
Article:
Bull, L., Worden, K., Fuentes, R. et al. (3 more authors) (2019) Outlier ensembles: A robust method for damage detection and unsupervised feature extraction from high-dimensional data. Journal of Sound and Vibration. ISSN 0022-460X

https://doi.org/10.1016/j.jsv.2019.03.025

Outlier Ensembles: a Robust Method for Damage Detection and Unsupervised Feature Extraction from High-Dimensional Data

L.A. Bull*, K. Worden, R. Fuentes, G. Manson, E.J. Cross, N. Dervilis

March 28, 2019

* Corresponding author: lbull1@sheffield.ac.uk

Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK

Abstract

Outlier ensembles are shown to provide a robust method for damage detection and dimension reduction via a wholly unsupervised framework. Most interestingly, when utilised for feature extraction, the proposed heuristic defines features that enable near-equivalent classification performance (95.85%) when compared to the features found (in previous work) through supervised techniques (97.39%) — specifically, a genetic algorithm. This is significant for practical applications of structural health monitoring, where labelled data are rarely available during data mining. Ensemble analysis is applied to practical examples of problematic engineering data; two case studies are presented in this work. Case study I illustrates how outlier ensembles can be used to expose outliers hidden within a dataset. Case study II demonstrates how ensembles can be utilised as a tool for robust outlier analysis and feature extraction in a noisy, high-dimensional feature-space.

Key words: damage detection; dimension reduction; outlier analysis; unsupervised feature extraction; vibration monitoring.

1 Introduction

Novelty detection algorithms that utilise outlier analysis have been used extensively for damage detection in practical applications of structural health monitoring (SHM) [1–3]. The problem is to identify, from the measured data, if a machine or structure has deviated from the normal condition, that is, if the data are novel [1]. Parametric, statistical methods were first introduced to SHM through case studies in [1]. It has been shown that using statistical outlier analysis not only allows for the diagnosis of novelty, but also a method for *dimension reduction*, as models look to define a single damage sensitive feature (novelty index), without losing efficiency of the diagnostic [1].

In SHM, the measured data are often high-dimensional (e.g. vibration observations). As a result, even large volumes of data records can be sparse in their feature space. This phenomenon is referred to as the *curse of dimensionality* [4]; for sparse data in high-dimensions, the distance measures used to define outliers may no longer be meaningful [4, 5]. Specifically, it has been shown that for sparse data, the magnitude of the distances between any pair of observations can become similar [4–6]; thus, any observation can be considered a potential outlier.

To combat issues of dimensionality, feature selection tools look to identify a low-dimensional subset of variables from the measured data that are sensitive to damage [7]. These low-dimensional data can then be used to describe outliers. Conventional engineering methods for *unsupervised* feature selection are effective when the data are relatively clean and consistent [1–3, 7] (see §2.4). However, it becomes infeasible to select representative features (by an automated framework) when the data are noisy [8–10]; furthermore, conventional distance metrics are highly sensitive to measurement noise within the chosen features [11].

While sophisticated feature selection tools can be utilised in such noisy/complex feature spaces, many of these methods require at least some supervision (labelled data) to inform the heuristic [8, 12]. As a result, these algorithms are acceptable for *supervised* leaning; but they are counter-intuitive when building *unsupervised* models (e.g. novelty detectors), as labelled data are not available. In the context of SHM, comprehensive training data, including observations from the damaged structure, are rarely available. Clearly it is impractical/infeasible to gather data from engineering structures (such as bridges, aircraft or wind turbines) for all the

expected operating and damaged conditions *a priori*. Therefore, the practice of using labelled data to inform feature selection is a *critical* issue for practical applications of SHM, despite regular use in the literature. Considering these issues, practical systems should be adaptive and capable of running online, incorporating any new classes (novel data-groups) as they are discovered; thus, data outside the normal condition must not be used to inform feature extraction. Specifically, unsupervised techniques are required for emerging *semi-supervised* and *active learning* methodologies [13], where labelled data are initially unavailable (or limited).

Another critical issue for novelty detection in SHM concerns *inclusive* outliers. Inclusive outliers are outlying groups, generally due to novelty rather than noise, hidden within the available data [1, 2]. These data can significantly influence model parameters, leading to *masking* or *swamping* effects. Masking is caused by inclusive outliers that lead to increased model variance; these data can mask their own presence [2, 14], and thus, the detection of future anomalies (i.e. leading to false negatives). Alternatively, outliers can shift the model location (mean), leading to *swamping*, causing normal data to appear as outlying (false positives) [14]. Tools that utilise *robust statistics* [2, 15–18] look to account for, and expose, inclusive outliers; these methods are summarised in §2.3.

This work utilises outlier ensembles as a simple but effective technique for *damage detection* and *dimension reduction* with various problematic engineering data. A group (ensemble) of novelty detectors are trained using either:

(a) random subsets of observations (*bagging*),

(b) or random subsets of features (*feature bagging*).

The outputs of each model are combined to provide an improved measure for damage detection (and thus dimension reduction) in a *wholly unsupervised* framework. Two different engineering applications are presented as case studies; the datasets are chosen to represent more practical examples of SHM.

Case study I applies ensemble analysis to expose inclusive outliers, hidden within the available data. Experiments empirically demonstrate that outlier ensembles can provide a comparable measure of novelty when compared to alternative methods (FAST-MCD), while offering an

intuitive framework and reduced computational cost.

Case study II demonstrates a novel framework for wholly unsupervised dimension reduction, applied to noisy high-dimensional data. The proposed heuristic provides a robust framework for outlier analysis in high-dimensional feature space; furthermore, when utilised for dimension reduction, unsupervised outlier ensembles can provide features that are comparable to those found in a supervised setting (in previous work) through manual/genetic algorithm feature selection.

The rest of this paper is structured as follows: Section 2 gives a review of outlier analysis from an engineering perspective, including conventional techniques; Section 3 provides an overview of the related work, and the theoretical reasoning for outlier ensembles; Sections 4 and 5 provide two engineering case studies to empirically support claims; finally, Section 6 offers concluding remarks.

2 Outlier Analysis

In an engineering context, outliers are suitably defined for *novelty detection* as:

> *'Data that deviate so much from other observations, as to arouse suspicions that they were generated by some different mechanism'* [19].

Specifically, outlying data should indicate a significant change in the underlying physics of the structure being monitored — rather than benign fluctuations in measurement noise. Although this description is conceptually simple, detecting informative outliers from noisy engineering data is a non-trivial task.

2.1 Various approaches & terminology

The use of outlier analysis for novelty/damage detection will be referred to as *unsupervised* learning [7]. The general, unsupervised terminology used here is justified by the reasoning that data labels, $\mathbf{Y} = \{y_i\}_{i=1}^{N}$, are not directly used with the input data, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, to learn the model. Instead, a model is built using using the input data alone; although, $\mathbf{X}$ can be

assumed to only represent the normal condition data. The unsupervised nature of outlier analysis (or novelty detection) can be conceptualised by considering that there is no ground truth, $\mathbf{Y}$, available to directly evaluate the quality of a model during training, and inform the learning process.

Problems

There are two basic scenarios that describe where outlier analysis is useful in an engineering context, these are *exclusive* or *inclusive* problems [1]. For exclusive analysis, it is assumed that only information relating to the normal condition is provided in the available data, $\mathbf{X}$, during training. This situation is common for SHM systems that are designed to run online, as data describing potential damaged states are rarely available *a priori* [1, 8]. Inclusive methods consider outlying/novel groups that are hidden within $\mathbf{X}$; these outliers remain unlabelled. The inclusive problem can occur when a large pool of SHM data becomes available, recorded over a range of operational/damage conditions, without descriptive labelling, $\mathbf{Y}$ [2]. This work addresses both inclusive and exclusive analyses, proposing two ensemble heuristics.

Models

Various frameworks for novelty detection have been proposed, based on different notions of what an outlier is, and dependent on the application [14]. The available techniques can be roughly divided into two general groups; *parametric statistical* approaches and *nonparametric* approaches [14].

Parametric, statistical approaches assume that the measured data can be represented by some d-dimensional random vector, $\mathcal{X}$, where each feature can be considered a random variable, $\mathcal{X}^{(i)}$, such that,

$$\mathcal{X} \in \mathbb{R}^d \quad \therefore \quad \mathcal{X} = \{\mathcal{X}^{(1)}, ..., \mathcal{X}^{(d)}\}. \tag{1}$$

The random vector $\mathcal{X}$ is assumed to be defined by some specific probability distribution function (p.d.f.) f, such that $\mathcal{X} \sim f$. Using these assumptions, the parameters of f can be estimated from the available data, $\mathbf{X}$, and the discordancy of any observation can be

determined as a measure of novelty [1, 14, 15]. The theoretical detection threshold (or critical value) to indicate novelty can be pragmatically determined for these models, as the form of f is predefined; for details, see §2.2.

Parametric statistical models are limited by the assumption that $\mathbf{X}$ is sampled from some (predefined/assumed) underlying density distribution, f — usually multivariate Gaussian. Furthermore, the estimated parameters of f (e.g. mean, standard deviation, covariance) are sensitive to the presence of measurement noise, as well as inclusive outliers [14]. Robust, statistical techniques hope to combat this issue, see §2.3.

The alternative *nonparametric* approach does not assume a specific distribution function for the data; instead, certain characteristics of the underlying distribution, f, are estimated [14]. These characteristics can be quantified with the use of distance and density-based methods; both offer a basic way to estimate the density of f around data points; this can be interpreted as kernel density estimation [14, 20]. Distance-based methods, using k-nearest neighbour graphs [21], find *global* outliers as those (roughly speaking) furthest away from the rest of the data [14]. Density-based methods, such as Local Outlier Factor (LOF) [22], identify *local* outliers, defined as records located in regions of apparent low-density [14].

A problem with such nonparametric methods is the (typically) high runtime for the algorithms, as computation includes finding k nearest neighbours for each data point [14]. Another limitation is that the *smoothness* (and therefore accuracy) of the density estimation is highly sensitive to parameter selection (e.g., the number of neighbours to consider, k [20]); inaccurate representations of f can lead to erroneous assessments. The definition of a statistically relevant detection threshold can be more problematic for nonparametric representations of f, due to the complexity of the density estimates and hyperparameter sensitivity.

An engineering perspective

For engineering applications, particularly SHM, it is common practice to use parametric statistical approaches over nonparametric methods [1–3, 7]. This is justified by the reasoning that practical measured data, from a mechanical system or structure, should remain (relatively) consistent over the normal condition — synonymous with the consistent *underlying physical*

properties of the structure being monitored.

For example, in the context of dynamics-based monitoring, it is expected that damage will manifest itself as alterations in the fundamental structural parameters; specifically, a reduction in stiffness [1, 3, 7]. Changes in structural stiffness will alter the dynamic characteristics of the system; therefore, frequency domain observations are regularly used to (indirectly) monitor any physical changes that could relate to damage.

Various sources of noise can obscure the underlying vibrational characteristics. These effects can be incorporated as variance within the statistical model f (provided that masking does not occur). The distribution function f is usually assumed to be multivariate Gaussian *over the normal condition*, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$,

$$f = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{such that} \quad \mathcal{X} \sim f. \tag{2}$$

Where each i^{th} feature in $\mathbf{X}$ is generated by some Gaussian distributed random-variable $\mathcal{X}^{(i)}$. The true distribution functions for most measured variables in $\mathbf{X}$ are unlikely to be Gaussian; the function f might have heavier tails, or could be multi-modal — relating to inclusive outlying data or different permitted normal conditions. Despite this, an ideal, representative feature is often assumed to be at least *approximately* Gaussian over the *normal condition* data [1–3, 7].

2.2 Mahalanobis squared distance

With the assumption of approximate Gaussian statistics, the p.d.f f for the d-dimensional random vector $\mathcal{X}$ has been predefined, where $f(\mathbf{x}) = P(\mathcal{X} = \mathbf{x})$, $\mathbf{x} \in \mathbb{R}^d$,

$$f(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}} \frac{1}{\mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} . \tag{3}$$

The *sample* mean $\hat{\boldsymbol{\mu}}$ and covariance $\hat{\boldsymbol{\Sigma}}$ can then be determined from the available data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, to give a maximum likelihood estimate of the underlying p.d.f, denoted by $\hat{f}$.

Using the parameter estimates, a discordancy test can be used to quantify the degree of novelty

(the novelty index). This is used to evaluate whether an observation is likely to have come from an alternative underlying distribution [2], and thus, if it has been generated by some different mechanism. A classic discordancy measure is the Mahalanobis squared-distance (MSD) [1, 2],

$$D_i^2 = (\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}), \tag{4}$$

where $\mathbf{x}_i$ is the observation considered as a potential outlier, and $\mathbf{x}_i \in \mathbf{X}$. The MSD can be interpreted as a covariance-weighted squared-Euclidean-distance, that is, if the covariance is equal to the identity, they become synonymous [20].

The distribution of distance measures

The sum-of-squares for k independent, standard Gaussian random variables (Z) is Chi-squared distributed, with k degrees of freedom, such that [23],

$$\sum_{i=1}^{k} {Z^{(i)}}^2 \sim \chi_k^2 \ . \tag{5}$$

Recall the assumption that $\mathbf{X}$ is sampled from the multivariate-Gaussian random vector, $\mathcal{X}$, such that $\mathcal{X} = \{\mathcal{X}^{(1)}, ..., \mathcal{X}^{(d)}\}$. As the MSD effectively standardises each feature (or random variable) in $\mathbf{X}$, the metric can, in theory, be considered a *similar* sum of squares — where each $Z^{(i)}$ is a standardised version of $\mathcal{X}^{(i)}$,

$$\mathcal{D}^2 \approx \sum_{i=1}^{d} {Z^{(i)}}^2 \sim \chi_d^2 \ . \tag{6}$$

It is important to note, however, that equations (6) and (5) are based on the *asymptotic* distribution of robust distances, $\mathbf{D}^2$ [18], such that the covariance and mean parameters are consistent estimators [16]. In practice, the empirical parameters $(\hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\mu}})$ used to calculate $\mathbf{D}^2$ are estimated from a finite sample, thus, they are inherently inaccurate. Further issues arise if $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\mu}}$ are estimated from a *subset* of the available data, which are assumed to be inlying. (This applies to robust methods, explained in the next section.) In this case, the distribution of $\mathbf{D}^2$ (for all N data) is shown to be better approximated by an F-distribution, such that

$\mathcal{D} \sim F_{d,N-d}$; for further details, a thorough analysis of the distribution of robust distances can be found in [24].

To summarise, $\mathbf{D}^2$ is unlikely to be distributed according to an exact distribution that is easily defined. Despite theoretical limitations, the use of critical values from the χ^2_d distribution has been shown to provide a simple yet effective approach to outlier analysis with MSD-based methods in the literature [16–18]. As a result, the approximations in this work (regarding the discussion of the distribution of outliers) are considered justified; that is, the vector of MSD values, $\mathbf{D}^2$, should be *approximately* χ^2_d-distributed for the *inlying* data.

Threshold calculation

The critical value, or threshold, must be defined in order to classify data as normal or novel. Considering the issues in assuming an exact form for the distribution of outliers, a general Monte Carlo (MC) method can be used to define a threshold for a (finite) Gaussian-distributed dataset and sample size [1]:

1. Construct a [$N \times d$] (observations $\times$ dimensions) matrix, with each element being sampled from a zero-mean, and unit-variance Gaussian distribution.
2. Calculate the sample mean and covariance $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$, then find the MSD values for each of the N observations, to give the vector $\mathbf{D}^2$, according to equation (4).
3. Repeat steps 1 and 2 for a large number of trials, storing the largest value from each $\mathbf{D}^2$ into an array, then sort this array in order of magnitude. The critical value for a 1% test of discordancy is given by the the value in the array above which 1% of the trials occur.

In this work, the threshold represents a 99% confidence bound for a [$N \times d$] data sample from the (assumed) Gaussian-distributed normal condition, f. Specific details of the MC implementation used to define the threshold is provided in each case study, § 4,5.

2.3 Inclusive novelty

Masking (or swamping) effects occur because novel groups contained *within* $\mathbf{X}$ invalidate the assumption that all the data in $\mathbf{X}$ are sampled from a uni-modal, Gaussian-distributed random

vector $\mathcal{X}$. Instead, inclusive outliers are being generated by some alternative distribution, f^*; thus, they are defined by an alternative random vector, $\mathcal{X}^*$. As a result, inclusive outliers significantly affect the parameter estimates ($\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$). Robust methods mitigate the influence of inclusive outliers, generated by f^*, and determine a more accurate estimate for the *embedded* normal condition, f.

Robust statistical methods

Robust statistical methods were introduced into the field of engineering/SHM in [2]. Roughly speaking, these algorithms accurately estimate f by finding which *h-subset* of observations, $\mathbf{H}$, (from the available data) have been generated by the normal condition; where $\mathbf{H} \subset \mathbf{X}$, such that the size of the set (cardinality) is $|\mathbf{H}| = h$. The optimal h-subset $\mathbf{H}$ can then be used to determine *robust* estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In the existing literature [15–18], the ways to define $\mathbf{H}$ consider the *minimum volume enclosing ellipsoid* (MVEE) or *minimum covariance determinant* (MCD). These techniques are summarised briefly; algorithm details can be found in their respective references, as well as engineering application papers [2, 3].

The MVEE approach defines $\mathbf{H}$ by searching for the *smallest volume ellipsoid* that encapsulates h observations in the feature space [15, 16]. Alternatively, MCD methods define $\mathbf{H}$ as the subset whose covariance matrix has the *minimum determinant* [17, 18]. Both definitions can be interpreted as a way to describe the (majority) h-subset ($\mathbf{H} \subset \mathbf{X}$) that is the most *concentrated* in the feature space [16]. Intuitively, this group is assumed to be generated by the same underlying mechanism, f. This intuition means that all the data in $\mathbf{X}$ no longer need to be Gaussian distributed; instead, only a majority of the observations need to be. In other words, robust statistical approaches only require that variables in $\mathbf{X}$ are approximately Gaussian *in their centre*; that is, excluding the outlying values [25].

The exact MCD or MVEE is hard to compute, as it requires the evaluation of all $\binom{N}{h}$ subsets in $\mathbf{X}$ [16, 18]. Solving this combinatorial problem is infeasible when datasets are large [17, 26]. As a result, approximate algorithms are applied to practical data. Various techniques can be used to search for the optimal h-subset; some examples include resampling algorithms [15–17, 27] and genetic algorithms [17].

When comparing frameworks, the MCD estimator has a better statistical efficiency because the parameter estimates are asymptotically normal [28], while the MVEE has a lower convergence rate [17]. (This implies that the MCD location estimate, $\hat{\boldsymbol{\mu}}$, is normally distributed around its *true* value, $\boldsymbol{\mu}$, with the standard deviation shrinking as the sample size, h, grows.) Therefore, discordancy measures that are based on MCD estimates are more precise [17]. Despite these advantages, MVEE estimators were generally preferred due to improved computational efficiency [16]. However, since the FAST-MCD heuristic was introduced by Rousseeuw and Driessen [17], the MCD method is now commonly used, particularly for large datasets [16, 17]. The FAST-MCD algorithm is applied in this work as a benchmark (§4). Algorithm details can be found in Appendix A.

2.4 Outlier Analysis in High-Dimensional Feature Space

As discussed, the curse of dimensionality is a significant issue for novelty detection as outliers become hard to define. Specifically, for the MSD, if the number of observations in $\mathbf{X}$ is too small, such that $N < (d+1)$, the estimated covariance $\hat{\boldsymbol{\Sigma}}$ will always be singular. This occurs when the data are too sparse to accurately represent f, and as a result, abnormally large measures of discordancy are predicted [17].

During vibration monitoring, the observations are regularly recorded at a high sample rate, to enable high-resolution measurements in the frequency domain; consequently, vibration data (i.e. transmissiblities or frequency response functions) are often high-dimensional. Considering the curse of dimensionality, an impracticable number of observations are required to build a reliable statistical model; therefore, the high-dimensionality of measured data remains a major issue for vibration monitoring [2, 7, 8] — particularly systems that hope to run online [29]. Conventional engineering techniques, applied to compress high-dimensional data, are summarised below.

Conventional methods: feature selection & dimension reduction

Various frequency-domain features are sensitive to damage; these might include phase information, modal properties, or characteristic operational frequencies (condition monitoring)

[7]. Most typically, the resonance frequencies of a system can be recorded over time and used as damage sensitive features. A critical issue with practical data, however, is that high levels of noise can make the identification of such representative features infeasible [8]. More importantly, with noisy/inconsistent data, selecting specific variables from the available feature-set can lead to important information being lost, while the compressed data become increasingly sensitive to noisy/abnormal behaviour.

Dimension reduction techniques offer another method for data compression, while retaining as much information as possible from the full feature space. Linear principal component analysis (PCA) is typically used; alternatively, nonlinear variations include kernel-PCA [30, 31] and auto-encoder networks [32, 33]. While these methods are highly effective, the resulting features can be interpreted such that variation within the available data is maximised. This can become an issue when observations that represent the normal condition are inconsistent, as the corresponding data-groups become dispersed across the feature space. In turn, this can lead to effects that resemble masking. Variation across normal condition data is typical with engineering data, and can occur following maintenance procedures, or environmental influences [7]. A detailed example of this problem is illustrated in Case Study II, §5.

Utilising a subset of labelled data to inform feature selection/extraction provides an effective alternative; these techniques are *supervised* methods. Sensitivity analysis [7] of variables over the input data can help identify representative features objectively [8]. Alternatively, the use of Genetic Algorithms (GA) has been shown to provide promising results when applied to vibration data [12]; details of this application are summarised in §5. Despite their success, a significant issue with supervised methods for feature extraction is their dependence on labelled data. This renders their application irrelevant for many practical SHM systems, which look to run online, with limited data, and in an adaptive manner [13, 29]. Furthermore, these problems highlight the need for an alternative approach to feature selection/extraction with high-dimensional engineering data, when conventional *unsupervised* techniques prove unsuitable.

3 Outlier Ensembles

Ensemble analysis is regularly applied in the machine learning literature to reduce the dependence of model prediction on a specific realisation of the potential data [10, 14]. In general terms, an ensemble refers to a weighted combination of M *diverse* base predictors, $\hat{f}_m$ [31], defining an ensemble output $\hat{f}_E$,

$$\hat{f}_E(\mathbf{x}_i) = \sum_{m=1}^{M} w_m \hat{f}_m(\mathbf{x}_i), \tag{7}$$

where the base predictor, f_m, refers to a machine learning model; typically, a supervised classifier is used [10, 31]. Ensemble analysis greatly increases the robustness of pattern recognition models [10], as the combined predictions are more immune to benign variations in the data that relate to noise, rather than novelty; theoretical justification is provided in §3.1.

For *outlier ensembles*, the base predictor is an unsupervised novelty detector. In this work, each member is an MSD novelty detector, defined by $\hat{\boldsymbol{\mu}}_m$ and $\hat{\boldsymbol{\Sigma}}_m$. It is important to note, however, that the ensemble framework in equation (7) is flexible; therefore, any model (that is appropriate for outlier analysis) can be used as the base predictor.

Introducing diversity

Successful ensemble analysis requires a diverse set of predictors/models [20, 34]; roughly speaking, there are two main approaches to introduce variability [10, 34]. Firstly, the base predictor can be varied across members in the ensemble (i.e. changing hyperparameters, or the algorithm itself). An issue with combining various models, however, is that the outputs can be incomparable, leading to calibration/normalisation issues when combining predictions [10, 34, 35]. Alternatively, *for the same model*, variability can be introduced through bootstrap samples of data from $\mathbf{X}$, i.e., sampling with replacement. Issues relating to calibration are greatly reduced when diversity is introduced through sampling, as the underlying base predictor remains constant. With these factors in mind, this work focusses on bootstrap methods, where samples of $\mathbf{X}$ are taken as either: (a) subsets of observations or (b) subsets of variables/features.

Considering the inclusive problem (presented in Case Study I, §4), the primary aim is to reduce the effect of abnormal *observations* within the data on the model estimate, $\hat{f}$. Ensembles learnt using bootstrap-sampled observations offer a solution; in this way, the set of predictors capture variability across observations, in an attempt to expose *inclusive novelty* and provide a *robust* estimate of f. Combining model predictions can reduce the influence of outlying groups, that might otherwise skew the estimated parameters of f.

The high-dimensional scenario is an important application of ensemble analysis for engineering data (presented in Case Study II, §5). The useful behaviour of measurements in high dimensional space is often described by a subset of dimensions, which are difficult to discover in practical settings [8, 10, 12]. The use of bootstrap-sampled features (feature bagging), introduced by Lazarevic and Kumar [36], has been shown to provide a novel, successful framework for outlier analysis in high-dimensional feature spaces [9, 35]. The resulting ensemble can provide a robust measure of novelty, as the combined outputs reduce the effect of any noisy/misrepresentative features. Thus, feature bagging can provide a more general, robust approach to feature selection, reducing the uncertainty associated with this inherently difficult process [10].

3.1 Theoretical justification

The general argument for ensemble frameworks is that all members are inaccurate and produce errors, but on different cases; if these errors are uncommon or independent, they should have a reduced effect on the combined output [14]. Bootstrap sampling is a way to realise this theory, as each subset can be viewed as a different sample drawn from the the underlying distribution f, as opposed to taking $\mathbf{X}$ as a single realisation [14].

The work by Zimek et. al. [14] does well to formalise this argument; if the majority of data in $\mathbf{X}$ (i.e. the normal/representative groups) were generated by an unknown distribution f, this majority can be viewed as a sample drawn from the true, but unknown density function f. Recall that parametric statistical novelty detectors estimate f from the available data, to give an empirical approximation, $\hat{f}$. For each member in the ensemble, the estimate $\hat{f}_m(\mathbf{x})$ can be

expressed as [14],

$$\hat{f}_m(\mathbf{x}) = f(\mathbf{x}) + \epsilon_m(\mathbf{x}), \tag{8}$$

where ϵ_m denotes the error of the estimate $\hat{f}_m$ at $\mathbf{x}$ [14]. Note, the quality of $\hat{f}_m$ determines the overall success of each member, and this depends on the error, ϵ_m. If it is possible to obtain multiple density estimates of f, learnt from different subsamples of data, a more reliable estimate for the underlying function might be available by averaging their *diverse* results [20]. Following a model-averaging approach, the multiple density estimates, $\hat{f}_m$, and associated errors, ϵ_m, can be considered as random variables, each with expected values [14],

$$E\big[\hat{f}_m(\mathbf{x})\big] = E\big[f(\mathbf{x})\big] + E\big[\epsilon_m(\mathbf{x})\big] \tag{9}$$

$$= f(\mathbf{x}) + E\big[\epsilon_m(\mathbf{x})\big]. \tag{10}$$

From this formulation, the benefits of ensemble analysis can be inferred; drawing multiple samples from $\mathbf{X}$ and averaging can reduce the influence of any randomness or unusual behaviour. Any predictions based on samples of normal/representative data, should contribute random errors that are relatively independent on $\mathbf{x}$, such that $E\big[\epsilon(\mathbf{x})\big] = E\big[\epsilon\big] = e$. Ideally, these errors have zero mean and are uncorrelated, such that they cancel out and $e \approx 0$. In reality, the associated error ϵ is often dependent on $\mathbf{x}$, skewing the average; particularly if misrepresentative groups within the data relate to (correlated) novelty, as opposed to random measurement noise. Despite this, the influence of any abnormal data should be reduced in the combined output, as they represent a minority, by definition.

In the scenario that an ensemble fails to capture any meaningful (general) behaviour, it is likely that the choice of base predictor, f_m, is inappropriate to represent the available data. If this is the case, a more suitable (or more flexible) model must be used to build the ensemble. As a result, when applying ensemble analysis with MSD novelty detectors (as in this work) it is being implied/assumed that there exists some consistent behaviour within the *inlying* data that can be modelled by a Gaussian-distribution.

Assessing model quality

While ensemble analysis has been applied extensively in a supervised setting (classification) [37], it is relatively unexplored for unsupervised techniques, particularly outlier analysis [14]. As suggested by Aggarwal [35], the most likely cause of this is a lack of *ground truth*, i.e. target labels of discordancy $\mathbf{Y}$. Therefore, it is not possible to externally evaluate the predictive performance of each novelty detector, due to the unsupervised nature of outlier analysis. This absence of information makes it hard to quantify the correctness or quality of a model in a statically robust way; consequently, various methods for learning model weights w_m (equation (7)) that are common to supervised ensembles (boosting, pruning, weighting), become hard to implement [34].

Several ideas to provide an *internal* (unsupervised) estimate of model quality have been suggested; these methods are based on an assumed ground truth [14]. Some examples include generating artificial outliers to approximately assess the performance of each member [38], the use of ROC curves [34], or density estimation methods [5, 35]. It is worth noting that the robust techniques, discussed in §2.3, follow a similar framework of internal evaluation. In this case, the quality of a model is assessed using the minimum volume enclosing ellipsoid (MVEE) or minimum covariance determinant (MCD); therefore, these approaches also use assumptions about an unknown ground truth, to approximately evaluate each potential model. Considering this, the MCD is suggested as measure of quality for ensemble analysis when exposing inclusive outliers, proposed in Case Study I, §4.

3.2 Model combination

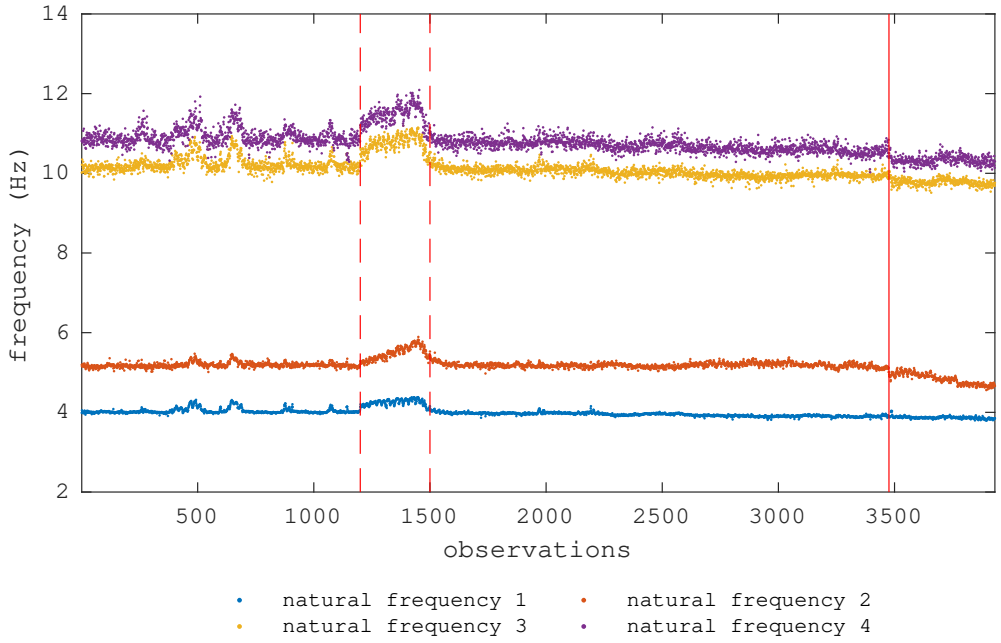
Now that a diverse ensemble can be built, a function for combining novelty detectors must be defined. There are various ways to do this; the best approach depends on the definition of outliers and the application. Any meaningful combination requires that the outputs are normalised [34, 35]. Thankfully, when using the Mahalanobis distance, output scores are effectively normalised by the covariance matrix, so the direct combination of outputs should not be problematic.

Suggested methods include: using the maximum output [34], output averaging [36], weighted averages [34] and various combinations [35]. Using the maximum output has a tendency to severely overestimate discordancy [34]; furthermore, it is counter-intuitive when considering the benefits associated with taking expectations. As a result, this work utilises averaging combination functions — defined by equation (7). Application specific variants of equation (7) are provided in each case study.

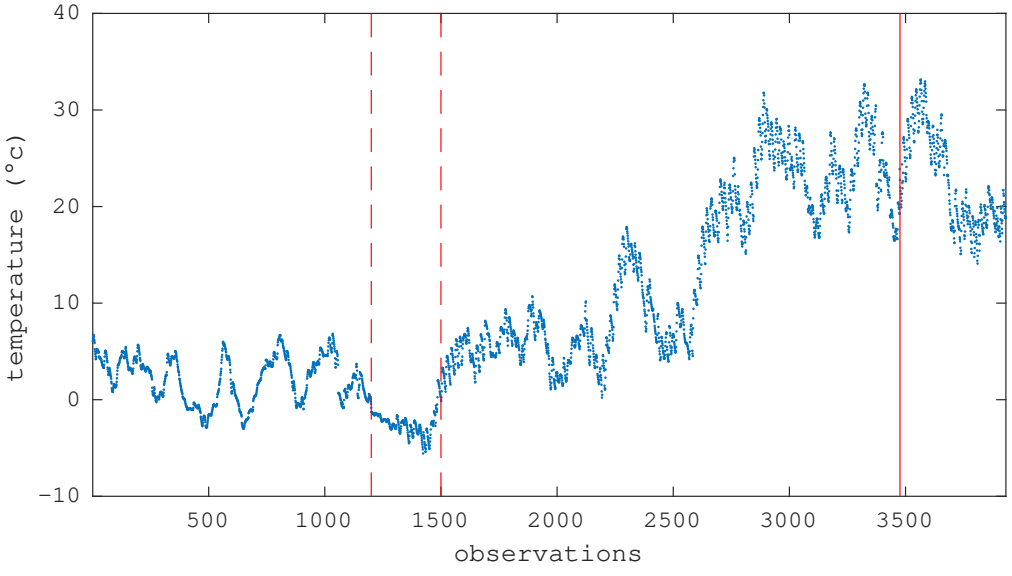
4 Case Study I: Inclusive Novelty — Z24 Bridge Data

The Z24 bridge was a concrete highway bridge in Switzerland, connecting Koppigen and Utzenstorf. In the late 1990s, before its demolition procedure, it was used for experimental SHM purposes under the SIMCES project [39]. Over a twelve month time period, a series of sensor systems were used to capture dynamic response measurements, in order to extract the first four natural frequencies of the structure. Environmental measurements were also recorded, including air temperature, soil temperature, humidity and wind speed [40]. This is a relatively large dataset, with 3932 observations in total. During the benchmark project, different types of real damage were artificially introduced towards the end of the monitoring year, starting from observation 3476 [2]. The natural frequencies, as well as soil temperature, are shown in Figure 1. These data are considered a benchmark dataset for SHM, and they are applied in the experiments for comparison to existing work in robust outlier analysis [2].

Figure 1 illustrates visible fluctuations between observations 1200 and 1500, while there is little variation following the introduction of damage at observation 3476. The visible fluctuations relate to periods of very low temperature in the bridge deck, which can be observed in the temperature plot, Figure 1b. It is believed that the asphalt layer in the deck experienced these very low temperatures during this time, leading to increased stiffness [17]. Therefore, as with damage, this significant change in the *underlying physics* of the structure should produce data that are *novel* (by definition).



(a)



(b)

Figure 1: (a) time history of natural frequencies, (b) time history of average deck temperature.

4.1 Building the ensemble

With these data, the outliers are contained within the dataset, and they are undefined; as a result, this is an *inclusive* problem. Keeping in line with the theory discussed in this work, it is assumed that there is some general behaviour within the data that can be estimated by a Gaussian distribution, i.e., the inlying data. For the proposed ensemble, each member is an MSD novelty detector with empirical parameters $\hat{\boldsymbol{\mu}}_m$ and $\hat{\boldsymbol{\Sigma}}_m$.

Considering the inclusive problem, the detection quality improves when reducing the negative effects of outlying *observations* (hidden within the data) on the parameter estimates; therefore, variation is introduced through bootstrap-sampled observations. In order to maximise the chance of a sample containing only normal-condition (Gaussian-distributed) inlying data, the subsample size, n_s, should be as small as possible [15, 17]. This is formalised, such that the probability of drawing a 'pure' sample is,

$$P_s = (1 - \alpha)^{n_s}, \tag{11}$$

where α is the assumed fraction of outliers within $\mathbf{X}$. Minimising n_s increases the probability of building an ensemble in which the majority of members are built on the normal inlying data. However, as discussed, if the number of observations in each sub-sample is too small, the estimated covariance, $\hat{\boldsymbol{\Sigma}}_m$, risks becoming singular, leading to abnormally large measures of discordancy. In the proposed heuristic, it is suggested to set $n_s = 3d$, for good generalisation across various engineering data. This value should prevent a singular (or near-singular) empirical covariance. The number of members in the ensemble, M, is set such that $(n_s \times M) = N$; as a result, the effective/sampled dataset is the same size as the original dataset, $\mathbf{X}$.

The authors acknowledge that the sample size, n_s, is data-dependent; therefore, some parameter tuning may be required to ensure a non-singular covariance while avoiding larger subsamples that risk high levels of outlier contamination. With the Z24 data, it was found that the range $3d < n_s < 5d$ provided reasonable (robust) outputs. These aspects of the heuristic are emphasised, as they highlight the need for parameter tuning with MSD outlier ensembles. While the heuristic is still unsupervised, it is necessary to define hyperparameters with some prior knowledge (and information in the available *measured data*) to ensure the results are

sensible. Importantly, these steps are carried out within the unsupervised framework.

Model combination by parameters

The quoted complexity of prediction for an MSD novelty detector is $\mathcal{O}(N \times d)$ [17], as the distance measures must be computed N times. An issue with many ensemble frameworks is that the base predictor is run M times for all N data, leading to increased complexity, $\mathcal{O}(N \times d \times M)$. For the inclusive problem, however, it is suggested that models are combined by averaging over the *parameter* estimates, rather than predictions of discordancy. Bootstrap sampling for parameter estimation is a common approach in the machine learning literature, and it is often used as a Monte Carlo technique to predict sampling distribution of the parameter estimates [31].

In the context of *inclusive* outlier analysis, bootstrap sampling the parameter estimates appears logical, as it is unnecessary to calculate N distance measures for M members when only the average of the outputs is ultimately used. This applies to the inclusive problem, as only the parameter estimates themselves need to be robust; in fact, the sensitivity of the output (discordancy) is desirable to some extent. (Note, this is not the case for Case Study II). Additionally, unlike supervised ensembles, the outputs of each model are not required for external evaluation of each member, as there is no ground truth available. Instead, the quality of each model can be estimated using properties of $\hat{\theta}$, defined below; specifically, the covariance determinant (see §4.1).

A weighted average can be applied to the committee of predictors, with M members,

$$\hat{\theta}_E = \frac{1}{M} \sum_{m=1}^{M} w_m \hat{\theta}_m, \tag{12}$$

where $\hat{\theta}_m$ denotes the parameter estimates, $\hat{\theta}_m = (\hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m)$, from the m^{th} member in the ensemble, and w_m is the associated weight. By applying equation (12), N output measures of discordancy do not need to be calculated M times; instead, the distances are only calculated once, after parameter averaging. This reduces the predictive complexity of the ensemble to a similar order of the base predictor, $\mathcal{O}([N + M] \times d)$.

Model combination — unweighted and weighted averaging

In the following experiments, the model weights w_m (equation (12)) are either:

1. set to unity for all members (unweighted average, simple mean); or,
2. defined according to an approximate measure of model quality (weighted average).

When applying method 2, samples with a smaller covariance determinant $\det(\hat{\boldsymbol{\Sigma}}_m)$ are assumed to represent the normal/inlying data, in agreement with the MCD estimator. Therefore, $\det(\hat{\boldsymbol{\Sigma}}_m)$ is used as the internal measure of model quality. In the proposed scheme, it is suggested that the weight vector, $\mathbf{w}$, is defined such that models with a large covariance determinant are *pruned* out of the ensemble.

The fraction of models to keep is set at P_s, because it is assumed that $P_s \times M$ of the sub-samples are likely to contain only 'pure' inlying data (see equation (11)). Therefore, $P_s \times M$ of the lowest determinant models are given a unit weighting, and contribute to the final output. In words, when following method 2, the weight vector, $\mathbf{w} = \{w_1, ..., w_M\}$, is defined in the following way: determine the empirical parameter estimates $(\hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m)$ for all M models in the ensemble, calculate and sort covariance determinants $\det(\hat{\boldsymbol{\Sigma}}_m)$, then sort them in ascending order; take the $P_s \times M$ of smallest covariance determinants and set their respective model weights in $\mathbf{w}$ to unity. The pseudocode for the ensemble heuristic applied to the inclusive problem is provided in Appendix B, Algorithm 1.

Ensemble thresholds

The threshold for the ensemble novelty index is defined by averaging, similar to output combination. The critical value, $\mathcal{C}_m$, is established for each member (according to §2.2) for the subsample size, $[\ n_s \times d\ ]$. The ensemble threshold, $\mathcal{C}_E$, is then set as the average of the thresholds found for each member,

$$\mathcal{C}_E = \frac{1}{M} \sum_{m=1}^{M} \mathcal{C}_m. \tag{13}$$

As each subsample contains the same number of observations, n_s, the thresholds will be approximately equivalent, $\mathcal{C}_1 \approx \mathcal{C}_2 \approx ... \approx \mathcal{C}_M$; thus, $\mathcal{C}_E \approx \mathcal{C}_m$. Therefore, the threshold only needs to be defined once, provided that there are a large number of trials in the Monte Carlo sampling regime §2.2 (10,000 trials are run in the experiments). This threshold is appropriate as the behaviour of the *combined* ensemble outputs is dependant on the sample size used to build each member, thus, a *robust* ensemble threshold should also change according to n_s.

4.2 Results & discussion

The ensemble heuristic (parameter averaging, Algorithm 1) is applied to the natural frequencies of the Z24 data, such that $\mathbf{X} \in \mathbb{R}^4$ and $N = 3932$. The novel framework is applied without pruning (method 1) and with pruning (method 2). The performance is compared to the standard Mahalanobis squared-distance (MSD) and the robust FAST-MCD (MCD). For the MCD estimator, the parameter h is set to its minimum value ($0.5 \times N$), in order to provide a sensitive measure of discordancy [17].

The results are provided in Figure 2. Each test was run 1000 times; the plots represent the discordancy values for one trial, drawn at random (there was little variation between trials). As expected, the standard Mahalanobis distance suffers seriously from *masking* effects, with very few data being flagged as outliers. In this case, the inclusive outlying data have significantly skewed the parameter estimates.

The robust MCD algorithm successfully eliminates the issue of masking, and it is more sensitive to inclusive outliers, particularly those relating to very cold temperatures. The discordancy could be *slightly* more sensitive for these data, as a significant proportion of the observations for early damage (observation 3470 – 3660) appear below the detection threshold (false negatives). Discordancy measures (consistently) pass the detection threshold shortly after the introduction of damage, around 3670 observations.

Outlier ensembles (method 1) provide similarly robust results; however, more of the data relating to novelty pass the detection threshold. While this means that outliers are flagged at a higher rate from the onset of early damage (3470 – 3660), there is a clear increase in the number of 'false positives', particularly for early observations. However, when observing

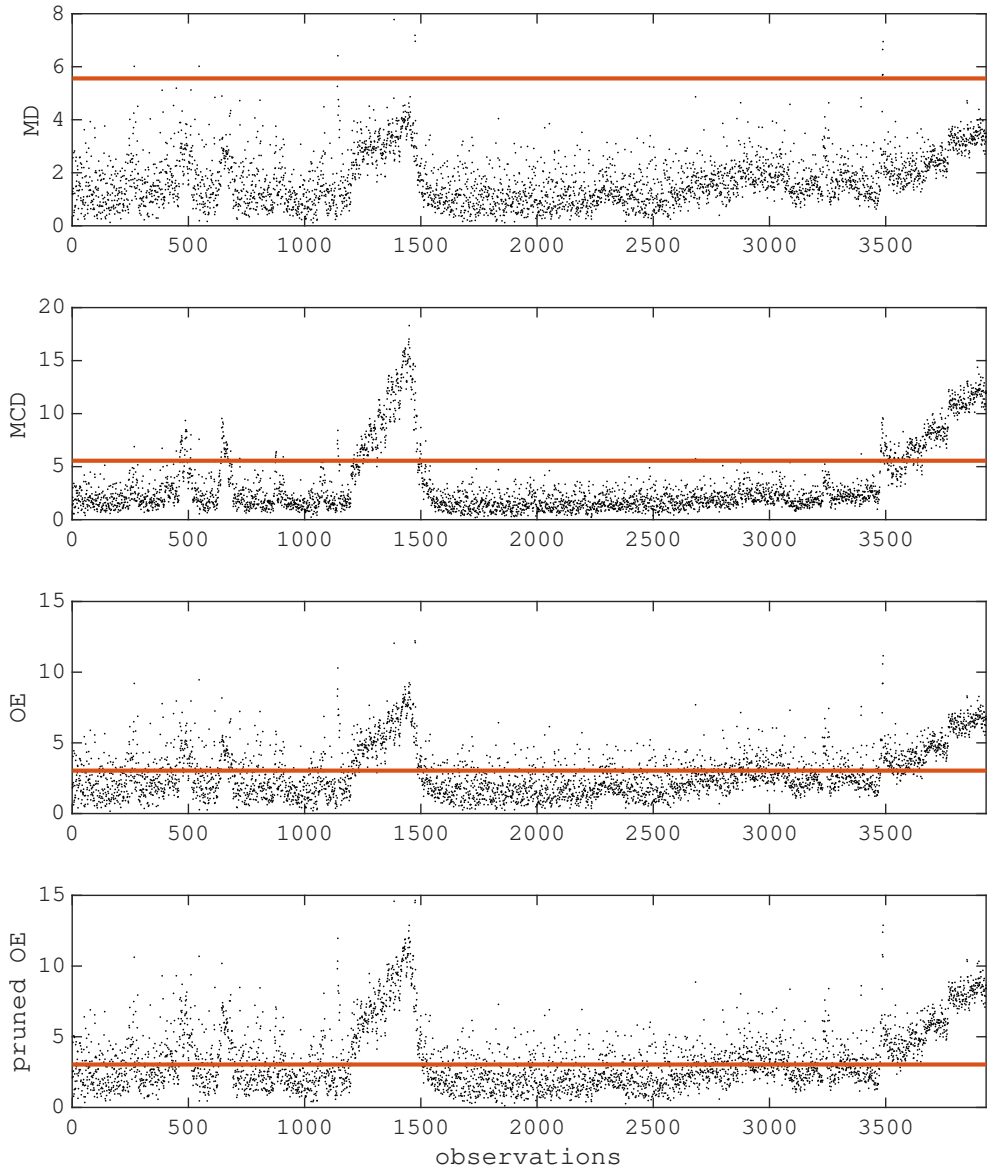


Figure 2: Novelty index for the standard Mahalanobis distance (MD), the minimum covariance determinant (MCD), outlier ensembles (OE) and pruned outlier ensembles (pruned OE).

Figure 1, it can been seen that some of these observations (500 – 750) correspond to cold temperature effects, so might be considered as outliers themselves. Considering the increase in the number of normal data flagged as outlying, a key benefit of this framework is that fewer application-specific parameters need to be defined — provided the subsample is large enough to ensure a non-singular covariance. The analysis naturally accommodates for various levels of contamination within $\mathbf{X}$, due to the way the ensemble is built and combined; in other words, the influence of the normal data should always be greater, as they represent a majority. While this capability suggests a more general/automated algorithm, the performance of the ensemble approach is more sensitive to the *proportion* of outliers within $\mathbf{X}$, as these will always influence the estimation of f. Therefore, the FAST-MCD algorithm is more likely to converge to the true underling distribution f for higher proportions of outliers, following parameter tuning of h; although, this approach is somewhat less automated.

When pruning methods are applied (method 2), the novelty index becomes more sensitive. With this ensemble, outliers are consistently flagged from the point that damage is introduced (3476), however, this comes with a further increase increase in the number of 'false' positives; as a result, it is the authors' opinion that this model flags too many data as outlying with these data. Additionally, an issue with this method is defining the contamination parameter, α, which was set to 0.1. This effectively 'tunes' the sensitivity of the ensemble, and introduces a more application specific parameter; however, if there is some prior knowledge of the contamination ratio, α, pruning offers a way to introduce this information.

Outlier ensembles provide a reduction in computational cost. When running the FAST-MCD with the LIBRA package [41], the average run-time over 1000 trials was 0.355s. Ensemble methods took 0.021s and 0.024s for unweighted and weighted methods respectively. This suggests that outlier ensembles are up to 15 times faster than the FAST-MCD algorithm.

5 Case Study II: Dimension Reduction — Gnat Aircraft Data

The Gnat data are an experimental SHM dataset, recorded using a network of sensors placed on the wing of a Gnat aircraft; schematics are provided in Figure 3. During experiments, the wing was excited using an electrodynamic shaker and white Gaussian excitation. Transmissibilities

associated with nine selected inspection panels (T1 – T9) were used as the main measurements — see [8, 42, 43] for further details. The sensor layout and transmissiblities are shown in Figure 3b. The transmissibility associated with each response transducer is obtained by taking the ratio of the acceleration response spectrum with the reference acceleration spectrum. In all cases, 1024 spectral lines were recorded, between 1024 and 2048Hz [12]. The logarithm of the transmissiblity magnitudes are used as the input measurements.

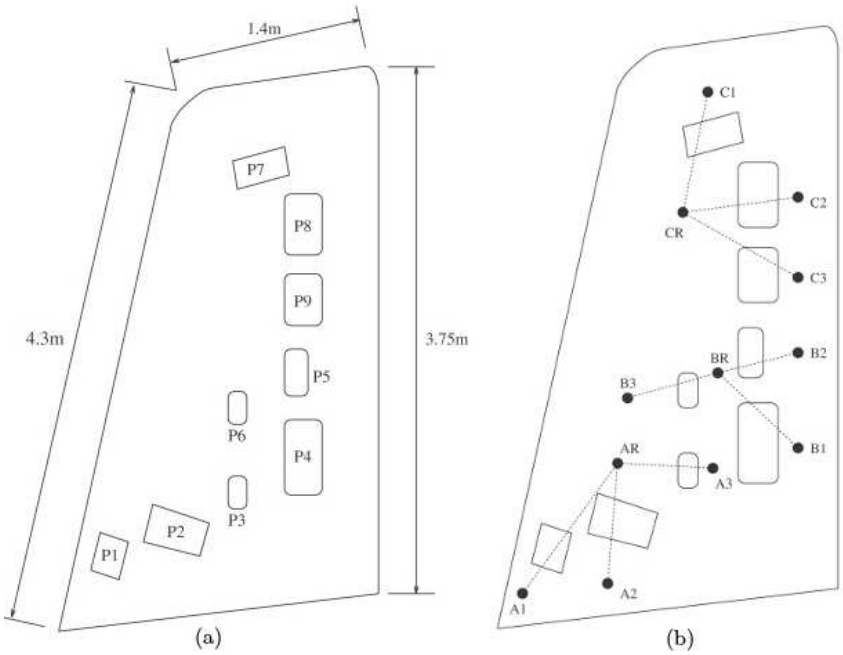


Figure 3: Schematics of the Gnat aircraft wing; (a) panel locations. (b) sensor groups and transmissibilities. Image Credit: [8]

During the experiments, artificial damage and maintenance procedures were simulated by sequentially removing or replacing each of the nine inspections panels. It should be considered that the removal of each panel imitates a fairly large, significant fault. Each panel is held in place with number of screws, ranging from 8 to 26. These were replaced using an electric screwdriver with controllable torque, in an attempt to keep constant boundary conditions [8]. It was estimated that panels 3 and 6 would cause the most problems for any pattern recognition techniques, as they are the smallest and placed relatively close together [8]. As a result, these data represent a 10-class problem; one class is associated with the normal condition (including repairs) and one class for each state of damage (nine in total). There are 2482 observations in the dataset; 700 one-shot measurements for the normal condition and 198 for each damage

condition [12].

The complete measured data have 9216 variables (1024×9). (If more transmissibilties are considered, the dimensionality can significantly increase.) With just 700 observations (of normal condition data) to inform dimension reduction, feature selection is clearly required. In the original papers [8, 12], these data were compressed to 9-dimensions using 9 MSD novelty detectors, one learnt from each transmissibility. In the proposed SHM framework, these discordancy outputs were initially used for *damage detection*; secondly, *damage location* was achieved using the discordancy measures as inputs to learn a classifier. A multilayer perception (MLP) was used [8].

During feature extraction, potential features for each novelty detector were selected by identifying regions from the spectrum that were observed to be unambiguously different to the normal condition when damage was simulated [43]. A total of 44 novelty detectors were trained via this semi-objective framework; then, the optimal subset of 9 damage-sensitive features was found using a Genetic Algorithm (GA) [12]. Briefly, the GA iterates though a population of different novelty detectors, represented by a set (vector of integers, ranging from 1 - 44). The fitness of each set is assessed using a simple multilayer perception, and the inverse classification error on a distinct *validation-set* [12]. The 'fittest' sets are passed on to the next generation by combining their solutions. Mutation is also included by the occasional random switch of a feature [12].

A large amount of 'engineering judgement' was used in the initial feature extraction steps, and the (damage location) labels were used informally to aid this semi-objective process. Furthermore, when applying the GA, a distinct subset of labelled data are used directly while optimising the set of representative features. As discussed, the use of data labels during feature extraction (both informally and directly) is a significant issue for practical applications of SHM. In real settings, it is infeasable to collect data relating to the damaged states before an SHM system is built. Furthermore, more practical methods (that run online, or use limited labelling [13, 29]) can only include the new information from novel data once they are discovered.

Application issues

Two characteristics of these data make them challenging to work with. Firstly, the data are noisy (see Figure (4)) making conventional methods for feature selection infeasible. Secondly, inconsistent data groups are present across the normal condition. It is believed that these inconsistencies occur due to repeated 'maintenance' procedures, which were simulated when replacing each panel using a screwdriver. Although a controlled torque screwdriver was used in an effort to keep the boundary conditions constant following each panel replacement, this created a lot of variability in the boundary conditions. In turn, this altered the dynamic characteristics of the wing, leading to 7 different groups (associated with each panel replacement) across the normal condition data, and 2 different groups associated with each class of damage (one for each panel removal).

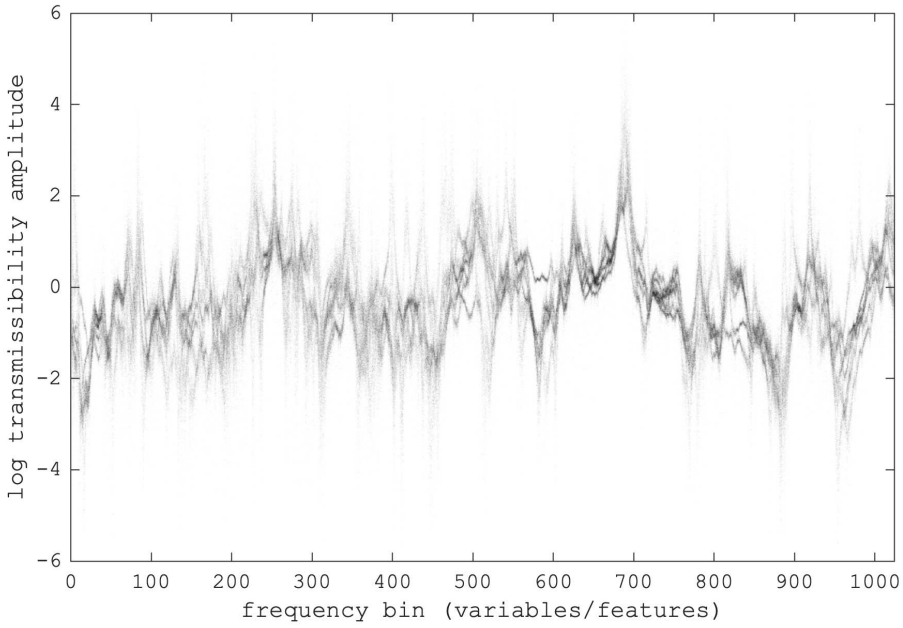


Figure 4: 700 transmissibility measurements across panel 1 for the normal condition data. These data are inconsistent and noisy; furthermore, the majority of features are clearly non-Gaussian. As a result, it is extremely difficult to represent the normal data as one class within the feature space.

The dispersion of the normal condition data mean that conventional methods for dimension reduction (§2.4) are impractical. Generally, these methods maximise variance, therefore, the projections lead to highly non-Gaussian (multimodal) features. As a result, the data become

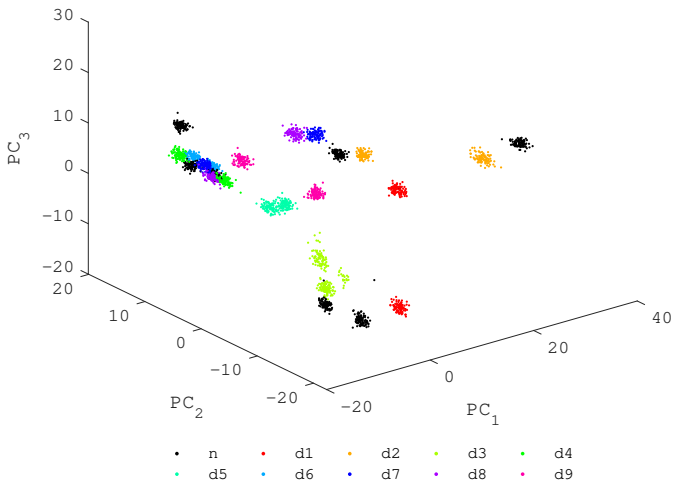
spread out across the feature space. For example, when using PCA with these data, the normal condition forms 7 disjoint clusters, while each state of damage forms 2 clusters, shown in Figure 5a; these data groups correspond to the changing boundary conditions. Clearly, conventional methods for dimension reduction are unsuitable when trying to represent any common/general behaviour across the normal condition data. Unsurprisingly, when using the principal components for MSD outlier analysis, masking occurs, illustrated in Figure 5b. Note, the MSD novelty detector is trained using a sample of 50% of the normal condition data. The training-sample (shown by a $\circ$ marker in all figures) is stratified, such that there are an equal number of data from each sub-group relating to the normal condition. Each sub group corresponds to changes following maintenance. The remaining 50% is used as a test set (shown by a $\bullet$ marker in all figures), to ensure model generalisation.

It could be argued that the use of a mixture model (or nonparametric approach) would provide a logical foundation for outlier analysis with the non-Gaussian features shown in Figure 5a. While this may succeed, if there is a way to represent the normal data as a single (ideally Gaussian) cluster, this can simplify any further modelling and outlier analysis in the SHM system.

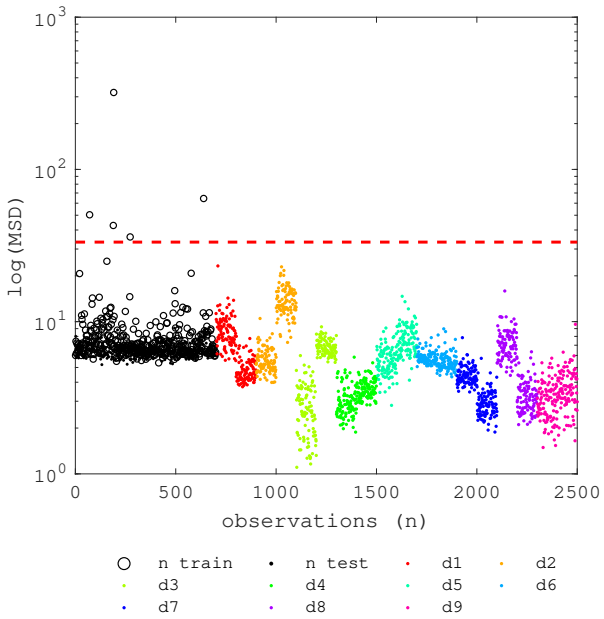
5.1 Building the ensemble

With these data, the primary aim is to alleviate the curse of dimensionality, allowing outlier analysis in high-dimensions, without significant loss of information. Secondly, the aim is to reduce the effects of noisy/misrepresentative features and capture any general/consistent behaviour across inconsistent normal condition data. In other words, represent the normal data as a single cluster within the feature space.

This has been achieved for the Gnat dataset in the past, but only through the use of labelled data [8, 12]. Considering these aims in an unsupervised setting, the problem can be considered as *exclusive* outlier analysis, such that the training data represent the normal condition data only. Using this approach, outlier ensembles are applied to define a single measure of novelty for each transmissibility, D_E^2; this produces a total of nine damage sensitive features, in line with previous experiments with these data. Bootstrap-sampled features (feature bagging) are



(a)



(b)

Figure 5: (a) Visualisation of the Gnat data using the first three principal components (tranmissibility across panel 1, T1). The normal condition data, n, define 7 separate clusters across the feature space. Damaged states d1-d9 are shown by colour markers. (b) MSD outlier analysis with the first 9 principal components as inputs. These features are clearly non-Gaussian over the normal data, therefore, significant masking occurs.

used to build each MSD base predictor. Similar to Case Study I, applying an MSD ensemble implies that there is some general behaviour *across the feature space* (for the normal data) that can be represented by a Gaussian-distribution.

In summary, feature bagging can provide an output novelty index that is robust; thus, it is not overly sensitive to abnormal behaviour within the feature-space. In other words, during exclusive analysis, outlier ensembles are utilised to reduce the abnormal effects (across the *normal* data) on the output measures of discordancy. This is in contrast to inclusive analysis, which looks to expose abnormal observations within the available data; therefore, the need for robust outputs, rather than parameters, is why parameter averaging is not suitable. As discussed, all ensembles are built with a stratified sample of 50% from the normal condition data (shown by a $\circ$ marker in all figures); the remaining 50% is included in the test-set (shown by a $\bullet$ marker in all figures) to ensure model generalisation.

Model combination

A standard weighted combination of outputs [20] is applied to the committee of models with M members,

$$D_E^2(\mathbf{x}_i) = \frac{1}{M}\sum_{m=1}^{M} w_m D_m^2(\mathbf{x}_i), \tag{14}$$

where D_m^2 denotes the discordancy from the m^{th} member in the ensemble, and w_m is the associated weight. Each member is an MSD novelty detector with empirical parameters $\hat{\boldsymbol{\mu}}_m$ and $\hat{\boldsymbol{\Sigma}}_m$. In this case study, the weight vector, $\mathbf{w} = \{w_1, ..., w_M\}$, is set to unity for all members. The discordancy measures for all the data in $\mathbf{X}$ are calculated for each of the M members, $\mathbf{D}_M^2$. For a single, robust measure of discordancy, the averaged output (for each observation) is used, $\mathbf{D}_E^2$, equation (14). Pseudocode is provided in Algorithm 2, Appendix B.

Ensemble threshold

Following a similar intuition to Case Study I, an ensemble threshold is found by applying the method outlined in §2.2, based on the size of the feature subsample used to build each member. Therefore, the threshold is defined, for an $[\, N \times n_f \,]$ dataset, where n_f is the number of features

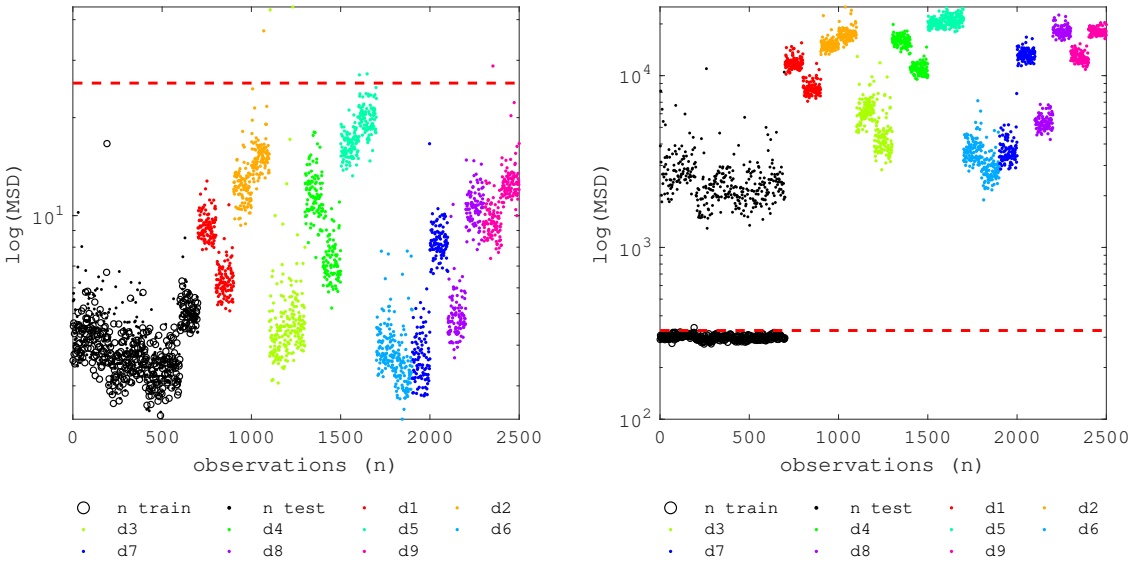
in each (random) subsample. An important distinction for this application is the reduction in dimensionality from d to n_f dimensions. As a result, when assuming Gaussian-distributed features, the combined output measures of discordancy should (theoretically) be approximately Chi-squared distributed, with n_f degrees of freedom, $\chi^2_{n_f}$, see equation (6).

It should be acknowledged that the normal condition data are clearly non-Gaussian for this dataset; therefore, the assumptions here might seem unreasonable. While this a valid statement, when using ensemble analysis, this approach is shown (in the experiments) to provide a successful method for *robust* damage detection with problematic engineering data. Furthermore, it provides a simple framework to find a representation of dispersed normal condition data as one cluster within the analysed feature space.

Subsample size

In contrast to Case Study I, the most representative feature subsample is more difficult to define. If the subsamples are too small, too much information is lost, leading to conservative measures of discordancy, see Figure 6a. If the subsamples are too large, the curse of dimensionality takes effect, leading to a near singular covariance, $\hat{\mathbf{\Sigma}}_m$. As a result, the novelty detectors over-train, and *all* new data are flagged as outlying, including those that represent the normal condition, see Figure 6b.

Generally, during feature bagging, the subsample size n_f is randomly selected between $d/2$ and $(d-1)$ for each member [35]. In this work, the size of each subsample is kept constant to aid defining the threshold (§2.2) and allow for a structured analysis the outlier distribution. For this application, $n_f = \sqrt{d}$, as this was found to generalise well across various transmissibilities with the Gnat data. As with Case Study I, this implies parameter tuning; therefore, n_f must defined to ensure that there is enough information in the subsamples, while avoiding a near-singular covariance matrix. This is not problematic, as the covariance determinant can be checked in an unsupervised setting. The number of members in the ensemble, M, is set such that the total number of sampled features is equal to the number of dimensions in the original dataset, i.e., $M = \sqrt{d}$.



(a) (b)

Figure 6: Example ensemble analysis with 50 members, applied to the data from the transmissibility across panel 1. (a) 4 features per subsample (b) 300 features per subsample.

5.2 Results & discussion

Novelty detection

One outlier ensemble is trained for each transmissibilty path, T1 – T9, shown in Figure 3. The combined outputs, or novelty indices (D_E^2) from each ensemble are used to compress these high-dimensional vibration-based measurements to 9 damage sensitive features. This follows the same SHM framework proposed in the original papers [8]; if one of the inspection panels are removed to simulate damage, at least one novelty index should pass the detection threshold, indicating novelty.

The combined ensemble outputs, D_E^2, for T1 – T9 are shown in Figures 7 – 11. Outlier ensembles are an appropriate tool for damage detection with these data, as novelty measures from the various damaged states generally pass (at least one) detection threshold, while few data from the normal condition are flagged as outlying. For each novelty index (T1 – T9), the false positive rate (FPR) for the normal data, and the false negative rate (FNR) for each

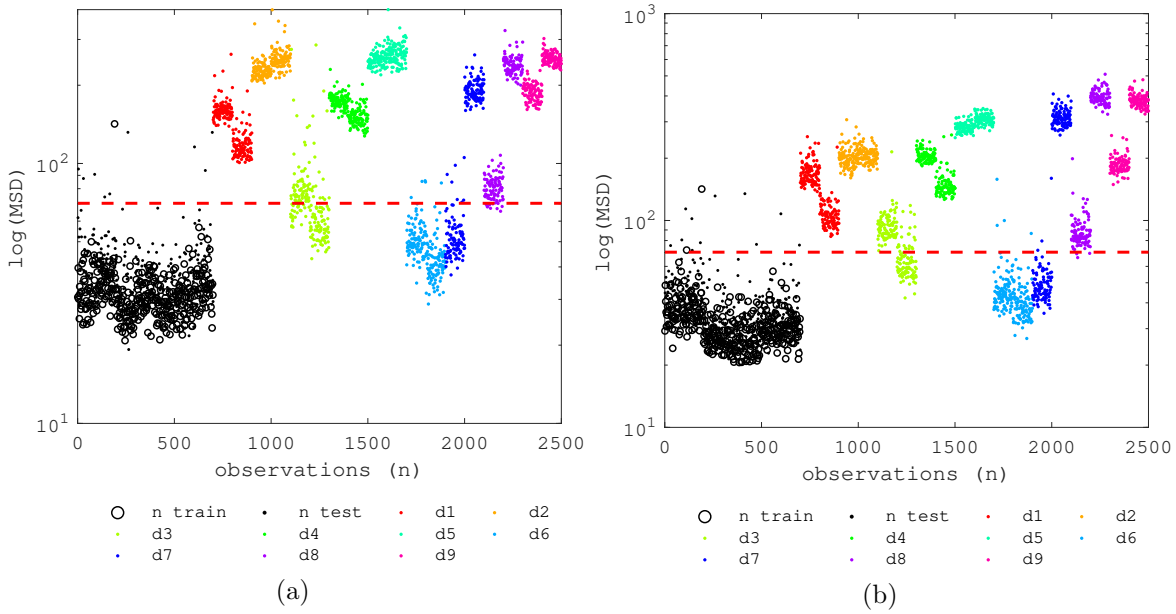


Figure 7: (a) Outlier ensemble novelty index (D_E^2) for data from transmissibility 1 (T1); (b) Outlier ensemble novelty index (D_E^2) for data from transmissibility 2 (T2).

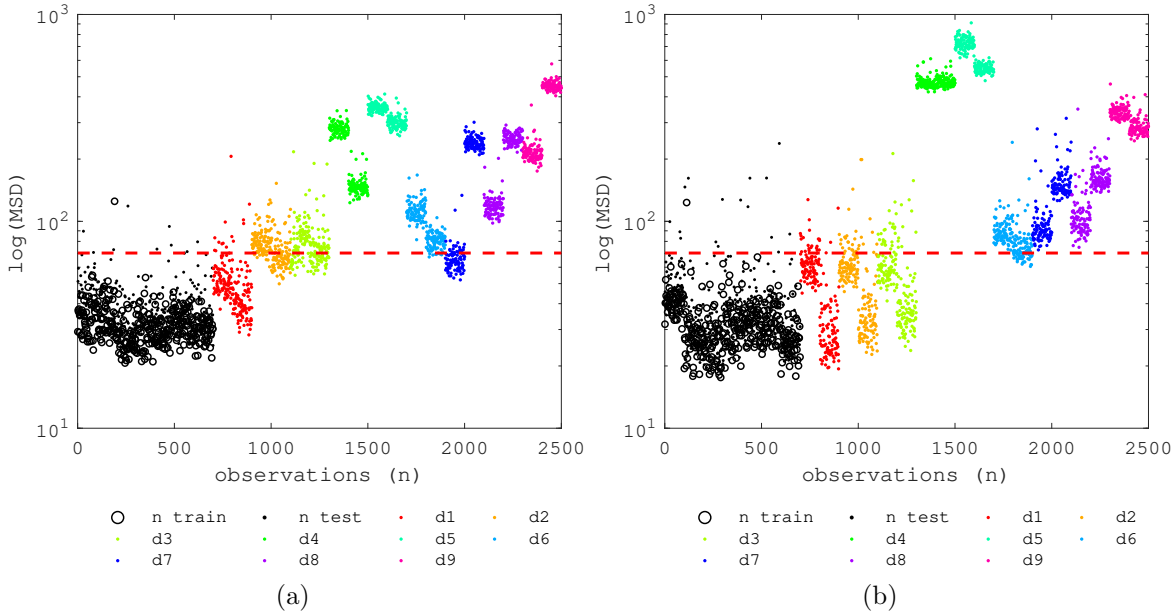


Figure 8: (a) Outlier ensemble novelty index (D_E^2) for data from transmissibility 3 (T3); (b) Outlier ensemble novelty index (D_E^2) for data from transmissibility 4 (T4).

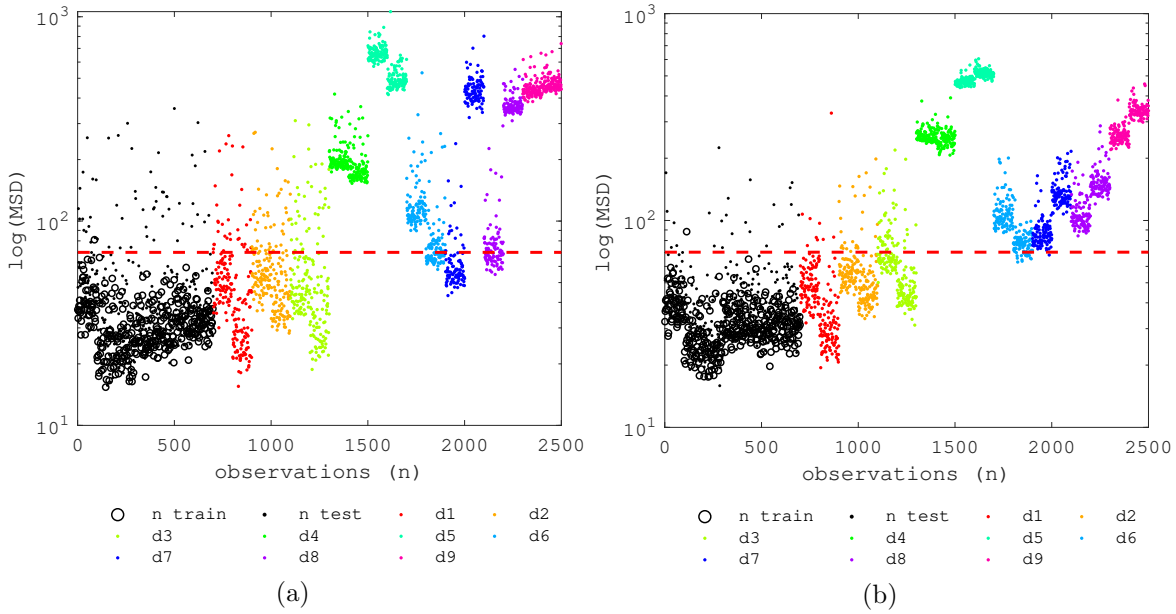


Figure 9: (a) Outlier ensemble novelty index (D_E^2) for data from transmissibility 5 (T5); (b) Outlier ensemble novelty index (D_E^2) for data from transmissibility 6 (T6).

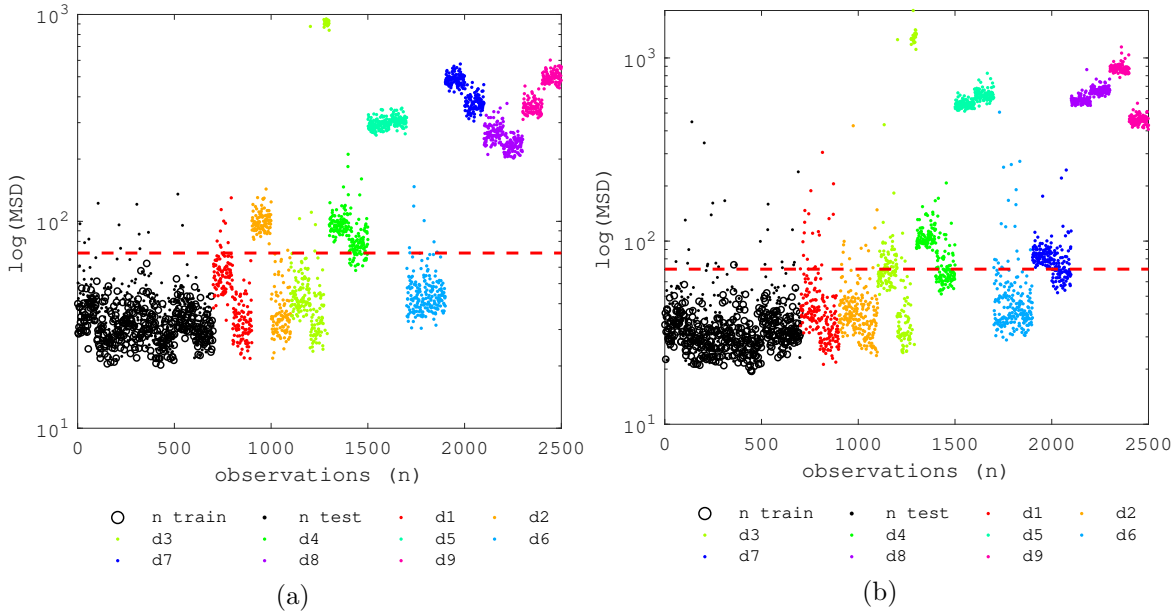


Figure 10: (a) Outlier ensemble novelty index (D_E^2) for data from transmissibility 7 (T7); (b) Outlier ensemble novelty index (D_E^2) for data from transmissibility 8 (T8).

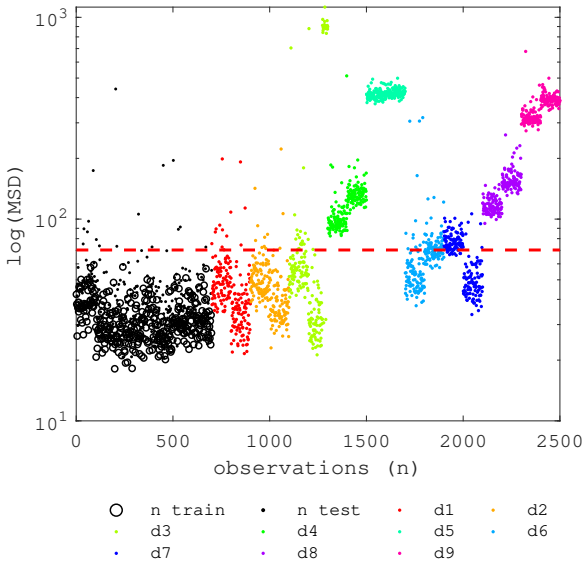


Figure 11: Outlier ensemble novelty index (D_E^2) for data from transmissibility 9 (T9)

Table 1: Normal condtion *false positive rate* (FPR, %) and damage state *false negative rate* (FNR, %) for combined ensemble outputs, T1 – T9.

	normal	d1	d2	d3	d4	d5	d6	d7	d8	d9
T1	2.57	**0.00**	**0.00**	55.00	**0.00**	**0.00**	97.50	46.50	5.00	**0.00**
T2	3.14	**0.00**	**0.00**	38.50	**0.00**	**0.00**	98.00	48.50	1.50	**0.00**
T3	2.57	93.00	39.00	38.50	**0.00**	**0.00**	4.50	38.50	**0.00**	**0.00**
T4	4.57	88.00	85.00	82.00	**0.00**	**0.00**	18.50	**0.00**	**0.00**	**0.00**
T5	*15.71*	83.50	76.00	80.00	**0.00**	**0.00**	25.00	42.50	25.50	**0.00**
T6	6.86	95.50	87.00	76.00	**0.00**	**0.00**	9.50	0.50	**0.00**	**0.00**
T7	3.43	95.00	49.50	84.50	12.00	**0.00**	96.50	**0.00**	**0.00**	**0.00**
T8	6.86	90.00	94.00	62.50	32.50	**0.00**	90.00	28.00	**0.00**	**0.00**
T9	4.57	93.50	94.50	77.00	**0.00**	**0.00**	67.00	53.50	**0.00**	**0.00**

damage state, are summarised in Table 1.

Table 1 highlights that the damaged states (other than d3 and d6) have at least one *ideal* feature; that is, a novelty index with a false negative rate of zero (bold, Table 1). It is suggested that the best feature FNR for d3 and d6 reaches 38.5% and 4.5% respectively as these measurements cover the two smallest panels, see Figure 3. It is hypothesised that these data are more affected by the changing boundary conditions following panel replacements due to size and location of panels 3 and 6, thus, the normal data become more difficult to represent. In support of this this theory, previous work with these data have found the classification of d3 and d6 problematic [8, 12].

The false positive rate for the normal data is acceptable (below 10%) for all ensemble outputs, *excluding* T5. Again, it is suggested that T5 shows inferior generalisation (italics, Table 1) as panel 5 is also relatively small and located in the centre of the wing, see Figure 3. While this inferior generalisation could be improved, there are three *ideal* damage sensitive features associated with panel 5; therefore, distinguishing between normal and novel data should not be problematic in practice.

Unsupervised feature extraction

According to the SHM system proposed in previous work, the 9 novelty indices can then be used as the inputs to a classification algorithm, to predict the location of damage. This 9-class dataset is visualised (*including* the normal data) via principal component analysis in Figure 12. It can be inferred that the use of outlier ensembles for dimension reduction has been successful, as each class forms clusters that are relatively separable and distinct. Furthermore, the inconsistent normal data are now represented as a *single* cluster within the feature space, while any novel measurements form separated groups.

The 9-class damage location problem (damaged data only) is visualised in Figure 13a. To assess these features against those found in a *supervised* setting, the alternative 9-class dataset (found using a genetic algorithm and semi-objective method [8, 12]) is visualised in Figure 13b. It can be observed in Figure 13a that each damaged class generally forms 2 separated groups; as discussed, this is expected as a result of the changing boundary conditions following panel

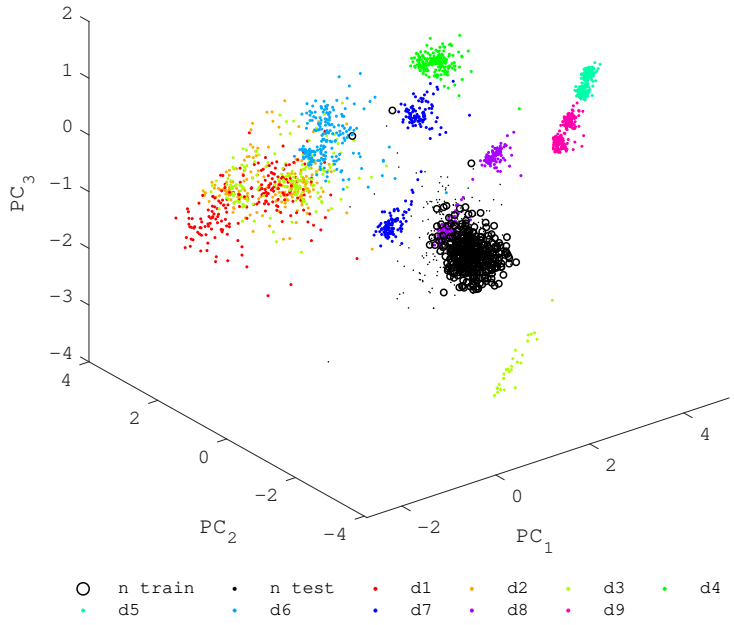
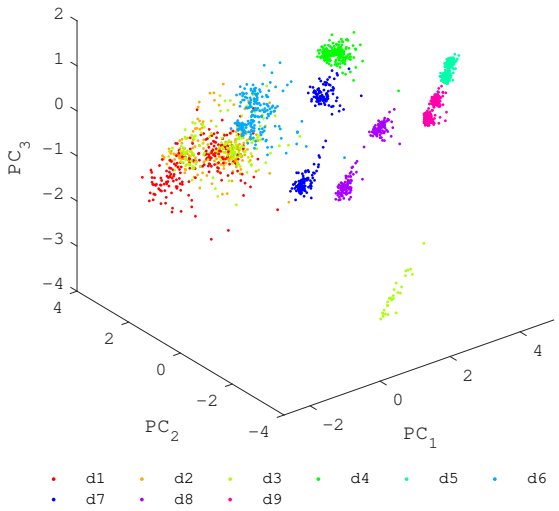


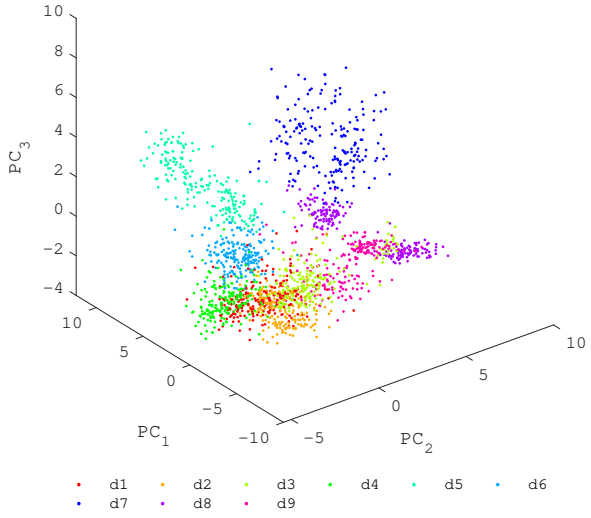
Figure 12: The compressed Gnat data; including the 9 damaged sates (colour markers) and the normal condition (black markers). The 9 damage sensitive features are visualised using PCA projection.

replacement. A multi-layer perception (MLP) is used to learn a mapping from the input measures of discordancy, to output labels of damage location. The MLP has one hidden layer (2-layers of weights); a bias is included with each layer of weights. The network has 9 inputs, and 9 outputs for classification. The *rectified linear unit* (ReLu) activation function is used in the hidden nodes and the Softmax activation function at the outputs. To limit the complexity of the network architecture and mapping, the maximum allowed number of hidden nodes is 10. (For discussions regarding over training/complexity, refer to [12]). The network is then trained according to the '1 of M' strategy [44]. The optimal number of hidden nodes ($1 - 10$) is determined using a distinct validation set, which is also used to prevent over-training via early stopping. For each network architecture, the weights are initialised 10 times; the weights with the best classification accuracy on the validation-set are used in the final model. The training set is built with a stratified sample, in which an equal number of data are randomly sampled from each damaged class. This approach differs to previous experiments with the Gnat data [8, 12], where the training-set was defined with a sample of equally spaced observations; therefore, slightly inferior generalisation is expected in the experiments here (for the same supervised features).

For these problematic engineering data, the use of outlier ensembles can produce damage-sensitive features with a classification accuracy that is *near identical* to the features extracted in a *supervised* framework. Specifically, the classification accuracy (on a distinct test-set) when using the features found via *unsupervised* outlier ensembles is 95.85%; when using the features found in a *supervised* setting [8, 12], the classification accuracy is 97.39%. The similarity in the classification performance is extremely significant for feature extraction from high-dimensional SHM data, particularly vibration-based measurements, as the use of outlier ensembles can produce robust damage sensitive features, *without* the need for *labelled data* or measurements of the system outside the expected normal condition(s). Additionally, outlier ensembles appear to successfully represent noisy, disjoint (non-Gaussian) features as a single cluster within the feature space, illustrated in Figure 12. This representation of measured data is a major benefit when applying any clustering/classification algorithms, later in the SHM framework.



(a)



(b)

Figure 13: 9-class classification problem for damage location. (a) PCA visualisation of the 9-dimensional dataset found using a genetic algorithm and semi-objective framework [8, 12] (supervised). (b) PCA visualisation of the 9-dimensional dataset found using outlier ensembles (unsupervised).

Analysis of outliers

According to the assumptions summarised in §2.2 (Gassian-distributed features, consistent parameter estimation) the discordancy measures across the normal condition data should be *approximately* Chi-square distributed with n_f degrees of freedom. A histogram of the empirical distribution of $\mathbf{D}_E^2$ (for all 9 transmissibilities, T1 – T9) is provided for the normal data in Figure 14.

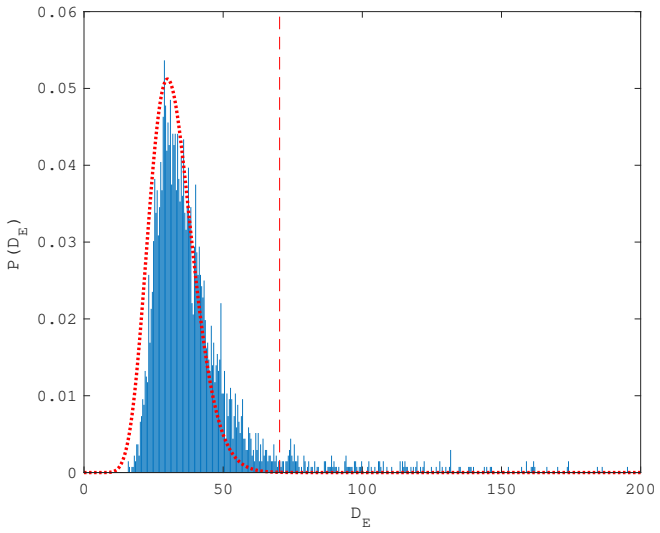


Figure 14: Normalised distribution of the *combined* output measures, D_E^2, for the normal condition test-data (T1 – T9). The red dotted line shows a Chi-squared distribution with n_f degrees of freedom. The red dashed line shows the 99% threshold found defined using the Monte Carlo sampling method.

While the distance measures are not Chi-square distributed, the outputs are considered *approximately* Chi-square distributed with n_f degrees of freedom, at least for the purposes of this duscussion. As the combined outputs ($\mathbf{D}_E^2$) are distributed in this way, it can be assumed, roughly speaking, that a *pseudo*-Gaussian representation of the normal data has been found, such that a uni-modal representation that captures the general behavior acorss the normal data. Therefore, considering the highly inconsistent nature of the data from the normal condition(s), feature bagging appears to capture any consistent/general behaviour over inconsistent training data, while producing a robust measure of novelty from a high-dimensional feature space.

An alternative way to view the output distribution is to analyse the discordancy measures from *all* M members before averaging, $\mathbf{D}_M^2$. The distribution of $\mathbf{D}_M^2$ for the training data (T1

– T9) is shown in Figure 15. In this analysis, the empirical distributions are better defined, as there are $M \times n \times 9$ measures of discordancy, where n is the number of observations. Using this approach, outlier ensembles offer a robust route to hypothesis testing with high-dimensional data. Note, hypothesis testing is feasibile becuase boot-strap sampling increases number distances which can be used to define the empirical distribution, as the outputs from all members of the ensemble are used. To illustrate the potential of this framework, the discordancy distribution for the normal condition *test-set* is shown in Figure 16a, and for the damaged data, Figure 16b. There is a clear difference between the output distributions following damage; specifically, the high-value tail of the distribution becomes much heavier. Hypothesis tests such as the Kolmogorov-Smirnov test (KS-test) or the maximum mean discrepancy (MMD) could be used to assess the difference between the output distributions. This should allow for the detailed analysis of outliers, and the potential to classify damage according to the distribution behaviour.

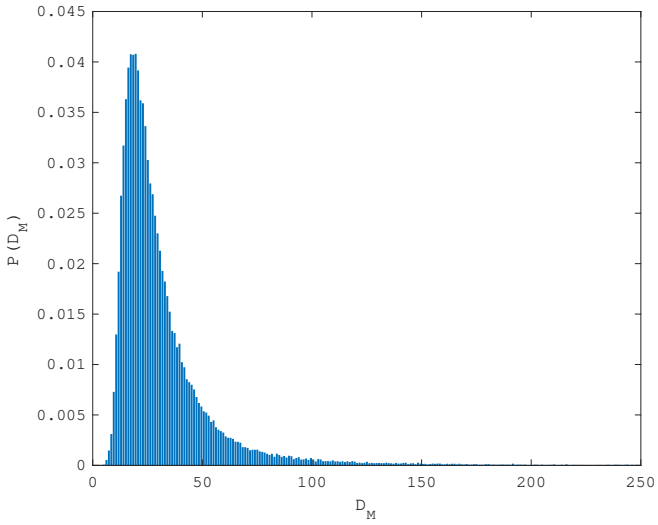
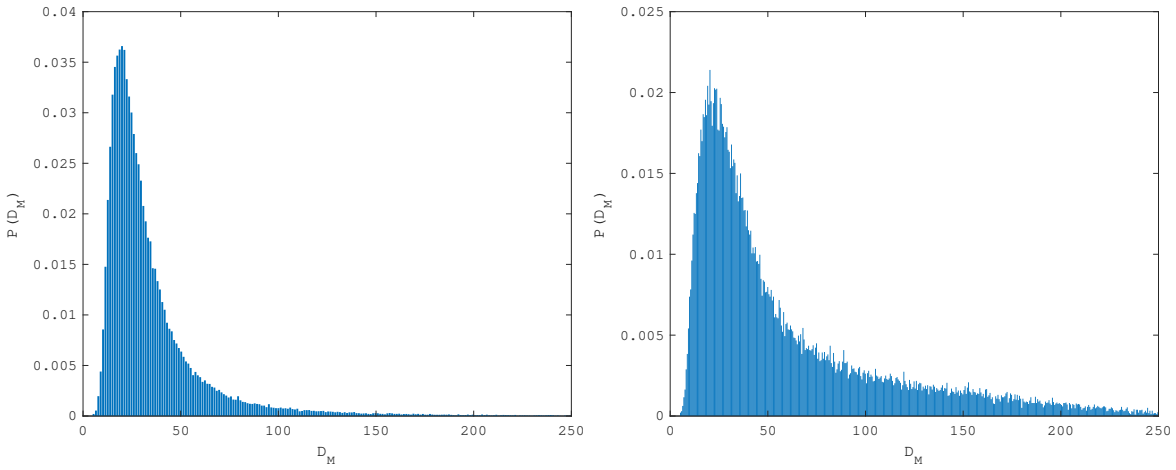


Figure 15: Distribution of the output measures of discordancy for *all members* in the ensemble $\mathbf{D}_M^2$ (normal condition data, T1 – T9).

The KS-test is applied to the results from the Gnat data for demonstration. In a two-sample test, the *test-statistic* is defined by the maximum difference between the two (empirical) *cumulative density functions*. If the data have been generated by the same underlying p.d.f, the test-statistic will tend to zero as the number of observations in each sample increases. For example, when compared the output distribution for the training data (Figure 15), the



(a) (b)

Figure 16: Output distribution from the outlier ensembles (T1 – T9): (a) normal test data, (b) damaged data d1.

normal condition test-data (Figure 16a) have a KS-statistic of 0.057, while the distribution for d1 damage (Figure 16b) gives 0.306. Similarly, the distribution of ensemble outputs *for each observation* can be compared to the distribution of the normal data. In this way, hypothesis tests can be used to provide another *robust* measure of novelty. Figure 17 shows the behaviour of the KS-statistic for each observation in the test data; these results are provided to highlight the potential for more detailed hypothesis testing using the distribution of outliers. Specifically, as the outlying data should appear in the tails of the distributions, it is suggested that outlier ensembles might be used as the foundation for extreme value statistical analysis [23].

6 Conclusions

Outlier ensembles have been introduced as a tool for robust statistical outlier analysis with practical examples of engineering data. A diverse ensemble of Mahalanobis squared-distance novelty detectors has been trained, using either bootstrap-sampled features (feature bagging) or bootstrap-sampled observations. Provided that there exists some consistent behaviour within the inlying data that can be modelled by a Gaussian-distribution, the ensembles outputs can be used as a tool for robust damage detection, as well as dimension reduction, in an unsupervised framework. In each of the case studies, while the heuristic has required some

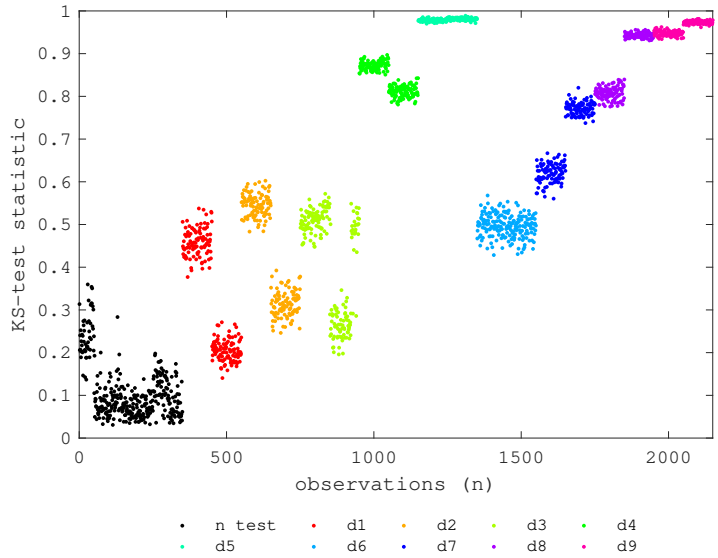


Figure 17: The two sample KS-test statistic comparing the distribution of D_E for each test-observation to the distribution of D_E from the normal condition training data.

parameter tuning, the demonstrated method can still be considered unsupervised, as it does not require any labelled data; specifically, the tuning of parameters is possible within the unsupervised framework.

In Case Study I, an alternative approach for robust inclusive outlier analysis has been proposed. The heuristic is compared to the benchmark FAST-MCD algorithm, and it is shown to provide a discordancy measure that is comparably sensitive to inclusive novelty within a practical engineering dataset. Additionally, the new framework runs up to 15 times faster than the FAST-MCD algorithm, and the (unweighted) method requires fewer application specific parameter to be set (provided that subsamples are large enough to avoid a singular covariance). This indicates an algorithm that should generalise well across various data, requiring less tuning of the parameters; it is acknowledged, however, that the success ensemble methods is inherently sensitive to the proportion of inclusive outliers within the dataset.

Case Study II demonstrates how outlier ensembles can be utilised as a tool for robust damage detection (and dimension reduction) with high-dimensional data in a wholly unsupervised framework. The analysis offers an effective method to represent noisy/inconsistent training data as one cluster within the feature space. Most significantly, when compared to features extracted in a *supervised* setting (via a genetic algorithm), the unsupervised features produce

a comparable classification accuracy in a damage location problem. Specifically, when using the features found via *unsupervised* outlier ensembles, the accuracy is is 95.85%, while the supervised features give an accuracy of 97.39%. This is of particular value to SHM systems, as labelled data are rarely available to inform feature extraction in practical applications; this includes any systems that look to run adaptively, online and with limited supervision.

Further work is suggested. Firstly, analysis should be applied to various datasets, with different levels of contamination to test generalisation. Alternative combination weightings could be explored to improve on the simple pruning (and unit weighting) regime suggested here. Finally, it is suggested that the output distributions from ensemble analysis offers an alternative foundation for extreme value statistical analysis.

7 Acknowledgements

The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC) through Grant reference numbers EP/R003645/1, EP/R004900/1.

References

[1] K. Worden, G. Manson, and N. R. Fieller. Damage detection using outlier analysis. *Journal of Sound and Vibration*, 229(3):647–667, 2000.

[2] N. Dervilis, E. Cross, R. Barthorpe, and K. Worden. Robust methods of inclusive outlier analysis for structural health monitoring. *Journal of Sound and Vibration*, 333(20): 5181–5195, 2014.

[3] M. Yeager, B. Gregory, C. Key, and M. Todd. On using robust mahalanobis distance estimations for feature discrimination in a damage detection scenario. *Structural Health Monitoring*, 2018.

[4] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In C. Beeri and P. Buneman, editors, *Database Theory — ICDT'99*, pages 217–235, Berlin, Heidelberg, 1999. Springer Berlin Heidelberg.

[5] C. C. Aggarwal and P. S. Yu. Outlier detection for high dimensional data. In *ACM Sigmod Record*, volume 30, pages 37–46. ACM, 2001.

[6] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In J. Van den Bussche and V. Vianu, editors, *Database Theory — ICDT 2001*, pages 420–434, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[7] C. R. Farrar and K. Worden. *Structural Health Monitoring: a Machine Learning Perspective*. John Wiley & Sons, 2012.

[8] G. Manson, K. Worden, and D. Allman. Experimental validation of a structural health monitoring methodology: Part III. damage location on an aircraft wing. *Journal of Sound and Vibration*, 259(2):365–385, 2003.

[9] H. V. Nguyen, H. H. Ang, and V. Gopalkrishnan. Mining outliers with ensemble of heterogeneous detectors on random subspaces. In *International Conference on Database Systems for Advanced Applications*, pages 368–383. Springer, 2010.

[10] C. C. Aggarwal. Outlier ensembles: position paper. *ACM SIGKDD Explorations Newsletter*, 14(2):49–58, 2013.

[11] A. Lazarevic, L. Ertoz, V. Kumar, A. Ozgur, and J. Srivastava. A comparative study of anomaly detection schemes in network intrusion detection. In *Proceedings of the 2003 SIAM International Conference on Data Mining*, pages 25–36. SIAM, 2003.

[12] K. Worden, G. Manson, G. Hilson, and S. Pierce. Genetic optimisation of a neural damage locator. *Journal of Sound and Vibration*, 309(3):529–544, 2008.

[13] L. Bull, K. Worden, G. Manson, and N. Dervilis. Active learning for semi-supervised structural health monitoring. *Journal of Sound and Vibration*, 437:373–388, 2018.

[14] A. Zimek, M. Gaudet, R. J. Campello, and J. Sander. Subsampling for efficient and effective unsupervised outlier detection ensembles. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 428–436. ACM, 2013.

[15] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*, volume 589. John Wiley & sons, 2005.

[16] S. Van Aelst and P. Rousseeuw. Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):71–82, 2009.

[17] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

[18] M. Hubert and M. Debruyne. Minimum covariance determinant. *Wiley interdisciplinary reviews: Computational statistics*, 2(1):36–43, 2010.

[19] D. M. Hawkins. *Identification of Outliers*, volume 11. Springer, 1980.

[20] C. M. Bihsop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[21] K. Zhang, M. Hutter, and H. Jin. A new local distance-based outlier detection approach for scattered real-world data. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 813–822. Springer, 2009.

[22] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander. LOF: identifying density-based local outliers. In *ACM sigmod record*, volume 29, pages 93–104. ACM, 2000.

[23] H. M. Wadsworth. *Handbook of Statistical Methods for Engineers and Scientists*. 1989.

[24] J. Hardin and D. M. Rocke. The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4):928–946, 2005. ISSN 10618600.

[25] P. J. Rousseeuw and W. V. D. Bossche. Detecting deviating data cells. *Technometrics*, pages 1–11, 2017.

[26] R. Cook, D. Hawkins, and S. Weisberg. Exact iterative computation of the robust multivariate minimum volume ellipsoid estimator. *Statistics & Probability Letters*, 16(3): 213–218, 1993.

[27] C. Croux and G. Haesbroeck. An easy way to increase the finite-sample efficiency of the resampled minimum volume ellipsoid estimator. *Computational Statistics & Data Analysis*, 25(2):125–141, 1997.

[28] R. Butler, P. Davies, and M. Jhun. Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, pages 1385–1400, 1993.

[29] T. Rogers, K. Worden, R. Fuentes, N. Dervilis, U. Tygesen, and E. Cross. A Bayesian non-parametric clustering approach for semi-supervised structural health monitoring. *Mechanical Systems and Signal Processing*, 119:100 – 119, 2019. ISSN 0888-3270.

[30] E. Reynders, G. Wursten, and G. De Roeck. Output-only structural health monitoring in changing environmental conditions by means of nonlinear system identification. *Structural Health Monitoring*, 13(1):82–93, 2014.

[31] K. P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT press, 2012.

[32] M. Silva, A. Santos, R. Santos, E. Figueiredo, C. Sales, and J. C. Costa. Deep principal component analysis: An enhanced approach for structural damage identification. *Structural Health Monitoring*, page 1475921718799070, 2018.

[33] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford university press, 1995.

[34] A. Zimek, R. J. Campello, and J. Sander. Ensembles for unsupervised outlier detection: challenges and research questions a position paper. *ACM Sigkdd Explorations Newsletter*, 15(1):11–22, 2014.

[35] C. C. Aggarwal and S. Sathe. *Outlier Ensembles: an Introduction*. Springer, 2017.

[36] A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 157–166. ACM, 2005.

[37] T. G. Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer, 2000.

[38] N. Abe, B. Zadrozny, and J. Langford. Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 504–509. ACM, 2006.

[39] G. D. Roeck. The state-of-the-art of damage detection by vibration monitoring: the simces experience. *Structural Control and Health Monitoring*, 10(2):127–134, 2003.

[40] B. Peeters and G. De Roeck. One-year monitoring of the Z24-bridge: environmental effects versus damage events. *Earthquake Engineering & Structural Dynamics*, 30(2): 149–171, 2001.

[41] S. Verboven and M. Hubert. LIBRA: a MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory Systems*, 75(2):127–136, 2005.

[42] K. Worden, G. Manson, and D. Allman. Experimental validation of a structural health monitoring methodology: Part I. novelty detection on a laboratory structure. *Journal of Sound and Vibration*, 259(2):323–343, 2003.

[43] G. Manson, K. Worden, and D. Allman. Experimental validation of a structural health monitoring methodology: Part II. novelty detection on a gnat aircraft. *Journal of Sound and Vibration*, 259(2):345–363, 2003.

[44] L. Tarassenko. *A Guide to Neural Computing Applications*. Butterworth-Heinemann, 1998.

Appendices

A The FAST-MCD algorithm

The heuristic is characterised by repeated C-steps (concentration steps) [17, 18]:

1. take subset $\mathbf{H}_1 \subset \mathbf{X}$, that is $|\mathbf{H}_1| = h$,
2. calculate the empirical estimates of the mean $\hat{\boldsymbol{\mu}}_1$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_1$ from the data in $\mathbf{H}_1$,
3. define relative distances for *all* N data:

   $D_1^2(\mathbf{x}_i) = (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)^\top \hat{\boldsymbol{\Sigma}}_1^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_1)$, $\forall\ \mathbf{x}_i \in \mathbf{X}$;
4. sort all distances in ascending order, such that

   $\{D_1^2(\mathbf{x}_i)_{(1)} \geq D_1^2(\mathbf{x}_i)_{(2)} \geq ... \geq D_1^2(\mathbf{x}_i)_{(N)}\}$, $\forall\ \mathbf{x}_i \in \mathbf{X}$;
5. now set $\mathbf{H}_2$ as the data with the smallest distances, such that:

   $\{D_1^2(\mathbf{x}_i) : i \in \mathbf{H}_2\} := \{D_1^2(\mathbf{x}_i)_{(1)}, ...D_1^2(\mathbf{x}_i)_{(h)}\}$, again $|\mathbf{H}_2| = h$.

Repeating these steps will always *concentrate* the determinant, such that $\det(\hat{\boldsymbol{\Sigma}}_2) \leq \det(\hat{\boldsymbol{\Sigma}}_1)$, with equality if and only if $\hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\Sigma}}_2 = \hat{\boldsymbol{\Sigma}}_1$ [17, 18]. C-steps can be iterated until a stopping criterion is met; for example, if $\det(\hat{\boldsymbol{\Sigma}}_{\text{new}}) = \det(\hat{\boldsymbol{\Sigma}}_{\text{old}})$ [18]. To construct the initial $\mathbf{H}_1$ subset, a small random $(d+1)$-subset is sampled and then enlarged to a h-subset with minimal discordancy measures [17] . This method yields better results than drawing a random h-subset directly, because the probability of drawing a outlier-free subset is much higher for smaller $(d+1)$-subsets [18]. The FAST-MCD algorithm has further improvements for computational efficiency, these include multiple initialisations and partitioning schemes for large datasets [18]; for details, refer to the original paper [17].

B Heuristics

The two variations of the algorithm are provided in pseudo-code, Algorithms 1 and 2.

Algorithm 1: *Averaged parameters*

Input : Available data $\mathbf{X}$

Output : Discordancy measures $\mathbf{D}^2$

$N :=$ number of observations in $\mathbf{X}$;
$d :=$ number of dimensions in $\mathbf{X}$;
$n_s = 3 \times d$ (*subsample size*);
$M = N/n_s$ (*number of members*);
$\hat{\boldsymbol{\mu}}_E \leftarrow \{M \times d\}$ (*initialise parameter*);
$\hat{\boldsymbol{\Sigma}}_E^{-1} \leftarrow \{d \times d \times M\}$ (*init. parameter*);
for $m = 1 : M$ **do**
    Random sample n_s observations from $\mathbf{X}$;
    Calculate empirical parameters from the random sample $(\hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m^{-1})$;
    Store estimates in the m^{th} location of the parameter arrays $\hat{\boldsymbol{\mu}}_E, \hat{\boldsymbol{\Sigma}}_E^{-1}$;
end
Combine parameter estimates by (a) or (b) averaging of $\hat{\boldsymbol{\mu}}_E$ and $\hat{\boldsymbol{\Sigma}}_E^{-1}$;
Calculate discordnacy measures (with averaged parameters) for all $\mathbf{x} \in \mathbf{X}$, $\mathbf{D}^2$;

Algorithm 2: *Feature bagging*

Input : Available data $\mathbf{X}$

Output : Discordancy measures $\mathbf{D}_E^2, \mathbf{D}_M^2$

$d :=$ number of dimensions in $\mathbf{X}$;
$n_f = \sqrt{d}$ (*subsample size*);
$M = \sqrt{d}$ (*number of members*);
$D_M^2 \leftarrow \{n \times M\}$ (*initialise outputs*);
for $m = 1 : M$ **do**
    Random sample n_f features from $\mathbf{X}$;
    Calculate empirical parameters from the random subsample $(\hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m^{-1})$;
    Calculate discordnacy measures $\mathbf{D}_m^2$ for all $\mathbf{x} \in \mathbf{X}$;
    Store $\mathbf{D}_m^2$ in the m^{th} column of the output array $\mathbf{D}_M^2$;
end
Combine outputs by averaging rows of $\mathbf{D}_M^2$, i.e. $\mathbf{D}_E^2 = \text{average}(\mathbf{D}_M^2)$;