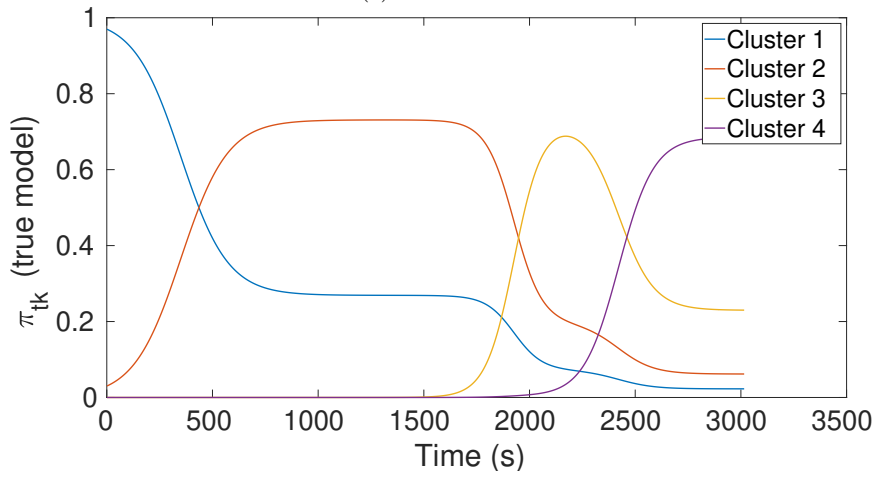
Data set description. The benchmark data set ORION-AE [51, 61] is used in this section to demonstrate the performance of the GMMSEQ method. The experiments were designed to reproduce the loosening phenomenon observed in aeronautics, automotive or civil engineering structures where parts are assembled together by means of bolted joints (Figure 6). The bolts can, indeed, be subject to self-loosening under vibrations. Consequently, it is of paramount importance to develop sensing strategies and algorithms for early loosening estimation [62].

The ORION-AE data set is composed of five parts collected during five measurement campaigns denoted as B , C , D , E and F in the sequel. ORION is a simple jointed structure made of two plates manufactured in a 2024 aluminium alloy, linked together by three bolts. The contact between the plates is done through machined overlays. The contact patches have an area of $12 \times 12 \text{ mm}^2$ and are 2 mm thick. The structure was submitted to a 100 Hz harmonic excitation force. The load was applied using a Tyra electromagnetic shaker, which can deliver a 200 N force. The force was measured using a PCB piezoelectric load cell and the vibration level was determined next to the end of the specimen using a Polytec laser vibrometer.

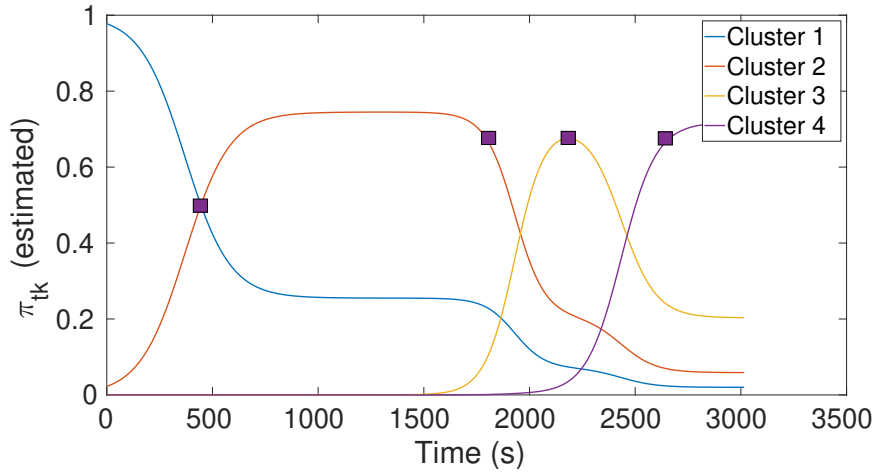
Seven tightening levels were applied on the upper bolt. The tightening was first set to 60 cNm with a torque screwdriver. After a 10 seconds vibration test, the shaker was stopped and this vibration test was repeated after a torque modification at 50 cNm. Torque modifications at 40, 30, 20, 10 and 5 cNm were then applied. Note that, for campaign C , the level 20 cNm is missing. All dimensions are detailed in Figure 6 to enable readers to reproduce the test.



(a) Initialization.



(b) True values.



(c) Estimated values.

Figure 5: Evolution of the proportions π_{ik} as a function of time (in arbitrary units and was generated randomly as explained in the text). Squared-shape markers represent the values used in Figure 2.

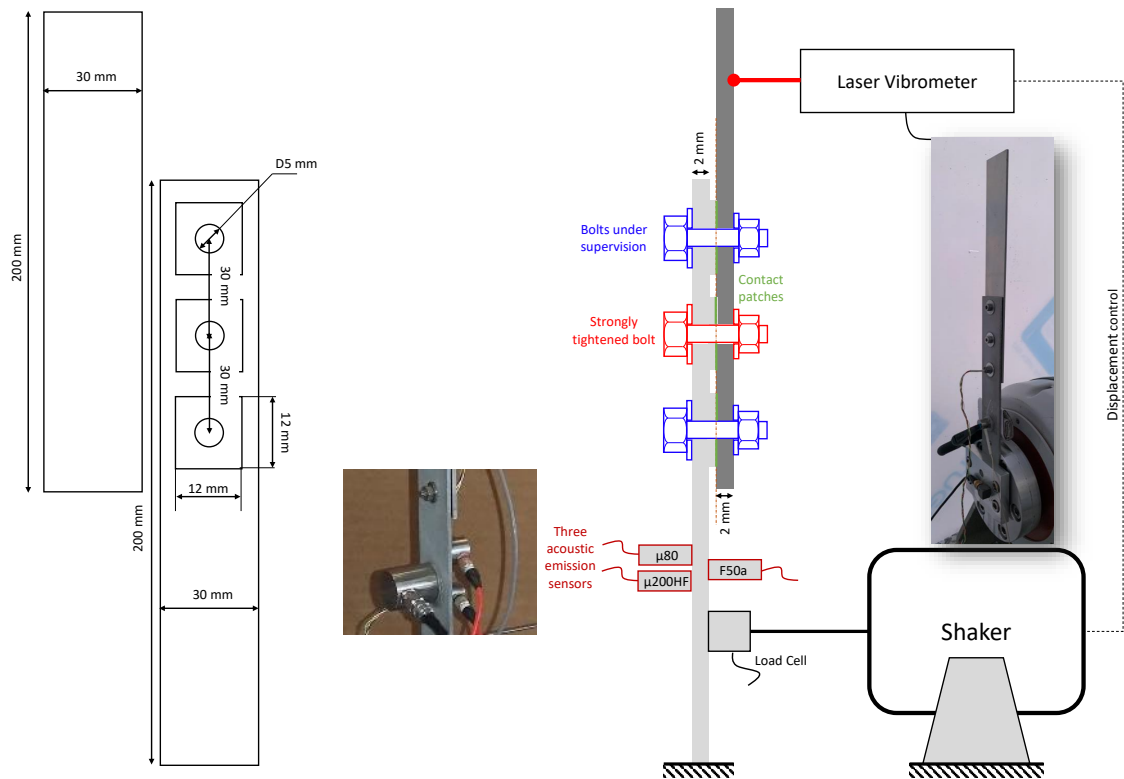


Figure 6: Setup description: part dimensions, sensors and bolts position (from [61]).

For each campaign, four sensors were used: a laser vibrometer and three different AE sensors (micro-200-HF, micro-80 and the F50A from Euro-Physical Acoustics) with various frequency bands were attached onto the lower plate. All data were sampled at 5 MHz. The velocimeter was used to control the amplitude of the displacement of the top of the upper beam so that it remains constant for all tightening levels. During vibrations, stick-slip transitions or shocks in the interface generate small AE events which are dependent on bolt tightening. These sources of AE signals have to be detected and identified from the data stream, which constitutes the challenge.

AE data sets are generally unlabeled because it is not possible to identify the AE source with certainty for all AE signals. However, the ORION-AE data set contains raw data for which the tightening levels are known. Therefore, it represents a good case study for performance benchmarking of clustering methods like GMMSEQ or existing methods like GMM, K-means, hierarchical clustering (HC) and Gustafson-Kessel (GK) algorithms. The data set is presented in a companion paper [61] with illustrations of raw data and signal processing for different campaigns. Figure 7 depicts the data in campaign F where the green curve represents the raw AE data (as used in this study). The staircase curve in blue represents the tightening levels. The red curve corresponds to the vibrometer data (reflecting harmonic vibration at 100 Hz with displacement control).

Signal processing. When using a feature-based clustering algorithm like GMMSEQ, the raw AE data must be first preprocessed by a hit detection procedure. This procedure is common to most AE data analyses and therefore, it is not often described in publications because commercial softwares are often used. In this work, the method used was described in [13]. It is summarized in Appendix B and some illustrations of this method applied to the ORION-AE data can also be found in [61]. The

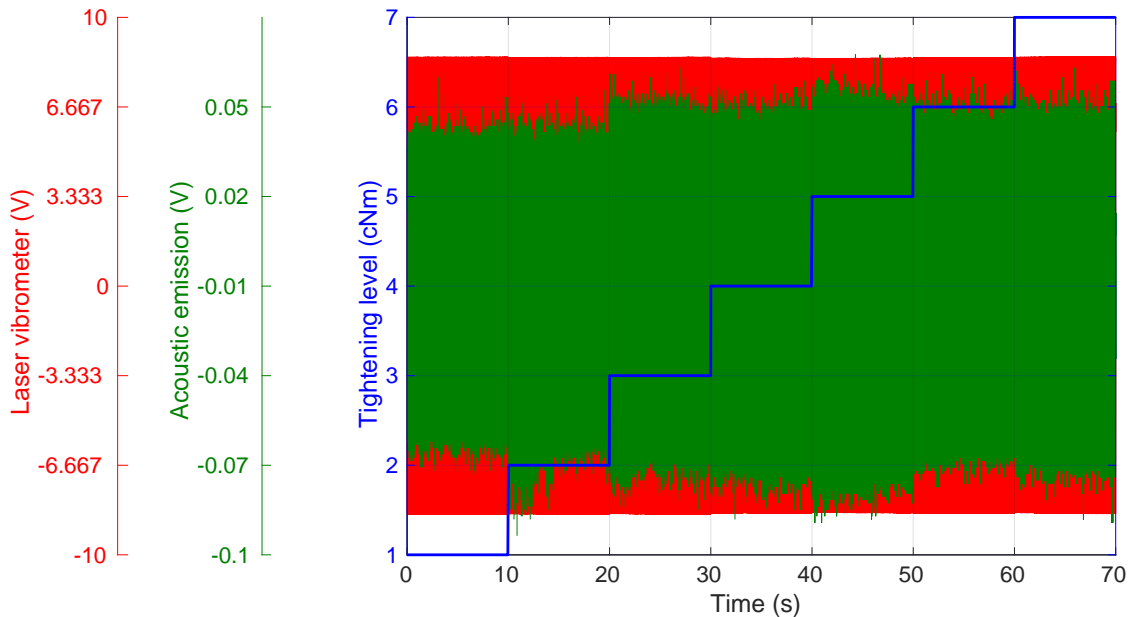


Figure 7: Tightening levels, acoustic emission and laser vibrometer data superimposed for measurements “F” and sensor micro-200-HF. The x-axis is here represented using the time of test (about 70 s), starting from 60 cNm from the left (around between $t \in [0, 10]$ s) to 5 cNm on the right (from $t > 60$ s).

215 method provides the following features, which are commonly used in AE literature [63, 40, 13, 16]: Rise time, counts, PAC-energy, duration, amplitude, average frequency, RMS, average signal level, counts to peak, reverberation frequency, initiation frequency, signal strength, absolute energy, partial power in the intervals [0, 20, 100, 200, 300, 400, 500, 600, 800, 1000] kHz, frequency centroid, peak frequency, weighted peak frequency. To this set of features were added the following ones:
 220 the Renyi number calculated from the scalogram as in [64] using a Morlet wavelet, as well as the frequency of the maximum of energy in the scalogram³. The set of feature vectors obtained in each campaign were then postfiltered using a 31-sample moving median applied to each dimension of the resulting feature matrix in order to ensure temporal coherence. Principal Components Analysis (PCA) was then used to extract the n first components explaining 99% of the variance. The value
 225 of n varies for the different campaigns as shown in Table 1, where sensor micro-200-HF was used. Figure 8 displays the first two components for campaign E and sensor micro-200-HF, where the colors are related to the level of loosening.

3.3. Results

230 We ran the algorithm 10 times for each of the same three initialization methods mentioned in Section 3.1. The number of clusters was varied from 4 to 14. For each number of clusters, the parameter estimates corresponding to the highest likelihood were selected.

In Section 1, we discussed the importance of the onsets in the analysis of AE data set (represented by τ_k in GMMSEQ). These values were stacked for each campaign, independently of the type of initialization or the number of clusters. Figure 9 shows how to interpret, in terms of timeline,

³The feature matrices for all campaigns are available at https://drive.google.com/drive/folders/1H413RxYu4ya7YMEgF_1Th_fHr7f1vv00?usp=sharing.

Table 1: Some statistics about the feature extraction step (sensor $\mu 200HF$).

Campaign	B	C	D	E	F
# of tightening levels	7	6	7	7	7
# features before PCA	32	32	32	32	32
# features after PCA	16	17	16	23	25
Total # of signals	10,866	9,461	9,285	15,628	17,810
Average # of signals per period	1.55	1.57	1.33	2.23	2.54

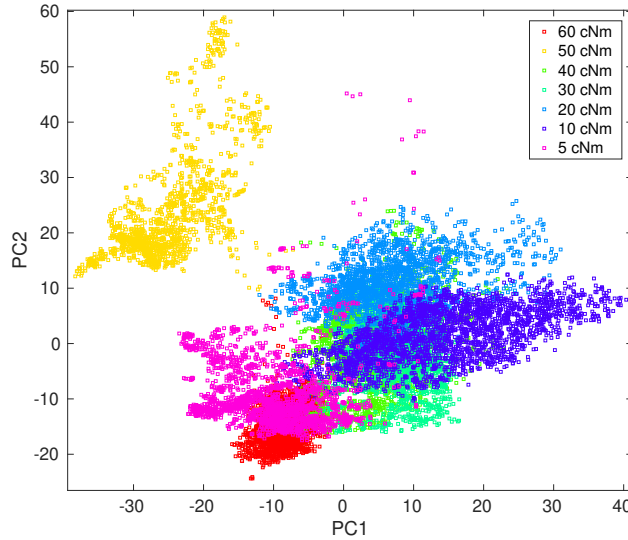


Figure 8: First two principal components of campaign *E* data.

235 Figures 10 to 14 representing, in **blue**, the normalised histograms of onsets estimated by GMMSEQ for each campaign. The dashed lines represent the instant when the tightening level was changed. These lines are separated by about 10 s (duration of each period) for each level; therefore, the horizontal axis, which represents time, can also be related to the tightening level: 0 s to 10 s corresponding to 60 cNm, 10 s to 20 s corresponding to 50 cNm, and so on until 60 to 70 s for 5 cNm (see Figure 9). The **red** bars represent the histograms when a prior on onsets is integrated through (17) for $K=7$ clusters.

245 In Figures 10 to 14, a peak in the histogram means that several models provided similar values for τ_k . Note that there are, for the **blue** bars, $\sum_{k=4}^{14} k = 99$ estimates of the τ_k values, whereas there are $K = 7$ values of τ_k for the **red** bars. In the latter case, we can observe that the prior on onsets allows us to obtain values of τ_k approximately equal to the ground truth (dashed lines) for all data sets and all tightening levels. The **red** bars depict a uniform distribution since the τ_k values are all different.

250 For the purely unsupervised setting (in **blue**), we can observe that the values of τ_k with the highest probability generally correspond to the instants when a change was made on the tightening level. This observation shows that GMMSEQ is able to discover the levels of tightening from the features. For each campaign we can make the following comments, remembering that the levels

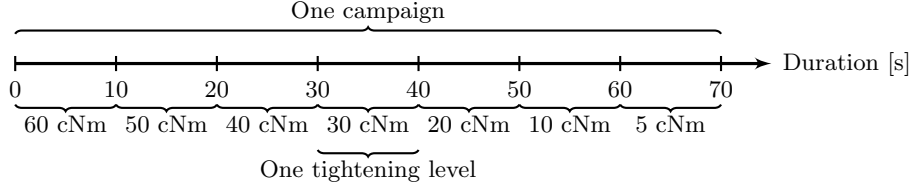


Figure 9: Timeline of tightening levels to interpret the next figures.

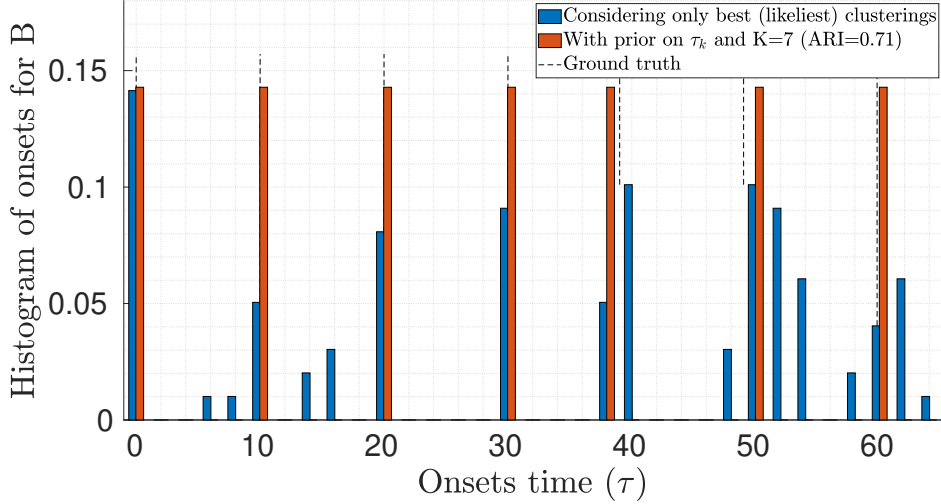


Figure 10: Campaign B: (blue) Histogram of τ_k estimates for the three initialization methods and K ranging from 4 to 14, and (red) histogram with prior on onsets with regularization.

were approximately equal to 60 (cluster 1), 50 (cluster 2), 40 (cluster 3), 30 (cluster 4), 20 (cluster 5), 10 (cluster 6) and 5 cNm (cluster 7):

- Campaign B (Figure 10): Levels 60, 40, 30, 20 are precisely detected with a clear peak centered at the correct place. Levels 50, 10 and 5 shows less noticeable peaks but the bins in the histogram show modes that are well positioned around the expected positions.
- Campaign C (Figure 11): For this campaign, a level is missing, which explains why there is no peak around 5 cNm (right-hand side); the levels were indeed shifted by one level in the figure due to one missing level. Figure 11 shows clear peaks at the correct positions. There are also two additional peaks at 25s (30 cNm) and (5 cNm), which may be due to a change in the level during vibration tests.
- Campaigns D (Figure 12) and E (Figure 13): All levels are precisely detected with a clear peak centered at the correct location.
- Campaign F (Figure 14): Levels 60, 50, 40, 20, 10, 5 are precisely detected with a clear peak (or two close peaks) centered at the correct location. For level 30 the peak is less noticeable. Some additional peaks appear around 5 s (middle of the period of 60 cNm) and 43 s (20 cNm), which may be due to a change in the level during vibration tests.

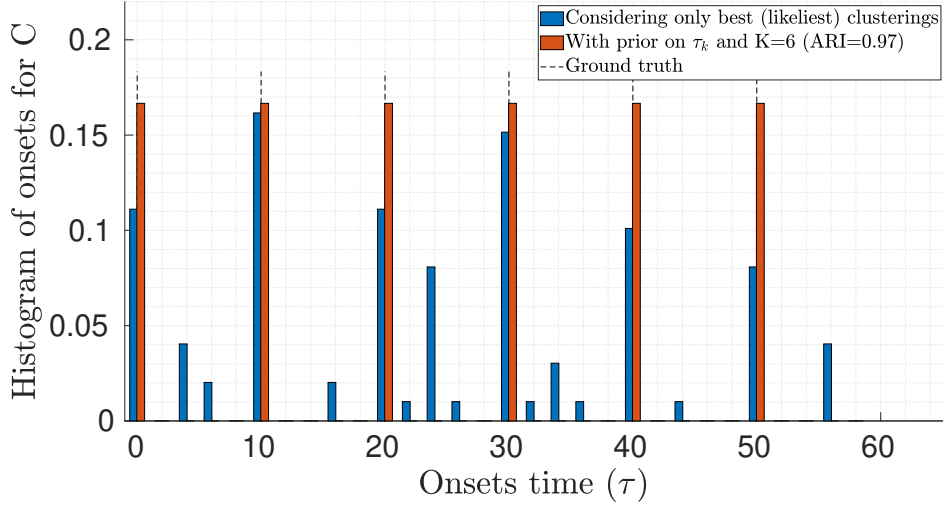


Figure 11: Campaign C: (blue) Histogram of τ_k estimates for the three initialization methods and K ranging from 4 to 14, and (red) histogram with prior on onsets with regularization.

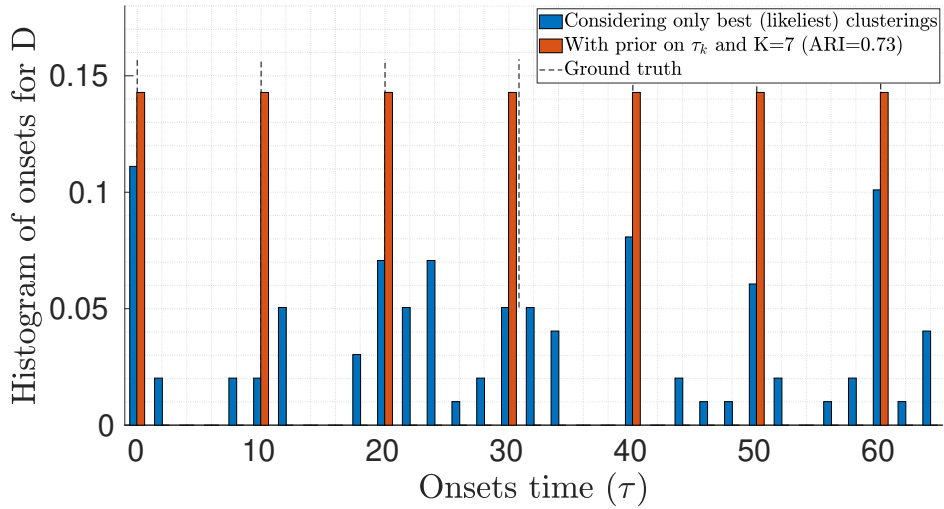


Figure 12: Campaign D: (blue) Histogram of τ_k estimates for the three initialization methods and K ranging from 4 to 14, and (red) histogram with prior on onsets with regularization.

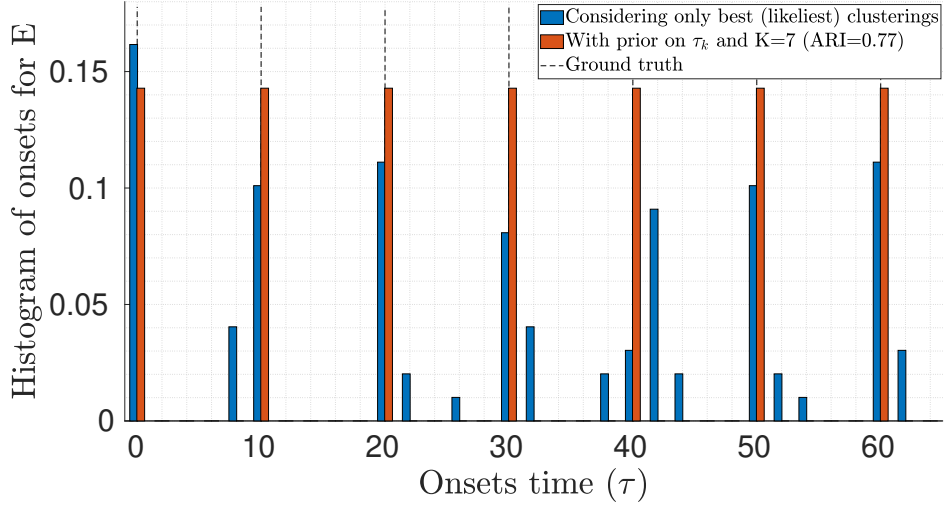


Figure 13: Campaign E: (blue) Histogram of τ_k estimates for the three initialization methods and K ranging from 4 to 14, and (red) histogram with prior on onsets with regularization.

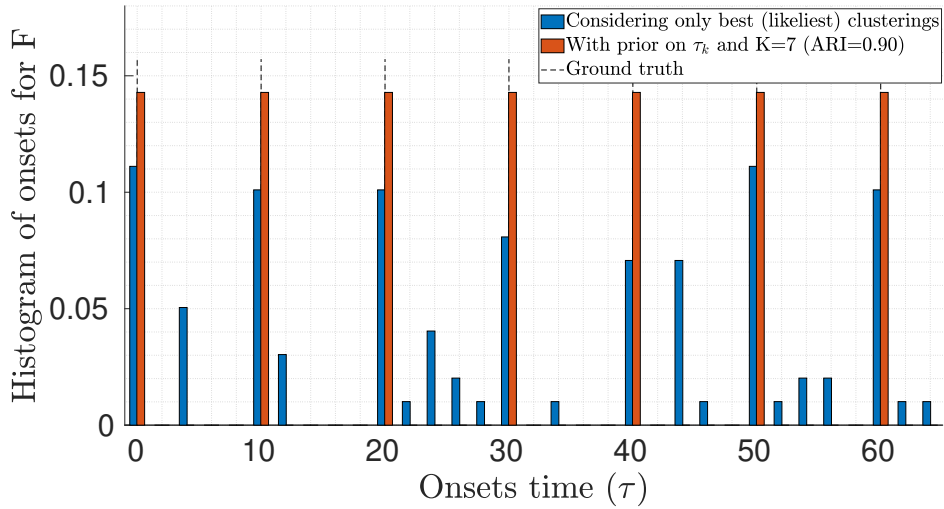


Figure 14: Campaign F: (blue) Histogram of τ_k estimates for the three initialization methods and K ranging from 4 to 14, and (red) histogram with prior on onsets with regularization.

Table 2: Performance of GMMSEQ with and without (“w/o”) prior on onsets, according to four measures of performance. Each measure is bounded in $[0, 1]$ (the higher the better).

Campaign	Precision		Recall		Entropy		ARI	
	with	w/o	with	w/o	with	w/o	with	w/o
B	1.00	0.123	1.00	1.00	1.00	0.944	0.708	0.731
C	1.00	0.146	1.00	1.00	1.00	0.992	0.974	0.966
D	1.00	0.098	1.00	0.857	1.00	0.852	0.733	0.842
E	1.00	0.156	1.00	1.00	1.00	0.955	0.772	0.774
F	1.00	0.143	1.00	1.00	1.00	0.961	0.899	0.847

From these figures, it can also be observed that GMMSEQ generated onsets with quite similar probability (except for campaign B). The bars in the histogram generally exhibit quite similar values for all tightening levels, which means that the number of AE signals for each tightening level are quite similar. Since the vibration has a fixed frequency independent of the tightening, the sources of AE signals seem to be activated in each cycle.

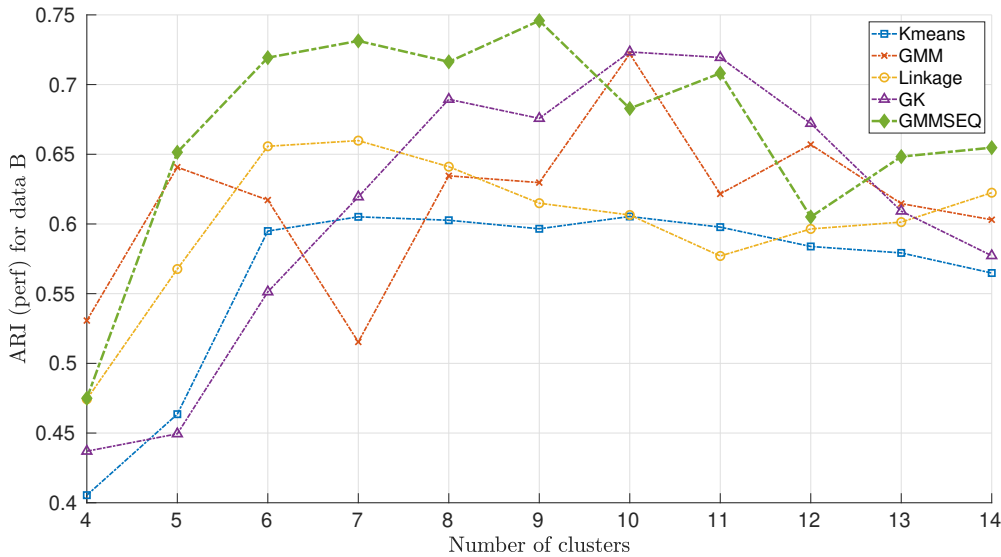


Figure 15: Campaign B: Comparison between standard clustering methods and GMMSEQ.

The evaluation of onsets was made by four criteria, each presenting a different view of the performance, namely precision, recall, entropy and ARI, detailed below. Each estimated onset was compared with the ground truth: If it remains within only ± 0.5 second around a true value then it is considered as a correct estimation and qualified as a true positive (TP). If the onset is outside the interval, it is counted as a false positive (FP). If no onset is found in a given interval, a false negative (FN) is counted. TP, FN and FP are counted for all onsets. Then the accuracy is given by

$$Pr = \frac{TP}{TP + FP},$$

which means that if a method has no false positive, the accuracy is 1. A low accuracy corresponds

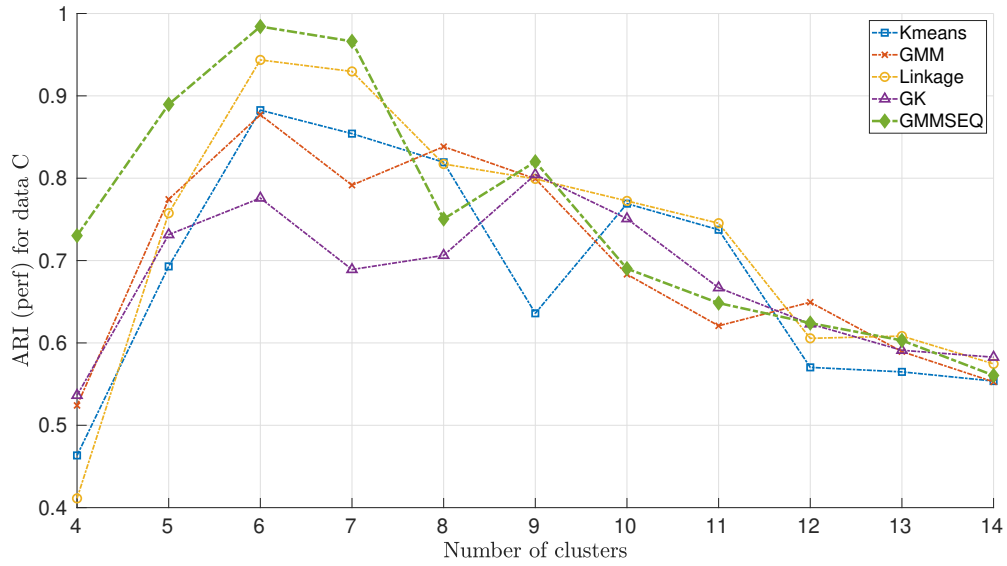


Figure 16: Campaign C: Comparison between standard clustering methods and GMMSEQ.

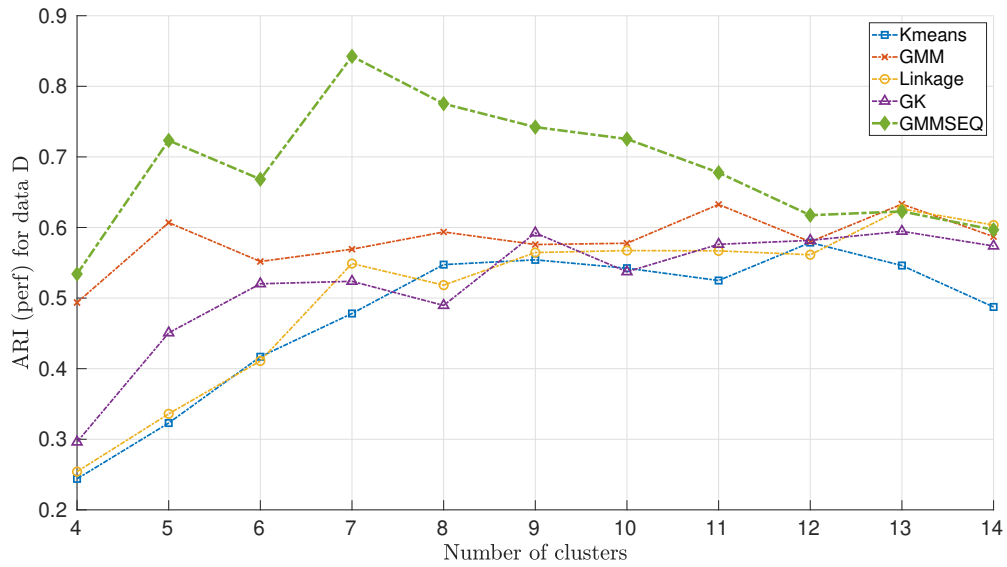


Figure 17: Campaign D: Comparison between standard clustering methods and GMMSEQ.

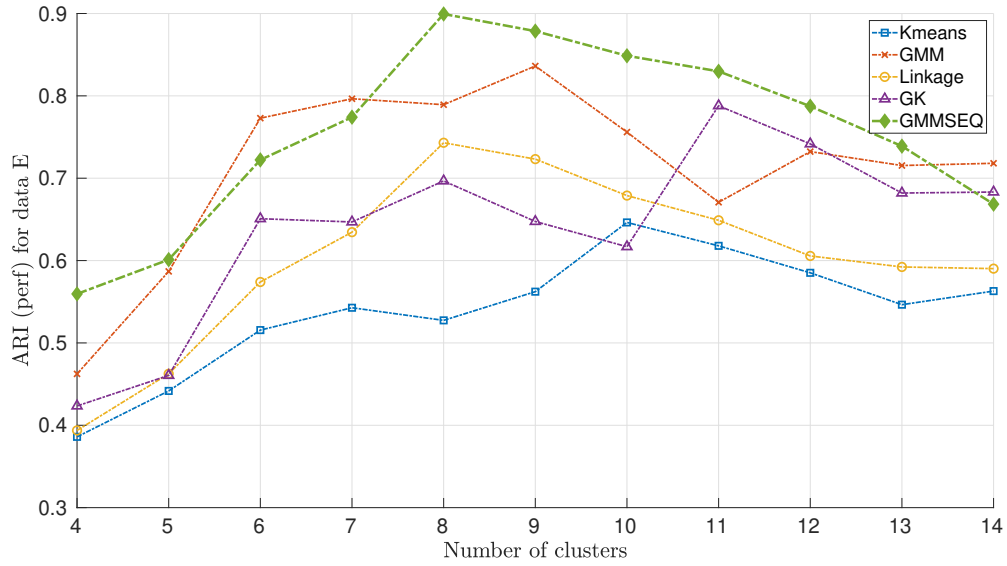


Figure 18: Campaign E: Comparison between standard clustering methods and GMMSEQ.

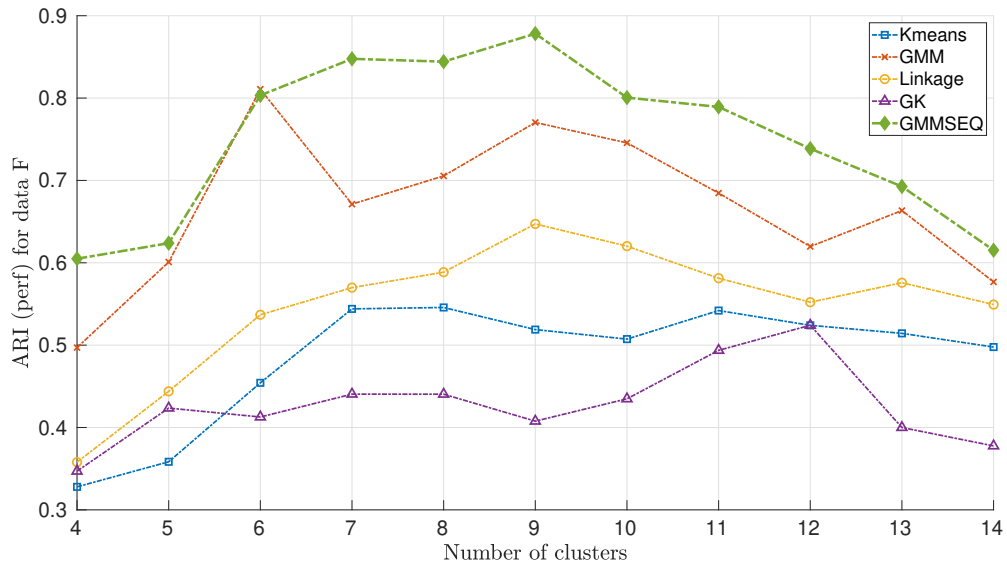


Figure 19: Campaign F: Comparison between standard clustering methods and GMMSEQ.

to many FP. Therefore, a complementary criteria, called recall, is also used:

$$Rec = \frac{TP}{TP + FN}$$

which means that if a method does not miss any onset, the recall is 1. A method with many FP can have a high recall if all expected onsets are found.

We also expect the methods to find onsets at similar locations when the number of clusters is changed (as explained in the analysis of histograms discussed above). For that purpose, the normalized entropy of the cumulated onsets is computed as

$$Ent = -\frac{\sum_{k=1}^{6 \text{ or } 7} p(k) \log_2 p(k)}{\log_2 K},$$

where the sum is over the number of levels (6 for campaign C and 7 for the others) and with

$$p(k) = \frac{\text{number of estimated onsets falling in } \pm 0.5 \text{ s around the truth}}{\text{total number of onsets estimated by the method}}$$

275 which means that if a method provides n onsets and there are $n/7$ onsets per level correctly located (or $n/6$ for campaign C), then the entropy is 1. An example of cumulated onsets with perfect location (leading to an entropy of 1) is given with the red bars in Figures 10 to 14.

The fourth criterion is the Adjusted Rand Index (ARI) [65]. ARI is a corrected-for-chance (and, thus, more severe) version of the Rand Index used for clustering evaluation in the presence
280 of a ground truth (in the present case, it corresponds to the tightening levels). The ARI is 1.00 for perfect clustering and 0 in case of random assignment of clusters or totally wrong clustering. The results are summarized in Tables 3, 4, 5, 6 and 7 for all campaigns (B to F respectively), all criteria (precision, recall, entropy and ARI) and several clustering methods (results will be described below).

285 For comparison purposes, four standard clustering methods were applied: K-means, a standard GMM, a hierarchical clustering (HC) using the Ward method and the Gustafson-Kessel (GK) algorithm (implementations are available in the MATLAB Statistics and Machine Learning (SML) toolbox and on MATLAB Central). The same matrix of features was used for all methods, including GMMSEQ. The K-means algorithm was run 10 times and the model with the smallest sum of
290 squared-distance was selected. GMM was run 10 times with full covariance matrices for each cluster and the model with the highest likelihood was selected. A GMM was used to initialize the Gustafson-Kessel algorithm with fuzziness degree set to 1.5 as proposed in [21]. For these three methods, the onsets are defined by the time of the first occurrence of every cluster. For GMMSEQ, onsets were given by τ values.

295 Figures 15 to 19 depict the ARI for all methods and all campaigns with respect to the number of clusters. In each figure, the curves with diamond markers represent the ARI for GMMSEQ. For this method, it can be observed that the maximum is generally observed for 6 to 9 clusters which is consistent with the expected number of tightening levels. These figures show that GMMSEQ globally outperforms K-means, GK, HC and GMM clustering methods. A more detailed analysis
300 is provided in Tables 3 to 7 using the four aforementioned criteria. The ARI is the average over 6 to 9 clusters (and 5-8 for campaign C). In the tables, the “**bold**” font is used for the best overall for a given criterion, while an “underlining” is used when comparing only GMMSEQ and the original GMM. Therefore, if one of these two methods is both underlined and bold then it represents the best of all methods for a given criterion.

305 The tables show that:

310

315

- For all campaigns and all methods, the precision is very low meaning that no method is able to detect all onsets at ± 0.5 second around the true value.
- For the recall criterion, GMMSEQ provides a value of 1.0 for four campaigns, while providing a value of 0.85 (against 0.714 only for GMM) for campaign D. It globally outperforms all methods. The recall of the original GMM is never better than the recall of GMMSEQ.
- According to the entropy, GMMSEQ provides the best performance for four campaigns, and is only outperformed by HC for campaign D (0.852 against 0.921). The entropy of the original GMM is never better than the entropy of GMMSEQ.
- Finally, GMMSEQ provides the best ARI for all campaigns. Compared to the GMM, we have +13.3%, +13.3%, +19%, +1.3% and +10.3% for, respectively, campaign B, C, D, E and F. The ARI of the original GMM is thus never better than the ARI of GMMSEQ.

These tables show that there is no best method for all campaigns and according to all criteria. However, from a quantitative point of view, GMMSEQ globally outperforms the GMM which shows the significance of the proposed model.

Table 3: Campaign B: Performance

Algo/perf	precision	recall	entropy	ARI
Kmeans	0.066	0.571	0.597	0.601
GMM	0.100	0.857	0.805	0.589
HC	0.067	0.571	0.712	0.652
GK	0.081	0.714	0.706	0.620
GMMSEQ	0.123	1.00	0.944	0.722

Table 4: Campaign C: Performance

Algo/perf	precision	recall	entropy	ARI
Kmeans	0.188	1.00	0.853	0.810
GMM	0.200	1.00	0.931	0.814
HC	0.118	0.667	0.739	0.877
GK	0.177	1.00	0.963	0.732
GMMSEQ	0.146	1.00	0.992	0.947

Table 5: Campaign D: Performance

Algo/perf	precision	recall	entropy	ARI
Kmeans	0.152	1.00	0.799	0.480
GMM	0.086	0.714	0.791	0.572
HC	0.212	1.00	0.921	0.493
GK	0.098	0.714	0.801	0.511
GMMSEQ	0.097	0.857	0.852	0.762

Table 6: Campaign E: Performance

Algo/perf	precision	recall	entropy	ARI
Kmeans	0.065	0.429	0.561	0.529
GMM	0.146	1.00	0.891	0.786
HC	0.069	0.571	0.600	0.651
GK	0.058	0.429	0.547	0.665
GMMSEQ	0.156	1.00	0.956	0.799

Table 7: Campaign F: Performance

Algo/perf	precision	recall	entropy	ARI
Kmeans	0.062	0.571	0.646	0.515
GMM	0.146	1.00	0.896	0.729
HC	0.078	0.714	0.774	0.565
GK	0.100	0.714	0.728	0.431
GMMSEQ	0.143	1.00	0.961	0.832

320 The previous results and analyses demonstrated the usefulness of considering onsets, τ , as
parameters to be identified from AE data together with clusters parameters. The simulated data
also allowed us to show that our model was able to perfectly recover not only the onsets but also
the levels of activation, β , and kinetics, γ for four clusters with different behavior. In order to
conclude this study, we propose an analysis of these parameters, at the core of GMMSEQ, on the
325 real data.

Figures 20 to 24 illustrate the degrees of activation, π_{tk} , in each campaign computed from the
GMMSEQ model with the highest ARI. The value π_{tk} is obtained from τ , β and γ using Eq. 8.
Onsets, kinetics and levels of activation are well illustrated on these figures. It is expected that
the levels of activation (the height of curves) should be similar for all tightening levels since we
330 have about the same number of cycles per tightening level. This is well illustrated on the figures.
Ideally, level 1 should tend to 1 due to the normalization (Eq. 8), which is the case for campaigns
C (Fig. 21), D (Fig. 22) and F (Fig. 24). For campaign B and E, three or more clusters coexist at
the start of the test (60 cNm) but the highest probability is still assigned to the correct cluster.

Another expectation is that the kinetics should be similar between levels and the slope of the
335 changes in the values π_{tk} should be quite steep because the tightening levels are modified abruptly
according to [61]. This is also well illustrated on the figures where the slopes are globally similar and
steep for each new onset detected. It is worth noting that gradual onsets are obtained for campaign
E characterized by lower γ values (estimated as [0, 3.8587, 2.9722, 1.2202, 2.4498, 3.4912, 1.9343]
seconds so an average of 0.719 ± 0.652 s at one standard deviation) than for the other campaigns
340 (with average of 1.298 ± 1.31 s for B, 2.77 ± 1.46 s for C, 2.07 ± 1.17 s for D and 2.28 ± 1.35 s
for F). As seen on Figure 21, the slopes for C are the steepest ones and the γ values are also the
largest as expected. The differences between values are sufficiently important to strongly modify
the slopes through the exponential in the sigmoid function (Eq. 9).

The onsets are not always sufficient to characterize the performance. Indeed, despite two
345 confusions illustrated for campaign E (levels 1/2 and 6/7), the clustering is close to be perfect with
an ARI around 0.975. It means that the mixture parameters (means and covariances of GMMSEQ)

also play a key role in the performance (as expected since these parameters pave the feature space).

Despite high ARI values compared to Tables 3-7 (in these tables, the best model was selected based on maximum likelihood, not based on the ARI), the location of the sigmoids is not perfect: For example, for campaign D, we can observe that the second level (50 cNm) is not detected properly with a confusion with level 1 (60 cNm). Level 2 is detected but its probability is around 0.2 against 0.8 for level 1. With these curves, the end-user is thus able to visually identify some confusions which is not possible with other methods. A similar behavior can be observed for campaign B. For this campaign, the best ARI is obtained with 11 clusters with several of them concentrated around levels 1 and 2. These confusions partly explain the result of 0.762 (resp. 0.722) for the ARI of campaign D in Table 5 (resp. campaign B in Table 3). It is worth noting that, in both cases, the main confusion arises between two consecutive levels (60 cNm and 50 cNm) which is more understandable than if the confusion were between very distinct levels.

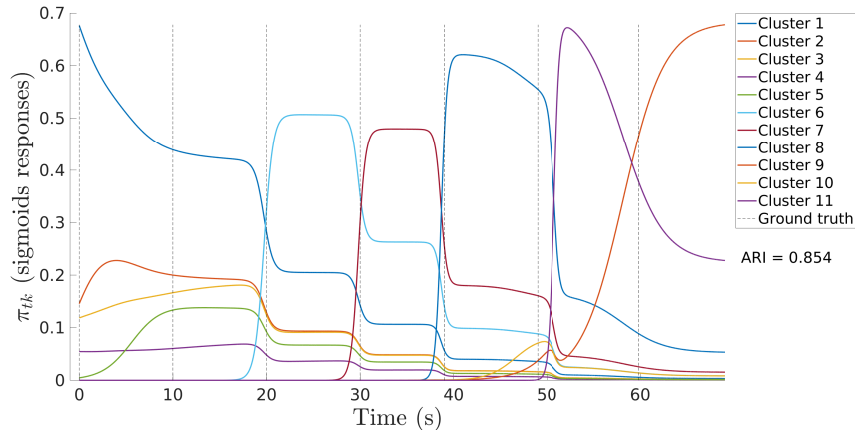


Figure 20: Campaign B: π_{tk} values provided by GMMSEQ.

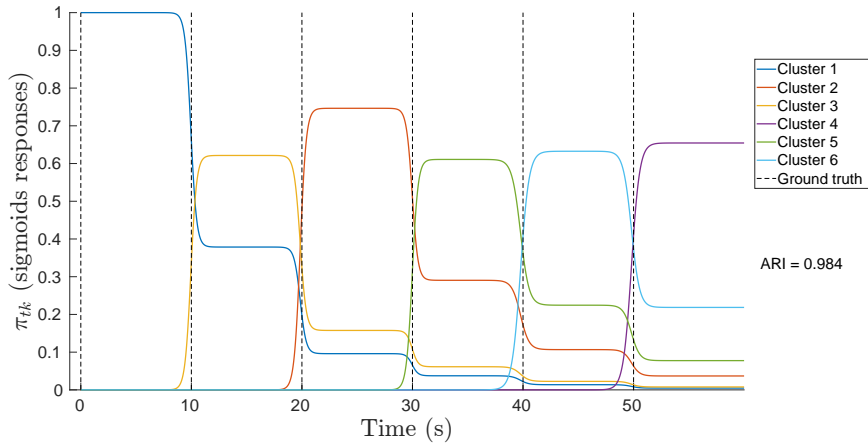


Figure 21: Campaign C: π_{tk} values provided by GMMSEQ.

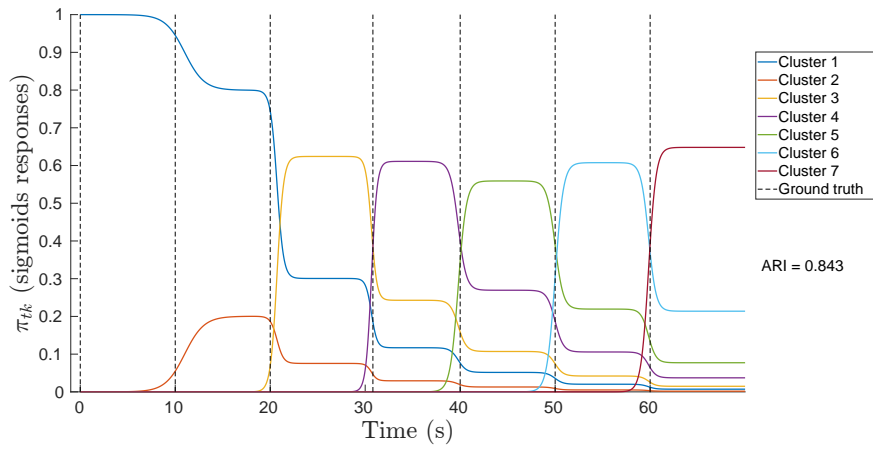


Figure 22: Campaign D: π_{tk} values provided by GMMSEQ.

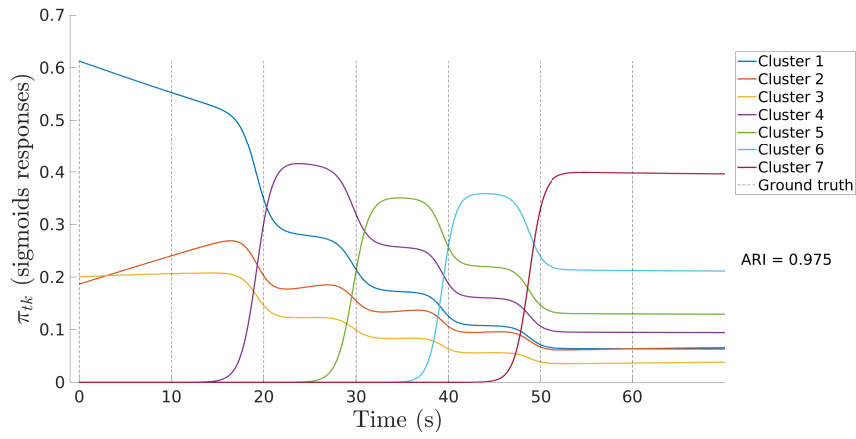


Figure 23: Campaign E: π_{tk} values provided by GMMSEQ.

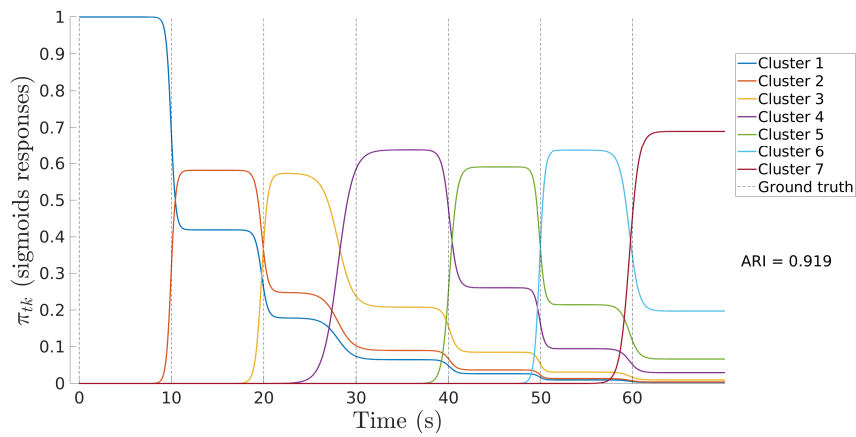


Figure 24: Campaign F: π_{tk} values provided by GMMSEQ.

4. Conclusion

360 GMMSEQ is a new clustering method introduced to manage continuous timestamps attached to AE signals. Timestamps are exploited during the clustering process allowing one to gain new insights into the AE data streams. In addition, cluster onsets, growth rates and levels of activation through time are estimated together with the parameters of their distributions in the feature space. Therefore, this new clustering method has unique characteristics that stand out from existing
365 methods for AE data interpretation.

GMMSEQ represents the first clustering method specifically developed for AE data, and it is the only method that allows AE users to estimate, directly from data, parameters related onsets, growth and kinetics. To our knowledge, there is no method with similar features in the literature.

The performance of the method has been demonstrated on simulated and real data sets. Close
370 to perfect results were obtained on simulated data with noise. On real data, we demonstrated the relevance of the clusters during loosening of bolted joints. The comparison with four standard clustering methods and according to different performance criteria shows that GMMSEQ not only provides useful qualitative indications about the timeline of clusters, but also has better performance in terms of cluster characterization.

375 This work opens up several perspectives. First, the approach can be extended to other mixture models considering non-Gaussian distributions. Our shared code includes, for example, an extension to a mixture of multivariate Student-t distributions, which is not described in this paper.

The possibility to include prior information about the cluster onsets was addressed in the paper using a regularization of the objective function. Prior information about some other parameters
380 such as the cluster centers could be exploited too, which could potentially improve the convergence of the algorithm to “better” estimates with a clear physical interpretation. The possibility of incorporating physical knowledge in the objective function could also be investigated.

Another direction to explore concerns the optimization procedure. In this paper, we have assumed that all data are available at once (offline or batch analysis). The extension to online
385 clustering with evolving parameters is a promising perspective for application, for example, to structural health monitoring or statistical process control.

Finally, while being developed with the goal to process AE data, this clustering method can be applied to other data sets containing data streams with gradually emerging clusters.

Acknowledgement

390 This work was partly carried out in the framework of the EIPHI Graduate school (contract ANR-17-EURE-0002) and the project RESEM-COALESCENCE funded by the Institut de Recherche Technologique Matériaux Métallurgie Procédés (IRT M2P) and Agence Nationale de la Recherche (ANR). The authors are also thankful to MIFHySTO and AMETISTE platforms. Finally, the authors thank reviewers for their effort to review the manuscript which helped us in
395 greatly improving its quality.

Codes and data

Codes to reproduce all results on clustering are available on Github at https://github.com/emmanuelramasso/MIXMOD_SEQUENTIAL. Data are also available in their raw format at

400 <http://dx.doi.org/10.7910/DVN/FBRDU0>. The features used as inputs for GMMSEQ can finally be downloaded at https://drive.google.com/drive/folders/1H413RxYu4ya7YMEgF_1Th_fHr7flvv00?usp=sharing.

Appendix A. Gradient of Q with respect to g_k , b_k and ξ_k

The gradients of the auxiliary function Q with respect to the variables of interest are given below. We first start by expressing the derivatives with respect to the instrumental variables, namely g_k, b_k, ξ_k for each component in the mixture:

$$\frac{\partial Q}{\partial g_k} = \frac{\partial Q}{\partial \gamma_k} \frac{\partial \gamma_k}{\partial g_k} = 2g_k \frac{\partial Q}{\partial \gamma_k}, \quad k = 2, \dots, K \quad (\text{A.1a})$$

$$\frac{\partial Q}{\partial b_k} = \frac{\partial Q}{\partial \beta_k} \frac{\partial \beta_k}{\partial b_k} = 2b_k \frac{\partial Q}{\partial \beta_k}, \quad k = 2, \dots, K \quad (\text{A.1b})$$

$$\frac{\partial Q}{\partial \xi_k} = \frac{\partial Q}{\partial \tau_k} \frac{\partial \tau_k}{\partial \xi_k} = \frac{\partial Q}{\partial \tau_k} \tau_k \left(1 - \frac{\tau_k}{T}\right), \quad k = 2, \dots, K, \quad (\text{A.1c})$$

where (A.1c) uses the following property of the logistic function $\Lambda(u) = 1/(1 - \exp(-u))$: $\Lambda'(u) = \Lambda(u)[1 - \Lambda(u)]$. The derivatives with respect to β_k, γ_k and τ_k are given by

$$\frac{\partial Q}{\partial \beta_k} = \sum_{l=1}^K \sum_{i=1}^N \frac{\partial Q}{\partial \pi_{il}} \frac{\partial \pi_{il}}{\partial \alpha_{ik}} \frac{\partial \alpha_{ik}}{\partial \beta_k}, \quad k = 2, \dots, K, \quad i = 1, \dots, N \quad (\text{A.2a})$$

$$\frac{\partial Q}{\partial \gamma_k} = \sum_{l=1}^K \sum_{i=1}^N \frac{\partial Q}{\partial \pi_{il}} \frac{\partial \pi_{il}}{\partial \alpha_{ik}} \frac{\partial \alpha_{ik}}{\partial \gamma_k}, \quad k = 2, \dots, K, \quad i = 1, \dots, N \quad (\text{A.2b})$$

$$\frac{\partial Q}{\partial \tau_k} = \sum_{l=1}^K \sum_{i=1}^N \frac{\partial Q}{\partial \pi_{il}} \frac{\partial \pi_{il}}{\partial \alpha_{ik}} \frac{\partial \alpha_{ik}}{\partial \tau_k}, \quad k = 2, \dots, K, \quad i = 1, \dots, N. \quad (\text{A.2c})$$

We have

$$\frac{\partial Q}{\partial \pi_{il}} = \frac{y_{il}^{(q)}}{\pi_{il}}, \quad l = 1, \dots, K \quad (\text{A.3})$$

and

$$\frac{\partial \pi_{il}}{\partial \alpha_{ik}} = \begin{cases} \frac{\sum_{q=1}^K \alpha_{iq} - \alpha_{il}}{\left(\sum_{q=1}^K \alpha_{iq}\right)^2} & \text{if } k = l \\ \frac{-\alpha_{il}}{\left(\sum_{q=1}^K \alpha_{iq}\right)^2} & \text{if } k \neq l \end{cases} \quad (\text{A.4})$$

for all k and l in $\{1, \dots, K\}$.

Using the equality $\alpha_{ik} = \beta_k \Lambda(\gamma_k(t_i - \tau_k))$, we have

$$\frac{\partial \alpha_{ik}}{\partial \beta_k} = \Lambda(\gamma_k(t_i - \tau_k)) = \frac{\alpha_{ik}}{\beta_k}, \quad (\text{A.5})$$

$$\frac{\partial \alpha_{ik}}{\partial \gamma_k} = \beta_k(t_i - \tau_k) \Lambda'(\gamma_k(t_i - \tau_k)) \quad (\text{A.6a})$$

$$= \alpha_{ik}(t_i - \tau_k) [1 - \Lambda(\gamma_k(t_i - \tau_k))] \quad (\text{A.6b})$$

$$= \alpha_{ik}(t_i - \tau_k) \left[1 - \frac{\alpha_{ik}}{\beta_k} \right], \quad (\text{A.6c})$$

and

$$\frac{\partial \alpha_{ik}}{\partial \tau_k} = -\gamma_k \beta_k \Lambda'(\gamma_k(t_i - \tau_k)) \quad (\text{A.7a})$$

$$= -\gamma_k \alpha_{ik} [1 - \Lambda(\gamma_k(t_i - \tau_k))] \quad (\text{A.7b})$$

$$= -\gamma_k \alpha_{ik} \left[1 - \frac{\alpha_{ik}}{\beta_k} \right]. \quad (\text{A.7c})$$

Remark 1. To impose the constraint $\gamma_1 = \dots = \gamma_K$, we simply sum the derivatives w.r.t. γ_k :

$$\frac{\partial Q_r}{\partial \gamma} = \sum_{k=1}^K \frac{\partial Q}{\partial \gamma_k},$$

where $\frac{\partial Q}{\partial \gamma_k}$ is computed using (A.2b).

When considering a prior on onsets, the derivatives w.r.t. τ_k (A.2c) become:

$$\frac{\partial Q_r}{\partial \tau_k} = \sum_{l=1}^K \sum_{i=1}^N \frac{\partial Q}{\partial \pi_{il}} \frac{\partial \pi_{il}}{\partial \alpha_{ik}} \frac{\partial \alpha_{ik}}{\partial \tau_k} - 2\lambda \left(\tau_k - \tau_k^{\text{prior}} \right), \quad k = 2, \dots, K, \quad i = 1, \dots, N. \quad (\text{A.8})$$

405

Algorithm 1 summarises the different steps of the gradient computation.

Appendix B. Summary of the hit detection and feature extraction steps

The AE signals were detected from the raw data stream using a method developed in [13]. This method is made of three steps: It starts by filtering the data stream, then applying a standard hit detection procedure, and finally feature extraction. The data were initially prefiltered using a high-pass filter of order 5 with a band-pass frequency set to 10 kHz and a band-pass ripple equal to 0.2 dB in order to remove the DC component of the data stream.

Step 1. The wavelet filtering step relies on wavelet denoising applied by frames of 250000 samples, practically shown in [13] to be a good compromise between computation time and quality of denoising with a better adaptation to noise and signals with varying properties. As advised in [13], the wavelet was set to a Daubechies “dB45” made of 90 coefficients, together with 14 levels of decomposition in order to detect the onset of AE signals. The soft Donoho-Johnstone universal threshold was applied on wavelet coefficients with a rescaling using a level-dependent estimation of level noise. A compensation of the group delay induced by the filtering was also performed.

415

Algorithm 1 Gradient computation.

Require: $\beta_k, \gamma_k, \tau_k, k = 2, \dots, K$
for $i = 1$ **to** N **do**
 for $k = 2$ **to** K **do**
 Compute $\frac{\partial \alpha_{ik}}{\partial \beta_k}$ using (A.5)
 Compute $\frac{\partial \alpha_{ik}}{\partial \gamma_k}$ using (A.6)
 Compute $\frac{\partial \alpha_{ik}}{\partial \tau_k}$ using (A.7)
 for $l = 1$ **to** K **do**
 Compute $\frac{\partial \pi_{il}}{\partial \alpha_{ik}}$ using (A.4)
 end for
 end for
end for
for $k = 2$ **to** K **do**
 Compute $\frac{\partial Q}{\partial \beta_k}$ using (A.2a)
 Compute $\frac{\partial Q}{\partial \gamma_k}$ using (A.2b)
 Compute $\frac{\partial Q_{ik}}{\partial \tau_k}$ using (A.2c) or (A.8)
 Compute $\frac{\partial Q}{\partial b_k}$ using (A.1b)
 Compute $\frac{\partial Q}{\partial g_k}$ using (A.1a)
 Compute $\frac{\partial Q}{\partial \xi_k}$ using (A.1c)
end for
Ensure: Gradient $\frac{\partial Q}{\partial \omega} = \left\{ \frac{\partial Q}{\partial b_k}, \frac{\partial Q}{\partial g_k}, \frac{\partial Q}{\partial \xi_k} \right\}_{k=2}^K$

Step 2. After filtering, a hit detection is applied. The goal is to find the start and end samples for each AE signal remaining after filtering. For that, samples are treated gradually. When one sample becomes above a threshold on amplitude (in our case 1.2 mV), samples are stored until one sample falls below the threshold. In this case, a counter is run. If the samples remain below the threshold during “HDT” microseconds (set to 1100 μs), then the end of the signal is found. If not, the counter is reinitialised to 0. When HDT is reached, a second counter, “HLT” (set to 80 μs), is run during which the detector is blind. It means that if the signal becomes above the threshold, it is not taken into account. This procedure is applied after filtering. However, the start and end are used to extract the AE signal in the raw data stream, not the filtered one (because the filtering alters the signals). This is one of the characteristics of the method.

Step 3. The features are finally extracted on each AE signal detected. Common features are described in [63, 40, 13, 16]: Rise time, counts, PAC-energy, duration, amplitude, average frequency, RMS, average signal level, counts to peak, reverberation frequency, initiation frequency, signal strength, absolute energy, partial power in the intervals [0, 20, 100, 200, 300, 400, 500, 600, 800, 1000] kHz, frequency centroid, peak frequency, weighted peak frequency. To this set of features were added the following ones: the Renyi number calculated from the scalogram [64] using a Morlet wavelet, as well as the frequency of the maximum of energy in the scalogram. For example features calculated in the frequency domain are well described in [63, 40]. Time-based features are described in [63, 13, 16].

440 *Illustration.* After AE hit detection, the start time of each AE signal is superimposed onto vibrometer data for measurement B for a few cycles in Figure B.25. We can observe that about 1-2 signals per cycle are found in this sample of about 1 s. This is confirmed in Table 1 that summarized the average over all campaigns and all files. We can also observe that the onsets are generally positioned on a similar displacement level (measured by the vibrometer) which means that the hit detection procedure finds signals mostly located at similar levels of displacement. This result demonstrates the reproducibility of the tests as well as the relevance of the hit detection.

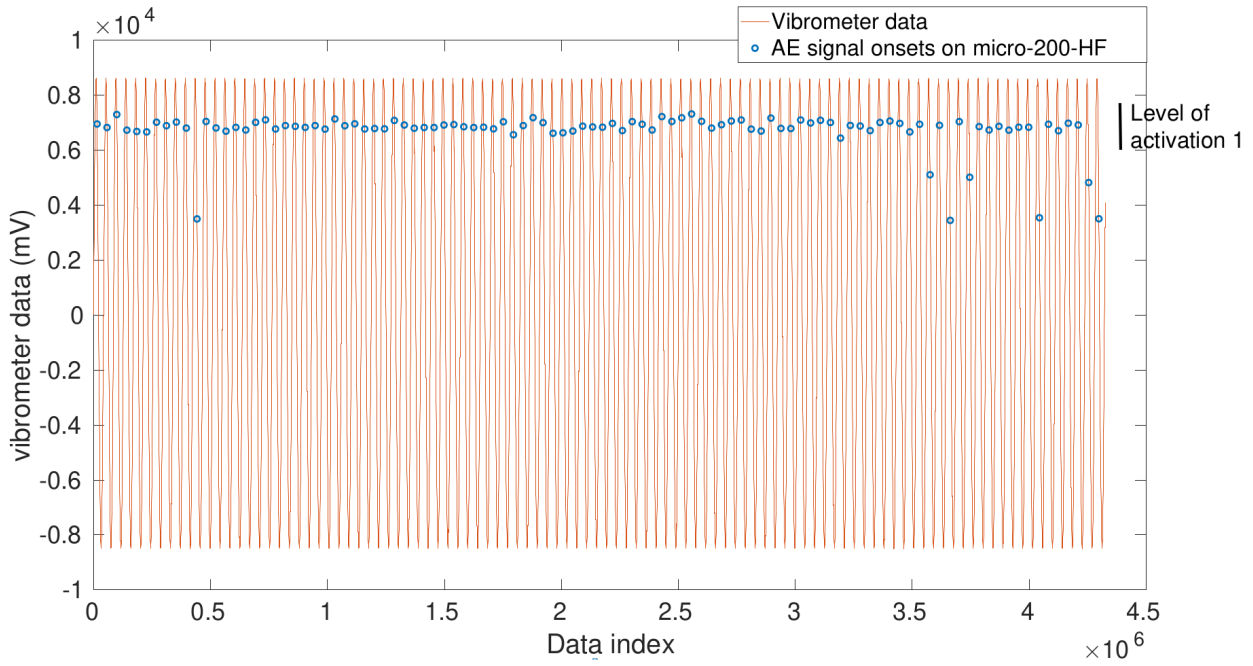


Figure B.25: Position of onset of acoustic emission signals onto vibrometer data (for measurements B, 60 cNm, first file of the data set, using micro-200-HF sensor). The horizontal axis represents the sample index.

445 **References**

- [1] Subcommittee E07.92, Standard terminology for nondestructive examinations, Standard, ASTM International, West Conshohocken, PA (2019).
- [2] Guidance notes on structural monitoring using acoustic emissions, Technical report, American Bureau of Shipping, City Plaza Drive, TX, USA (2016).
- 450 [3] C. Scruby, An introduction to acoustic emission, *Journal of Physics E: Scientific Instruments* 20 (8) (1987) 946.
- [4] C. Farrar, K. Worden, *Structural Health Monitoring: A Machine Learning Perspective*, John Wiley & Sons, Ltd, 2013.
- [5] J. Awerbuch, F. Leone, D. Ozevin, T.-M. Tan, On the applicability of acoustic emission to identify modes of damage in full-scale composite fuselage structures, *Journal of Composite Materials* 50 (4) (2016) 447–469.
- 455 [6] M. Y. Bhuiyan, J. Bao, B. Poddar, V. Giurgiutiu, Toward identifying crack-length-related resonances in acoustic emission waveforms for structural health monitoring applications, *Structural Health Monitoring* 17 (3) (2018) 577–585.
- [7] Y. He, M. Li, Z. Meng, S. Chen, S. Huang, Y. Hu, X. Zou, An overview of acoustic emission inspection and monitoring technology in the key components of renewable energy systems, *Mechanical Systems and Signal Processing* 148 (2021) 107146.
- 460 [8] O. AlShorman, F. Alkhatni, M. Masadeh, M. Irfan, A. Glowacz, F. Althobiani, J. Kozik, W. Glowacz, Sounds and acoustic emission-based early fault diagnosis of induction motor: A review study, *Advances in Mechanical Engineering* 13 (2) (2021) 1687814021996915.
- [9] J. H. Kurz, C. U. Grosse, H.-W. Reinhardt, Strategies for reliable automatic onset time picking of acoustic emissions and of ultrasound signals in concrete, *Ultrasonics* 43 (7) (2005) 538–546.
- 465 [10] E. Pomponi, A. Vinogradov, A. Danyuk, Wavelet based approach to signal activity detection and phase picking: Application to acoustic emission, *Signal Processing* 115 (2015) 110 – 119.
- [11] D. Bianchi, E. Mayrhofer, M. Gröschl, G. Betz, A. Vernes, Wavelet packet transform for detection of single events in acoustic emission signals, *Mechanical Systems and Signal Processing* 64 (2015) 441–451.
- 470 [12] T. W. Liao, Feature extraction and selection from acoustic emission signals with an application in grinding wheel condition monitoring, *Engineering Applications of Artificial Intelligence* 23 (1) (2010) 74 – 84.
- [13] M. Kharrat, E. Ramasso, V. Placet, M. Boubakar, A signal processing approach for enhanced acoustic emission data analysis in high activity systems: Application to organic matrix composites, *Mechanical Systems and Signal Processing* 70 (2016) 1038–1055.
- 475 [14] R. Madarshahian, P. Ziehl, J. M. Caicedo, Acoustic emission bayesian source location: onset time challenge, *Mechanical Systems and Signal Processing* 123 (2019) 483–495.
- [15] E. Ramasso, P. Butaud, T. Jeannin, F. Sarasini, V. Placet, N. Godin, J. Tirillò, X. Gabrion, Learning the representation of raw acoustic emission signals by direct generative modelling and its use in chronology-based clusters identification, *Engineering Applications of Artificial Intelligence* 90 (2020) 103478.
- 480 [16] S. Kattis, Noesis: Advanced data analysis, pattern recognition & neural networks software for acoustic emission applications, in: *Kolloquium Schallemission, Statusberichte zur Entwicklung und Anwendung der Schallemissionsanalyse*, Vol. 12, Fulda, 2017, pp. 1–8.
- [17] G. Manson, K. Worden, K. Holford, R. Pullin, Visualisation and dimension reduction of acoustic emission data for damage detection, *Journal of Intelligent Material Systems and Structures* 12 (8) (2001) 529–536.
- 485 [18] M. Li, J. H. Yang, Feature selection of acoustic emission signal for the slow-speed and heavy-load equipment, in: *Applied Mechanics and Materials*, Vol. 110, Trans Tech Publ, 2012, pp. 3199–3203.
- [19] D. Doan, E. Ramasso, V. Placet, S. Zhang, L. Boubakar, N. Zerhouni, An unsupervised pattern recognition approach for AE data originating from fatigue tests on polymer–composite materials, *Mechanical Systems and Signal Processing* 64 (2015) 465–478.
- 490 [20] M. G. Sause, *In Situ Monitoring of Fiber-Reinforced Composites: Theory, basic concepts, methods, and applications*, springer series in materials science Edition, Vol. 242, Springer Int. Publishing, 2016.
- [21] E. Ramasso, V. Placet, M. Boubakar, Unsupervised consensus clustering of acoustic emission time-series for robust damage sequence estimation in composites, *IEEE Trans. on Instr. and Meas.* 64 (12) (2015) 3297–3307.
- [22] S. Martin-Del-Campo, F. Sandin, Online feature learning for condition monitoring of rotating machinery, *Engineering Applications of Artificial Intelligence* 64 (2017) 187 – 196.
- 495 [23] R. Fuentes, R. Dwyer-Joyce, M. Marshall, J. Wheals, E. Cross, Detection of sub-surface damage in wind turbine bearings using acoustic emissions and probabilistic modelling, *Renewable Energy* 147 (2020) 776 – 797.
- [24] C. Wang, Z. Bao, P. Zhang, W. Ming, M. Chen, Tool wear evaluation under minimum quantity lubrication by clustering energy of acoustic emission burst signals, *Measurement* 138 (2019) 256–265.

- 500 [25] S. K. Chelliah, P. Parameswaran, S. Ramasamy, A. Vellayaraj, S. Subramanian, Optimization of acoustic emission parameters to discriminate failure modes in glass–epoxy composite laminates using pattern recognition, *Structural Health Monitoring* 18 (4) (2019) 1253–1267.
- [26] W. Zhou, W.-z. Zhao, Y.-n. Zhang, Z.-j. Ding, Cluster analysis of acoustic emission signals and deformation measurement for delaminated glass fiber epoxy composites, *Composite Structures* 195 (2018) 349–358.
- 505 [27] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: *Proc. of the Fifth Berkeley Symposium on Math., Stat. and Prob., Vol. 1, 1967*, pp. 281–296.
- [28] M. Chai, J. Zhang, Z. Zhang, Q. Duan, G. Cheng, Acoustic emission studies for characterization of fatigue crack growth in 316ln stainless steel and welds, *Applied Acoustics* 126 (2017) 101–113.
- [29] J. C. Dunn, A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters, *Journal of Cybernetics* 3 (1973) 32–57.
- 510 [30] S. N. Omkar, S. Suresh, T. R. Raghavendra, V. Mani, Acoustic emission signal classification using fuzzy c-means clustering, in: *Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02., Vol. 4, 2002*, pp. 1827–1831 vol.4.
- [31] D. E. Gustafson, W. C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, in: *1978 IEEE conference on decision and control including the 17th symposium on adaptive processes, IEEE, 1979*, pp. 761–766.
- 515 [32] G. J. McLachlan, K. E. Basford, *Mixture models. Inference and applications to clustering*, Vol. 84, Marcel Dekker, Statistics: Textbooks and Monographs, New York, United States, 1988.
- [33] H. A. Sawan, M. E. Walter, B. Marquette, Unsupervised learning for classification of acoustic emission events from tensile and bending experiments with open-hole carbon fiber composite samples, *Composites Science and Technology* 107 (2015) 89–97.
- 520 [34] M. Kaminski, F. Laurin, J. Maire, C. Rakotoarisoa, E. Hémon, Fatigue damage modeling of composite structures: the onera viewpoint, *AerospaceLab* 6 (9) (2015) 1–12.
- [35] A. Saxena, K. Goebel, C. C. Larrosa, V. Janapati, S. Roy, F.-K. Chang, Accelerated aging experiments for prognostics of damage growth in composite materials, *Tech. rep., NASA, Moffett Field CA Ames Research* (2011).
- 525 [36] T. W. Liao, Clustering of time series data - a survey, *Pattern recognition* 38 (11) (2005) 1857–1874.
- [37] T.-C. Fu, A review on time series data mining, *Engineering Applications of Artificial Intelligence* 24 (1) (2011) 164–181.
- [38] A. Belhadi, Y. Djenouri, K. Nørnvåg, H. Ramampiaro, F. Masegla, J. C.-W. Lin, Space–time series clustering: Algorithms, taxonomy, and case study on urban smart cities, *Engineering Applications of Artificial Intelligence* 95 (2020) 103857.
- 530 [39] C. S. Möller-Levet, F. Klawonn, K.-H. Cho, O. Wolkenhauer, Fuzzy clustering of short time-series and unevenly distributed sampling points, in: *International symposium on intelligent data analysis, Springer, 2003*, pp. 330–340.
- 535 [40] M. Sause, A. Gribov, A. Unwin, S. Horn, Pattern recognition approach to identify natural clusters of acoustic emission signals, *Pattern Reco. Lett.* 33 (2012) 17–23.
- [41] N. Godin, P. Reynaud, G. Fantozzi, *Acoustic emission and durability of composites materials*, ISTE-Wiley editions, 2018.
- [42] A. Monti, A. El Mahi, Z. Jendli, L. Guillaumat, Mechanical behaviour and damage mechanisms analysis of a flax-fibre reinforced composite by acoustic emission, *Composites Part A: Applied Science and Manufacturing* 90 (2016) 100–110.
- 540 [43] T. Shiraiwa, K. Ishikawa, M. Enoki, I. Shinozaki, S. Kanazawa, Acoustic emission analysis using bayesian model selection for damage characterization in ceramic matrix composites, *Journal of the European Ceramic Society* 40 (8) (2020) 2791 – 2800.
- 545 [44] J. Carmichael, R. Julius, Finding natural clusters, *Systematic Biology* 17 (2) (1968) 144–150.
- [45] V. Placet, E. Ramasso, L. Boubakar, N. Zerhouni, Online segmentation of acoustic emission data streams for detection of damages in composites structures in unconstrained environments, in: *11th Int. Conf. on Structural Safety & Reliability, 2013*, pp. 1–8.
- [46] A. Hohl, A. D Griffith, M. C. Eppes, E. Delmelle, Computationally enabled 4d visualizations facilitate the detection of rock fracture patterns from acoustic emissions, *Rock Mechanics and Rock Engineering* (2018) 1–14.
- 550 [47] I. Rastegaev, D. Merson, A. Danyuk, M. Afanasyev, A. Vinogradov, Using acoustic emission signal categorization for reconstruction of wear development timeline in tribosystems: Case studies and application examples, *Wear* (2018) 83–92.
- [48] L. Li, Y. Pu, J. Chen, Maximum likelihood parameter estimation for armax models based on stochastic gradient algorithm, in: *2018 10th International Conference on Modelling, Identification and Control (ICMIC), 2018*, pp.
- 555

1–6.

- [49] I. Rastegaev, D. Merson, I. Rastegaeva, A. Vinogradov, A time-frequency based approach for acoustic emission assessment of sliding wear, *Lubricants* 52 (7) (2020) 1–24.
- [50] N. Chandarana, Combining passive and active methods for damage mode diagnosis in tubular composites, Ph.D. thesis, Manchester University, Faculty of Science and Engineering, Departement of Materials (11/18/2019).
- 560 [51] B. Verdin, G. Chevallier, E. Ramasso, ORION-AE: Multisensor acoustic emission datasets reflecting supervised untightening of bolts in a jointed vibrating structure (Harvard Dataverse, 2021). doi:10.7910/DVN/FBRDU0.
- [52] E. P. Frigieri, P. H. Campos, A. P. Paiva, P. P. Balestrassi, Joao, A mel-frequency cepstral coefficient-based approach for surface roughness diagnosis in hard turning using acoustic signals and gaussian mixture models, *Applied Acoustics* 113 (2016) 230 – 237.
- 565 [53] P. R. Prem, A. R. Murthy, Acoustic emission monitoring of reinforced concrete beams subjected to four-point-bending, *Applied Acoustics* 117 (2017) 28 – 38.
- [54] R. V. Sagar, J. Srivastava, R. Singh, A probabilistic analysis of acoustic emission events and associated energy release during formation of shear and tensile cracks in cementitious materials under uniaxial compression, *Journal of Building Engineering* 20 (2018) 647 – 662.
- 570 [55] A. K. Das, D. Suthar, C. K. Leung, Machine learning based crack mode classification from unlabeled acoustic emission waveform features, *Cement and Concrete Research* 121 (2019) 42 – 57.
- [56] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society* 39 (1) (1977) 1–38.
- 575 [57] G. J. McLachlan, T. Krishnan, *The EM Algorithm and Extensions*, 2nd Edition, Wiley, New York, 2008.
- [58] S. Chatterjee, O. Romero, S. Pequito, Analysis of a generalised expectation–maximisation algorithm for gaussian mixture models: a control systems perspective, *International Journal of Control* 0 (0) (2021) 1–9.
- [59] C. Biernacki, G. Govaert, Choosing models in model-based clustering and discriminant analysis, *Journal of Statistical Computation and Simulation* 64 (1) (1999) 49–71.
- 580 [60] J.-P. Baudry, A. E. Raftery, G. Celeux, K. Lo, R. Gottardo, Combining mixture components for clustering, *Journal of computational and graphical statistics* 19 (2) (2010) 332–353.
- [61] E. Ramasso, B. Verdin, G. Chevallier, Monitoring a bolted vibrating structure using multiple acoustic emission sensors: A benchmark, *MDPI DATA* 7 (2022) 31–45.
- [62] Z. Zhang, Y. Xiao, Z. Su, Y. Pan, Continuous monitoring of tightening condition of single-lap bolted composite joints using intrinsic mode functions of acoustic emission signals: a proof-of-concept study, *Structural Health Monitoring* 18 (4) (2019) 1219–1234.
- 585 [63] M. of MISTRAS Holdings Group, *Pci-2 based ae system user’s manual rev. 3* (2007).
- [64] S. Gonzalez Andino, R. Grave de Peralta Menendez, G. Thut, L. Spinelli, O. Blanke, C. Michel, T. Landis, Measuring the complexity of time series: an application to neurophysiological signals, *Human brain mapping* 11 (1) (2000) 46–57.
- 590 [65] N. Vinh, J. Epps, J. Bailey, Information theoretic measures for clusterings comparison: Is a correction for chance necessary?, in: *Proc. of the 26th Annual Int. Conference on Machine Learning*, ACM, New York, NY, USA, 2009, pp. 1073–1080.