On Bayesian Networks for Structural Health and Condition Monitoring

A Thesis submitted to the University of Sheffield

for the degree of Doctor of Philosophy in the Faculty of Engineering

by

Jose Ramon Fuentes Esquivel

Department of Mechanical Engineering

University of Sheffield

June 2017

Acknowledgements

My initial motivation for pursuing a PhD was an interest in doing research. While that has and will remain interesting, it turned out that it was the people who I have, and continue to work with, that have made this experience truly worthwhile. I could have not asked for a more positive, fun, encouraging, interesting, environment than the Dynamics Research Group at Sheffield. In particular, I have Dr. Elizabeth Cross and Prof. Keith Worden to thank for this. I do not think better PhD supervision is possible. I am most grateful for the fact that they have been able to put up with me, my distractions, digressions, and varied interests but also for providing me with an excellent research environment. Starting this PhD would not have been possible without Keith, and finishing it would have out of the question if it were not for Lizzy.

I also have to thank Dr. Andrew Halfpenny and Rob Plaskitt, from nCode who were, for various reasons, instrumental in making this happen in the first place. Their technical and professional mentoring, and friendship has been extremely valuable these last few years.

Some parts of this thesis include work I did whilst working together with the Tribology group at Sheffield. For this I have Prof. Rob Dwyer-Joyce and Dr. Matt Marshall to thank. Also, the collection of wind turbine data used towards the last chapter would have been impossible without the help and friendship of Dr. Tom Howard. I have also to thank Dr. Jon Wheals and Tom Huntley from Ricardo as well as Oddbjørn Malmo, Dr. Rune Harald Hestmo and Ove Sagen Asden from Kongsberg Maritime for funding and supporting the wind turbine acoustic emission work, and ultimately for letting me use that work towards this thesis.

Last but no least, I am ever grateful to all my family and friends, you all know who you are. Everyone has contributed in one way or another to my continued happiness and well-being throughout this last four years.

Abstract

The first step in data-driven approaches to Structural Health Monitoring (SHM) is that of damage detection. This is a problem that has been well studied in laboratory conditions. Yet, SHM remains an academic topic, not yet widely implemented in industry. One of the main reasons for this is arguably the difficulty in dealing with Environmental and Operational Variation (EOV), which have a tendency to influence damage-sensitive features in ways similar to damage itself. A large number of the methods developed for SHM applications make use of linear Gaussian models for various tasks including dimensionality reduction, density estimation and system identification. As highlighted in [1], a wide range of linear Gaussian models can be formulated as special cases of a general class of probabilistic graphical models, or Bayesian networks. The work presented here discusses how Bayesian networks can be used systematically to approach different types of damage detection problems, through their likelihood function. A likelihood evaluates the probability that an observation belongs to a particular model. If this model correctly captures the undamaged state of the system, then a likelihood can be used as a novelty index, which can point to the presence of damage.

Likelihood functions can be systematically exploited for damage detection purposes across the vast range of linear Gaussian models. One of the key benefits of this fact is that simple models can easily be extended to mixtures of linear Gaussian models. It is shown how this approach can be effective in dealing with operational and environmental variabilities. This thesis thus provides a point of view on performing novelty detection under this wide class of models systematically with their likelihood functions. Models that are typically used for other purposes can become powerful novelty detectors in this view. The relationship between Principal Component Analysis (PCA) and Kalman filters is a good example of this. Under the graphical model perspective these two models are a simple variation of each other, where they model data with and without time dependence. Provided these models are trained with representative data from a non-damaged system, their likelihood function presents a useful novelty index. Their limitation to modelling linear Gaussian data can be overcome through the mixture modelling interpretation. Through graphical models, this is a straightforward extension, but one that retains a probabilistic interpretation.

The impact of this interpretation is that environmental and operational variability, as well as potential nonlinearity, in SHM features can be captured by these models. Even though the interpretation changes depending on the model, the likelihood function can consistently be used as a damage indicator, throughout models like Gaussian mixtures, PCA, Factor Analysis, Autoregressive models, Kalman filters and switching Kalman filters. The work here focuses around these models. There are various ways in which these models can be used, but here the focus is narrowed to exploring them as novelty detectors, and showing their application in different contexts. The context in this case refers to different types of SHM data and features, as this could be either vibration, acoustics, ultrasound, performance metrics, etc.

This thesis provides a discussion on the theoretical background for probabilistic graphical models, or Bayesian networks. Separate chapters are dedicated to the discussion of Bayesian networks to model static and dynamic data (with and without temporal dependencies, respectively). Furthermore, three different application examples are presented to demonstrate the use of likelihood function inference for damage detection. These systems are a simulated mass-spring-damper system, with varying stiffness in its non-damaged condition, and with a cubic spring nonlinearity. This system presents a challenge from the point of view of the characterisation of the changing environment in terms of global stiffness and excitation energy. It is shown how mixtures of PCA models can be used to tackle this problem if frequency domain features are used, and mixtures of linear dynamical systems (Kalman filters) can be used to successfully characterise the baseline undamaged system and to identify the presence of damage directly from time domain measurements. Another case study involves the detection of damage on the Z-24 bridge. This is a well-studied problem in SHM research, and it is of interest due to the nonlinear stiffness effect due to temperature changes. The features used here are the first four natural frequencies of the bridge. It is demonstrated how a Gaussian mixture model can characterise the undamaged condition, and its likelihood is able to accurately predict the presence of damage. The third case study involves the prediction of various stages of damage on a wind turbine bearing. This is an experimental laboratory investigation - and the problem is also tackled with a Gaussian mixture model. This problem is of interest because the lowest damage level seeded in the bearing was subsurface yield. This is of great relevance to the wind turbine community, as detecting this level of damage is currently not feasible. Features from Acoustic Emission (AE) measurements were used to train a Gaussian mixture model. It is shown that the likelihood function of this model can correctly predict the presence of damage.

Contents

Acronyms

ADC Analogue-to-Digital. 147

AE Acoustic Emission. iii, 6

AIC Akaike Information Criterion. 165

AR Auto-Regressive. 83

ARMA Auto-Regressive Moving Average. 115

BIC Bayesian Information Criterion. 172

CDF Cumulative Distribution Function. 45

DBN Dynamic Bayesian Network. 82

DOF Degree of Freedom. 10

EM Expectation Maximisation. 22, 31, 36, 59

EMA Experimental Modal Analysis. 19

EOV Environmental and Operational Variation. ii

EVS Extreme Value Statistics. 22, 49

GMM Gaussian Mixture Model. 62

HMM Hidden Markov Model. 62

IEPE Integrated Electronic Piezo-electric. 154

IR Inner Raceway. 154, 166

MCD Minimum Covariance Determinant. 35, 138

MCMC Markov Chain Monte Carlo. 31

MDOF Multi-Degree-of-Freedom. 19

ML Maximum Likelihood. 31, 37

MSD Mahalanobis Square Distance. 34

NDT Non Destructive Testing. 13

NLL Negative Log-Likelihood. 121

OC Outer Casing. 153, 154, 166

OLS Ordinary Least Squares. 24

OMA Operational Modal Analysis. 20

PCA Principal Component Analysis. ii

RMS Root Mean Square. 159, 160

SDOF Single-Degree-of-Freedom. 19

SHM Structural Health Monitoring. ii, 1

SLDS Switching Linear Dynamical System. 120

SSI Stochastic Subspace Identification. 20

STFT Short Time Fourier Transform. 111

SVD Singular Value Decomposition. 21, 65

Chapter 1

Introduction

1.1 Modern-day SHM Challenges

SHM refers to the assessment of structural integrity of an engineering system. It achieves this through continuous monitoring of relevant measurements of the system. There may be different motivations for doing so, and they will vary between industries; however, the common factor is the desire to predict the potential failure of a system with a prescribed lead time. This lead time will be dictated by the industry and the ultimate goal of the monitoring. There are many different goals for implementing an SHM strategy:

1. Increasing economic output by minimising the number of failures in systems.
2. Reducing maintenance by performing smart diagnosis based on data,
3. Preventing catastrophic failure that can lead to loss of life.

The motivations will not be discussed here in detail, as such discussions can be found elsewhere [2], but the focus here will be in the challenges involved in implementing SHM strategies, and how this thesis contributes towards tackling those challenges. In broad terms there are two approaches that can be taken to perform SHM

- Model-based: a physical model of the structure is constructed, and this is used as a reference to compare in-field measurements against

- Data-driven: no physical model is built. Instead, measurements are used to establish a baseline condition of the undamaged model

There are clear pros and cons for both approaches. The major benefit of the model-based approach is the flexibility to model any kind of structural state, while the disadvantage is the accuracy that can be achieved in doing so. Clear examples of this are dynamic Finite Element (FE) models. On the other hand, the clear advantage of data-driven methods is that relatively complex structural states and their response to the environment can be captured, provided they can be measured. The disadvantage is thus the impracticality and cost incurred in obtaining data from real world damage scenarios. The work presented in this thesis deals purely with the data driven approach.

SHM in all relevant industries including civil, aerospace, automotive, and power generation are concerned with damage detection, diagnosis and prognosis in structures where the operating conditions change over time, the input excitations are not necessarily known, the structures might not behave in a linear fashion and their dynamics might also change according to the environment they operate in. These conditions call for algorithms that are robust to these changes. A well accepted hierarchical structure for damage identification is the Rytter scale [3], which breaks the problem down to four levels:

1. Detection: Is damage present?
2. Localization: What is the physical location of the damage?
3. Severity: What is the extent of the damage?
4. Prognosis: What is the remaining useful life in the structure?

The damage identification problem will normally become more difficult as the diagnosis level increases in Rytter's scale. In laboratory conditions, where the excitation may be stationary and the environment unvarying, a level 1 diagnosis is now a well-understood problem (in the case of linear structures) [2]. However, achieving a level 1 damage identification is a challenge if the structure operates within a changing environment, operation and/or exhibits nonlinear dynamics. This is precisely the focus of this thesis.

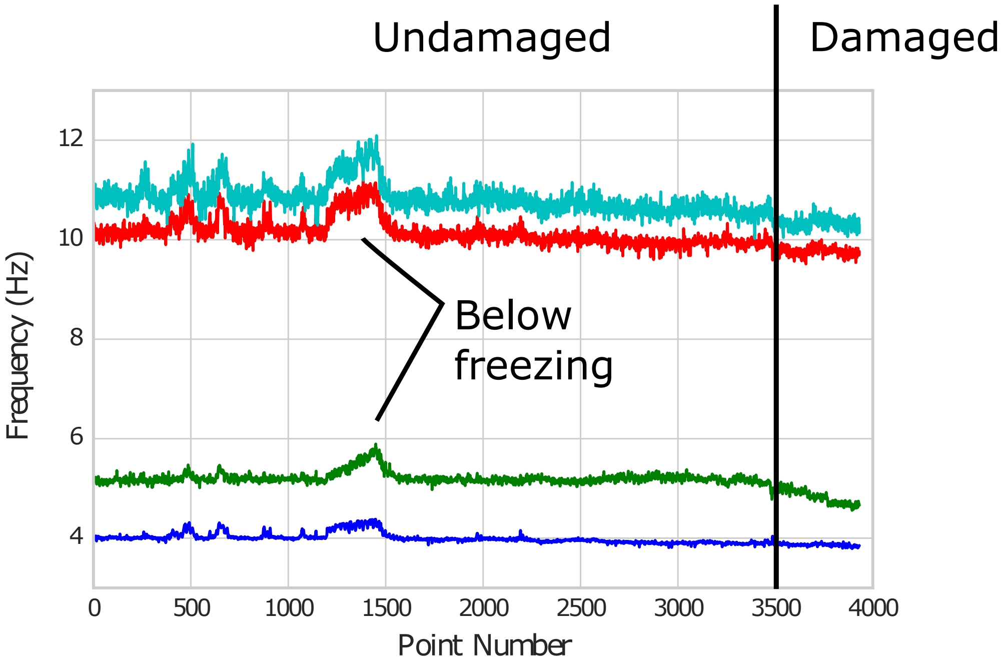

Figure 1.1: First four natural frequencies of Z-24 bridge. Note that after point number 3500 damage was introduced to the bridge girder, and so a slight decrease in the second natural frequency is evident. Note also the abrupt changes between points 1000-1500, these correspond to temperature dropping below freezing

It is a well-established principle that one cannot measure damage directly with a sensor, but its presence can be inferred from features derived from raw sensor data [4]. A statistical pattern recognition algorithm is bound to be more discriminative of damage if the raw data gathered from a structure is first pre-processed to generate features that are sensitive to damage. As an example, consider the case of the Z-24 bridge, which will be used as a case study later in Chapter 7. The natural frequencies for this bridge are shown in Figure 1.1, where the effect of damage and Environmental and Operational Variabilities (EOVs) is highlighted. Given that there are a total of 4000 observations in this case, applying a statistical pattern recognition algorithm on these four natural frequencies, per observation, instead of the original thousands of time domain points is more sensible given the lower number of dimensions.

1.2 EOV Challenges

Performing levels one to four of damage identification is a solved problem in certain materials with carefully controlled conditions; when the structural excitation is constant and when the structure behaves in a linear regime (whether this concerns structural vibration or ultrasound wave propagation). The literature surrounding this type of problems is now vast; however, relatively little of this research concerns performing damage detection under changing environments and operations. Changes in environment and operation, relevant to SHM, include anything that affects the dynamic response of the system, and this will be vary across industries.

In an aerospace setting, operational changes are possibly the most prominent, and they can be a result of:

- Physically changing the configuration of the vehicle. Changes as simple as changing payload may modify the mass and stiffness characteristics. This implies a sudden discrete change in the dynamics.
- Fuel consumption will change the total mass, and automated balancing systems may change the mass distribution; this presents a slowly changing trend in the dynamics.
- Changing loading conditions will affect the dynamic response. This can be particularly severe in rotary wing aircraft, where harmonic excitation from blade passage may be subject to small changes, and may also be close to natural frequencies. Varying loading presents a challenge in the aerospace community where significant effort has been paid to develop Operational Modal Analysis (OMA) methods to identify modal parameters from operational or natural loading [5, 6] and some of these have focused on fault detection [7].

Furthermore, aerospace structures can be subjected to temperature changes, resulting from the different atmospheric conditions at ground level compared to high altitude. The effect of temperature changes is aggravated by the fact that critical structural components in most modern aircraft use plastic reinforced composites. It is well known that their mechanical properties are subject to change with respect to temperature, and this will change the dynamics. A common well-understood stiffness change is the stiffness change resulting from the glass-rubber transition of some polymers. The EOVs emulated in Chapter 6 using a simulated dynamic system with changing stiffness were designed with these types of material behaviour in mind.

Civil infrastructure could also see significant benefits from the development of SHM algorithms that can deal with EOV's. There have been a range of research efforts in understanding, characterising and removing the effects of EOVs in civil infrastructure for SHM applications [8, 9, 10, 11]. Within Civil domain, one of the most critical, yet difficult problems, is that of bridge monitoring. Potentially large amounts of traffic can flow through a bridge on a daily basis; the loading is complex and sometimes uncertain (wind, waves and traffic), and the environment can cause a significant change in the system dynamics. Investigating the application of data-driven SHM algorithms in the case of bridges is further complicated by the requirement for data. Validating an SHM algorithm would require data gathered from a damaged condition, in order to demonstrate the capability to detect damage. Nevertheless, test campaigns have been conducted where a bridge is intentionally damaged during whilst data is being collected. This thesis uses data gathered on the Z-24 bridge in Switzerland; it has been the subject of numerous studies into SHM for civil infrastructure, as data was gathered for an entire year, in which varying temperatures caused the stiffness of the concrete deck to freeze. Both the introduction of damage, and the freezing temperatures result in a change of the natural frequencies, and hence the interest in the civil SHM research community in this dataset [12]. The first four natural frequencies of the Z-24 bridge are shown in Figure 1.1, where natural frequency changes due to temperature and due to damage are highlighted. The Z-24 data set has already been extensively used in academic papers for the proposal and demonstration of new SHM algorithms [13, 14, 15, 16, 17, 18]. In this work it is also used to demonstrate the use of likelihood inference from probabilistic graphical models. The Z-24 dataset presents a good case study in SHM as the features extracted from it are simple (natural frequencies) but their relationship with temperature has proven challenging to model. The Z-24 problem will be discussed with greater detail in Chapter 7. Some of the Z-24 data is also used as an illustrative example when describing Bayesian networks for modelling static data in Chapter 4.

EOVs also introduce a barrier for SHM in other industries, and one key industry is wind energy. There are clear economic motivations using advanced monitoring systems in wind turbines that would be capable of detecting faults in various components at an early stage. The economic argument does not form part of this work, but it is worth noting that, according to the European Wind Energy Association

(EWEA); in 2015, wind overtook hydro-power in the European Union (EU), and accounted for 15.6% of total energy production [19]. There was also a 6.3% increase in wind energy installations. Meanwhile, the demand for global energy is expected to rise 21% annually until 2021 [19]. Even though wind has not overtaken gas and coal, wind installations are on the rise, and manufacturers are responding by investing heavily in high-capacity turbines. Currently, gearbox failures amount for the highest downtime (not necessarily failure rate) in current wind turbines, although that is likely to change in the future with the introduction of direct drive-turbines. However, that still means that there is an interest in monitoring bearings, as in current turbines, bearing subsurface damage is often the root cause for gearbox degradation and failure. Monitoring technologies developed for bearings are also useful for next generation direct-drive turbines, as they still require bearings for the main rotating components.

Shifting the attention back to EOVs in wind turbines; of the critical components that need to be monitored, blades and gearboxes are both susceptible to dynamic responses that are dependent on temperature and other external variables albeit in different ways. The problem of data-driven damage detection on wind turbine blades has been studied in laboratory conditions [20], although implementations in operation have yet to see successful application. The blade operates over a range of possible pitch angles, which are adjusted according to the wind speed. The loading conditions and therefore the dynamic response is dependent on this very basic quantity. Furthermore, blades are prone to icing of the leading edge, which not only reduces their efficiency but also adds mass, and therefore changes the dynamic response. Whether vibration-based or other methods such as AE or Ultrasound are used, the data-driven algorithms used downstream must be able to account for this changing environment.

The general point being made here is that across industries, the level 1 damage identification process of Rytter's hierarchy is significantly complicated by EOVs introduced in the real world. In general these changes manifest themselves as a modification to the dynamic response of the system. To put it in simple terms, this may confuse a data-driven algorithm, and the solution is to create algorithms that can carefully extract the difference between an EOV trend and damage. Various methods have been suggested for separating the effect of EOVs from the effect of damage. The current engineering practice for dealing with this is often ad-hoc, and involves an expert engineer splitting data according to different operating and environmental regimes. In the modal example, this could involve manually splitting natural frequency data according to different temperature ranges, or even performing a different modal analysis strategy according to the environment. In an aerospace example this could involve an engineer sifting through data to identify different flight regimes, or creating a look-up table.

From a machine learning point of view, if measurements from the EOVs are available this is a supervised learning problem. Regression algorithms could be used, with EOVs as inputs, to remove their effect on the dynamic response. The drawback is that this requires measurements of the variables influencing the dynamic response, and these are often not available. The alternative is to treat this as an unsupervised learning problem, where the probability density of the data is used to model the trends in the dynamic response alone.

Some methods suggested to remove environmental variation in this context include the use of Principal Component Analysis (PCA). This linear transformation reduces a multivariate data set to a smaller number of variables that explain most of the variance in the original data. It has been suggested that removing the principal components that explain lower variance levels and then reconstructing an inverse PCA transform to reconstruct the data may sometimes remove the effect of environmental variation [21]. The success of this approach depends directly on the variance of the environmental effects compared to the variance of the regular dynamic response.

Another approach suggested for removing EOV trends is the use of cointegration [8, 22]. This method, rooted in econometrics, uses regression to identify common linear trends in variables within a dataset. Within the econometrics community, this is a popular way to determine the degree of cointegration between variables. Cointegration reduces a nonstationary data set, to a stationary white noise process. In the context of SHM, if the effect of the environmental or operational variables is incremental, such as changes in temperature or the decrease in mass from fuel usage, their effects can be removed using cointegration [22]. The classical formulations of cointegration [23, 24] use linear regression to model relationships between variables, however this may not be suitable for some SHM problems. In fact it has been shown that linear cointegration is not suitable for the removal of environmental trends from the Z-24 dataset, where the different temperature regimes require different cointegration models [8]. There are currently nonlinear extensions to this approach being suggested [25], to deal with this issue and to make this cointegration more flexible.

While these approaches are well suited to the removal of environmental and operational effects with incremental changes, they would fail to capture trends in dynamic responses containing multiple regimes. For this, a clustering approach is more suitable. One of the most straightforward ways of clustering data together is with the k-means algorithm [26], but this yields hard boundaries between data sections; it is not probabilistic. Having a probabilistic model allows one to quantify the level of uncertainty of a prediction. The Gaussian mixture model is a natural extension of this that yields soft decision boundaries between clusters and is, as such, a probabilistic clustering algorithm. Some studies have looked into the use of Gaussian mixtures for characterising environmental variability in SHM [11, 27]. However, they do not see the Gaussian mixture as under a general framework of Bayesian networks. One general approach to novelty detection (not in an SHM context) that makes use of this general density estimation interpretation is [28]. However,this view can be extended to a wide range of models. This thesis makes the argument for doing so in SHM.

1.3 Scope of this Thesis

The value of algorithms such as Gaussian outlier detection, PCA, Factor Analysis, Kalman filters and others for damage detection has been recognised in SHM research [29, 30, 31, 32, 33, 34, 20]. In their simplest forms, they can all be seen as special cases of linear Gaussian models, which have been unified in a general framework by Roweis and Ghahramani [1]. These models share various common aspects. They are all *generative* latent variable models, so that the data is seen as being generated by a set of latent, unobserved variables. In this context the “data” refers to features extracted from measurements of the dynamic response of the structure, and the latent variables could either be mathematical abstractions or actually have physical meaning, such as the state space of a Kalman filter which could relate to velocities and accelerations. Seeing these algorithms as generative models allows for the use of the machinery of Bayesian networks to be used for the computation of probabilities. These probabilities are not being exploited for SHM purposes. The reason this is desired is because it provides another level of insight as to how the model represents the data. It also simplifies the extensions of models to account for more complex data. In the case of SHM, this complexity comes from EOVs. The contribution of this work to the pool of SHM research is to show how Bayesian networks can be used in the general setting of damage detection. By doing so, simple extensions to existing models provide a means for dealing with data from multiple dynamical regimes that arise from EOVs. The focus is narrowed to the application to damage detection, even though in reality there is more information one could extract, such as localisation and damage severity. Furthermore, the focus is limited to extensions of well-known linear Gaussian models, setting as a motivation the investigation of more complex models under this viewpoint.

The first part of this thesis, Chapters 2 and 3 are introductory. The current problems facing SHM, namely environmental variability, have already been outlined in this Chapter. Chapter 2 provides an outline overview of the current standards of data and features used for SHM inference. It also provides a brief overview of machine learning concepts that are key to the work presented in this thesis. Chapter 3, introduces the use of the likelihood function for novelty detection, and some important aspects to consider, such as the setting of thresholds. This is a key chapter as it provides both a background to the interpretation of thresholds, and a discussion of the different ways of setting this critical decision boundary.

Chapters 4 and 5 of this thesis will then take a deep dive into Bayesian networks and their application to SHM. Various well-known algorithms have a probabilistic interpretation, captured very well by their graphical model representation. A strong focus is thus placed in evaluating the probability of data given a model, but the treatment is split into two chapters. Chapter 4 discusses models for *static* data, where no temporal relationships exist between variables. This constitutes the first part. Chapter 5 discusses inference when the data is better explained as having temporal relationships, or dynamics.

Two real-world case studies are examined. The first is the well-studied Z-24 bridge, presented in Chapter 7, for which natural frequencies have been recorded for several months, including periods of freezing temperatures as well as damage. The second case study is the fault detection of a wind turbine gearbox using AE data, in Chapter 8. The Z-24 bridge data set has been extensively studied, so the approach taken here is put in the context of this well known data set. On the other hand, the wind turbine AE data consists of a new experimental investigation and presents novel results, namely the ability to detect subsurface damage on a wind turbine bearing. For this reason, Chapter 8 includes a description of the experimental procedure.

In order to explore the application of these methods to a dynamical system under changing loads and operation, Chapter 6 presents a case study using simulated data from a three Degree of Freedom (DOF) mass-spring-damper system. The reason for presenting a numerical simulation instead of a laboratory, or field investigation, is because effects such as nonlinearities and changing global stiffnesses (akin to those observed with temperature changes) are easier to explore in this setting. The chapter focuses on comparing the use of static data models such as PCA, factor analysis and mixtures of these, with their dynamical counterparts: Kalman filters, and mixtures of Kalman filters.

Throughout, attention is placed on the use of the likelihood function as an assessment of data probability against learned models, as a damage detection strategy, albeit with different models and features. This message will be repeated throughout, but it highlights the overall strategy being explored, using these models for what they are: simple variations of each other.

Chapter 2

Overview of data-driven Structural Health Monitoring

2.1 SHM data

The work presented in the next few chapters presents a particular viewpoint for performing data driven SHM. While the core of this work relates to the probabilistic interpretation of a certain class of models, the data that these models use plays a central role. This section will discuss the different types of data that may constitute “SHM data”.

As discussed in [4], sensor data cannot indicate damage directly, but features derived from such data can. The following subsections thus will give a brief review of SHM data as well as some of the common damage-sensitive features that can be derived from this data.

2.1.1 Vibration

Vibration-based SHM is now a very popular technique for assessing the state of an engineering system, with a significant part of SHM research and industrial applications being devoted to using vibration measurements to infer the presence and location of damage; its development dates back to the 1990’s [35].

In SHM, vibration monitoring is based on the premise that an adverse change to the structure will cause a change in the dynamic response, which should be quantifiable using vibration measurements [4]. This change in the dynamics often manifests itself in changes to measurable parameters such as mode shapes, natural frequencies or damping. The Z-24 bridge data shown in Figure 1.1 (dealt with in more detail in Chapter 7), which consists of four natural frequencies is a good example of such a process. There are a several steps required to go from a raw vibration data stream to a natural frequency, and there are many ways one could go about it. In general, this step should involve some form of modal analysis, which is a system identification process and can be done either in the time or frequency domain.

Vibration data can cover a wide range of frequencies, and this is application dependent. Structures that vibrate at very low frequencies include most civil infrastructure, such as bridges, buildings, stadia, as well as offshore oil rigs and wind turbines. The excitation source for most civil structures tends to be its own environment. Wind, traffic loading, earth movement, are common excitation sources in civil infrastructure. In wind turbines, the low frequency loading comes predominantly from the aerodynamic loading caused by the rotation of the blades, and the effects that blades have on each other. This loading tends to be transmitted through the main driven shaft into the gearbox and subsequently into the tower.

In other industries, such as aerospace, loading and resonance frequencies tend to be of much higher frequencies. In rotary wing aircraft, the main excitation source, coming from blade rotation, tends to be low to medium frequency, in the 2Hz-50Hz range where the loading is almost purely harmonic. Fixed wing aircraft tend to be excited by a number of factors, but the excitation tends to be broadband, and is typically approximated by Gaussian white noise, or other coloured noise for analysis purposes. Unless there is an aeroelasticity issue, the physical excitation sources in fixed wing aircraft come predominantly from aerodynamic loads, such as buffeting, friction, and gusts. The frequency ranges involved in this can range from medium frequency narrow-band buffeting on the 50Hz-1000Hz range, to much higher frequency broadband noise involving several kiloHertz.

Although the analysis of the condition of rotating machinery is strictly classed as Condition Monitoring (CM) [2], rotating components still generate an excitation source, which can be significant in various SHM-related problems. In rotary wing aircraft, for example, the vibration caused from gearbox and blade components has been known to cause failure of primary structural parts. In a wind turbine, the excitation generated from the blades can contribute significantly to the overall fatigue life of the tower, and other components. In general, rotating machinery, mounted on or near an engineering structure will tend to excite it harmonically, which involves more energy concentrated on a few frequencies. From the point of view of this work, the resulting vibration can be useful for SHM, by assessing the change in which a structure transmits such vibration. Chapter 6 will discuss some damage detection strategies on a simulated structure excited by white noise, but these are equally applicable to structures excited harmonically.

The piezoelectric accelerometer is arguably the most popular and practical instrument for measuring vibration. The shapes and sizes vary according to application and cost, but they all generally rely on the same principle: to transmit a mechanical acceleration into an electrical signal through the piezoelectric effect. An alternative method for measuring vibration is a laser vibrometer. Laser vibrometry relies on the Doppler effect on the reflection of a laser beam on a moving surface, to measure velocity normal to a surface. This technique solves some of the main drawbacks of accelerometers. Being non-contact, it does not add mass to the system and does not require a mounting point. Furthermore, large areas of a structure can be scanned with minimal setup. This can save large amounts of test time compared with setting up accelerometers, if the channel count is high. The main drawbacks of laser vibrometry are the fact that the surface has to be reflective and in line of sight, and the laser setup is more suitable for laboratory than operational environments.

2.1.2 Acoustic Emissions

Acoustic Emissions (AE) are high frequency stress waves released from a material when the internal structure undergoes a change. These waves are recorded as bursts, and can be generated by processes such as friction, corrosion, stress, and growing cracks. Because the application of stress leads to the generation of AE, this is a popular technique in Non Destructive Testing (NDT) for assessing the loading history of a structure thanks to the Kaiser effect [36]. Kaiser discovered, in the 1950's [36] that a structure will emit AE if loaded up to a stress that it hasn't been loaded to before. Any subsequent application of stress will result in much reduced AE levels.

Before continuing, and to avoid confusion, it is worth establishing some contrast between what is classed as structural vibration, and AE. Physically they are both the same phenomena, but the excitation source and frequency are much different. In theory, structural vibration can go up to any arbitrarily high frequency, but practical constraints limit vibration analysis to the tens of kilohertz range. Unless a structure is unreasonably stiff and lightweight, its first few natural frequencies will lie within the 0-10kHz range. Mechanical excitation within this frequency range tends to be associated with environmental loads, shock, and rotational motion. This is thus classed as structural vibration.

Going into the tens of kiloHertz range, one finds that natural mechanical excitation arise from completely different sources. To generate a wave at this frequency, an impulse would have to be much shorter than the average impulse used, for example, in modal hammer testing. Micro-cracks tend to generate very short impulses, sending mechanical stress waves across the material, and this is what is referred to as AE. When generated from material dislocation, AE can be observed as discrete bursts, or as referred to in the AE literature as *hits*. Figure 2.1 illustrates a series of AE hits generated from a yielding steel sample (taken from [37]).

A lot of research followed on from Kaiser's original thesis, and it is now a well understood fact that stress causes AE. This fact is particularly useful given that a crack introduces a discontinuity in a material, and therefore cause stress concentration. On the other hand, crack growth estimation methods rely heavily on stress concentration factors to estimate residual life. There is a strong link between AE and crack growth; this has been made a long time ago [38]. The strain energy release rate from a crack can manifest itself as stress waves. This quantity is thus well correlated to the count of discrete AE hits and their energy.

Because AE is related to high frequency stress waves generated by micro-cracks, it is also a useful diagnostic tool. This is application area of Chapter 8, where AE is used to detect cracks at the onset stage on a wind turbine bearing. This is particularly challenging due to the background noise present in gearboxes in operation. However, in less noisy applications AE has been shown to be very useful technique at detecting damage, especially when combined with good signal processing strategies and machine learning algorithms. Some good examples of this can be found in [39, 40, 41].

Compared to vibration-based damage detection, there are few examples of the use of machine learning methods to make diagnostics from AE data. Even though

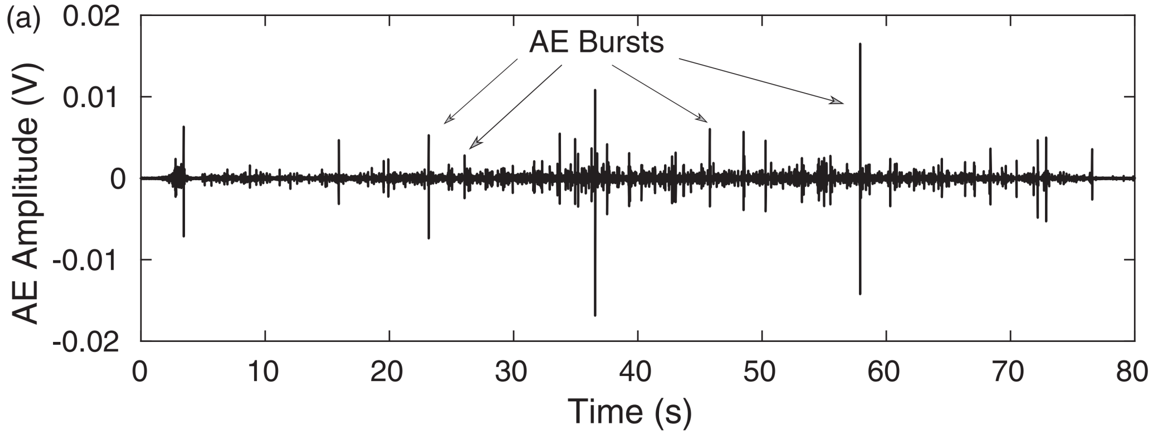

Figure 2.1: Illustration of AE bursts, generated from a yielding steel sample [37].

in principle AE has the capability of discerning damage at the micro-crack stage, the sampling rate required to to acquire these waveforms is normally in the order of 1-2MHz. This presents a data storage challenge, and is possibly the largest contributing factor for the gap in machine learning research into the subject. Most methods (including this work) capture AE hits via some form of threshold and store only features based on these. One interesting approach suggested in [42] has been to view the problem from the point of view of epidemology. It uses a spatial scanning statistic to detect abnormal activity. The approach is to model the rate of incidence of AE hits; in [42] this is done using a Poisson distribution, which is appropriate for modelling processes where an incidence rate is involved.

One of the advantages of using AE over vibration, is that the source of damage can be located much more accurately. If multiple sensors are used, the time-of-flight difference between different sensors can be used to find the spatial location of the acoustic source. If the geometry of the material is simple and the material properties isotropic, then all that is required is the wave propagation speed and a triangulation scheme. If the geometry is complex, a look-up table approach has been suggested [43] called Delta-t mapping. Another, better method based on Gaussian Process regression has been suggested in [44]. Both of these methods, require an example data set of time of arrivals with known source locations.

It is also possible to approximately detect the source location using a single sensor. This approach relies on the fact that acoustic waves will often propagate in different modes through the material. In a solid, sound is likely to generate both bulk (longitudinal) shear (transversal) and surface waves. All wave modes travel at different speeds and carry different amounts of energy. Most AE damage sources are likely to generate bulk, shear and surface waves. Out of these three modes, bulk waves travel fastest with the least energy, followed by shear waves and surface waves, which go slowest but carry most of the energy. This difference in speed makes it possible to compute roughly how far a wave has travelled.

The quantification of the distance travelled by an acoustic wave is carried out implicitly in the wind turbine bearing damage detection procedure outlined in Chapter 8. This is done merely by including a particular AE feature into the analysis: hit rising time. The time it takes from the first arrival of the wave (longitudinal mode) to the highest peak (usually shear or surface wave) gives a measure of how far the wave has travelled.

2.1.3 Time Series Models

Time series models provide a means for extracting features, and forecasting time series purely in the time domain. Chapter 5 deals with a particular interpretation of time series models as temporal Bayesian networks which allows their likelihood function to be used as a damage detection index in a consistent manner across different models. Arguably the most popular time series model is the linear Auto Regressive (AR) formulation, which views a signal $y(t)$ as a linear function of its previous p values

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \eta \tag{2.1}$$

where $\theta_{AR} = \{a_1, ...a_k\}$ are autoregressive coefficients, which effectively encode a spectral representation of $y(t)$, and η is a noise term. This is essentially a linear regression problem on lagged versions of the signal, so the AR coefficients can be estimated with Ordinary Least Squares (OLS) regression. Just as in the case of a frequency domain representation, these coefficients can highlight a change in the dynamics, and so have been extensively studied as damage sensitive features [45, 35, 46, 47, 48]. Figure 2.2 illustrates the AR coefficients of a 3-DOF mass-spring-damper system with and without damage.

Time series models distinguish themselves from spectral models since they can be used as predictors, and this can be readily used by SHM algorithms. A prediction is a one (or multiple) step ahead forecast of where the signal will be in the future, so one could use the model residual, ϵ, as a damage sensitive feature. The premise being that a change in the underlying dynamics would cause a change in the "true" AR coefficients of the system. Any predictions on this, using a baseline undamaged AR model, will result in an increased model discrepancy, ϵ.

Most of the early studies into the use of auto-regression focus purely on the damage detection and localisation problem with no external influences [35, 48]. However, attention has shifted towards SHM under changing environmental and operational conditions. In this instance, a residual on a simple linear AR model is not suitable.

Cointegration has been suggested and demonstrated as a method for removing cumulative trends in SHM data [8], and this method could be readily applied to AR coefficients or to residuals of model predictions. This is ideal when the external influence changes slowly and somehow affects all of the damage-sensitive features. An ideal example of this is temperature, which exerts its influence on structural stiffness slowly and globally. The gradual decrease in mass due to fuel usage on an aircraft is another good example of where cointegration may succeed. However, cointegration is not well suited to the problem of discrete changes, caused by operational variation. For example, if a structure were to change its dynamic characteristics due to abrupt changes in loading, mass or stiffness. A mixture modelling framework is better suited for modelling this type of EOVs, and this point is emphasized in this thesis. However, this will be saved for Chapter 5.

Extensions of the linear AR model exist to account for Moving Average (MA) terms. For example, a linear mass-spring-damper system excited by unknown Gaussian noise can be captured by an ARMA model. These can be extended to account for external or eXogenous inputs (ARMAX) [49]. However, a linear AR, ARMA, or ARMAX model would fail to model nonlinear dynamics, and thus some extension of this model is required if the system in question behaves nonlinearly. If the nonlinear parametric form of the underlying system is known, then this should be used. One of the most general models is the Nonlinear Autoregressive Model with eXogenous inputs (NARMAX) [50, 51]. The exogenous or MA terms can be dropped if it is apropriate for the application. The functional mapping between a signal $y(t)$ and its lagged values $y(t-p)$ for a NARMAX model can really be anything from nonlinear polynomials to more complex Neural Network [26] or other nonparametric forms such as Radial Basis Function networks or Gaussian Process regression [52, 53].

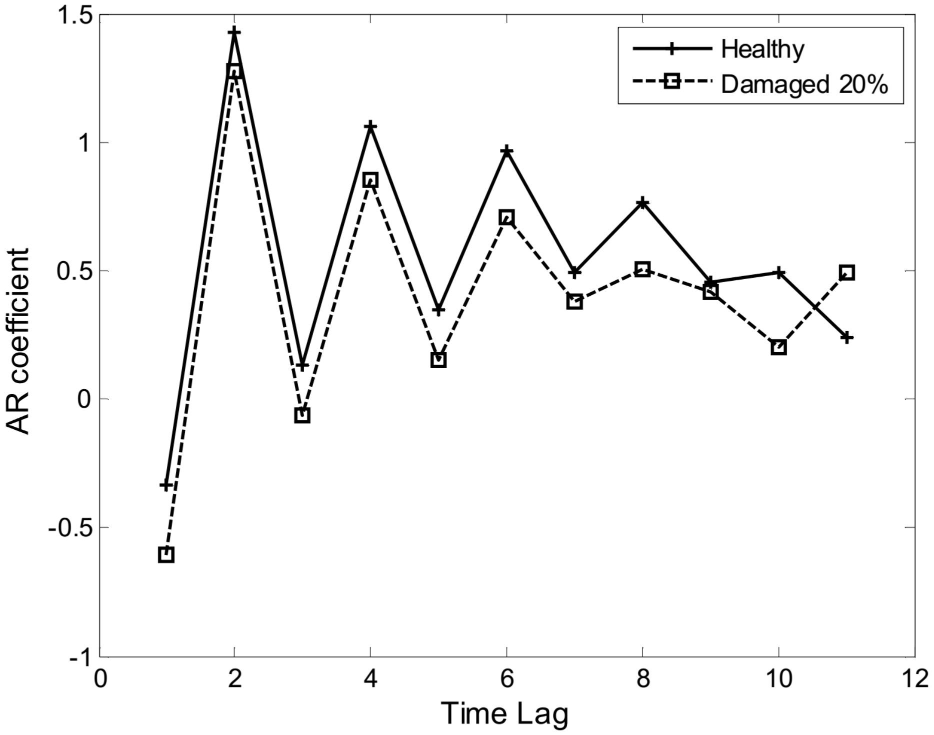

Figure 2.2: Illustration of AR coefficients from an undamaged and damaged system, where damage is represented by a stiffness reduction on a 3-DOF mass spring damper system.

Some examples of applications of black box NARX models to SHM can be found in [54, 55]. The key difference between the general NARX framework and the dynamic Bayesian network implementation is the way in which the state estimates are propagated forward. Dynamic Bayesian networks rely on Bayes' theorem to propagate the probability distribution of the state of the system forward, and this probabilistic interpretation can be exploited for SHM purposes. NARX and NARMAX models have a close relationship to the dynamic Bayesian networks of Chapter 5. In effect, there is nothing preventing the use of a Bayesian network to propagate the probabilities of a general NARMAX state space model. The next section deals with the basic formulation of modal analysis, a classical structural dynamics tool, in a state space form.

2.1.4 Modal Analysis

Modal analysis is one of the most popular techniques to analyse and understand the dynamics of engineering structures. It is particularly well suited to the analysis of Multi-Degree-of-Freedom (MDOF) systems. The underlying idea behind modal analysis is the principle of linear superposition, and the goal is to represent the MDOF response of the system as a linear combination of Single-Degree-of-Freedom (SDOF) systems. The advantage of doing so is that the response of a (linear) system can be fully described by a set of mode shapes, natural frequencies and damping ratios. This is a result of orthogonality between the mode shape vectors.

Modal analysis falls under the general category of system identification techniques, and there are various ways of performing modal testing to extract the modes. Experimental Modal Analysis (EMA) is now a mature field, used in industry primarily to understand and evaluate structural dynamics with the purpose of avoiding certain resonance frequencies, managing structural damping to control fatigue life, and to validate and update Finite Element Models (FEM) [56].

The usefulness of modal analysis in damage detection and localisation problems has been identified a long time ago [57, 58]. Early studies have focused on examining the link between modal parameters and the structural degradation process. Natural frequencies extracted through EMA have been identified as a primary feature, as they will tend to decrease as the structure is degraded. The curvature of mode shapes has also been identified as a feature that is useful for localising damage [57].

Furthermore, FEM model updating techniques have also been explored as a means of detecting and locating damage [59].

One of the key aspects of EMA is that input excitations to the system are available for modelling and analysis. This is relevant, as extracting modal parameters involves finding a set of parameters for the system equations of motion that agree in some way with the data being measured. This is termed *system identification*, and it is (now) relatively simple to do this using Frequency Response Functions (FRFs) which measure the input-output relationship between the forcing and the acceleration response. Modal parameter estimation is often done by finding a set of analytical FRFs that provide a good fit to the measured FRFs.

In order to compute an FRF, a measurement of the input force is normally required. However, this input loading may be very difficult to measure in most practical engineering applications, which is the main reason why EMA is usually confined to laboratory settings. Methods have been developed to estimate modal parameters without the need for the measurement of input loads, and they are classed as Operational Modal Analysis (OMA). One of the most popular techniques for doing this is Stochastic Subspace Identification (SSI) [60], which fits a state space model to the vibration response of the system. One of the key assumptions of this method (and most OMA methods) is that the loading can be approximated by white Gaussian noise.

The concept of OMA is very relevant to this thesis (hence the attention). The Bayesian networks discussed in Chapter 5 are in fact state space models, and one of the key points in this work is highlighting the value of prediction error in these models for SHM. Furthermore, Chapter 7 makes use of modal parameters extracted from the Z-24 bridge, using SSI, to illustrate a damage detection procedure based on Bayesian networks. A linear Gaussian state space model is defined as

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w} \quad && \mathbf{w} \sim \mathcal{N}(0, \mathbf{Q}) \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{v} && \mathbf{v} \sim \mathcal{N}(0, \mathbf{R}) \end{aligned} \tag{2.2}$$

where $\mathbf{x}$ is a vector defining the state of the system and $\mathbf{y}$ is the vector of observations (typically accelerations, in OMA). The terms $\mathbf{v}$ and $\mathbf{w}$ define the observation and process noise variance, which are modelled as zero-mean Gaussian noise processes.

The square matrix **A** defines the time evolution of the state vector **x**, while **C** relates the observations **y** to **x**. If the equations of motion for a system are given, it is a straightforward procedure to assemble **A** and **C**. The eigenvalues of the state transition matrix **A** are related to the natural frequencies of the system, and the eigenvectors can be mapped into the mode shapes through **C**. SSI belongs to the general class of subspace identification methods. In broad term, the procedure is to first find the observation matrix **C**, through theSingular Value Decomposition (SVD) of a block-Hankel matrix formed using stacked observation vectors **y**. Once **C** is known, the state vector (for a given observation set) can be computed, and **A** can be solved for using linear least- squares regression.

A lot of attention will be devoted in Chapter 5 to the *Kalman filter*. This is an algorithm for making optimum estimates of the state of a system. State estimation plays a central role in many engineering fields, and the Kalman algorithm is rooted in automatic control engineering. However, in SHM, its value often lies in its prediction errors. These highlight discrepancies between the model, which represents an undamaged condition, and the measured data.

There are not a lot of publications on the use of Kalman filtering in SHM. The current literature on Kalman filtering and SHM focuses on its use as a state estimator, to perform tasks such as tracking feature vectors. For example, [33, 61] both discuss the use of a Kalman filter, and some nonlinear extensions, to estimate and track structural parameters such as stiffness and damping. In order for a Kalman filter, or a general recursive Bayesian filter to achieve this it requires the observation matrix, **C**, to represent the model structure. This is a powerful interpretation that allows for the recursive estimation of system parameters, but requires that at least the model structure be known. This is a strong assumption, and assembling these matrices will require a significant amount of knowledge about the system, which may not always be readily available.

An alternative interpretation is to *learn* the system matrices from the data, so that the error on state estimates provided by the Kalman filter give an indication of the discrepancy between model and data. This idea will be discussed in much more detail in Chapter 5. It is worth mentioning here, that there have been some similar attempts to use the Kalman filter in this fashion for SHM [17]. However, at the moment, applications of Kalman filtering within the general remit of novelty detection offer a richer description of the approach, see for example [62]. However most studies focus solely on the inference task and do not treat the overall problem of identifying an appropriate model from the data. Lee et al. [62] proposes the use of Extreme Value Statistics (EVS) (these will be discussed in Section 3.2.5) and Kalman filter inference in the performance of novelty detection. In this study, the Kalman filter is set-up to only perform inference, and an observation matrix is assembled that infers the autoregressive coefficients from the observations. Novelty detection is then carried out with the aid of EVS on the residuals of this process. Hayton takes a full approach to the work presented here, of learning the state space parameters using the Expectation Maximisation (EM) algorithm, and then using Kalman filter inference to do novelty detection [63]. This is a applied to performance data from a jet engine, with successful results. While the approach taken in [63] is similar in principle to the work presented in the later chapters of this thesis, here it is shown how this fits within a wider general class of models, which can also be extended relatively easily to allow for varying environmental and operational conditions. More details will be given in Chapter 5.

2.2 Machine Learning in SHM

This section will provide a general overview of machine learning, as well as a review of relevant and/or recent developments in the field, and in the context of monitoring.

Machine learning deals with the task of *learning* models from data and performing *inference* on data using those models. Learning is referred to as the task of identifying a suitable model, $\mathcal{M}$, and model parameters $\boldsymbol{\theta}$, that represent a data set Y. The task of model selection and parameter identification are separate, yet related. Parameter identification alone, assumes that one already knows a model that is responsible for generating Y, and can proceed to find suitable parameters for it. This, in some cases, may be a strong assumption. Model selection involves not only finding suitable parameters, but a model that best explains the observed data. Model selection thus must include parameter estimation steps and is, as such, a more complex problem.

Inference, is generally referred to as the task of making predictions with a model [1]. The type of prediction, will depend on the model, but it will generally be one of

[1] Note that inference can also refer to predictions about models and their parameters, and this terminology is often used in *Bayesian inference*

1. Classification: predict a discrete class label, given other continuous or discrete variables

2. Regression: predict a continuous variable, given other continuous variables

3. Density Estimation: predict the probability density of a data set

There are models, discussed later, that are non-parametric, and so strictly speaking, do not require a parameter identification step. Parametric and non-parametric models have both advantages and disadvantages. The chief advantage of non-parametric models is that no assumption is made about a physical model that generates the data. This is desirable, because more often than not, the physical models available to explain a data set are unsuitable, or not complex enough. This is the case, for example, if one were to fit vibration data originating from system operating in two different conditions, to a single model. However, a non-parametric model will not be very interpretable. The parameters identified from a parametric model will not only be useful for prediction, but will also give some engineering insight into the system in question. For example, SSI, discussed above for performing Operational Modal Analysis, could learn a set of modal parameters, which give the user a lot of information about the behaviour of the system, but may not be suitable if the system is nonlinear, or operating in multiple conditions.

Machine learning algorithms also are split into two classes: supervised and unsupervised. Supervised algorithms involve learning some function between an example set of inputs and outputs. Unsupervised algorithms on the other hand try to perform inference about the relationships between inputs only.

Regression and classification both involve supervised learning, as they need example data to learn a model. The principal goal in both regression and classification is to determine the relationship between a given set of inputs and outputs. The discussion below will focus on regression as it is more relevant to this work, as mostly continuous data will be dealt with

2.2.1 Regression

Regression seeks to find a mapping

$$\mathbf{Y} = f(\mathbf{X}) \tag{2.3}$$

where $\mathbf{X}$ and $\mathbf{Y}$ and continuous inputs and outputs respectively. Note that $\mathbf{X}$ and $\mathbf{Y}$ are matrices with rows of observations and columns of dimensions. Learning the functional map between $\mathbf{X}$ and $\mathbf{Y}$ given a set of training inputs, allows one, in principle, to make inference or predictions over new inputs. Linear regression is the simplest, and arguably one of the most useful regression tools in machine learning. It assumes a linear relationship between $\mathbf{X}$ and $\mathbf{Y}$ through a weighing matrix, of the form $\mathbf{Y} = \mathbf{XW}$. Ordinary Least Squares (OLS) presents an elegant closed form solution for the weighting matrix $\mathbf{W}$.

Linear regression plays an important role in SHM. A trend in the literature is to use linear regression to remove or investigate environmental trends, in particular in long term monitoring of civil infrastructure [64, 65, 8, 13]. Some of these studies extend the use of linear regression to robust implementations to visualise and separate the effect of environmental variations from the effect of damage on modal parameters [13].

Also, OLS plays a central role in time series modelling. Recall the linear autoregressive model discussed previously, where a point in a time series $x(t)$ is cast as a linear function of its previous p values $x(t-p : t-1)$ [2]. This is effectively a problem of linear regression, and the weights derived from this are the autoregressive coefficients, which have been shown to be damage sensitive features in numerous publications [45, 35, 46, 15, 14, 47, 66].

Neural networks are a popular choice of model when nonlinear regression is required [67]. The idea is to treat the regression problem the same way that a brain processes information, by passing data through a network of nodes, each of which have an activation function, and apply a weighting factor to the data. The user would normally have to select a network topology, and training involves adjusting the weights of the connections between the nodes $\mathbf{W}$ to minimise a cost function, typically a squared error. Back-propagation is a popular learning procedure, details of which can be found in [67, 26]. One of the issues of neural networks is that it may be easy to generate models that overfit. This can be avoided by cross-validation and regularisation, but this may lead to high training times and require a significant amount

[2] A semi-colon inside the argument of a vector will be used here to denote a range

of training data. Bayesian learning of network weights [68, 69] is a solution to this problem. Neural networks have been widely studied in SHM. Some key applications are their use in NARX models (discussed in section 2.1.3) where they have been used to detect damage, or changes in structural parameters in nonlinear systems [70, 71]

Alternatively, one could turn to non-parametric regression models for this. Popular non-parametric models for regression, also used in SHM include Support Vector Machines (SVM) [26], Relevance Vector Machines (RVM) [72], and Gaussian Process (GP) regression [52]. As opposed to parametric models, kernel methods rely on the idea of holding training data in memory and comparing any new test inputs to all the previously seen training points, to find the closest example in the training set. These are kernel-based methods: they rely on the use of a kernel function to impose a certain smoothness on the data.

2.2.2 Density Estimation

Density estimation is an unsupervised learning problem. There are no inputs or outputs to a model, but rather one wishes to characterise the probability density of a set of variables. Learning in this case involves estimating a probability density function $p(x)$, so that the probability of x taking on a value between a and b s given by the integral of the density function over that range[3]:

$$P = \int_a^b p(x)dx \tag{2.4}$$

In a similar fashion to regression, the function $p(x)$ that describes the probability density can be either parametric or non-parametric. Arguably the most popular form of parametric density estimator is the Gaussian distribution; in its univariate form it takes the form,

$$p(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2} \tag{2.5}$$

where μ and σ^2 are the mean and variance of the distribution respectively, and the only two parameters required to make predictions of probabilities. Taking the

[3] Note P is used here to denote a probability, while p is used to denote a probability density

Gaussian to multiple dimensions yields

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})^T\} \tag{2.6}$$

where $\boldsymbol{\Sigma}$ is now a covariance matrix, $\mathbf{x}$ is a vector of the d-dimensional data points, and $\boldsymbol{\mu}$ is a vector of means. The Gaussian distribution is just one of many distributions, albeit one with particularly useful mathematical properties. The fact that the Gaussian distribution is parametrised by $\boldsymbol{\mu}$ and Σ is, in a sense, limited; it imposes certain structure to the data and thus limits its ability to model complex data sets.

There exist various methods for performing non-parametric density estimation, such as the histogram, K Nearest Neighbours (KNN) and Kernel density estimators [73]. They do not impose a particular strict structure to the data set, but rather let the data "speak for itself". These non-parametric methods for density estimation are very useful, in particular, for exploratory data analysis purposes. Kernel density estimators are used throughout this thesis whenever a density needs to be illustrated as they impose smoothness constraints in the density function. The use of kernel methods can go beyond exploratory data analysis, as they too have a probabilistic interpretation. They have been used, for example in the context of monitoring wind turbine power curves [74].

2.3 Dimensionality Reduction

A damage-sensitive feature vector can range from low-dimensional representations, such as modal parameters, to very high dimensions, such as FRF's, power spectra, Lamb wave responses and wavelets. Dealing with high dimensions often presents a problem due to the *curse of dimensionality*, which refers to the general sets of problems introduced when analysing high dimensional spaces. One such problem is the possible sparsity that a high dimensional space may introduce when the feature of interest in the data manifests itself in only a handful of dimensions amongst a very high-dimensional space. There is thus a strong motivation for projecting high dimensional feature vector into lower dimensional embeddings. Arguably the most popular technique for this is Principal Component Analysis (PCA). PCA is a linear transformation from an n-dimensional feature vector $\mathbf{Y}$ to a lower, p-dimensional vector $\mathbf{X}$, through a rotation of the principal axes of the data that seeks to maximise the explained variance along the dimensions of $\mathbf{X}$. In other words, PCA represents a high-dimensional data set through a lower-dimensional one that explains most of the variance in the data. One of the interesting things, highlighted in [1, 75] is that PCA can be represented by a probabilistic graphical model, and this interpretation of PCA is thoroughly exploited in this thesis.

While PCA is a useful analysis tool, it also play a role in exploratory data analysis. It is always easy to visualise the relationship between two variables, through a scatter plot. Exploratory data anlysis of three, four or five variables may be accessible through such visualisation tools such as a scatter-matrix: a combination of scatter plots where all variables are plotted against each other. When the dimensions are high, and in particular when one is not certain of which variables/dimensions play an important role in the analysis, producing scatter or correlation plots may be difficult. Visulisation of the principal components of the data can solve this problem. If one arranges the principal components by order of decreasing variance, plotting the first two in a scatter can yield an informative visualisation of the patterns in the data.

Factor analysis is a related tool to PCA, with the key difference being that factor analysis is not an orthogonal transformation, so it imposes less structure on the data. Depending on the context of the analysis, this could be an asset. Factor analysis assigns individual noise variances to model the observation noise, so it is more suitable for modelling problems where different dimensions vary according to different variances. A good example of this could be a vibration process where the low frequencies undergo significant changes, while the higher frequencies remain fairly constant.

PCA and factor analysis are both linear transformations, and this could be a major drawback. One could visualise learning a PCA as an embedding of the data into a linear (flat) manifold in a state space. This represents a major restriction, if the data could be better represented by a curved, or arbitrary manifold. There are several techniques for performing a nonlinear embedding of the data, used mostly within the context of data visualisation. Kernel PCA is on popular, and relatively simple technique for achieving this [76]. It has seen some applications in fault detection, such as [77] where its ability to embed multivariate data with nonlinear relations is leveraged. In [77], fault detection is performed by comparing the low dimensional manifolds resulting from KPCA instead of generating a baseline undamaged model and comparing reconstruction residuals. This thesis, on the other hand, takes the approach of using the framework of mixtures of PCA when a nonlinear embedding is required, and focuses on the use of reconstruction errors, via the use of the likelihood function of the model.

Another notable method for nonlinear dimensionality reduction is the Gaussian Process Latent Variable Model (GP-LVM) [78]. This is an extension of PCA that uses Gaussian Process (GP) regression to generate the map between the latent space and the measured data. Because both the GP and the latent variables are unknown, an optimisation is required to uncover an optimal latent space that accurately represents the data through a GP. The problem is therefore very computationally expensive, especially considering that GP regression has a cubic complexity in the number of training data points. Unless a sparse approximation to this problem is used for the GP regression (for example [79]), this is not a practical way of visualising high-dimensional data due to the computation cost. Not a lot of investigations of GP-LVMs in the context of SHM have been carried out. An application of GP-LVMs to the problem of detection damage in wind turbine bearings using AE data is presented in [80] by this author. The likelihood function of the GP-LVM is used to successfully detect small surface damage in bearings. However, this approach is impractical due to its computational cost; the methods presented in Chapter 8 based on Gaussian mixture likelihoods are more practical whilst yielding successful results.

The use of ANNs and linear PCA for novelty detection has been thoroughly explored in [81]. Their use is demonstrated in the context of damage detection of a nonlinear system; as expected, ANNs are shown to perform better at this task than linear PCA. The approach explored in Chapter 4 is similar in principle, but uses mixtures of PCA models to enhance the linear restriction, and treat this as a divide-and-conquer problem. The GP-LVM model is also suitable for novelty detection, provided a suitable sparse approximation is used. This is a largely unexplored application of the GP-LVM model; the only publications so far this author is aware of that treat this problem are [80] and [82]. Kernel PCA is also a viable tool for novelty detection through a low-dimensional embedding. This has been demonstrated rather recently in [76], although this has not resonated in the SHM community.

2.4 Bayesian inference

There are two schools of thought for interpreting probability, the frequentist and the Bayesian. Frequentists view statistics and probability purely in terms of probability distribution functions. They interpret the probability of a as the integral under the curve of its PDF (equation (2.4). In this view, the probability of discrete event is related only to frequency of its occurrence on a previously measured trial.

The Bayesian viewpoint enhances that of the frequentist by interpreting probability as a belief. At the centre of the Bayesian school of thought lies Bayes' theorem. It relies on conditional probability $p(a|b)$, which encodes any conditional relationship between a and a separate variable. Bayes' theorem updates the belief in a given new knowledge about b,

$$p(a|b) = \frac{p(b|a)p(a)}{p(b)} \tag{2.7}$$

It is a simple but powerful relationship. It could also be described as shrinking uncertainty about a using knowledge of b, together with the relationship between a and b. The probabilities in Equation (2.7) can be either discrete or continuous, though this work will be dealing mostly with continuous probabilities. The conditional probabilities in Equation (2.7) are illustrated in Figure 2.3, for the case when both a and b are Gaussian distributed, and correlated. Bayes' theorem provides a useful tool for using conditional probability to make inference about variables, given observations on other variables. More importantly, this inference is done with probabilities, not point estimates, so a measure of the uncertainty around variables is retained. It is important here to understand the engineering applications of Bayes' theorem. This idea can be applied in a multitude of contexts, but two are of particular interest in engineering, and specifically in SHM:

- In the system identification case, when the parameters of a dynamical system have to be inferred from a set of measurements.
- In generative models, when unobserved variables need to be inferred from those observed.

The two cases are related, but some clear distinctions must be made between them. This thesis deals with the latter case, as Bayesian inference of parameters is not the topic of this thesis. The focus here is on the use of Bayes' theorem in the context of a generative models. The author's interest in generative models arises from their usefulness in capturing a wide range of data scenarios. Generative models are the

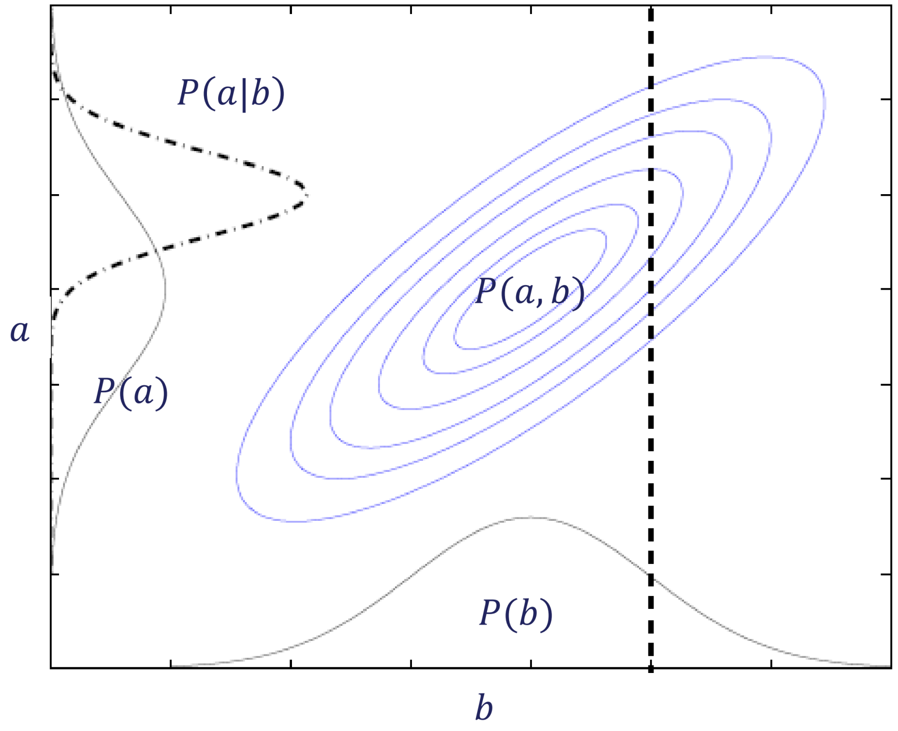

Figure 2.3: Illustration of the shrinkage of uncertainty that Bayes theorem provides, using two Gaussian distributed, correlated variables. Note how the variance of $p(a|b)$ is lower than $p(a)$, due to the evidence (dashed line) given about b.

topic of Chapters 4 and 5. In this brief section, the objective is to put the use of Bayes' theorem in the context of this work.

When applied to system identification, Bayes theorem yields probability densities for the system parameters. The parameters can be represented in vector form as $\boldsymbol{\theta} = \{\theta_1, ..., \theta_n\}$ for a system with n parameters. If the observed data is denoted as $\mathcal{D}$, the formulation in this case requires one to solve for $p(\theta|\mathcal{D})$. In other words, the probability (distribution) of the parameters given the data. If Bayes' theorem is to be used to solve for this, a prior distribution for the parameters is needed $p(\boldsymbol{\theta})$ as well as a likelihood function $p(\mathcal{D}|\boldsymbol{\theta})$. In an engineering context, the prior would ideally encode any knowledge about the system. The likelihood function gives a measure of error. Applying equation (2.7) yields

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D},\boldsymbol{\theta})} \tag{2.8}$$

The denominator can be solved for by using the sum rule of probability expressed as the integral $\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$. The chief benefit of casting the problem of learning as Bayesian inference is that it yields uncertainty over parameters. The disadvantage, however, is that for certain choices of prior distributions $p(\boldsymbol{\theta})$ the integral that describes the joint distribution becomes analytically intractable and one needs to resort to methods such as Markov Chain Monte Carlo (MCMC) or variational approximations to find it [83, 26]. Sampling methods such as MCMC can be computationally very expensive, while variational techniques are often mathematically very intricate, and offer only an approximation.

Bayesian learning of parameters solves some of the problems of traditional parameter optimisation, often based on maximising the likelihood term, $p(\mathcal{D}|\boldsymbol{\theta})$ (which encodes the model error). Maximum Likelihood (ML) parameter estimation methods could easily suffer from overfitting, and a popular approach is to add regularisation terms, such as those employed by Akaike and Bayesian information criteria methods [84, 85].

The Bayesian interpretation of learning as inference is put aside in this thesis, and instead the focus is placed on the use of the likelihood function $p(\mathcal{D}|\boldsymbol{\theta})$ as a measure of novelty, and thus damage. Part of the reason this is done is because this approach has not generally been explored so far. Likelihood inference is still interesting because it applies to a wide class of models.

Learning of $\boldsymbol{\theta}$ in this thesis is done using the now well-known EM algorithm [86], which deals well with the problem of optimising $\boldsymbol{\theta}$ as well as missing variables. In the context of generative models, the missing variables are the latent variables. This will be discussed in some more detail in Chapter 4. For now, it serves to say that latent variable models help one represent data efficiently by imposing certain structure on it. Bayesian inference lies at the centre of the formulation of generative models, through the use of the conditional probability relations between the observed data, and the latent variables.

The reader is reminded that the ultimate objective in this work is the use of the likelihood function to infer the presence of damage, so before discussing the models in greater detail, the next chapter will provide an overview of likelihood functions.

Chapter 3

Novelty detection and the Likelihood function

Novelty detection could be seen as one of the goals of performing density estimation on a data set, where one seeks to find subsets of $\mathbf{Y}$ that fall outside the nominal density of $\mathbf{Y}$. The key point from an SHM perspective is that the density estimate of the data, $p(\mathbf{Y})$, correctly captures the normal operating condition of the structure, in an undamaged state. Note here the emphasis on $\mathbf{Y}$ being a matrix, to denote multivariate observations. Data-driven SHM relies heavily on the use of novelty detection as a damage identification step. In theory, this step is made simple through the principled use of a density estimation method and an appropriate distance metric to create a novelty index. In practice, damage detection is complicated by several factors. The presence of Environmental and Operational Variations (EOVs) remains as the main obstacle to applying novelty detection in SHM. This thesis introduces the general use of probabilities in Bayesian networks to tackle this problem; they provide a convenient way of modelling probabilistic relationships between variables, making them particularly useful for dealing with multivariate data, typical of structural dynamical systems. Bayesian networks provide graphical means of deriving likelihood functions for a wide range of models.

A likelihood function, effectively captures the discrepancy between a model and an observed data set (or point). One of the arguments this thesis is trying to make, is that the likelihood function of a Bayesian network can be systematically used as an index of novelty, which is not the current practice in SHM. The formulation of several well-known models as Bayesian networks is discussed in detail in Chapters 4 and 5. This chapter does not discuss Bayesian networks. Instead, it introduces likelihood functions as a novelty index. Some of the key issues when using a novelty index are addressed, namely the underlying principle and the use and setting of thresholds.

3.1 An SHM background to Novelty Detection

A now well-established method for performing novelty detection in SHM is based on Gaussian outlier analysis, as described by Worden [29]. The idea is to use a Mahalanobis Square Distance (MSD) metric as a *novelty index*. The MSD is effectively the term that establishes the deviation between a (multivariate) measured point $\mathbf{y}_i$ and the mean of the process $\boldsymbol{\mu}$:

$$\mathbf{y}_i = (\mathbf{y}_i - \boldsymbol{\mu})\mathbf{S}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})' \tag{3.1}$$

The reader should recognise this as the term inside the exponent in the multivariate Gaussian distribution of Equation (2.6). Once one has a suitable novelty index, such as an MSD, the next step is to decide on a threshold, above which a measurement is considered to be abnormal.

There is substantial shared ground between the problem of outlier analysis and novelty detection. Both tasks deal with the problem of identifying an observation that has not been generated by the same mechanism as the rest of the observations. The same definition is true for novelty detection. The difference (at least from this author's point of view) between the two tasks lie with *which data* contains the outliers. In the machine learning context, one requires a training data set from which to build a model, and a test or validation set to prove that the model predictions generalise well and do not overfit. An outlier inside the training data set presents a different challenge to an outlier outside this set. Outside of the training set, outlier analysis and novelty detection methods perform exactly the same task (to find unlikely data points). However, if outliers are present inside the training set this could easily lead to biased model parameters. This leads to the question; what exactly is an outlier? Though several definitions exist [87, 88], it could be generally agreed upon that it is a set of abnormal measurements that constitute a minority of the data set.

Two scenarios could be considered to contain outliers, and these are illustrated in Figure 3.1. The first consists of outliers that manifest themselves as a second mode on the probability density. An example of this is given in Figure 3.1a. The physical meaning of this could be a result of a secondary mechanism generating a separate set of features, and the Z-24 data is a classical example of this. A second way in which outliers could manifest themselves is through extreme values, and this would result in a distribution with heavy tails. This type of outliers is illustrated in Figure 3.1b. There may not necessarily be a second physical mechanism generating these outliers; This kind of outlier is more likely to come from noisy observations introduced either through the data collection process or data corruption. Because heavy tails are characterised by a small amount of very large values, this is more likely to cause a small increase in variance, as well as shift in the mean towards the direction of the outlier.

Figure 3.1 shows a density estimate of the data as well as the density of a Gaussian distribution fitted to all the data (dashed blue), and the data set excluding the outliers (dashed green). This highlights the biases introduced in the parameters of a Gaussian distribution.

Some interesting work has been done to make use of robust statistical measures to take EOVs into account within a novelty detector for SHM. Dervilis, [81], argues that EOVs could be treated as inclusive outliers, and shows how one could use the Minimum Covariance Determinant (MCD) as a robust method for estimating a multivariate Gaussian distribution. The application of robust regression is also demonstrated as a robust means of removing EOVs in [13]. The philosophy of this approach is to treat those features that are generated through an EOV mechanism as outliers and to fit a simpler model to these if it is appropriate. A similar idea, but based on nonlinear robust regression, to deal with more complex SHM data has been suggested in [53]. These robust methods are potentially capable of dealing with both types of outliers presented in Figure 3.1 in a multivariate scenario.

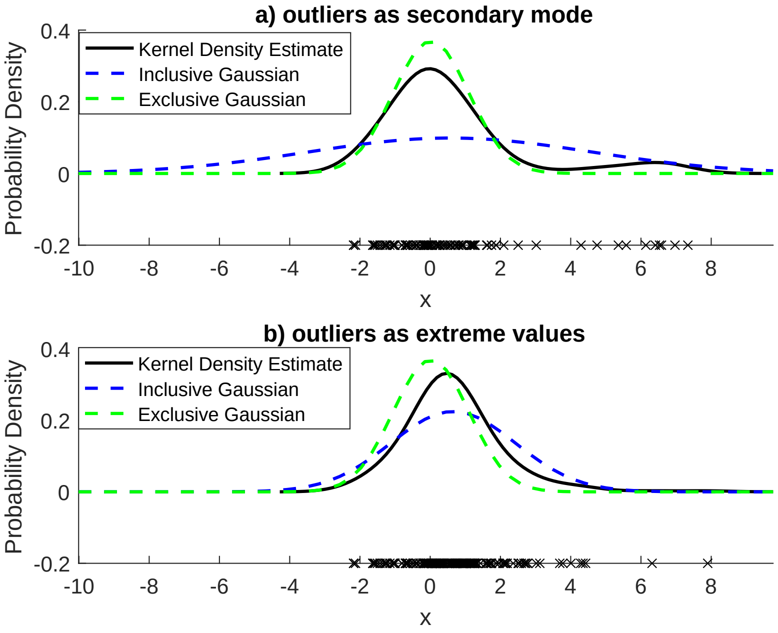

Figure 3.1: Illustration of two types of outliers. Data points are shown as well as true density and estimated Gaussian and kernel densities. a) shows data points and densities from outliers belonging to a secondary mode and b) shows outliers generated as extreme values, inducing heavy tails

3.2 The Likelihood function

The likelihood function has already been briefly introduced as a component of Bayes' theorem, but this will be reviewed in a little more detail in here. A likelihood can be interpreted as the probability density of a data set $\mathbf{Y}$ given a model parametrised by $\boldsymbol{\theta}$: $p(Y|\boldsymbol{\theta})$. This can also be interpreted as the probability density of $\mathbf{Y}$ as a function of its parameters. The likelihood function is thus a useful tool for optimising parameters and comparing models. Maximisation of the likelihood can be done either analytically or numerically, if no analytical solution can be found. This thesis resorts to the EM algorithm [86] as a numerical alternative to finding maximum likelihood solutions for models where an analytical optimum is hard or intractable. In some cases, however, the maximum likelihood solution is straightforward to derive.

When the fit of an entire data set is being questioned, it is helpful to work with the complete data likelihood. This is, in other words, the total probability of the data, under a given model (and parameters), for independent observations $\mathbf{Y}_i$,

$$\mathcal{L} = \prod_{i=1}^{N} p(\mathbf{y}_i) \tag{3.2}$$

where $\mathcal{L}$ will be used herein to denote a likelihood. In this case, the function $p(\mathbf{y})$ is simply the probability density of $\mathbf{y}$, using the desired parametric distribution, or a non-parametric estimate. Note that probability densities are *not* probabilities, and do not have to scale between 0 and 1. For large N, the product in (3.2) will be either a very large or very small number, so working with the log likelihood is helpful, in order to work with a sum over $\log(p(\mathbf{y}))$:

$$\log(\mathcal{L}) = \sum_{i=1}^{N} \log(p(\mathbf{y}_i)) \tag{3.3}$$

As an example, Figure 3.2 shows the total log likelihood evaluated on samples of a Gaussian random variable with $\mu = 2$ and $\sigma = 4$. As expected, the likelihood peaks at the correct parameters. This is a good illustration, but a rather inefficient way of actually finding the right parameters, in the case of the Gaussian distribution. The ML solution for the mean and variance as

$$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i$$
$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \mu)^2$$

are in fact the result of differentiating the Gaussian likelihood function with respect to μ and σ^2 and equating to zero. While in the case of the Gaussian distribution this is straight-forward to do analytically, more complex distributions may require numerical optimisation.

The likelihood ratio between two models is a simple and effective way of providing an overall measure of best fit, within some probabilistic context. One of the drawbacks is that a pure likelihood does not naturally penalise models of higher complexity (more parameters). One is left with rather heuristic methods for artificially penalising for model complexity such as the Akaike and Bayesian information criteria.

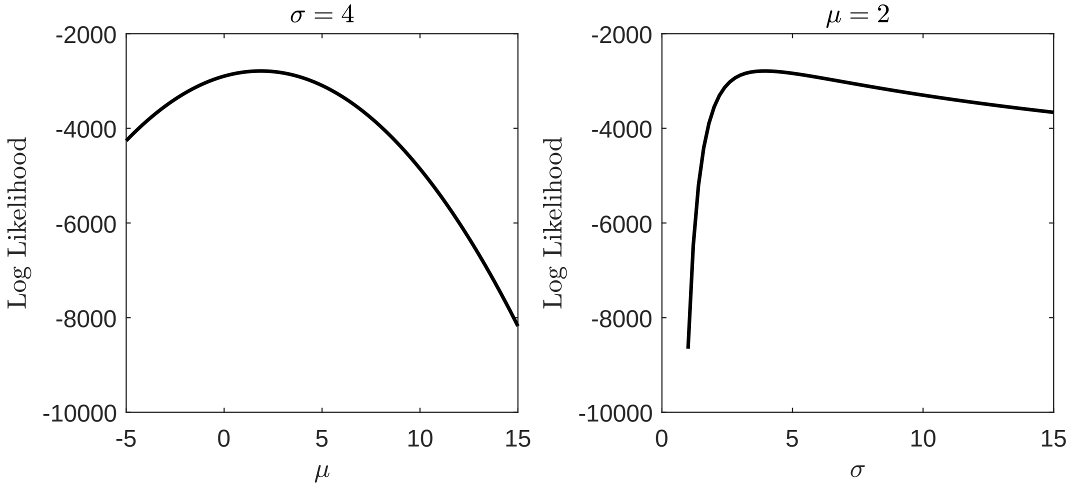

Figure 3.2: Illustration of log likelihood function as a function of parameters on a Gaussian random variable

Several things must be done correctly for a successful damage detection methodology to be in place. First, the probability density model must capture the process that generated the data. Secondly, a threshold must be selected that minimises the number of false positives and maximises the true positive rate. The correct threshold is largely dependent on the probability distribution of the novelty index, in this case, the likelihood function. For this reason, it is worth examining the distribution of likelihood estimates for Gaussian models. This makes sense at this point given that the rest of this thesis largely discusses the use of linear Gaussian models via Bayesian networks.

3.2.1 Threshold selection on likelihoods

Given the problem of wanting to use likelihood estimates as measures of novelty, the next question is the appropriate strategy for defining a threshold above which the data has changed enough so as to arise suspicion that damage may be present. Before investigating the point of thresholds, which is given in the next section, it is worth exploring in a little more detail how likelihood estimates are distributed.

The point here is that the usage of the likelihood is slightly different from the use given to it when performing parameter estimation or model comparison and selection. Much like the use of other distance metrics, such as the MSD, one wishes to:

- Identify the ML model parameters that fit the data well, with good generalisation, and with no overfitting.
- Evaluate the likelihood function on the training data set and set a threshold.
- Validate the threshold on a validation undamaged condition data-sets.
- Classify any further measurements as either normal or abnormal according to whether they fall inside or outside the threshold.

Using a likelihood in this setting is simply turning it into a distance metric, to evaluate the discrepancy of each data point against a given model of normality. The success or otherwise of the approach can be judged, overall, by the rate of true and false positives, and this in turn will depend on the correct setting of a threshold.

Now, exactly how the likelihood is evaluated, depends on the nature of the data. The complete likelihood, given by Equation (3.3) assesses the total probability of the data fitting the model. However, for the purposes of SHM, and general novelty detection, a better approach is to consider the log likelihood of each individual measurement $\log(p(\mathbf{y}_i|\theta))$, where $\mathbf{y}_i$ is the multivariate i_{th} measurement vector of Y.

If a log likelihood is evaluated over every measurement point, more points will be available to establish a threshold, and this is a desirable aspect. However, if windowed statistics such as averages or maxima are used instead, one could make use of the central limit theorem or extreme value statistics (EVS) to more easily establish a reliable threshold. Depending on the nature of the data, some compromise is needed between these two. This point is illustrated well in Chapter 6, where under a PCA model of frequency spectra, the log likelihood is evaluated for every feature vector. This is appropriate given that a Fourier transform already summarises information across a time window. This contrasts sharply against time-domain models, in particular when these are fitted to raw data. In this scenario, a subtle change in the system will induce a change in the residual (and thus the log likelihood) on average, rather than on every point. The details of this are left to Chapter 6.

Doing this also has the effect of “whitening” the distribution of the log-likelihood, which is expected from the central limit theorem. This is illustrated in Figure 3.3, where the distribution of negative log-likelihoods on a Gaussian random variable are shown; a moving window was used to evaluate the probability densities, with

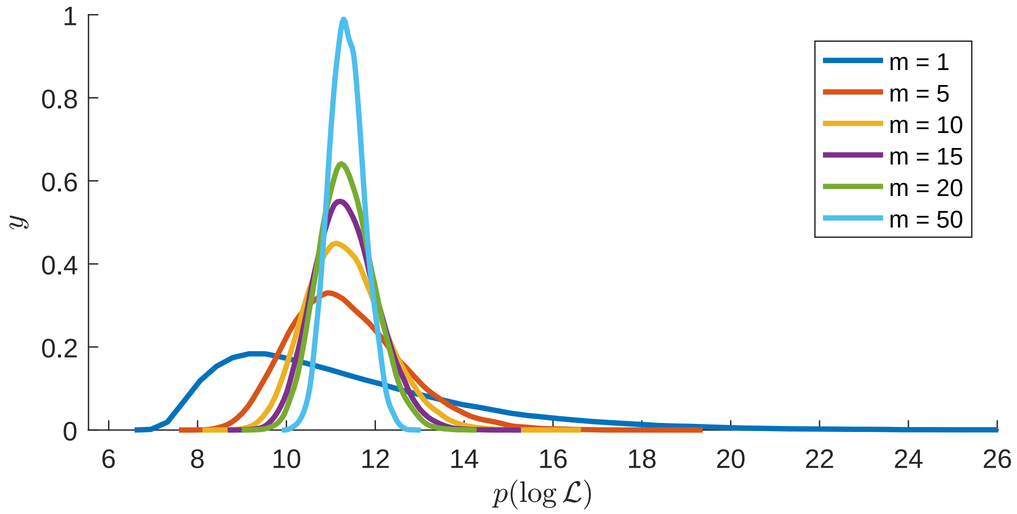

Figure 3.3: Illustration of the distribution of negative log likelihood on a multivariate Gaussian random variable, averaged with increasing window sizes, m.

increasing number of samples, m, in each window. For a Gaussian distribution an MSD is just a scaled version of a log-likelihood

The densities over the negative log likelihood (Figure 3.3) would look the same as that for a MSD measure except for a horizontal shift in the density. For a window size of 1 (no averaging), the distances have very heavy tails on the right. The log-likelihoods, evaluated with the ML parameters, can be shown to be asymptotically distributed as a $\mathcal{X}^2$ distribution with p degrees of freedom, where p is the number of parameters in the distribution, in the case the multivariate Gaussian. This is a result of Wilks' theorem [89].

So far, it has been discussed that for the case of a simple Gaussian distribution, the log-likelihood is an equivalent measure to the MSD, but offset by the determinant of the covariance matrix and $\frac{1}{2\pi^{d/2}}$ (where d denotes data dimension). However, likelihood inference can be extended to much more complex models, as the next chapters will demonstrate. This thesis discusses the use of linear Gaussian models and their extensions as mixtures to deal with more complex, multi-regime data. The simplest of these cases is a simple mixture of Gaussian distributions, so this will be considered here for a novelty detection example. The maximum likelihood parameter estimation strategies for this model are discussed in Chapter 4. Here, it will be assumed that the parameters are known and the focus will be on the use of its likelihood function.

For any multivariate measurement $\mathbf{y}_i$, the log-likelihood function for a Gaussian mixture distribution is simply a weighted sum of the individual component probability densities evaluated on that data point, using the mean μ_k and covariance $\mathbf{S}_k$ for each of the k components as [26]:

$$\log p(\mathbf{y}_i) = \log \left\{ \sum_{k=1}^{K} \boldsymbol{\pi}_k \mathcal{N}(\mathbf{y}_i | \mathbf{S}_k, \boldsymbol{\mu}_k) \right\} \tag{3.4}$$

where $\mathcal{N}(\mathbf{y}_i | \mathbf{S}_k, \boldsymbol{\mu}_k)$ evaluates the probability density of $\mathbf{y}_i$ for every k_{th} Gaussian component, and $\boldsymbol{\pi}_k$ is the mixing proportion of the k_{th} component with respect to the complete density. It is desirable to use a log-likelihood in order to have numerically tractable measures of probability. With exponential probability densities, such as the Gaussian, the probability density assigned to data points lying far from the data mass can be low enough for numerical precision to matter. In some cases this may not matter, but in this case it does as the task here is to identify and quantify outlying points.

Even though the log likelihood function for the Gaussian mixture model now departs from the MSD, it is built from the same building blocks. The usefulness of a Gaussian mixture is that any arbitrarily complex density function can be approximated using a weighted combination of Gaussians. In fact, it can be shown that a (Gaussian) kernel density estimate is simply a Gaussian mixture model, in the limit of placing a Gaussian component on every observation. In fact, the idea of using a Gaussian mixture to perform novelty detection is not new. However, the approach of some studies is to fit a Gaussian mixture to the data set (using EM, covered in the next chapter) and to then check for outlying points by using the MSD of each component, and either monitoring all of them, or picking the one with the minimum value as a novelty index. An example of this can be found in [27]. This approach is not necessarily wrong, but it can suffer from various drawbacks. The first obvious one is that the novelty index does not have a probability interpretation. This may not be an issue if the data clusters are well separated. If they are not, then the points that could belong to either cluster will have a hard assignment to the cluster with the lowest distance. In contrast, the Gaussian mixture density function (equation (3.4)) uses the mixing proportions, $\boldsymbol{\pi}$ which introduce a prior probability over the observations. A Mahalanobis distance could be enhanced with prior probabilities too, but then it would quickly become a scaled version of an *expected* log likelihood function (which can be formulated in terms of sums of log-Gaussian densities [26]).

The use of GMMs for damage detection described in [11] is closer to this interpretation since it weights the distance of each point to the centre of the assigned cluster, weighted by the variance of each individual clusters.

Ultimately, the most natural argument for using a log-likelihood function in a novelty detection context is that it represents the probability density of the data. In other models, such as regression, it represents the probability density of the residual. If the parameters of a model are correctly optimised, then monitoring the resulting likelihood of new points is bound to indicate discrepancy from the reference model. The applicability of this idea is not restricted to parametric models; this idea has been extended to non-parametric models such as Gaussian Process regression, in elegant manners. One notable example of this is extreme function theory [90], which uses Extreme Value Theory (EVT) to determine appropriate thresholds on the likelihood function of a nonlinear and nonparametric regression model. The application of novelty detection in this case focuses on biomedical signal processing applications. This leads to the next point, which is how to determine appropriate thresholds over likelihood functions.

3.2.2 Determination of Thresholds

This discussion is provided here to give the reader a background in the techniques available for establishing thresholds, and to provide a solid illustration of how these may be applicable to likelihood functions derived from Bayesian networks. It has been discussed that a likelihood function is a novelty index that quantifies model discrepancy against an observation (or sets of). It is commonly used as an objective function in parameter optimisation, but the interest here is to use it to determine when an observation deviates significantly from a reference condition, which would normally belong to an undamaged state.

The correct determination of a threshold is crucial to the success of a damage detection strategy. This can be a daunting task when faced with multiple, possibly hundreds of variables. Thankfully, likelihood functions reduce the complexity of the threshold setting problem to a single (meaningful) variable. However, the question of setting a threshold over this likelihood still remains. Here three methods will be reviewed: empirical distribution percentiles, Monte Carlo sampling based percentiles, and Extreme Value Statistics. In order to compare strategies, a toy problem has been devised. It consists of a three variable and two component Gaussian mixture distribution, with closely spaced means, where one component has a much higher proportion. The outliers are introduced as a much more closely-spaced (low variance) single component Gaussian, close to the data mass but with a different mean vector. The data is illustrated in Figure 3.4

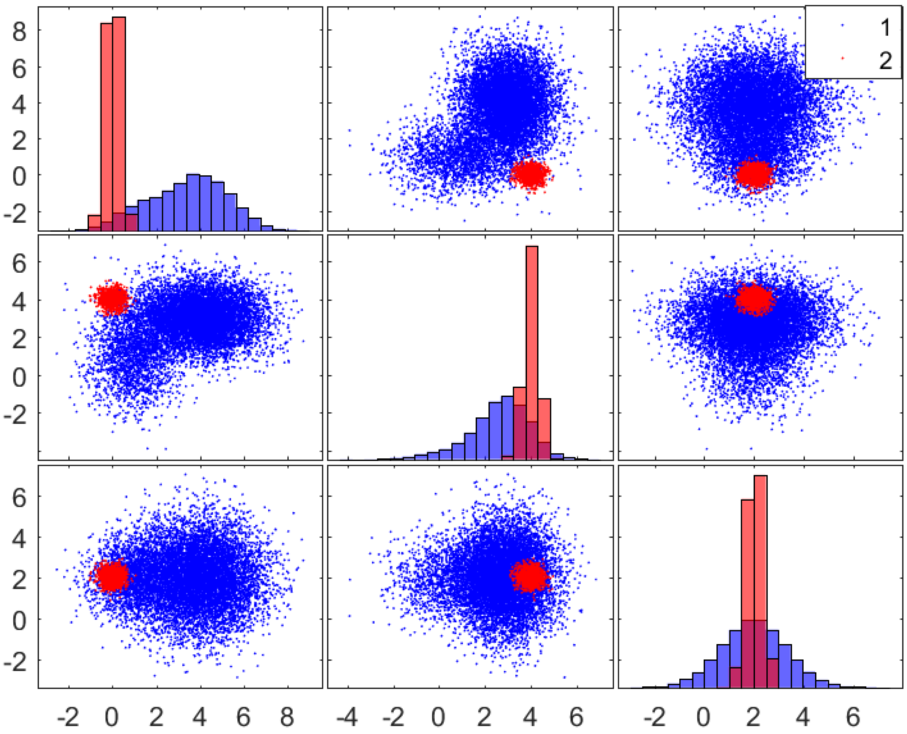

Figure 3.4: Scatter matrix showing the normal condition data (blue), and the outliers (red). The diagonals of the plot show the relative density of each dimension.

The idea of this example is to illustrate a problem in SHM where the normal condition data contains two regimes, with one dominating regime and an alternative regime. The data shown in Figure 3.4 has been generated with $\pi = \{0.8, 0.3\}$ so that the first component has high weighing in the true underlying density compared to the second mode.

The mean vectors were selected such that the outliers lie within the data mass in some dimensions but not in others, as should be clear from Figure 3.4. To make this illustration slightly more realistic, the maximum likelihood parameters were fitted using the standard EM algorithm for Gaussian mixtures [26] to a 5000 observation sample from the underlying distribution. In this case, a two-component Gaussian

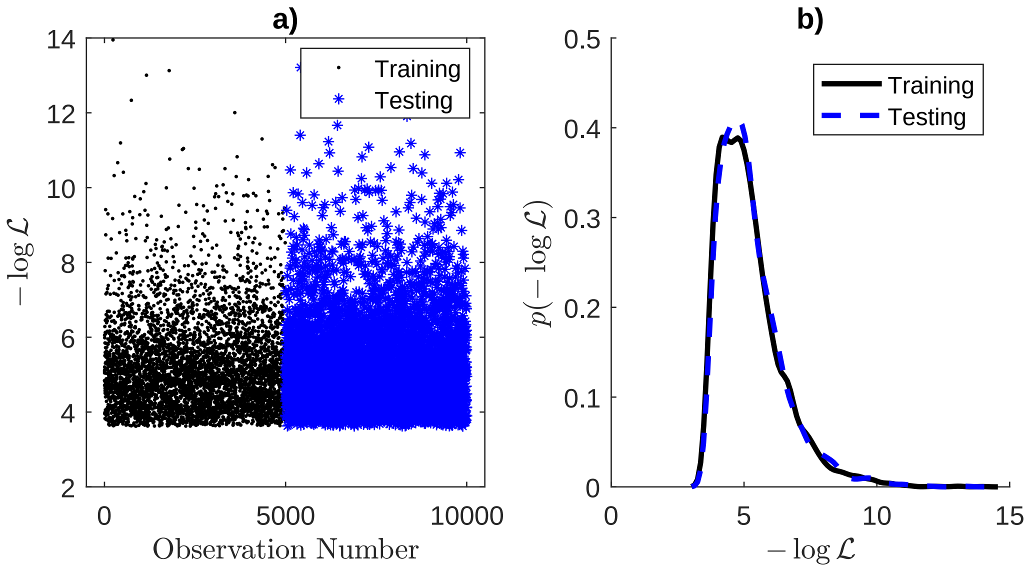

Figure 3.5: a) Negative log likelihood of Gaussian mixture model on training and testing data, b) Probability density of negative log likelihoods for both training and testing sets. Note these almost the same.

Mixture was assumed as it is the distance measure not the training procedure that is being illustrated here. Note that the data used for training does not contain outliers; it represents a case in SHM where feature data from an undamaged condition has been obtained and is free from measurement-related outliers. The negative log likelihoods for training and test sets are shown in Figure 3.5. Note that Figure 3.5b shows the probability density of the resulting $-\log(\mathcal{L})$ in order to compare how this distance measure is distributed for both training and testing sets. The first point to note is that the distribution of both negative log likelihoods are alike. This is not unexpected given that this is a toy problem. The distribution of the negative log likelihoods resembles that of a $\mathcal{X}^2$ distribution as expected. It is at this point that one would like to establish thresholds on the negative log-likelihood, such that an exceedance of such a threshold raises an alarm. The discussion will start with order statistics and move on to Monte Carlo sampling and EVS. It should be noted that all of these methods are all inter-related; they all seek to establish decision boundary above which the negative log likelihood will not cross, with a prescribed confidence or probability. Order statistics perhaps offer the most fundamental view in this respect. In the proceeding discussion, when referring to “observations”, it should be clear to the reader that in this case the negative log likelihood is being referred to, as this is the observation of interest for the purpose of establishing a threshold.

3.2.3 Order Statistics

A percentile is one of the simplest ways of establishing a threshold, and it makes use of the concept of order statistics. When observations are ordered from smallest to largest, the k^{th} order statistic is simply the k_{th} element in this ordered list. The concept of a percentile is a normalised version of an order statistic, where the largest value corresponds to the 100^{th} percentile. An n^{th} percentile can also be interpreted as the value below which n percent of the data lies. Percentiles are closely tied with the Cumulative Distribution Function (CDF), which normalises the values so that the largest order equates to 1. A CDF established using order statistics would constitute an empirical, or non-parametric CDF, as it is established purely using measured data. For this reason, percentiles are a simple and robust way of establishing decision thresholds. They are robust in the sense that order statistics deal well with the adverse effect of extreme values. A median, which is a robust analogue for a mean, is simply the 50^{th} percentile of the data (the value below (and above) which 50% of the observations lie). To establish a threshold above which an observation would be classed as abnormal, a high percentile is required, typically in the range between 95% to 99.99%. A percentile indicates the probability that an observation will lie under a prescribed value.

While a percentile is simple to compute it has the downside that, being related to an empirical distribution, it is sensitive to the number of points in the training observations. If the empirical CDF from which the percentile was derived is not captured accurately, a percentile will fail to capture the true probability of exceeding a certain threshold. This is aggravated by the fact that novelty detection seeks to accurately model the tails of the distribution. In the case of the negative log-likelihood, the interest is on the right tail, as this is where outliers will manifest themselves. Now, the tails of the distribution represent values that occur with very low probability, so a sample must contain a very high number of samples in order to "contain" the required number of values in the tail, so they can be well represented in the CDF.

Figure 3.6 shows the upper tail of the empirical CDF of the negative log likelihood of random samples of the toy Gaussian mixture model (with no outliers). CDFs are shown for increasing numbers of samples, in order to illustrate the sensitivity of a high upper percentile (99^{th} is shown) on the resulting threshold. Note that, based on a 99^{th} percentile, 100 and 500 draws show the most disparity in the threshold,

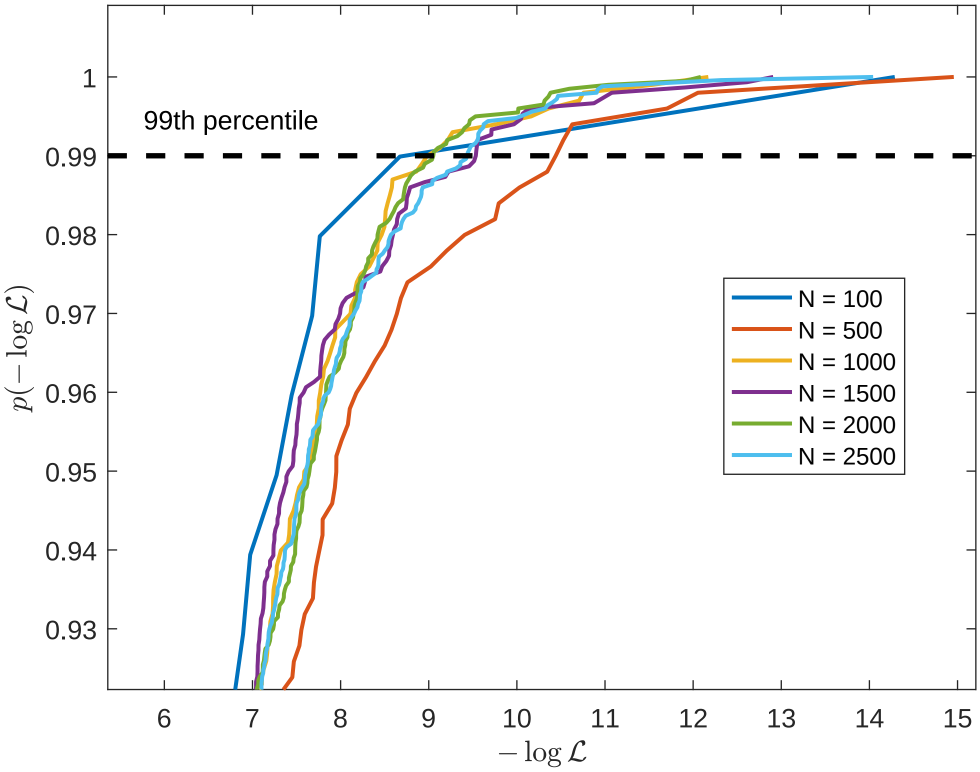

Figure 3.6: Empirical CDF for negative log likelihood of random draws from the illustrative 2-component Gaussian mixture model with increasing number of samples, N

while increased sample sizes narrow down into a threshold range between 8.5 and 9.

So, in summary, a percentile based on an empirical cumulative density function is a relatively simple and straightforward way of setting a threshold, but it should be kept in mind that it is derived from an emprical CDF, which may not correctly capture the tails of the distribution. In general, a percentile may be appropriate where a large number of samples exist for a training set.

In the examples given in the rest of this thesis, percentiles will be used to provide decision thresholds. This is because they are simple to derive and in the applications considered in this thesis a large amount of training data was available to render a percentile a reliable threshold. Nevertheless, it is still important to discuss other approaches.

3.2.4 Monte Carlo sampling

Empirical distributions may be doomed under low sample numbers. However, the samples could be increased if one were able to first capture the underlying distribution using a parametric model. Once this is done, infinitely many samples could be drawn from this parametric distribution, which acts as a proxy to the true distribution, and a threshold can be derived based on the results of many subsequent random draws. This is a Monte Carlo sampling approach. It was originally suggested as a means for selecting thresholds in SHM problems in [29], and has been a popular method for finding appropriate thresholds where the sampling distribution is a single multivariate Gaussian.

However, the technique is equally applicable to samples from more complex distributions or models. Here, the approach is demonstrated with the toy Gaussian mixture model, but the procedure, with the likelihood function as a novelty index, is equally applicable to other models discussed in the preceding chapters: PCA, Kalman filters and mixtures of these.

A Monte Carlo method was used to arrive at the threshold value and this may be summarised by the following steps:

1. Establish a generative model using a maximum likelihood method; in this case it is a Gaussian mixture model.
2. Sample m observations from the baseline generative model established using data from undamaged structure.
3. Compute negative log likelihoods for each observation (the form of which depends on the particular model), and store the maximum value.
4. Repeat steps 1-2 a large number of times and evaluate the empirical CDF of the array containing the maximum values for each set. Compute the percentile of this empirical CDF, which defines the threshold.

This procedure is more conservative when selecting a threshold compared to a simple empirical CDF because one is free to draw as many samples from the Monte Carlo trials as required, and if the residuals of the model are Gaussian, this will tend to generate large outlying points given a large enough draw size. Note that the procedure outlined here is slightly different from that described in [29], to accommodate for the fact that the sampling is done from a generative model as opposed to a single multivariate Gaussian. The reader will note that the sample size, m, from which the maxima are selected is now user selected. The choice of m will in fact dictate the threshold this procedure will yield. The empirical CDF and a kernel density estimate of the sample maxima is shown in Figure 3.7, for increasing m. Recall that the threshold is selected as a percentile of this CDF (Figure 3.7a). In this case, the appropriate threshold is a function of the sample size, and the threshold on the negative log likelihood clearly increases unboundedly with an increasing sample size. Because the threshold being sought is on a probability, this implies that if one draws enough random samples from a Gaussian distribution, arbitrarily unlikely values will be observed. There are practical limitations to how large a negative log likelihood one will eventually observe, the most obvious being machine precision.

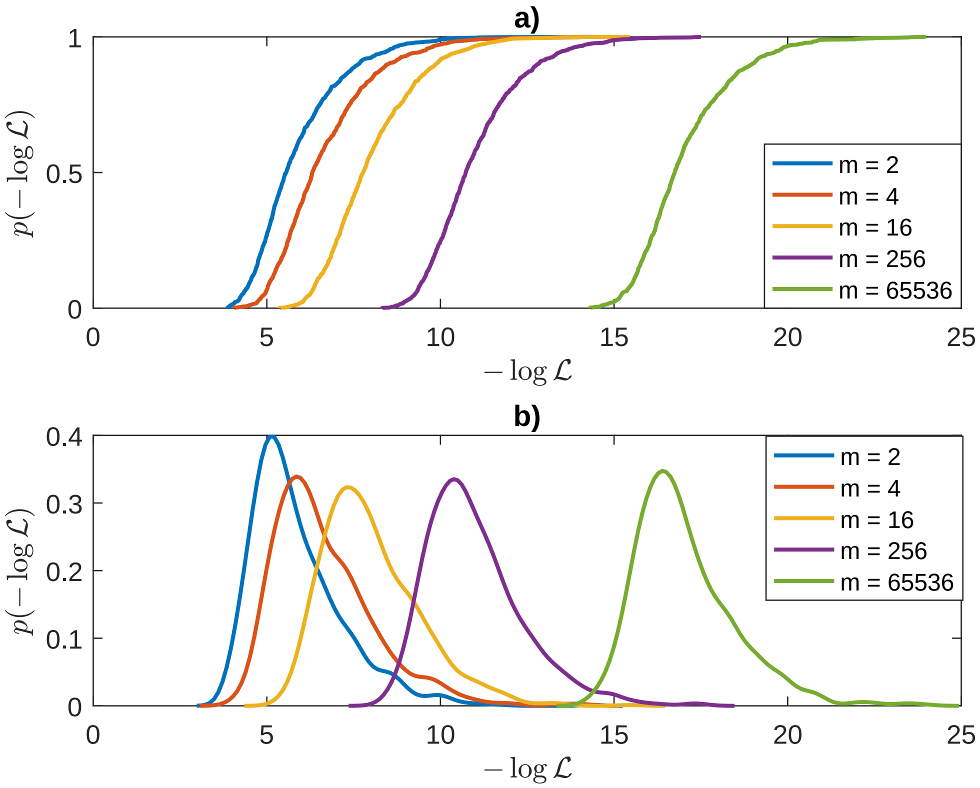

Figure 3.7: a) Empirical CDF and b) kernel density estimate of the maxima of the negative log likelihoods evaluated on random draws of size m, from the Gaussian mixture model

The observation that, for Gaussian (or mixtures thereof) random variables as m increases, the maximum negative log likelihood increases, is a result of the fact that the support of a Gaussian distribution is unbounded; the tails decrease exponentially to infinity. This is arguably not an acurate representation of reality, in particular when considering the dynamic response of an engineering system. Engineering experience, and physical constraints suggests that the behaviour of physical systems is bounded.

However, the focus here is narrowly the problem of selecting decision boundaries with a low probability of false positives, and there is no room for philosophical digressions. The practical implication of this tendency for the "optimum" threshold to grow with the sample size (in this Monte Carlo context), is that if the Monte Carlo scheme is to be used to establish a decision threshold, the value of m should be representative of the problem. This is a simple thing to achieve, as one often knows the approximate sample size of an engineering problem.

3.2.5 Extreme Value Statistics

EVS is a branch of statistics that deals with the distributions of extreme values. An extreme value is either the minimum or maximum of a sample. There is a close link between EVS and order statistics. EVS could be considered as a branch of order statistics dedicated specifically to the modelling of lowest, and highest orders: the maxima and minima of data-sets. EVS is founded on the Fisher-Tippet theorem [91], which describes distributions of maxima and minima of sets of a random variables, $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, where each subset, $\mathbf{x}_i$ is size m. If a centering and scaling of x by μ and ψ is considered

$$\mathbf{y} = \frac{\mathbf{x} - \mu}{\psi} \tag{3.5}$$

then the maxima of $\mathbf{x}$ is described by either one of three distributions, whose cumulative distribution functions, $F(\mathbf{y})$ are given by

$$\text{Gumbel: } F(y) = \exp(-\exp(-y)) \tag{3.6}$$

$$\text{Frechet: } F(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ \exp(-y^{-\xi}) & \text{if } y > 0 \end{cases} \tag{3.7}$$

$$\text{Weibull: } F(y) = \begin{cases} \exp(-(-y)^{\xi}) & \text{if } y \leq 0 \\ 1 & \text{if } y > 0 \end{cases} \tag{3.8}$$

where ξ is a parameter to be estimated. Because these EVS distributions model extrema of data sets, they are effectively modelling the tails of the distributions. Note that EVS also describes three distributions for the minima of data, but here only the maxima are treated given that it is negative log-likelihoods that are of concern here.

Note that these three distributions over maxima have similar form, but they differ in their approach to modelling of maxima. The Gumbel distribution has unbounded support, while the Frechet has a lower bound on its support and the Weibull has an upper bound. The Weibull distribution is popular in reliability analysis; it is rooted in the analysis of uncertainty of material strength [92], and it is widely use in survival analysis [93].

The EVS approach to describing the distribution of maxima is arguably better suited for the problem of threshold estimation than performing Monte Carlo trials based on a Gaussian distribution. The exponentially decaying tails of the Gaussian are not well suited for modelling heavy tails. There is no issue with using Gaussians and mixtures of Gaussian models for modelling the areas close to the data mass. However, as has been pointed out before, deciding what threshold to place on the novelty distance measure (negative log likelihoods in this case) requires an assesment of the probability of extreme events. It should be clear now, that in the context of the models discussed in this work, an “extreme” event constitutes an observation with very low probability and high negative log likelihood (under the given model). A percentile, on the other hand, may compare more favourably against EVS, because an empirical CDF lets the data “speak for itself”, provided there are enough observations to define the tails of the CDF accurately.

Moving on to determining the right EVS distribution and parameters the reader may have already noticed that one needs to select from one of the three distributions above. Which distribution correctly describes the tails depends entirely on the data. It can be shown that the extreme values of random variables originating from unbounded exponential distributions are described by a Gumbel distribution (Equation (3.6)) [91]. In the specific case of the Gaussian, the translation and scaling parameters, μ and ψ, can be described by [94]

$$\mu = \sqrt{2\ln m} - \frac{\ln\ln m + \ln 4\pi}{2\sqrt{2\ln m}}$$
$$\psi = \frac{1}{\sqrt{2\ln m}}$$

where m is, as before, the sample size from which the maxima are drawn.

Although EVS was developed to model phenomena such as tides, floods, and meteorological data where the extremes are of special interest its value has been recognised in engineering [95]. Their use in the context of general novelty detection has been studied in [96], and more recently in [28]. Applications to threshold setting in damage detection have been provided in [97]. These studies all use “classical” EVS, where a choice has to be made regarding the type of distribution that better describes the extreme values.

The type of distribution that best fits the the maxima of interest can be determined using probability papers, which scale the vertical axis on a cumulative distribution plot to yield a straight line under either of the distributions. The most suitable distribution is the one that provides the best fit to the data in this sense.

A more principled approach would be to use a Generalised Extreme Value (GEV) distribution, for which maximum likelihood methods are available for parameter estimation. The cumulative distribution function for the GEV is

$$F(x) = \exp\left\{-\left(1+\xi\frac{x-\mu}{\psi}\right)^{-1/\xi}\right\}, 1+\xi\frac{x-\mu}{\psi} > 0 \tag{3.9}$$

It can be shown that the Gumbel, Frechet and Weibull distributions are special cases of Equation (3.9) [98]. If $\xi \to 0$ then the GEV distribution is equivalent to the Gumbel distribution. Similarly, if $\xi \leq 0$ or $\xi \geq 0$, it corresponds to Frechet and Weibull distributions respectively. A GEV is useful because the choice of the extreme value distribution is implicit in the parameter estimates. Whilst optimisation of the likelihood function is not possible analyitically, numerical solutions have been suggested such as [98, 99] based on moment matching and Newton-Rhapson methods. More recently, differential evolution and linear quadratic programming have been suggested as an approach to optimising the GEV parameters [100], with SHM applications in mind.

As an example, Figure 3.8 shows the maximum likelihood fit for the training set of the toy Gaussian mixture problem. To perform the GEV fit, the data is first pre-processed by a moving window, where the maximum amplitude is recorded. This windowed maxima variable is denoted as z. For this data set, the maximum likelihood solution for ξ yielded 0.03 (relatively close to zero), indicating that this Guassian log likelihood has a domain of attraction to a Gumbel distribution. This is expected from the fact that the log likelihood is derived from a Gaussian distribution, which is exponential and unbounded. Note that there still remains a choice of window sizes, m, from which to select the maximum values of the negative log likelihoods. If the limiting extreme value distribution is Gumbell, then the possible damage thresholds that could result from this is still unbounded. However, this is not as much of a problem as in the Monte Carlo sampling case, because the GEV is being fitted to actual observed values of negative log likelihoods. In the real world, it is unreasonable to expect larger negative log likelihoods as m is increased. The natural frequencies of an aircraft will vary due to temperature, vehicle mass distribution, fuel, etc. It may even be reasonable to model these variations as Gaussian, but this does not mean that tails of the distribution will be strictly correct. Even in extreme changes to an aircraft, such as a change in its configuration, there are physical bounds to how much the natural frequencies will change. Therefore, thresholds set using extreme value distributions simply set reasonable bounds based on maximum observed values of the system.

3.3 Illustrative comparison on synthetic data

This final short section presents an application of the three threshold methodologies outlined in the previous sections to outlier detection on the synthetic, or toy problem, of the tri-variate, two-component Gaussian mixture distribution.

In this example, training and testing data were drawn from the reference distribution, shown in blue in Figure 3.4, while the outliers to be detected are shown in red. The maximum likelihood parameters for the Gaussian mixture were found

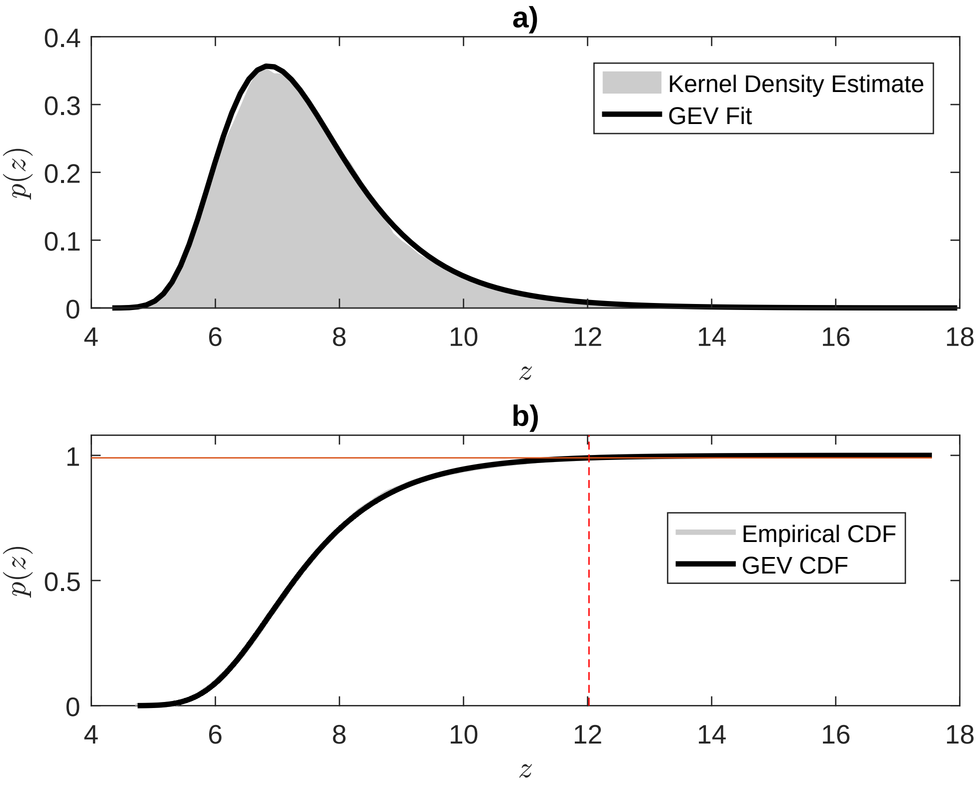

Figure 3.8: Generalised extreme value fit of windowed maxima of negative log likelihood (denoted as z) as well as their empirical a) probability densities and b) cumulative probability densities. The vertical red line denotes a 0.99 probability threshold.

using EM optimisation, which will be described in more detail for this model in Chapter 4. The key point of this illustration is to demonstrate both the use of the likelihood function as a novelty index, and the application and comparison of the three different threshold selection methodologies described in the previous sections.

This illustration is important, as throughout this thesis, different flavours of the same basic problem will be presented. Gaussian likelihood functions for all the linear Gaussian models presented in the next chapters will have the form shown in Figure 3.5. Whether one establishes the damage decision boundary directly on the negative log likelihood, or on a post-processed version of this, depends entirely on the context.

Figure 3.9 shows the resulting negative log likelihood, as well as thresholds derived from the training data set, using increasing levels of moving average window sizes. The feature that emerges from the three plots in Figure 3.9 is that the detection gets better and better as the moving average window size increases. The three thresholding methods also yield decision boundaries closer to each other.

This is easy to explain. The outliers drawn for this example all lie inside the data mass of the reference “undamaged” distribution. When projected through the negative log likelihood of the model, very few outliers cross the 0.99 confidence boundaries established through either of the three threshold methodologies. However, the outliers clearly form a second mode of the distribution of the negative log likelihood. This is akin to an engineering system that undergoes a change, or that develops an additional abnormal behaviour. Such is the case, for example, of the time series models, considered in Chapters 5 and 6. Even if individual observations don’t point to a change in the system, they do so on average. This is exactly the point being made with Figures 3.9b and c. Because the outliers lie in the data mass, the average negative log likelihood performs much better at identifying change.

Examples where this averaging may not be appropriate are rich features that already summarise time domain changes, such as Fourier and wavelet coefficients or natural frequencies. This is demonstrated in later chapters.

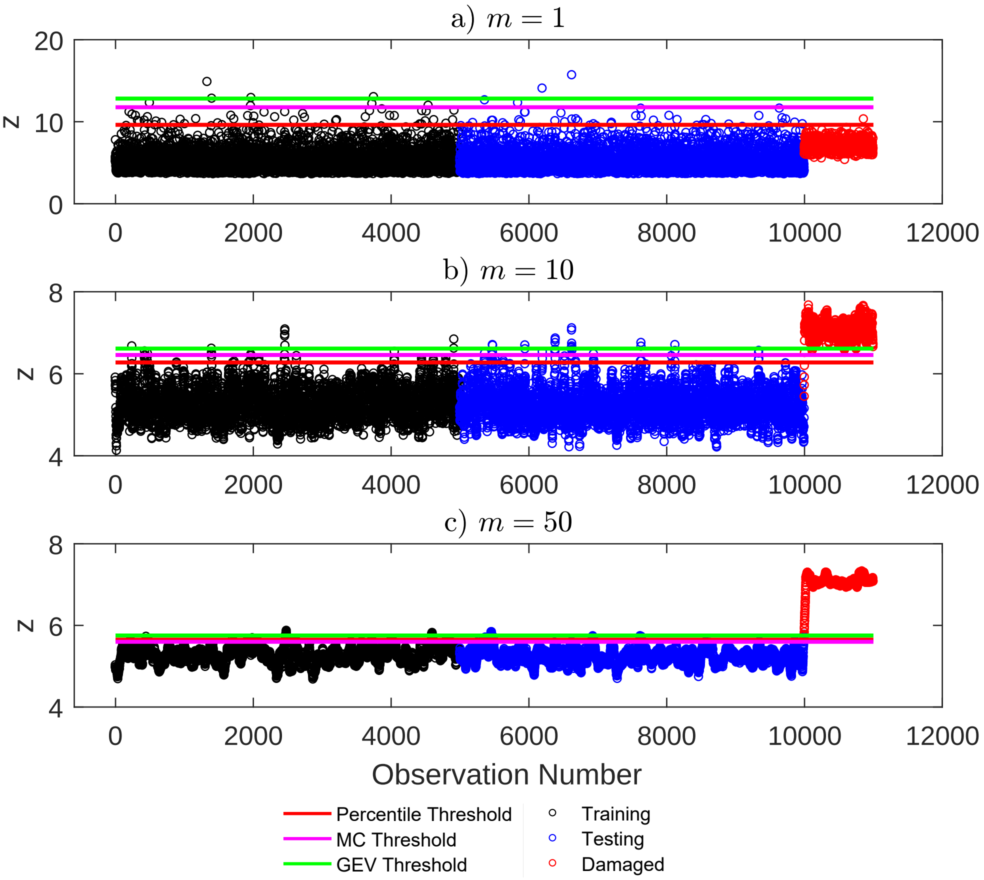

Figure 3.9: Negative log likelihood, with increasing moving window average lengths, m, for the illustrative Gaussian mixture model. The points related to damaged states are the outliers shown in Figure 3.4, and thresholds for the three different strategies are shown for each m

3.4 Chapter summary

In summary, this chapter has presented an overview of novelty detection in SHM, and has worked through the use of likelihood functions for this task. The remaining chapters deal with using likelihood functions for damage detection in different models and contexts. This chapter used a simple example of performing novelty detection on a synthetic multivariate Gaussian mixture distributed data set. The focus was placed on the usage of the likelihood function as a novelty measure, and so its distribution has been examined.

When using any distance measure for the purpose of deciding whether a system has changed or not, it is important to establish a reliable boundary for this decision.

This can be done by selecting appropriate maximum thresholds on the negative log likelihood of a statistical model. Three different thresholding strategies have been discussed, which have shown to be relevant in other SHM literature. These were order statistics, Monte Carlo sampling and extreme value theory. In the remainder of this thesis, order statistics will be used to define thresholds, based on 99^{th} percentiles of negative log likelihoods. This is due to their simplicity, and applicability when there are enough observations in the training set to define the tails of the cumulative distribution function well. This is also a result of the observation that if one whitens the negative log likelihood through a moving average, the thresholding methods yield answers closer to each other.

Chapter 4

Bayesian Networks for Novelty Detection Part 1: Static Data Models

This chapter presents the Bayesian network interpretation of the common linear Gaussian models introduced in the previous chapters.

The focus here is on the use of PCA, and Gaussian mixture models, while the next chapter will discuss extensions of these models that consider temporal relationships in data: Kalman filters and Hidden Markov Models. The obvious limitation of these models is their inability to characterise non-linear and non-Gaussian data, so the mixture-modelling framework will also be reviewed as a means of extending the capability of each model to data with greater complexity. The notion that a wide range of linear Gaussian models are simple variations of each other, under a Bayesian network framework, was systematically reviewed by Roweis and Ghahramani in 1999 [1]. Here, the systematic application of these models is reviewed with an emphasis in using their likelihood functions as novelty indexes. This is an important observation to be made, as the relationship between these models has not really been put into focus for an SHM application. Models such as PCA, Gaussian outlier analysis and Gaussian mixture models (all static data models) have been used before for damage detection [20, 29, 27], but their relationship has not been exploited in terms of using a consistent error function for damage detection that can be interpreted across models. The viewpoint provided here allows for a systematic extension to simple models such as PCA and factor analysis that has not been explored in SHM.

Roweis and Ghahramani's 1999 paper presents linear Gaussian models as generative, or latent variable models; the observed data $\mathbf{y}$ has an underlying cause, the latent variables $\mathbf{x}$, which generate $\mathbf{y}$ through some model $\mathcal{M}$. Inference is viewed as the task of estimating the probability distribution of $\mathbf{x}$. Learning is presented as the task of maximising the likelihood of the observed data, given the latent variables $p(\mathbf{y}|\mathbf{x})$. This is done using the Expectation Maximisation (EM) algorithm [86]. In the case of generative models, the probability of observed data given the latent variables is the likelihood function of the model, and it provides an estimate of how much the observed data deviates from the baseline condition. This thesis, in general, explores the use of generative models in SHM through the systematic use of likelihood functions across different models. This chapter presents the details of the derivations of the likelihood functions, and doing so is facilitated through their Bayesian network interpretation.

Linear Gaussian models represent probabilistic relationships between variables with the assumption that the measured variables, the latent variables and the uncertainty in the model are all Gaussian distributed. Probabilistic graphical models, as the name suggests, provide a graphical representation of dependencies embedded in probabilistic models; there are two main motivations for using them. The first motivation is that it is easier to visualise the structure of the reasoning that the probabilistic model represents by means of examining a graph. It may sometimes be easier for a user of an SHM system to interpret complicated conditional relationships between variables through a visual representation rather than one written in equations. The second motivation is that all the necessary mathematical manipulations for deriving likelihood functions are implicit in the graphical representation. This aids with the objective here, of deriving likelihood functions useful for novelty detection applications in SHM.

Linear Gaussian models can be viewed as a particular class of probabilistic graphical models. Although the graphical representation of these models is not essential to their understanding, they highlight (just as much as the equations do) how linear Gaussian models are simple variations of one another. They also help one to understand the task of inference and learning. This section will introduce the basic concepts and rules for probabilistic graphical models. Enough theory will be given in order to be able to "read' the models presented here from their graphical representations. This is not intended as a full discussion on the topic, good references on this can be found in [26, 101, 83].

There are two types of probabilistic graphical models: Bayesian networks and Markov random fields. In Bayesian networks, the links between nodes are specified by arrows, which have a specific direction in order to show conditional dependencies between variables. These directed arrows show causal relationships between variables. The links for Markov random fields on the other hand do not specify a direction. This section will focus on Bayesian networks, since linear Gaussian models can all be viewed as a sub-class of Bayesian networks, and these networks can be used to perform inference and learning for all of the models being considered. The idea behind Bayesian networks is very simple; each random variable is represented by a node, and their links represent conditional dependence between them. The graph represents the joint probability distribution over the set of random variables. Bayesian networks are mathematically *directed acyclic graphs*, since they are specified by directed arrows, and they also have the property that one must not be able to trace the path of a set of links back to the original node/variable (thus the term acyclic). Figure 4.1a shows an example of a simple Bayesian network/directed acyclic graph. Taking that network as an example, through the use of the product rule of probability, it is possible to factorise the joint distribution over three variables as[1]

$$p(x, y, z) = p(z|x, y)p(x, y) \tag{4.1}$$

which can be further factorised as

$$p(x, y, z) = p(z|x, y)p(y|x)p(x) \tag{4.2}$$

Some terminology from graphical models is required at this point. Node x is said to be the *parent* of nodes y and z. Node z is a *child* of both x and y. The graphical model contains qualitative information, but there are probability distributions implicit in each link which can be quantified exactly. It will be shown later that the problem of finding the precise parameters for the conditional probability distributions connecting a given graphical model is the problem of learning (in machine learning terms) or system identification (in the dynamics terminology), and this will be discussed later in Section 4.3.1 using the EM algorithm for a specific case. Con-

[1] The reader should note that x, y, z are used here as illustrative variables, and do not relate to the later convention of x being a latent variable and y being an observation

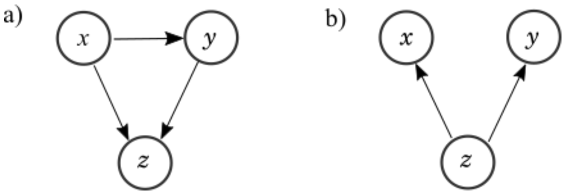

Figure 4.1: a) Bayesian network corresponding to equation (4.2) and b) Bayesian network corresponding to factorisation in equation (4.4) where not all nodes in the graph are connected.

tinuing with the example, Equation (4.2) can be generalized to n random variables

$$p(x_1, ..., x_n) = p(x_n|x_1, ..., x_{n-1})...p(x_2|x_1)p(x_1) \tag{4.3}$$

The factorisation above represents a graph that is fully connected since all pairs of nodes are linked (as in Figure 4.1a).

4.1 Conditional Independence

A fully connected graph will not be very interesting in general as its factorisation is nothing more than a statement of the product rule of probability. It is the lack of connections that makes graphical models more interesting and useful as their factorisation convey information about *conditional independence* between variables. Consider instead the following factorisation, corresponding to the graph in Figure 4.1b

$$p(x, y, z) = p(z)p(x|z)p(y|z) \tag{4.4}$$

In general, the joint probability for a network that is not fully connected is given by

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i|\pi_i) \tag{4.5}$$

where π_i is simply the set of parent nodes of x_i. The factorisation of the joint distribution encodes all the conditional independence relationships between the variables.

A random variable x is conditionally independent from y given the value of z if

$$P(x, y|z) = P(x|z)P(y|z) \tag{4.6}$$

This is simply the mathematical way of saying that two random variables are conditionally independent given the value of z if their joint probability, conditioned on z, factorises to the product of their conditioned marginal probabilities. It is possible to check, for a specific factorisation of variables, whether there is conditional independence by means of the sum and product rule to determine whether the variables in question satisfy the condition of equation (4.6). A useful key property of graphical models is that any variable is independent of its ancestors given the value of its parents. This is called the causal Markov assumption, and leads to a property that can be used to infer conditional independence much more easily through systematic visual examination of the graph due to a very useful property called *d-separation* [101]. This allows one to establish conditional independence between two variables given a third one, provided certain conditions about their connections are met. The way in which the paths of two variables intersect through a third one, can be used to infer whether conditional independence exists. The directions of the arrows through the intersecting node are key here. These directions can be either head-to-head or tail-to-head or head-to-tail, as illustrated in Figure 4.2.

The d-separation property states that sets of variables x and y d-separate (are conditionally independent) if every possible path connecting them (that is directed and undirected paths) is blocked by another variable z, such that the connection through z is [101]:

- Tail-to-tail or tail-to-head and the value of z, or its descendants is observed.

- Head-to-head and neither z nor its descendants have been observed.

If these conditions are met then x and y are conditionally independent given z. These conditions are illustrated in Figure 4.2. Note that the greyed-out nodes indicate *evidence*, in other words that the variable is observed. The principle of d-separation is particularly useful when considering the conditional independence of variables in time series models; given an observation, past values can be modelled as conditionally independent on future values. This will be seen in more detail in Chapter 5

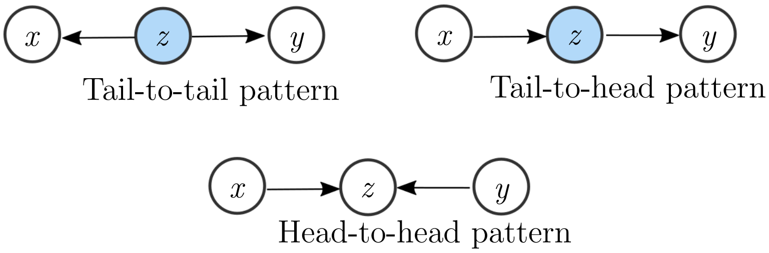

Figure 4.2: D-separation illustrations showing conditional independence of x and y given z.

4.2 Generative Models

The underlying idea of generative modelling is that the observations y, can be explained by a set of hidden or latent variables x. In other words, $y = f(x)$. Expressing generative models as probabilistic graphical models, allows one to use the graphical model machinery to extract conditional and marginal probability relations for y and x as well as to learn the corresponding mapping between them.

Evaluating the conditional probability of x given y as well as the marginal over y, the observed data, is a problem of *inference*, which can be carried out systematically with algorithms such as junction-tree for generic graphs or with more model-specific alrorithms such as Principal Component Analysis (PCA) or Gaussian Mixture Model (GMM). Other algorithms discussed later, that model temporal relations within data are autoregressive models, the Kalman filter and Hidden Markov Model (HMM) (discussed in Chapter 5). There is a clear distinction in inference algorithms for generative models: those which model temporal relations and those that don't. It is worth starting the discussion with the non-temporal, or static networks. This chapter will only deal with static data models, while Chapter 5 will discuss models for dynamic data.

4.2.1 Principal Component Analysis

Principal Component Analysis (PCA) is a common and simple example of generative modelling of "static" data. PCA is traditionally a dimensionality reduction technique which makes it particularly useful for visualising patterns in data with high dimensionality.

The previous chapter discussed PCA and some applications, which tend to focus

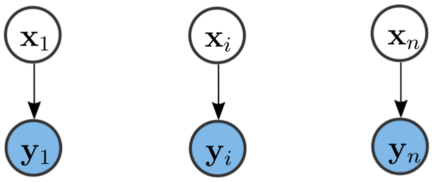

Figure 4.3: Bayesian network representation of PCA and factor analysis

around visualisation of high dimensional data. This chapter discusses the probabilistic interpretation of PCA, and how one might use the reduced dimensionality to compute a likelihood function on data. *Probabilistic* PCA (PPCA) has been introduced as a method for learning the coordinate transformation using maximum likelihood (EM) as well as modelling noise within the model [102, 75]. Note that for clarity of presentation, when referring to PCA, it will be assumed that its probabilistic interpretation is being referred to, unless otherwise specified. Assuming that the dataset, $\mathbf{Y}$ has been centred by its mean, $bv\mu$, both PCA and factor analysis can be described by the following equation

$$\mathbf{Y} = \mathbf{C}\mathbf{X} + \eta, \quad \eta \sim \mathcal{N}(0, \mathbf{R}) \tag{4.7}$$

where, $\mathbf{Y}$ represents the observations, $\mathbf{X}$ are the latent variables, and $\mathbf{C}$ is the mapping between them. The difference between the ”classical” and the probabilistic PCA definition is the introduction of the noise term η. This Gaussian distributed term with zero mean and covariance matrix $\mathbf{R}$, also determines whether equation 4.7 describes PCA or factor analysis. More specifically, if the covariance term is diagonal, learning with EM will lead to the standard factor analysis model. If $\mathbf{R}$ is an identity matrix scaled by a variance term, so that $\mathbf{R} = \sigma\mathbf{I}$, then this corresponds to PCA. This graphical model is illustrated in Figure 4.3. The graphical model is relatively simple; it simply specifies that the observations are dependent on the latent variables.

PCA works by extracting the n largest eigenvectors of the covariance matrix of the observations $\mathbf{Y}$, which define a rotation and matrix projection into a lower-dimensional subspace which maximises the variance of $\mathbf{Y}$ explained by the principal components $\mathbf{X}$. A clear drawback of this classical definition of PCA is the lack of a probabilistic interpretation (a density model), which is straight-forward in the graphical model interpretation. The graphical model allows the computation of a likelihood function, which views the probability of observations belonging to the model, as a function of the parameters. This is useful for a number of reasons, including the ability to compare models and to directly perform novelty detection, which is the focus of this work. The likelihood function is normally of interest in order to optimise the parameters of the model as well as to compare how different models fit a data-set. Here the focus is on its use for novelty detection, so it is worth exploring a simple derivation for the PCA case. The following is mostly borrowed from [75] as it provides a clear and concise derivation of the marginal likelihood $p(\mathbf{y})$. A small guidance to the notation being used should be given here. A column vector, denoting a multidimensional variable is denoted as $\mathbf{y}$, while an ensemble of these variables is collected into a matrix $\mathbf{Y}$, such that $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$. PCA arises by constraining $\mathbf{R}$, the covariance matrix of the noise process, η, in Equation 4.7 to be isotropic ($\eta \sim \mathcal{N}(0, \sigma^2\mathbf{I})$). The probability density of $\mathbf{y}$ given $\mathbf{x}$ is thus

$$p(\mathbf{y}|\mathbf{x}) = (2\pi\sigma^2)^{-d/2} \exp\{-\frac{1}{2\sigma^2}|\mathbf{y} - \mathbf{C}\mathbf{x} - \boldsymbol{\mu}|^2\} \tag{4.8}$$

where an isotropic Gaussian prior has been assumed over the latent variables $\mathbf{x}$, defined by,

$$p(\mathbf{x}) = (2\pi)^{-q/2} \exp\{-\frac{1}{2}\mathbf{x}'\mathbf{x}\} \tag{4.9}$$

where q denotes the dimensions of the vector of latent variables, and $'$ denotes vector transposition. The marginal distribution $p(\mathbf{y})$ can be derived as,

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \\ &= (2\pi)^{-d/2}|\mathbf{R}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu})\} \end{aligned}$$

where the covariance of $p(\mathbf{y})$ can now be expressed as,

$$\mathbf{R} = \sigma^2\mathbf{I} + \mathbf{C}\mathbf{C}' \tag{4.10}$$

For novelty detection, the log-likelihood of a data set $\boldsymbol{Y}$ can be expressed conveniently in terms of the model covariance matrix $\mathbf{R}$, and the sample covariance of new observations, $\boldsymbol{\Sigma}$ [75]

$$\begin{aligned} \mathcal{L} &= \sum_{n=1}^{N} \ln\{p(\mathbf{y}_n)\} \\ &= -\frac{Nd}{2}\ln(2\pi) - \frac{N}{2}\ln|\mathbf{R}| - \frac{N}{2}tr\{\mathbf{R}^{-1}\boldsymbol{\Sigma}\} \end{aligned}$$

A point-by-point log-likelihood can also be computed for $\mathbf{y}_i$ simply by noting that the data covariance matrix is modelled using Equation (4.10), which results in the (log) normal distribution

$$\mathcal{L}_i = -\frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{R}| - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\mathbf{R}^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{4.11}$$

In this case, the probabilistic PCA interpretation effectively yields a density model, and readily allows application of novelty detection for SHM applications. The applications of PCA in this context are scarce, but include [81].

In order to find the model parameters, one can find a matrix $\mathbf{C}$ that maximises the log-likelihood $\mathcal{L}$ on normal condition training data using EM, or other optimisation schemes. In the case of PCA, there are analytical solutions for the maximum likelihood parameters, achieved with techniques such as SVD. In practice, if the number of dimensions of the data set $\mathbf{Y}$ is very large, then using SVD or other solutions to the eigenvalue problem may be computationally expensive, and the EM algorithm has been shown to converge to the maximum likelihood solution with much less computations than an eigenvalue solver.

In terms of computational efficiency, one advantage of the EM approach for PCA is that the number of principal components, p, can be specified exactly, by specifying the dimension of $\mathbf{C}$. As pointed out in [102], direct diagonalisation of the covariance matrix, has a complexity of $O(d^3)$, (where d is the dimension of $\mathbf{Y}$). In contrast, EM does not require computation of the sample covariance and the complexity of the algorithm is limited to $O(qnp)$ (where q is the number of eigenvectors and n is the number of observations in $\mathbf{Y}$). However, this does not solve the problem of estimating the appropriate number of principal components and methods such as the Bayesian or Akaike information criteria could be used, or a fully Bayesian implementation of PCA could be adopted [103, 104, 75]. One further advantage of using EM for learning PCA parameters is that recursive learning is possible, with point-by-point parameter updates. There exists in fact an online implementation of the EM algorithm [105], which is fully applicable not just to PCA but to all the latent variable models discussed in this thesis. This is particularly helpful in applications with large amounts of data, where storing all of the training data in memory in order to compute $\mathbf{\Sigma}$ (the sample covariance) is simply not practical. That is not to say that there aren't online algorithms for computing the SVD of a covariance matrix.

The probabilistic interpretation of PCA does not, as such, make it a good *density estimator*. However, simple extensions via a mixture modelling framework make it much more powerful, as one is then able to use this model directly as a density estimator and use $-\mathcal{L}$ to perform damage detection when the dataset considered contains operational and environmental variations, which may lead to multiple densities. Alternatively, the dimensionality reduction ability of PCA can be used for purposes beyond data visualisation, so that principal components could be used as features within a mixture modelling framework such as a Gaussian mixture model. In fact, one could define a proper mixture of PCA or factor analysis models; this is discussed in Section 4.3.4.

4.3 Mixture Models

Linear Gaussian models, such as PCA are constrained in terms of the complexity of the data and probability distributions they can model. As the name implies, they model linear relationships between variables, and define the probability distributions as Gaussian, hence the limitation. A simple, yet powerful extension is to consider mixtures of distributions. The simplest example of this is the well known Gaussian mixture model. However, as with PCA, the mixture framework can be extended to any latent variable model. Moreover, a mixture model can be defined as a probabilistic graphical model, allowing one to make use of the standard machinery for deriving likelihood functions once again. In a probabilistic graphical modelling framework, the mixing variables are also defined as hidden variables; this means that one does not know a-priori which model each data segment belongs to. A standard solution in the mixture model framework is to use EM to optimise over latent variables and therefore to segment the data into different regions.

The mixture modelling framework fits well with a novelty detection framework based on probabilistic graphical models, as it is relatively straight-forward to derive and evaluate a likelihood function, giving a probability density over observed data points given a model. A general mixture probability density for K components can be defined as

$$p(\mathbf{y}) = \sum_{k=1}^{K} \pi_k p(\mathbf{y}|k) \tag{4.12}$$

where π_k represents the mixing proportion of each component, k represents the component index, and $p(\mathbf{y}|k)$ represents the probability distribution of each component. The number of components is not normally assumed to be known a-priori and needs to be estimated from the data.

The mixing coefficient π_k is a prior probability that a data point $\mathbf{y}_i$ belongs to model k; thus, they must satisfy $0 \leq \pi_k \leq 1$, and since this is a finite mixture, they must also all sum to unity: $\sum_k \pi_k = 1$.

One useful functionality of the mixture framework is being able to apply Bayes' rule to update the probability of an observation $\mathbf{y}$ belonging to model k. This effectively yields posterior probabilities over the mixing coefficients

$$p(k|\mathbf{y}) = \gamma_k = \frac{\pi_k p(\mathbf{y}|k)}{\sum_j \pi_j p(\mathbf{y}|j)} \tag{4.13}$$

where γ_k is defined here as the component *responsibility*, for an given observation. This result can be interpreted intuitively, as the contribution of each component normalised by the total contribution of all models. This posterior probability over the models readily allows model segmentation[2], and also allows for the computation of a likelihood for the data point, given a model, while taking into account model segmentation.

The log likelihood function for this generic mixture can be written, for a data set of N points, as a function of the parameters

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k p(\mathbf{y_i}|k) \right\} \tag{4.14}$$

While this may be of practical use for some inference purposes, the sum inside the logarithm presents a problem when maximising the log-likelihood. A better and more common approach would be to use the generative model corresponding to the mixture model, where a set of hidden variables, $\mathbf{z}$ with K dimensions, dictate which component the measurement $\mathbf{y}_i$ corresponds to. The $\mathbf{z}$ vector in this case is a binary indicator variable, where $z_{ik} = 1$ indicates that the i^{th} observation ($\mathbf{y}_i$)

[2] Throughout this thesis, model segmentation refers to the separation of observations, within a data-set into different sub-models, or components, which belong to a mixture. In other words, clustering

belongs to component k. The indicator variable thus contains zeros elsewhere for each column. If one were given labelled data (knowledge of which class a particular measurement belongs to), then finding the component parameters may be more straightforward, and standard learning methods could be applied to the segmented dataset. The interest here is in the case where $\mathbf{z}$ is unknown, and must be found. If the generative graphical model interpretation, shown in Figure 4.4 is taken, where $\mathbf{z}$ is the latent variable generating $\mathbf{y}$, the joint probability $p(\mathbf{y})$ can be easily derived in terms of $\mathbf{z}$ from examination of the graph in Figure 4.4

$$p(\mathbf{y}) = p(\mathbf{z})p(\mathbf{y}|\mathbf{z}) \tag{4.15}$$

Using equation (4.12) and defining the probability of the latent variable as

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{4.16}$$

the log-likelihood for such a model can be written as

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \ln\{\pi_k p(\mathbf{y}_i|\mathbf{z}_k)\} \tag{4.17}$$

where the conditional probability $p(\mathbf{y}|\mathbf{z})$ is model dependent. This form is useful, since all the sums over components and observations are outside the logarithms.

The implication of using a mixture for damage detection is that one can use a combination of probabilistic models to fit the data. When computing the likelihood function, one does not have to explicitly choose which model the data point should belong to, if any. $\mathcal{L}$ will weigh more heavily the models that are more likely to explain the data, and weigh down those that do not explain the data well. One is essentially performing density estimation of data with arbitrarily complex underlying densities, while still retaining interpretability because each model component represents a different "state" of the data. Ultimately, for successful novelty detection, the mixture model must be able to assign a low likelihood to data points that do not fit any of these models, which should happen naturally if the data is segmented into correct models.

One of the things that makes this framework powerful is the systematic application of Bayes' rule to derive posterior probabilities over model segmentation, and marginal distributions over observed data. The usefulness is vast in problems where

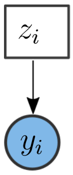

Figure 4.4: Bayesian network representation of a mixture model. The indicator variable, z_k is a discrete variable, encoding the mixture component a data point belongs to.

data is generated from engineering systems with changing environments, loading conditions and possibly changing dynamics due to nonlinearities; it leaves the user with a probability measure of how much the system has changed, taking into account different states, while segmenting those states without prior knowledge. The next section discusses the mixture modelling framework based specifically on Gaussian distributions.

4.3.1 Gaussian Mixture Models

A Gaussian mixture model is defined as a weighted sum of Gaussians, by taking the general mixture probability density of equation (4.12) and setting $p(\mathbf{y}|k)$ to be Gaussian; the marginal distribution over $\mathbf{y}$ becomes

$$p(\mathbf{y}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{y}|\mu_k, \mathbf{S}_k) \tag{4.18}$$

where, $\mathcal{N}(\mathbf{y}|\mu_k, \mathbf{S}_k)$ represents the normal Gaussian distribution of the k^{th} component, with mean vector μ_k and covariance matrix $\mathbf{S}_k$. This is rather intuitive definition. However, it is important to note that equation (4.18) is a result of setting the conditional density of observations on the hidden variables, $p(\mathbf{y}|\mathbf{z})$, to be

$$p(\mathbf{y}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}_k, \mathbf{S}_k)^{z_k} \tag{4.19}$$

which is simply a statement that the density of $\mathbf{y}$, given knowledge that $z_k = 1$, is just the density of the k$^{\text{th}}$ Gaussian component. This is an important point to make, because even though equation (4.12) is readily useful as a density estimator (provided one has a suitable set, θ, of parameters), what makes a Gaussian mixture powerful is a suitable learning strategy to find θ. Without any prior knowledge of $\mathbf{z}$ (the segmentation of data), EM is a well suited, elegant approach for finding the maximum likelihood data segmentation. Another approach to the parameter learning problem could be to parametrise the covariance matrices $\mathbf{S}_i$ using Cholesky decomposition, and optimise against the likelihood using a gradient method. The EM approach is favoured in this work because of its simplicity, and its ability to deal with different models with hidden, or missing, variables. The EM algorithm for Gaussian mixtures is relatively simple. It consists of iterating over an Expectation step, where the expectations for the hidden variables are computed, and a Maximisation step, which updates the parameters according to the expectations evaluated. The expectations for $\mathbf{z}$ are in fact evaluated by taking the responsibilities, $\boldsymbol{\gamma}$, for each mixture parameter, using equation (4.13). Once these are found the updates to the parameters are computed as follows:

$$\boldsymbol{\mu}^{new} = \frac{1}{N_k}\sum_{i=1}^{N}\gamma(z_{ik})\mathbf{y}_i \tag{4.20}$$

$$\mathbf{S}^{new} = \frac{1}{N_k}\sum_{i=1}^{N}\gamma(z_{ik})(\mathbf{x}_i - \boldsymbol{\mu}_k^{new})(\mathbf{x}_i - \boldsymbol{\mu}_k^{new})' \tag{4.21}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{4.22}$$

where N_k is given by the summation of responsibilities over the component k: $N_k = \sum_{i=1}^{N}\gamma(z_{ik})$. Each M step is guaranteed to increase the log-likelihood (equation (4.17)) at every iteration [86]. Note that the updates to the mean and covariance are just the regular computations of a mean and covariance, but weighted according to the posterior probability of each point belonging to the k^{th} component. The mixing proportion updates are simply the proportion of points that belong to that component. EM normally yields good results (correct segmentation) with a few caveats [26]. EM is guaranteed to converge to a local maximum of the log-likelihood, but not a global one. Because the parameters have to be initialised at the first iteration of EM, the algorithm is sensitive to the initial selection. A robust and sensible choice is to initialise θ using k-means clustering [75], or simply to have multiple initialisations and to pick the one with the best overall likelihood. The EM algorithm for Gaussian mixtures is also prone to finding singular solutions; if EM chooses an update to the mean such that $\mu_k = \mathbf{y}_i$, then this component will contribute a term to the likelihood function that goes to infinity as the variance of the component goes to zero. In other words, EM fits one component, to a single point, which is infinitely probable but also likely to be just noise in the data. A pragmatic approach to solving this would be to add a regularisation on the variance terms on every M step, or to add a step within the algorithm that restarts whenever this occurs. A more pressing issue is that of choosing the right number of components, K, that dictate the model order. As discussed in the previous chapter, too many components will overfit, while too little may fail to generalise. Ideally, one would fit enough components for them to highlight a physical context. A common approach to model selection is to fit the best possible model for different K, and to select an order where the increase in likelihood with respect to K becomes marginal. This approach is the popular Akaike Information Criterion (AIC), but it is still prone to selecting model orders that overfit. A similar approach, the Bayesian Information Criterion (BIC) attempts to solve this by penalising the likelihood with the number of parameters, so that the optimum K will be indicated by a minimum of the BIC index. An even better approach would be to use a prior over the parameters $p(\theta)$ in order to have a Bayesian treatment of the problem. One of the pitfalls of Bayesian methods is that under certain prior distributions, the likelihood function is often specified in terms of intractable integrals. Approximate methods are required in this case, to optimise over the parameters. The two general approaches for this are to use sampling methods, such as Markov Chain Monte Carlo (MCMC) sampling [106, 107] or variational approximations [26]. The MCMC approach tends to be computationally heavier, while variational approximations tend to be mathematically much more complex [26]. The key advantage of the Bayesian approaches is that they offer a good solution to model order selection, and thus avoid over-fitting. The variational and sampling approaches may offer better model order selection; however, for the problem at hand, of using likelihood functions of generative models to detect damage, the simpler AIC and BIC approaches suffice, provided they are used correctly.

To summarise, the procedure for using a Gaussian mixture for damage detection suggested by this author is as follows:

1. Select a (damage sensitive) feature vector $\mathbf{y}$ to represent the data from a healthy condition. Any operational and environmental changes should be captured in $\mathbf{y}$.

2. Select a subset of $\mathbf{y}$ to use for training the model, $\mathbf{y}_{train}$. Select another subset

to test the model predictions on: $\mathbf{y}_{test}$.

3. Decide whether the model should have physically meaningful constraints. Should the mixing proportions be equal? Is there a-priori knowledge of the number of mechanisms generating the data?

4. Choose an initial set of means, covariances and mixture proportions (with the above constraints in mind).

5. Iterate EM steps until convergence of likelihood function (equation (4.18)) to optimise the parameters, using equations (4.13) to evaluate the component responsibilities and equations Equations (4.20) to (4.22) to update the parameters. To avoid local maxima, restart the EM algorithm multiple times, and select the model parameters with the highest (log) likelihood.

6. Evaluate negative log-likelihood function on $\mathbf{y}_{train}$, point-by-point, using equation (4.18), and set a threshold $\mathcal{T}$(see discussion on previous chapter on thresholds) under which the majority of $-\log\mathcal{L}$ for the training set lies.

7. Evaluate $-\log\mathcal{L}$ on the testing data set using one of the methods discussed in Section 3.2.2, and ensure that it remains within the $-\log\mathcal{L}$, threshold

8. If the model correctly captures the variability of the healthy condition, exceedances of $\mathcal{T}$ indicate damage, or other previously unseen (and possibly benign) changes.

4.3.2 Illustration on Z-24 bridge data

As an example, a Gaussian mixture model is fitted to the Z24 bridge data using the above procedure. This dataset is discussed in more detail in Chapter 7. It consists of a time series of four natural frequencies of the Z-24 bridge between Bern and Zurich. The natural frequencies of the bridge were extracted using Stochastic Subspace Identification (SSI) for a period of approximately a year. During that period there were severe temperature variations that led to an observed bilinear stiffness behaviour. Damage to a girder was introduced towards the end of the measurements. In this example, only the first two natural frequencies were used in order to make visualisation of the Gaussian mixture model easy. The fit of the model to the two natural frequencies is shown in Figure 4.5. For now, the relevant point to note is that the threshold set on $-\log \mathcal{L}$ defines a contour in the multivariate data space within which the model classifies the data to belong to a normal condition. If novelty detection were done with a Gaussian distribution, the shape of those contours would be restricted to an ellipse, however by introducing a mixture one is able to model arbitrarily complex contours in the space of $\mathbf{Y}$. In this case, two Gaussian components were used for this illustration, Chapter 7 discusses the use of BIC for this problem in more detail. The objective of doing this is that of damage detection. Figure 4.6 shows the negative log likelihood for this model, evaluated on data corresponding to the training, testing and damaged conditions, with an empirical 99^{th} percentile threshold shown. It is clear that this simple, yet powerful model can capture the multimodal density inherent in this natural frequency dataset, given that most test data points lie under the threshold, established using the training likelihoods. This is given here for illustrative purposes, hence the use of only two natural frequencies as feature vectors, so that the negative log likelihood contours can be highlighted in two dimensional space. Chapter 7 will focus on the Z-24 case study in more detail.

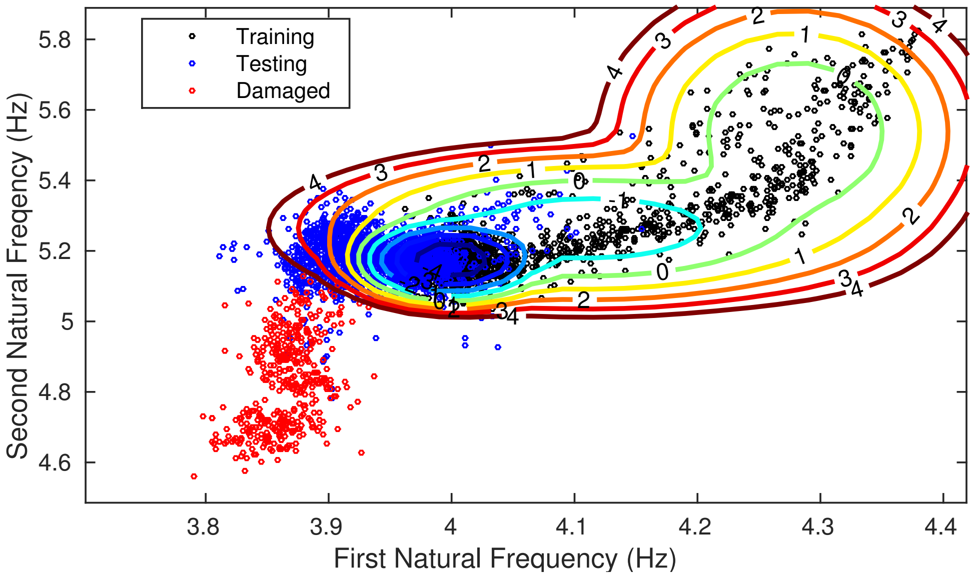

Figure 4.5: Log Likelihood contours for a 3-component Gaussian mixture model fit to the first two natural frequencies of the Z-24 bridge dataset. The last contour plotted corresponds to the 99th percentile threshold on training data points.

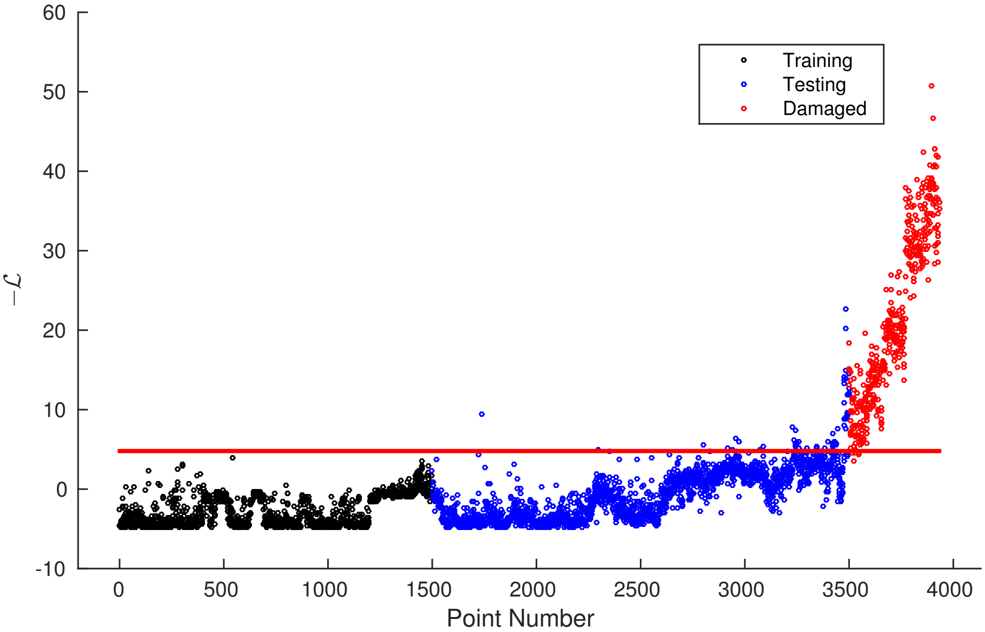

Figure 4.6: Negative log-likelihood for a 3-component Gaussian mixture model fit to the first two natural frequencies of the Z-24 bridge dataset. Note the points up to 1500 were used as a training set, and damage occurs after point 3500.

4.3.3 Fitting Gaussian Mixtures to lower-dimensional data representations

A small digression is made here to discuss why, and how one could combine PCA and Gaussian mixture models for SHM on high-dimensional data. The previous section discussed how one might make use of the density model provided by a Gaussian mixture in order to perform novelty detection using a likelihood function. This approach is well suited to problems where the data $\mathbf{y}$ contains a number of dimensions commensurate to the number of training data points. In general, EM will not work well if the number of observations is not much greater than the number of dimensions. This is because the necessary covariance matrices may not be well defined. Even in the case of a single Gaussian distribution, the covariance matrix may not be well defined if the number of dimensions is disproportionate to the number of observations. This is exacerbated by a mixture model because the number of effective training points is split between the different components; this is a problem if the feature vectors are high-dimensional, but scarce. There is one relevant engineering application where one might encounter such a problem: when using frequency domain feature vectors. The reason the curse of dimensionality is particularly problematic here comes from the inherent trade-off between time and frequency when performing a Fourier transform. If a large window is used when performing a Fast Fourier Transform (FFT), then there will be high frequency resolution, and therefore high dimensionality, but poor time resolution and hence less observations. One may not want to trade time resolution for frequency resolution when monitoring a problem where the shift in natural frequencies due to damage, operational changes, or the environment may be small. In this case, reducing the dimensions of the problem is desirable, and PCA is well suited for this problem.

A nonlinear mass-spring-damper, 3-DOF system is considered here as an illustrative example. This is the same system used later in Chapter 6 as a case study. The nonlinearity is introduced between the first mass and the ground. As discussed above, frequency spectra are a popular damage sensitive feature for this type of system [2], which represents a wide class of systems in structural dynamics. However, these presents a challenge due to the high number of dimensions resulting from a Short Time Fourier Transform (STFT), compared to the low number of observations it yields. Figures 4.7 and 4.8 show the FFT and average FFT vectors for accelerations recorded on the 3rd mass for three different excitation levels: Gaussian noise with $\sigma = (1, 4, 8)$. The feature vectors resulting from these excitation are shown sequentially in the form of a spectrogram in Figure 4.7. The idea of this numerical example is to show how one could use the idea of using PCA to reduce the dimensions of a high dimensional feature vector, fit a Gaussian mixture, and perform novelty detection on the likelihood of this model.

The nonlinear 3-DOF system is used because it presents a challenge if varying load levels are considered. Note that not only do the natural frequencies change at different load levels, but extra harmonics are evident towards the higher loads. In this case, the first two principal components were used in order to visualise the results later. This exercise will be repeated in a later chapter including more dimensions. A four component Gaussian mixture was then fit to the two principal components, and likelihoods were computed (using equation (4.18)) for a training set of 330 points, and a test set of 440 points where the last 110 points come from a damaged system. The damaged system in this case consists of a 10% stiffness reduction on the second spring. The results are presented in Figure 4.9. Note that the contours represent the negative log-likelihood ($-\log \mathcal{L}$) of the Gaussian mixtures, and the last contour shown represents a 95th percentile threshold of $-\log \mathcal{L}$ from the training set. This

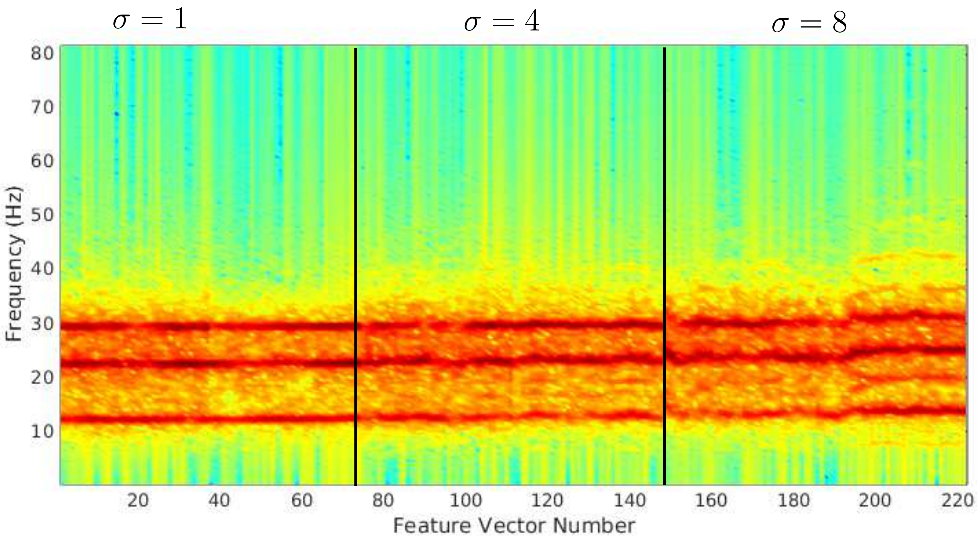

Figure 4.7: Training data set comprising of FFT vectors from nonlinear 3-DOF system excited at increasing loads. Note the natural frequency increases slightly at different levels, and harmonics are evident at the highest loading. The averages of these FFT vetors are shown in Figure 4.8

approach can be successful at segmenting the data in a lower dimensional space, and subsequently detecting damage by setting a threshold on $-\log \mathcal{L}$ which defines contours over the normal condition of $\mathbf{Y}$. This is an illustrative example, so only two principal components were used, to make visualisation easy. However, this does not define a strict probabilistic model, even though this is straight-forward to do through the mixture framework of Bayesian networks; this yields a mixture of PCA models, which is a proper probabilistic model for which a likelihood can be derived directly.

4.3.4 Mixtures of PCA models

The previous example demonstrated how PCA can be used to reduce the dimensions of a high-dimensional feature space, in order for a Gaussian mixture to model the problem. Here, an alternative procedure is presented, by considering a mixture of PCA models. Though the idea of mixture of PCA models has been used in many forms, Tipping introduced the application of an EM algorithm to segment probabilistic PCA models [108]. The approach is analogous to the Gaussian mixture model and fits within the probabilistic mixture modelling framework presented in this chapter. The mixture of PCA models fits a probability model to $\mathbf{Y}$ using equa-

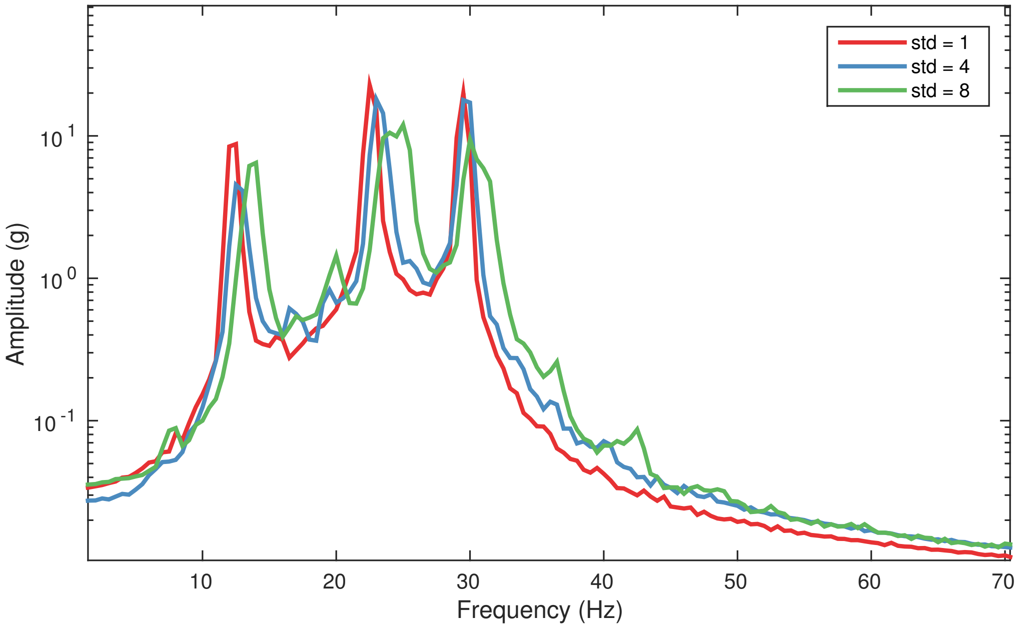

Figure 4.8: Average FFT for the response to three excitation levels on a nonlinear 3-DOF system, grouped by excitation level. The system was excited with white Gaussian noise with standard deviations of 1,4 and 8 N on the first mass.

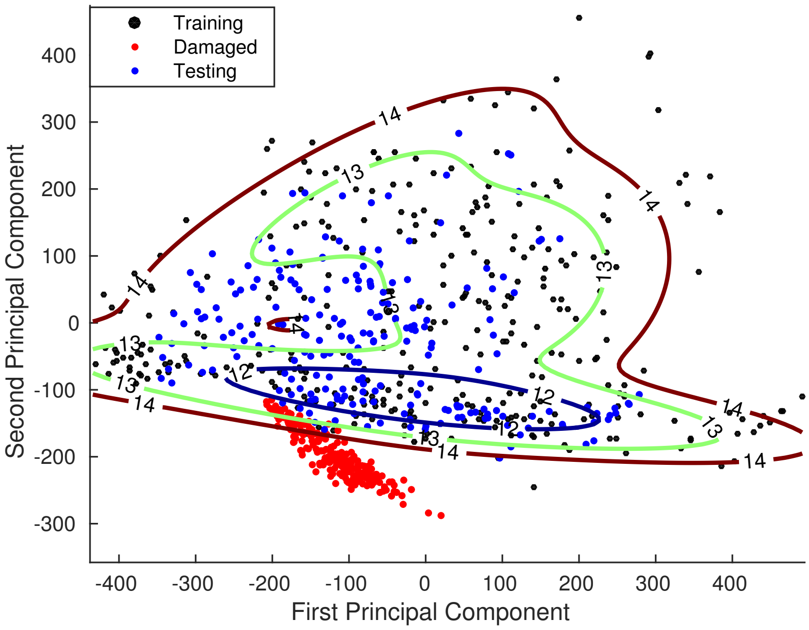

Figure 4.9: First two principal components of FFT from example acceleration of 3DOF nonlinear system with varying loads. The contours show the negative log likelihood of a Gaussian mixture model fit to a training set. The data is coloured by training, testing and damaged condition sets

tion (4.12), from which the likelihood function can be derived, and it fits individual component densities using the covariance for a probabilistic PCA model, described by equation (4.10). It should be stressed that this procedure is different from that of using PCA for dimensionality reduction, and then applying a Gaussian mixture model to the resulting latent variables. In the mixture PCA model, a different PCA model is assigned to different data points.

Note that PCA centres the data, $\mathbf{Y}$, by removing the mean before EM, or SVD are applied. The PCA mixture models the means for data segmentations, as well as their separate covariances. The parameter vectors for a PCA mixture for K components thus consists, for each k component, of a vector of means $\boldsymbol{\mu}_k$, a covariance matrix $\mathbf{R}_k$ and a projection into a lower dimensional space $\mathbf{C}_k$.

The EM algorithm is very similar to that for the Gaussian mixture model. At every iteration, the E step performs likelihood inference on the probabilistic PCA model, and evaluates the posterior distribution over the latent variables $p(\mathbf{x}|\mathbf{y})$ for every k model. In addition, the E step also evaluates the expectation of component responsibilities γ_{ik} for each i point and k component. This is evaluated using equation (4.12), where $p(\mathbf{y}_i|z_{ik} = 1)$ is the probability model for the k^{th} PCA model. The M updates for $\{\boldsymbol{\mu}_k, \mathbf{R}_k, \pi_k\}$ are exactly those for the Gaussian mixture model, in equation (4.20), except that the covariance matrix for the PCA density is denoted $\mathbf{R}_k$ instead of $\mathbf{S}_k$. The updates in the M step for each $\mathbf{C}_k$ (the projection matrix) can be done using the regular probabilistic PCA M step, or can be evaluated using SVD or an eigendecomposition of the covariance matrix $\mathbf{R}_k$, weighed by the responsibility of the k^{th} component (see equation (4.20)).

The log-likelihood derived from a mixture of PCA models, is now simply a sum of individual model likelihoods, weighed by the mixing proportion. This can be written down using equations (4.17) and (4.11) as

$$\mathcal{L}_i = \sum_{k=1}^{K} z_{ik}(\ln \pi_k - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{R}_k| - \frac{1}{2}(\mathbf{y} - \boldsymbol{\mu}_k)'\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}_k)) \qquad (4.23)$$

This likelihood function is a result of placing the likelihood for the simple PCA model, from equation (4.11) into the log-likelihood for a general mixture model (equation (4.17)). PCA mixtures provide a simple, yet powerful way of modelling data, that is inherently multi-regime. Note that the means, $\boldsymbol{\mu}_k$ used in equation

(4.23) denote the individual component means. While PCA provides a linear projection of the observed data into a lower-dimensional manifold, PCA mixtures achieve a projection into a nonlinear manifold, through a divide and conquer approach. The likelihood function of the model (equation (4.23)) can be used for novelty detection in SHM. Once again, this is relevant when the damage sensitive features being used are sensitive to EOVs. The application of PCA mixtures is discussed in detail in Chapter 6, with a simulated system under changing (undamaged) stiffness as well as nonlinearities. While the probabilistic PCA model has been used in an SHM context [81], the PCA mixture model has not, let alone the use of its likelihood function for damage detection.

There is a major caveat with PCA mixtures, related to selecting the number of mixture components. It has been discussed that for a PCA model, the number of principal components can be selected using their contribution to the overal variance (a more classical approach), or Bayesian or Akaike information criteria. Furthermore, Bayesian extensions exist for the probabilistic PCA model [75]. When a mixture framework is applied to mixtures, an important question arises: Should different components be allowed to have a different number of latent variables? If so, the problem becomes very hard. Bishop did consider mixtures in his Bayesian PCA paper [75]; however, only the order of $\mathbf{x}$ (the latent variables) was given a Bayesian treatment, while the number of mixture components was given a BIC-style (penalised likelihood) approach.

A pragmatic approach to this problem would be to compute BIC or AIC functions for different combinations of K (number of mixtures) and P (order of $\mathbf{x}$) where any model only allows an order P for all its components. If different P were allowed for each mixture component, for each model considered the problem complexity grows combinatorially and becomes impractical, at least for the current objective.

4.4 Chapter Conclusions

This chapter has presented an approach to damage detection using linear Gaussian models in the context of Bayesian networks variables. Doing so, it provides a consistent approach to examining likelihood functions with unknown variables, and to estimating those variables, together with model parameters using the EM algorithm. The mixture modelling framework for linear Gaussian generative models is well suited for parameter learning with EM. The models presented here allow modelling of static data, with no temporal relationships. When modelling vibration data, these models are well suited when a feature vector is used that takes into account such temporal relationships (such as Fourier or wavelet coefficients). Two examples were presented in this chapter, that use PCA and mixtures of Gaussians to identify damage in two systems with changing environments and loading conditions; the Z-24 bridge, and a simulated nonlinear 3-DOF system. It was shown that the likelihood functions derived from different combinations of models can be used effectively as density estimators. Within a mixture model framework, the contours of the (feature) data density can be modelled with arbitrary complexity in the underlying density. From this, it can be established that a threshold placed on the data likelihood (or in this case, the negative log-likelihood) defines a contour on the density space within which data from the regular condition lies, including operational and environmental changes, and outside of which the data can be classified as abnormal. The following chapters will present applications of this methodology to various relevant SHM problems and datasets. The next chapter explores an extension of the methodology presented here that accounts for temporal relationships in the data.

Chapter 5

Inference and Learning in Bayesian Networks, Part 2: Dynamic Data

The previous chapter discussed how graphical models can be used to represent data and perform damage detection. The data in this case consist of damage sensitive feature vectors. Natural frequencies and Fourier coefficients were used as examples, but the applicability of the methodology extends to many types of features (recall discussion of appropriate features from Chapter 1). The aspect that features appropriate for this type of modelling have in common is that they already take into account temporal relationships in the dataset. Fourier, wavelet, Hilbert, modal, and other types of domains all encode information contained in raw waveforms and represent the dynamics in a more succinct manner, so that these features could be assumed to contain no temporal correlations. However, in many instances it may be more efficient or beneficial to infer the state of the structure using raw measurements directly. This motivates the contents of this chapter, which discusses the use of *dynamic* Bayesian networks; these extend the concept of modelling data with graphical models to account for temporal relationships between variables. The interesting thing is that most well-known models have a Bayesian network interpretation: Auto Regressive (AR) models, Hidden Markov Models (HMM), Kalman filters, particle filters, and many more. The graphical model interpretation of these models is not necessary for their use within a monitoring context. In fact, some of these models were born out of the need to track and monitor. The Kalman filter was first used to track the position of the Apollo spacecraft to the moon [109]. HMMs have been developed for applications in speech and text recognition. AR models are widely used for for prediction of financial time series data. So, although the Bayesian network interpretation is not necessary for monitoring applications, it offers one key benefit; different models can be viewed as simple variations of each other, and a likelihood function can be derived for each model which can be interpreted in the same way. In other words, the interpretation of the probability of a new data point, given a particular model, is the same across models, and this probability can be used to monitor a process. Moreover, extensions of these models that can deal with more complex, or multiple regime data, are readily available via the mixture framework, introduced in the previous chapter. Naturally, this chapter will discuss how dynamic Bayesian networks fit within the mixture framework, and some interesting extensions of these are discussed.

The information that one can extract from modelling a dataset using Dynamic Bayesian Network (DBN)s is vast, but this chapter is strictly focused on answering one question: how can a model be used to compute the probability that a new measurement is abnormal. In effect, this can be done by computing model negative log likelihoods which give an indication of novelty and in the context of SHM, indicate damage.

5.1 Dynamic Bayesian Networks as Generative Models

The concept of generative, latent variable models, introduced in the previous chapter for static data, can be extended directly to a dynamic model. If one considers for example a first order Markov process, as shown in Figure 5.1 (an n^{th} order Markov process assumes that a temporal variable is dependent at most on the previous n values), the probabilities for dynamic data models follow directly from the tail-to-head pattern shown in Figure 4.2. Following directly from the d-separation property, it can be seen that given the value of y_t, y_{t-1} and y_{t+1} are conditionally independent.

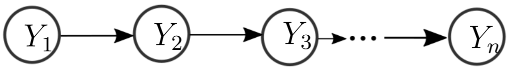

Figure 5.1: First order Markov process

Their joint probability distribution is then

$$p(y_1, y_2, ..., y_n) = p(y_1) \prod_{t=2}^{n} p(y_t|y_{t-1}) \tag{5.1}$$

Recall that the key idea behind latent variable models is to express the observed variables by a hidden layer of unobserved/latent variables. This means modelling each observation (y_t) as temporally independent of the others, and to model the underlying dynamics through the latent variables, x. The Bayesian network shown in Figure 5.2 illustrates this idea; in this case, $Y = (y_1, y_2, ..., y_n)$ are the observed variables, and $X = (x_1, x_2, x_n)$ are the hidden or latent variables, which encode the temporal relationship. Using d-separation, given the value of x_t, y_t is conditionally independent of all previous and future x_n, so the joint probability factorises to

$$p(\mathbf{X}, \mathbf{Y}) = p(x_1)P(y_1|x_1) \prod_{t=2}^{n} P(y_t|x_t)P(x_t|x_{t-1}) \tag{5.2}$$

This joint distribution will be returned to later, in Section 5.3.1, when discussing inference for Kalman filters, Hidden Markov models, and variations of these. However, it is worth beginning the discussion of inference in DBNs with a popular model that is non-generative: linear autoregressive models. They represent probably the simplest kind of DBN, but their application is extensive in SHM and other engineering fields.

5.2 Auto-Regressive modelling

Auto-Regressive (AR) models are now very popular in the SHM field because of their flexibility in modelling signals, their interpretability, and their ease of extension to

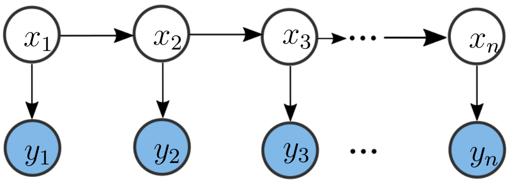

Figure 5.2: Probabilistic graphical model describing a latent variable model with underlying dynamics. If the relationships between the nodes are linear, this corresponds to the state space model of equation (5.8)

nonlinear models. The purpose of this section is simply to formulate these well-known models in terms of a dynamic Bayesian network, to motivate this treatment.

The idea behind linear AR models is very simple; they expresses a signal y_t as a linear combination of its previous p values, through a weighting vector $\mathbf{a} = \{a_1, a_2, ..., a_p\}$

$$y_t = \sum_{l=1}^{p} a_l y_{t-l} + \epsilon_t \tag{5.3}$$

where a_l is the AR coefficient corresponding to the l_{th} lag of y_t. The residual term, ϵ is assumed to be Gaussian distributed with zero mean and variance σ^2. The graphical model for an AR process of order p is shown in Figure 5.3. The joint probability of all observations under this model follow from the graph, and also by noting that this is a p^{th} order Markov process. Extending equation (5.1) to include p lags yields

$$p(y_{1:T}) = \prod_{t=2}^{n} p(y_t|y_{t-1}, ..., y_{t-p}) \tag{5.4}$$

where the values y_t at index points $t \leq 0$ can be set to zero. The conditional probability of an observation given p lagged values is a Gaussian distribution, with the mean given by the predicted value of y_t evaluated using equation (5.3). More formally

$$p(y_t|y_{t-1}, ..., y_{t-p}) = \mathcal{N}(y_t| \sum_{l=1}^{p} a_l y_{t-l}, \sigma^2) \tag{5.5}$$

Learning the AR parameter vector, $\mathbf{a}$, is straightforward using a maximum likelihood approach, and this fits well within the context of graphical models. It can be shown (see for example [83, 49]) that the maximum likelihood estimate for the AR parameters is essentially a standard least squares solution

$$\mathbf{a} = \Big(\sum_t \mathbf{y}_l \mathbf{y}_l'\Big)^{-1} \sum_t (y_t \mathbf{y}_l) \tag{5.6}$$

where $\mathbf{y}_l$ denotes the vector of p lags of y, corresponding to time t, $\mathbf{y}_l = \{y_{t-1}, ..., y_{t-p}\}$ and y_t denotes the observation at time time t.

In a monitoring context, an AR diagnostic process can be implemented in two ways; by monitoring the residuals of model predictions, or by tracking the model parameters. The former approach, is similar in principle to extracting a frequency spectrum from a signal. In fact, a spectral representation of the signal can be recovered from the AR coefficients. The vector of AR parameters $\mathbf{a} = \{a_1, ...a_p\}$ can be a useful feature vector, and if the process generating the data is nonstarionary (or even non-linear), and enough measurements are available, the density of $\mathbf{a}$ could be modelled with a Gaussian mixture model, decomposed with PCA and/or modelled with a mixture of PCA models, to name a few options. This is the approach suggested in this work, if AR coefficients were used as feature vectors. The case study in Chapter 6 uses this principle but with Short Time Fourier Transform (STFT) coefficients. However, depending on the context it may be more appropriate to use AR coefficients. The only reason this is highlighted here is because AR models are an important sub-class of state-space models, and it should be clear that there are various ways one may implement them in a monitoring context.

If AR coefficients were to be used as feature vectors, a training dataset would have to be partitioned into windows, of at least the length of the parameter vector. This would be in a similar fashion as the windowing applied to a signal during a Short-Time Fourier Transform. Alternatively, $\mathbf{a}$ could be estimated sequentially using a method such as Recursive Least Squares, or as will be shown later, with a Kalman filter.

As an alternative to tracking the AR parameter vector, the residuals of the AR process can be monitored. However, a linear AR model, as presented in equation (5.3) is limited to characterising stationary processes. Unless some form of flexibility is added to the model, such as a mixture or a switching variable, a linear AR model may be too restrictive. These extensions will be discussed in the context of state space models later in this chapter. Furthermore, if an AR model is to be used to monitor the output of a vibrating system, the excitation source must be considered

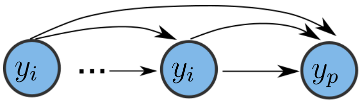

Figure 5.3: Graphical model representation of an AR model of order p

beforehand. It is a well known result that the output of a linear time-invariant second-order dynamical system, excited by white noise, can be expressed as an Autoregressive Moving Average (ARMA) model:

$$y_t = \sum_{j=1}^{p} a_j y_{t-j} + \epsilon_t + \sum_{j=1}^{q} m_j \epsilon_{t-j} \tag{5.7}$$

where m_j represents the moving average coefficient of the j^{th} lag of the white noise residual ϵ_t. The model can be used to make predictions; these predictions can be compared with the measurements to generate a residual, and an abnormal condition can be defined as an exceedance of the residual variance. If the residual grows, then something has changed in the process that generates the data. An issue with this process is that the AR model makes an assumption of stationarity in the data, and this may not always be the case. Depending on the type of nonstationarity present in the data, the problem could still be tackled with AR models in different ways. AR and ARMA models are relevant to the problem of structural vibration because the response of a linear dynamical system to a sinusoidal excitation can be described by a linear AR model. Conversely, the response of a linear dynamical system excited by white noise can be described by an ARMA process.

5.3 State Space models

State-space models have already been introduced in Section 2.1.3 in the context of Operational Modal Analysis (OMA). Here, they are revisited in more detail, with an emphasis on the formulation of their likelihood function and their use in an SHM context.

When monitoring a dynamical system, one is often restricted to making measurements that do not necessarily measure the state of the system directly, but are related to the underlying process driving those measurements through some function. A state space model provides a solution to this type of problem; it models the observations $\mathbf{y}_t$ as some function of the underlying dynamics of $\mathbf{x}_t$. If the relationship between $\mathbf{x}$ and $\mathbf{y}$ is linear, then this can be represented by the standard linear state space formulation:

$$\begin{aligned} \mathbf{x}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t \quad \mathbf{w}_t \sim \mathcal{N}(0, \mathbf{Q}) \\ \mathbf{y}_t &= \mathbf{C}\mathbf{x}_t + \mathbf{v}_t \qquad \mathbf{v}_t \sim \mathcal{N}(0, \mathbf{R}) \end{aligned} \tag{5.8}$$

where $\mathbf{C}$ represents the linear function linking observations to underlying dynamics, and $\mathbf{A}$ represents the linear dynamics; the time evolution of $\mathbf{x}$. The observation model, and the dynamics are both modelled with white Gaussian noises, $\mathbf{v}_t$ and $\mathbf{w}_t$, with zero mean, and covariance matrices $\mathbf{R}$ and $\mathbf{Q}$ respectively [1]. State space models have applications spanning various areas of science and engineering, and are now implemented in other fields such as financial time series modelling. They are useful whenever one makes use of measurements that can be somehow related to the underlying state of a system, and those measurements are corrupted by noise. It is easy to relate this to a structural dynamics context. A MDOF linear vibrating system can be described by a linear superposition of SDOF systems; this is the foundation of modal analysis. The underlying driving functions are the SDOF oscillators, each of which has a characteristic natural frequency, and they are related to a physical location on the structure via a mode shape. Any good structural dynamics textbook will contain methods for representing MDOF systems in state space form. There may be more than one valid state space representation of a system and one must adopt the one that is most suitable to the problem at hand. In SHM, there could be two different contexts where one might seek a state space representation:

1. The state vector $\mathbf{x}$ represents the parameters of a model, and one is interested in estimating those parameters, as they evolve through time.
2. The state vector $\mathbf{x}$ represents some hidden variables that better model the underlying linear dynamics of a set of measurements $\mathbf{y}$.

In the first case, the state space modelling approach is cast as a parameter identification problem. In this case, a state space model could be seen as an alternative

[1] Standard linear state-space models also include additive terms to account for control inputs. These are omitted here for simplicity, and since control inputs are not used in this work

view to the popular Recursive Least Squares (RLS) algorithm (created by Gauss himself!), which essentially solves the standard OLS problem, point-by point. Section 5.3.2 discusses the application of a state space model to the identification of the AR parameter vector discussed above. In this interpretation of state space models, the state vector will usually have some direct physical meaning, and it will often be easy to interpret the results. In the second interpretation, the meaning of $\mathbf{x}$ can be a lot more subtle, in fact it does not necessarily need to have physical meaning at all. In this interpretation, if one possesses a physical model of the structure being considered (say, via an FE or an analytical model), then $\mathbf{C}$ and $\mathbf{A}$ can be derived, and one is interested in inferring the state $\mathbf{x}$ and in monitoring the residuals $\mathbf{v}_t$ and $\mathbf{w}_t$. An estimate of $\mathbf{A}$ and $\mathbf{C}$ that correctly describes the measured data is therefore required, and for this it is useful to turn to the Bayesian network interpretation of a state space model as it readily provides estimates of $\mathbf{x}$ as well as parameter identification methods for all the state space model parameters, through the use of EM, in the same way that has been done for the PCA and mixture models described in Chapter 4.

Both of these approaches are discussed in more detail in the Sections 5.3.2 and 5.3.3. Even though in both cases $\mathbf{x}_t$ represents something fundamentally different, both have useful SHM applications, and most importantly, both require a method for providing an estimate for $\mathbf{x}_t$. The graphical model of Figure 5.2 represents the state space model of equation (5.8). All the necessary probability relationships can be evaluated using this graphical model, but if linear Gaussian assumptions are made, the Kalman filter algorithm provides an efficient and intuitive solution for the inference problem, and so it is applicable to both the parameter estimation, and latent variable approach to implementing state space models for novelty detection in SHM.

5.3.1 Inference via the Kalman filter

The key point of the graphical model interpretation of a linear dynamical system is that the observations are modelled as independent of each other, and dependent only on the state vector at time t, while the state vector is dependent on all previous values on the Markov chain. From examination of the graph in Figure 5.2, two conditional probability relationships are evident: the probability of the observation vector given the state vector $p(\mathbf{y}_t|\mathbf{x}_t)$, and the probability of the state vector given the same state vector at a previous point in time $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. As previously discussed, if these two densities are assumed to be linear Gaussian they can be written down as:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\mathbf{A}\mathbf{x}_{t-1}, \mathbf{Q}) \tag{5.9}$$

$$p(\mathbf{y}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{y}_t|\mathbf{C}\mathbf{x}_t, \mathbf{R}) \tag{5.10}$$

Note that these probabilities convey effectively the same relationships as the classical definition of a linear state space model, described by equations (5.8). One quantity of interest is the complete data probability, or complete data likelihood. This can be written, using the sum rule of probability, as:

$$p(Y) = p(\{\mathbf{y}_1, ...\mathbf{y}_T\}) = \int_X p(X, Y)dX \tag{5.11}$$

The joint probability $p(X, Y)$ can be evaluated using the joint probability relationship for a dynamic generative model, given in equation (5.2), where the conditional probabilities have already been shown in equations (5.9) and (5.10). Summing over the whole data set in log-form yields [110]:

$$\begin{aligned}\log p(X, Y) = & -\frac{1}{2}\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{C}\mathbf{x}_t))'\mathbf{R}^{-1}(\mathbf{y}_t - \mathbf{C}\mathbf{x}_t) + \log|\mathbf{R}| \\ & -\frac{1}{2}\sum_{t=2}^{T}(\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1}))'\mathbf{Q}^{-1}(\mathbf{x}_t - \mathbf{A}\mathbf{x}_{t-1}) + \log|\mathbf{Q}| \\ & -\frac{1}{2}(\mathbf{x}_1 - \mu_1)'\mathbf{Q}_1^{-1}(\mathbf{x}_1 - \mu_1) - \log|\mathbf{Q}_1| - \frac{T(p+k)}{2}\log 2\pi\end{aligned}$$

This complete data log-likelihood is particularly important for learning purposes; it is required to perform learning using the EM algorithm.

For inference purposes, the interest is in estimating the unknown state vector, $\mathbf{x}_t$. Because $\mathbf{x}_t$ is not observed, the true state of the system will never be known, but a probability density can be estimated, and if Gaussianity is assumed on the residuals of the dynamics of the state, the distribution is fully specified by the mean and variance of the state vector at every point in time. There are three probability distributions for $\mathbf{x}_t$ of interest, and they are commonly referred to (in the statistical time series communities) as prediction, filtering and smoothing, where each of them compute the following conditional probabilities:

1. Prediction: Probability of state vector, $\mathbf{x}_t$ at time t given observations up to time $t-1$, $p(\mathbf{x}_t|\mathbf{y}_1, ..., \mathbf{y}_{t-1})$.
2. Filtering: Probability of state vector, $\mathbf{x}_t$ given observations up to time t $p(\mathbf{x}_t|\mathbf{y}_1, ..., \mathbf{y}_t)$.
3. Smoothing: Probability of state vector, $\mathbf{x}_t$ given *all* the observations available $p(\mathbf{x}_t|\mathbf{y}_1, ..., \mathbf{y}_T)$.

Prediction filtering and smoothing relationships can all be derived from the conditional probabilities encoded in the graphical model of the specific dynamic Bayesian network. For the state space model (the network in Figure 5.2) the filtering distribution can be shown to be:

$$p(\mathbf{x}_t|\mathbf{y}_{1:t}) \propto \int_{x_{t-1}} p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}_{1:t-1}) \tag{5.12}$$

The filtering distribution is important as it is central to deriving an algorithm that computes the probability $p(\mathbf{y}_t|\mathbf{y}_{1:t})$, which is ultimately required for damage detection.

So far this does not assume linearity or Gaussianity. One is free to model the conditional probabilities inside equation (5.12) with any arbitrarily complex distribution. However, the assumption of linearity and Gaussianity simplifies things significantly because of Gaussian identities; the product of two Gaussians is itself a Gaussian, and the integral of a Gaussian is Gaussian too. These properties are what make the representation of the conditional densities as Gaussians (equations (5.9) and (5.10)) so efficient. The Kalman filter algorithm effectively solves equation (5.12) for the linear Gaussian case. These probabilities could also be evaluated directly using other algorithms for Bayesian network inference, such as variable elimination (this would be highly inefficient) [83] or the standard junction tree algorithm [101]. The approach of using the junction-tree algorithm together with EM to do inference and learning on a large class of dynamic Bayesian networks has been extensively explored by Murphy [111], in the context of speech processing. The Kalman filter approach, however is arguably one of the most efficient (bearing in mind that there are optimised versions of it, such as the square root filter [112]) when dealing with a linear Gaussian dataset, and so its use is discussed extensively here. Other kinds of models exist, of course, but the argument being made here is that a dynamic Bayesian network is viable tool for SHM and so the Kalman filter is thus a natural place to start. Even though it is now over 40 years old, it is a well studied algorithm and widely used in engineering industry. It gives an efficient representation for an ARMA process, and provides a systematic and efficient way of computing data probabilities without the need for a specific choice of features. Furthermore, it can be easily extended to model more complex data.

The evaluation of the three different state probabilities (prediction, filtering and smoothing) depends on the context. The first two are provided in the first and second steps of the Kalman algorithm, while computing the "smooth" probabilities usually involves running the Kalman algorithm forwards and backwards. The smoother step not only yields lower error predictions of the state, but is also necessary to formulate an EM algorithm for parameter learning. The Kalman filter algorithm is described in Appendix A.

5.3.2 Inference for parameter identification

The Kalman filter is similar in nature to the Recursive Least Squares (RLS) algorithm. RLS has been extensively investigated in the structural dynamics community as a method for identifying system parameters in real-time [113, 114, 56]. A sequential or recursive form for a linear AR parameter estimate, mapping y_t (note it is one -dimensional) to its lagged version, can be described by letting the state space vector $\mathbf{x}_t$ be the AR parameter vector $\mathbf{a}$ (as defined as in Equation (5.6)):

$$\mathbf{x}_t = \mathbf{x}_{t-1} - K_t(y_t \mathbf{x}_t - y_t) \tag{5.13}$$

where K_t is a function of the filtered variance, v_t, at time index $(t-1)$, which represents the confidence in the state vector (in this case the parameters)

$$K_t = v_{t-1} y_t (1 + v_{t-1} y_t^2)^{-1} \tag{5.14}$$

and the update to the variance is:

$$v_t = v_{t-1} - v_{t-1}^2 x_t^2 (1 + v_{t-1} y_t^2)^{-1} \tag{5.15}$$

These recursive parameter estimates can be shown to give the same solution that OLS provides, once all the time steps have been processed. The equations above show the RLS parameter updates for a single variable case, but the generalisation to a multivariate form is essentially a matrix version of equations (5.13), (5.14) and (5.15). This version is in fact provided by the Kalman filter recursions (described in Appendix A). Using the Kalman filter recursions, RLS can effectively be achieved for a variety of problems. The state vector $\mathbf{x}_t$ represents the parameters, and the observation matrix $\mathbf{C}$ would represent the observations. The state transition matrix $\mathbf{A}$ can have different forms, but a popular choice is to set it to be an identity matrix; this assumes that the parameters follow a random walk, the volatility of which is provided by the state-transition noise model, $\mathbf{w}_t \sim \mathcal{N}(0, \mathbf{Q})$. Different choices of the state transition matrix could put different constraints on the parameter updates. One could solve for a variety of problems through a careful choice of $\mathbf{C}$ and $\mathbf{A}$. An application of interest in SHM is to estimate the AR coefficient vector recursively. This could be modelled by setting the observation matrix $\mathbf{C}_t$ in every Kalman filter recursion to be the lags of the signal of interest y. In other words, the observation matrix varies with time, and is:

$$\mathbf{C}_t = \{y_{t-1}, ..., y_{t-p}\} \tag{5.16}$$

for an AR model with p lags. Because $\mathbf{A}$ is defined as an identity matrix, the state transition reduces to specifying that the AR parameters should be close to the previous ones in time, to within a specified variance:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{w} \tag{5.17}$$

As an illustration of this approach, Figure 5.4 shows the fit of an AR model with 40 lags to the response of the second mass of the 3-DOF nonlinear system, while it undergoes a step change in the response due to the nonlinearity. Note that the estimates for both the AR parameters ($\mathbf{x}_t$) and their variances, $\mathbf{w}$ change as the system changes its dynamics. It is useful to be able to infer parameters in real-time, and as pointed out before, these could be used as features in novelty detection based on static data. The Kalman filter here is merely just providing estimates for those parameters, but nevertheless they need to be chosen, through the structure of $\mathbf{C}$. This is shown here for illustrative purposes, since the approach used here uses inference for monitoring of residuals, through the model likelihood function. However, it should be noted that this approach to tracking AR parameters in SHM has not currently been explored.

Evaluating the probability of observations given the model (the likelihood function) for this model does not yield a particularly useful novelty detection method, because the model of the data is embedded in the parameter vector $\mathbf{x}_t$, which is constantly changing with time. A better approach in this case is to fix the model parameters, and evaluate the likelihood of new observations against this model. This fits with the approach taken in this thesis and is discussed in the following section.

5.3.3 Inference for novelty detection using likelihood function

While using a dynamic Bayesian network for parameter inference may sometimes be beneficial, in particular when interpretability is important, inference can be applied directly on observed vibration (or other) data if the state transition matrix $\mathbf{A}$ is allowed some meaningful structure. Novelty detection is performed in this case using the residuals of the predictor, filter or smoother. This is not a foreign concept in SHM; several studies have been carried out where a time series model is used to make forward predictions in time using a variety of models [115, 47, 35, 2]. Linear AR and ARMA models have been extensively used as predictors as well as their nonlinear extensions; some base the predictions on nonlinear models such as neural networks, Gaussian Process (GP) regression [55] or Support Vector Machines [54]. The idea presented here is more subtle; the likelihood function of a dynamic Bayesian network can be systematically used as a measure for damage detection, as it represents the probability of the data being generated by a given model. This could arguably be the best measure of change in a system, if for instance, one is dealing with a structure known to be linear (or operating within its linear regime), and with stationary known (or Gaussian) forcing. In practice, not all structures will behave like this, and even the structures that operate within their linear regimes may have environmental and operational changes that mean a single predictor, no matter how complex it is, may never be able to predict the response of the structure

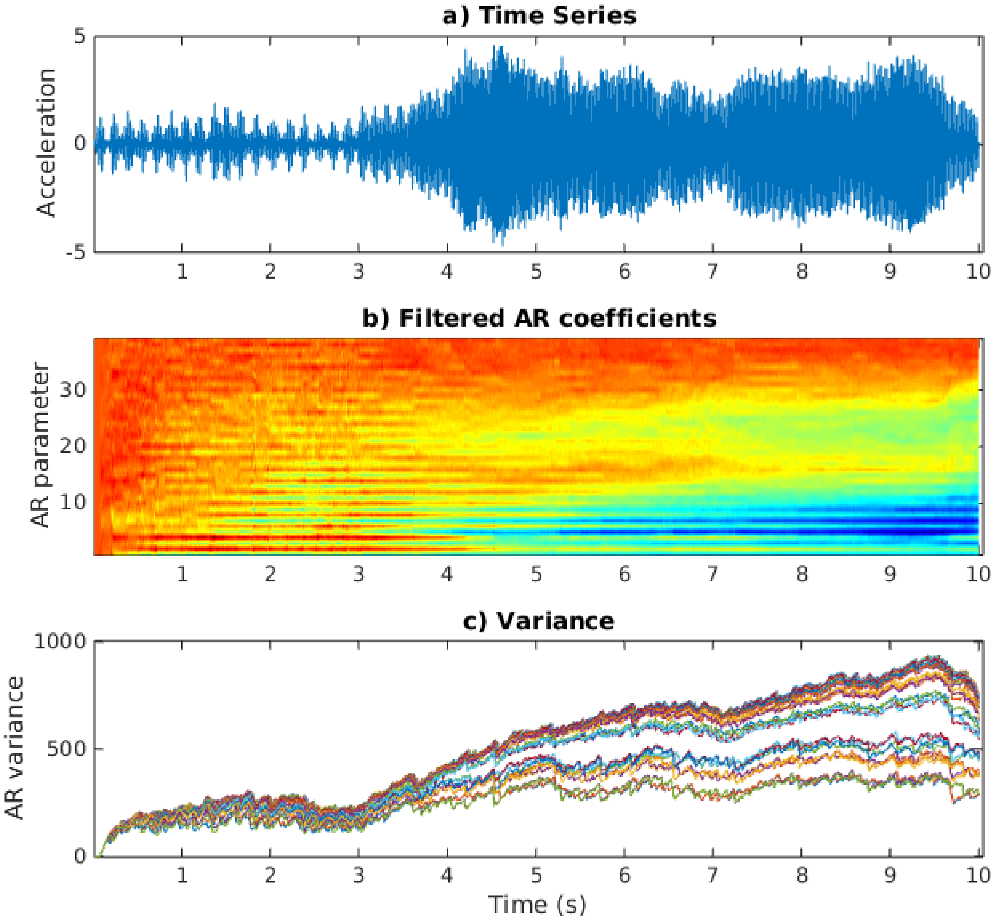

Figure 5.4: Sequential estimation of AR parameters using a Kalman filter, a) shows the response of a nonlinear 3-DOF system to white noise, with a step change in response. b) shows the state vector $\mathbf{x}_t$, containing the AR parameters, and c) shows the variance of each dimension of $\mathbf{x}_t$.

across all its regimes. The usefulness of the Bayesian network interpretation is the flexibility it adds. Examples of this flexibility would be the relative ease with which external inputs to a system could be added, and the natural extensions for modelling more complex data via mixture distribution frameworks.

Bayesian networks allow for a principled way of adding domain knowledge into a system. The likelihood function may be easily derived for a large class of models, and the necessary integrals for prediction, filtering and smoothing recursions (such as equation (5.12)) may be simple to derive. The linear Gaussian case is one of such cases. In nonlinear and non-Gaussian instances, the integrals may be analytically or numerically intractable, and one could resort to other methods for their solution. Approximate methods such as the Unscented Kalman filter, and sampling approximations such as particle filtering can be used for this, but these are discussed at the end of the chapter.

The quantity of interest, for novelty detection is the marginal probability of the observation, $p(\mathbf{y}_t)$. The previous chapter showed how this likelihood function could be evaluated point-by-point. This is sufficient when the data being modelled does not contain any dynamics. The gain of augmenting a generative model for static data, such as PCA (see Figure 4.3) is that one can evaluate the conditional density $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$[2], the predictive distribution of $\mathbf{y}_t$ conditioned on its previously observed values. The likelihood of the complete data Y can be evaluated using this conditional

$$p(Y) = p(\mathbf{y}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \tag{5.18}$$

In the linear Gaussian case, the Kalman filter recursions let one evaluate this quantity exactly, for every point in time. The implication of this is big, for novelty/damage detection; not only the mean of the predictions over $\mathbf{y}_t$ is being updated recursively, but the uncertainty over those predictions is also being updated recursively. This approach contrasts significantly with simply taking residuals from a time series model such as AR, or its various extensions: NAR, NARMAX, etc.

[2] The notation $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ is used here as a short-hand to denote the probability of $\mathbf{y}$ given all measurements from one to $t-1$.

In the linear Gaussian case, evaluation of equation (5.18) yields

$$\log p(Y) = -\frac{1}{2}\sum_{t=1}^{T}\Big((\mathbf{y}_t - \boldsymbol{\mu}_t)\boldsymbol{\Sigma}_t^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_t)' + \log 2\pi|\boldsymbol{\Sigma}_t|\Big) \tag{5.19}$$

where the mean, μ_t is the projection of the *predicted* state vector $\mu_t = \mathbf{C}\mathbf{x}_{t:t-1}$, and the covariance matrix $\boldsymbol{\Sigma}_t$ is the projection of the predicted state uncertainty $\mathbf{V}_{t:t-1}$ onto the measurement space, given by

$$\boldsymbol{\Sigma}_t = \mathbf{C}\mathbf{V}_{t:t-1}\mathbf{C}' + \mathbf{R} \tag{5.20}$$

Note that this is projection is used when computing the Kalman gain in equation (A.3) (recall that $\mathbf{R}$ is the covariance of the measurement noise) In effect, $\boldsymbol{\Sigma}_t$ encodes the uncertainty of the measurements at time t given the model parameters $\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{C}, \mathbf{Q}, \mathbf{R}, \mathbf{P}_0, \mathbf{x}_0\}$. The approach to monitoring now depends on the nature, and availability of the data. If a data stream is readily available, then a good approach would be to evaluate $\log p(Y)$ for sections of data, noting that the value is dependent upon the number of points, T. This is easily solved by using the expected log-likelihood (normalising over T). If the data are more scarce, or real-time monitoring is necessary, point estimates of the uncertainty over $\mathbf{y}_t$ are required. Using the conditional probability $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$, point-by-point would be a better approach:

$$\log p(\mathbf{y}_t|\mathbf{y}_{1:t-1}) \propto -(\mathbf{y}_t - \mu_t)\boldsymbol{\Sigma}_t^{-1}(\mathbf{y}_t - \mu_t)' \tag{5.21}$$

The reader familiar with SHM or machine learning would recognise that this is a (negative) Mahalanobis distance over the residuals of model predictions. The contrast between monitoring residuals of a time series model with a Mahalanobis distance, and the approach presented here based on a linear Gaussian dynamic Bayesian network, is that the uncertainty is updated recursively; Bayes' theorem is used to sequentially estimate the uncertainty of the measured data, taking into account overall measurement and model uncertainty, albeit with Gaussian assumptions which significantly simplify treatment. The correct learning of the parameter vectors is clearly important. The EM algorithm offers a solution here for maximum likelihood parameter estimation, although it is prone to getting stuck in local minima. The EM iterations for a linear dynamical system are derived in [110], and not presented to keep the discussion concise. The underlying assumption throughout this thesis is that the system excitation is unknown; if this is the case, Stochastic

Subspace Identification (SSI) offers an alternative solution to parameter learning, although it does not explicitly solve for the system covariances $\mathbf{Q}$ and $\mathbf{R}$, but SSI is not prone to local minima.

To summarise, the process for performing damage detection using a Bayesian network, as applied to vibration signals is as follows:

1. Select a training data set, Y_{train} consisting of an MDOF vibration response from the structure in question, in an undamaged state.
2. Select a subset of vibration data available from an undamaged state to use for testing/validation Y_{test}.
3. Iterate EM steps until convergence of the likelihood function. To avoid local sub-optimal optimisation, restart the EM algorithm multiple times, and select the model parameters with the highest negative log-likelihood. Alternatively, start the model with an SSI.
4. Evaluate the conditional density $p(\mathbf{y}_{1:T})$, either point by point or for groups of data and derive the negative log likelihood,$-\log \mathcal{L}$, and set a threshold $\mathcal{T}$ that bounds $-\log \mathcal{L}$ for the training set.
5. Evaluate $-\log \mathcal{L}$ on the testing data set, and ensure that it remains within $\mathcal{T}$, the threshold with a low level of false positives.
6. If the model correctly captures the variability of the healthy condition, exceedances of $\mathcal{T}$ indicate damage, or other previously unseen changes.

Note that this is very similar to the procedure given in the previous chapter for static data models, except that one does not need to choose a damage sensitive feature vector; inference is carried out directly on the measured data. This takes away a major decision from the user, and lets the model represent the data directly. As an illustrative example, the procedure outlined above is applied to vibration data from the simulated linear dynamical system. As before, damage is simulated by introducing a 10% reduction in the stiffness of the second mass. The resulting point-by-point evaluation of the negative log likelihood is shown in Figure 5.5. This approach to novelty detection is very useful, as it is inherently suitable for real-time monitoring applications, the obvious limitation is the ability to only model linear Gaussian systems. There are however, numerous extensions of this model that deal well with nonlinear and non-Gaussian data; some of these are reviewed in the next section.

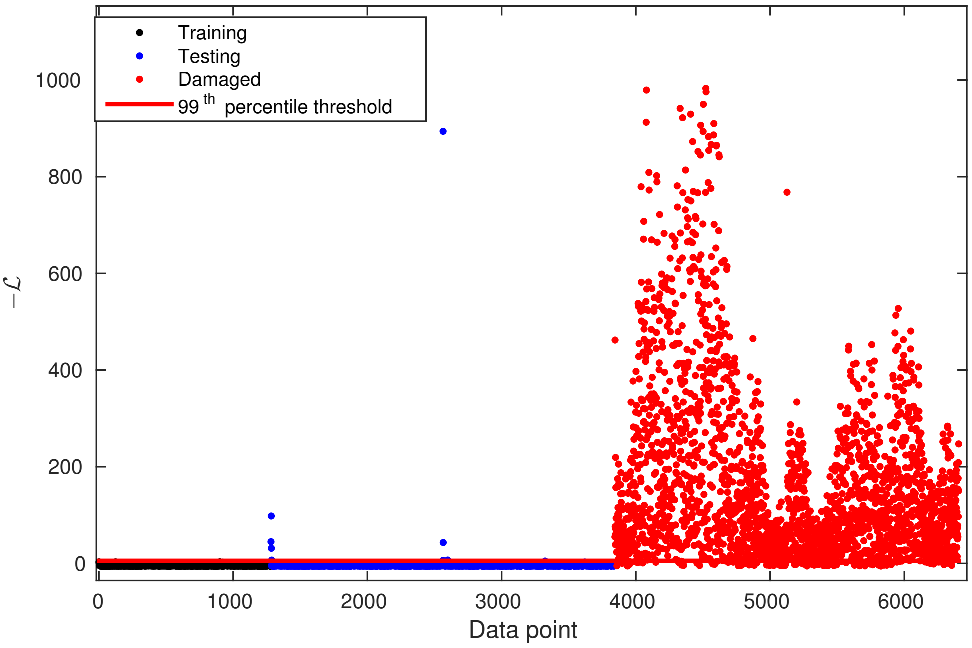

Figure 5.5: Negative log likelihood, evaluated point-by-point for the 3-DOF linear system example. The negative log likelihoods are shown for the training, testing, and a damaged case consisting of a 10% stiffness reduction on mass 2.

5.4 Dealing with environmental and operational variability

The framework of dynamic Bayesian networks has been introduced within a damage detection context, and it has been discussed that $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ is a useful measure of novelty in a data point (or vector), being a building block of the complete data likelihood of a dynamic Bayesian network. The true power of dynamic Bayesian networks is the flexibility they allow. In this section, extensions of the basic (generative) dynamic Bayesian network, of Figure 5.2 are discussed; these enhance the basic model in various ways. The extensions can allow one to model nonlinear and/or non-Gaussian data, and they can also allow the introduction of prior engineering knowledge about the system in question, into the model.

Modelling multiple linear regimes can be achieved with simple extensions to the

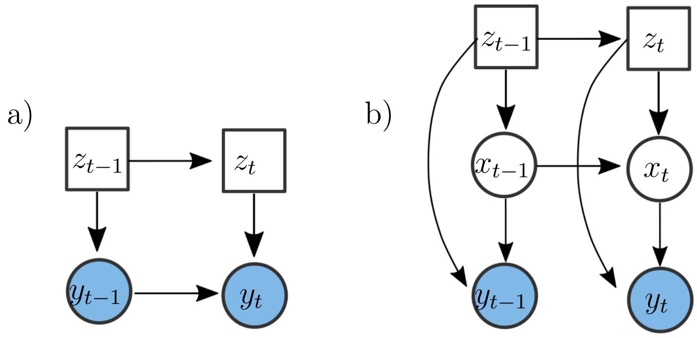

Figure 5.6: a) Bayesian network representation of a 1st order switching autoregressive model and b) Bayesian network representation of a switching linear dynamical system.

autoregressive (Figure 5.3) and the linear dynamical system models (Figure 5.2). In the framework of dynamic Bayesian networks, this can be done by using a discrete switching variable; this is shown for an AR and a linear dynamical system in Figures 5.6a and 5.6b respectively. Note that the switching variable $\mathbf{z}_t$ in this case is conditioned on $\mathbf{z}_{t-1}$. This results in a model with transition probabilities between the switching of states. Inference for this type of model effectively yields a Hidden Markov Model (HMM) on top of the original dynamic Bayesian network. A HMM is basically the discrete counterpart of the linear Kalman filter, where the state vector $\mathbf{x}$ is not continuous, but discrete. The conditional probabilities ($p(\mathbf{y}_t|\mathbf{x}_t)$ and $p(\mathbf{x}_t|\mathbf{x}_{t-1})$), used for inference in the Kalman filter are now discrete probability tables for the HMM. Inference for the HMM will not be discussed at length here as it would digress from the main point, but the interested reader is referred to [49, 83, 26]

In general, the switching variable is modelled as unobserved (one may not know a-priori segmentation due to EOVs) so one must learn the segmentation of the data and the individual dynamical models in the training step, and infer the most likely model, during inference. Inference for the case of the switching AR model is relatively simple and well studied, so it will be discussed here briefly. It has also been used in SHM applications before [35, 2]. From Figure 5.6, it can be seen that the joint distribution for an order p AR model is given by:

$$p(\mathbf{y}_{1:T}, \mathbf{z}) = \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{y}_{t-1}, ..., \mathbf{y}_{t-p}, \mathbf{z}_t)p(\mathbf{z}_t|\mathbf{z}_{t-1}) \tag{5.22}$$

where the conditional density is given by an AR model, with a changing parameter vector, $\mathbf{a}_k$, governed by $\mathbf{z}$. Note that the the switching variable has a transition probability, $p(\mathbf{z}_t|\mathbf{z}_{t-1})$, which dictates the likelihood of transition from model k to model j. The probability of observation $\mathbf{y}_t$ given the model component $\mathbf{z}_t$ is thus:

$$p(\mathbf{y}_t|\mathbf{z}_t) = \mathcal{N}\Big(\mathbf{y}_t|\mathbf{y}_{t-1:t-p}\mathbf{a}(\mathbf{z}_t), \sigma^2(\mathbf{z}_t)\Big) \tag{5.23}$$

On its own, this effectively describes the conditional density corresponding to a mixture of AR models, and this could define a model on its own. However, because a probability has been placed over the component transition, the filtering recursion over the discrete switching variable, derived through a HMM can be written as:

$$p(\mathbf{z}_t|\mathbf{z}_{t-1}) = \sum_{\mathbf{z}_{t-1}} p(\mathbf{y}_t|\mathbf{y}_{t-1:t-p}\mathbf{a}(\mathbf{z}_t), \sigma^2(\mathbf{z}_t)\Big) \tag{5.24}$$

As an illustration, an AR-HMM has been fit with EM (using the approach described in [83]) to the response of the 3-DOF numerical model, with a nonlinear stiffness. The result is shown in figure 5.7, to illustrate the segmentation of a time series with complex dynamics by this model. This type of switching behaviour lets the network model physically meaningful relationships. For example, if one was interested in monitoring the structural dynamics of an aircraft dropping a store, one would allow the dynamics to change from a state of having one (or several) store, to having none, and a very low probability could be placed on the dynamics changing from a state of no store, to containing one. Likewise, the dynamics of an aircraft or vehicle operating with changing mass properties due to fuel usage could be approximated with a large number of components, with transition probabilities that only allow it to go from a state of high mass to a state of lower mass. In this case, the user places a large amount of prior domain knowledge into the network. If the transitions between model components are unknown and example training data is available, then there may be an EM algorithm to learn the model components and their transitions.

The concept of the switching model can be extended to modelling switching between linear dynamical systems, such as the one in Figure 5.6b. In general, when a transition probability is placed on the switching variable, the inference problem can easily become numerically intractable. Exact inference for this problem scales with K^{t-1}, where K is the number of components in $\mathbf{z}$. To illustrate this, recall the integral in equation (5.12), which the Kalman filter solves efficiently, and consider

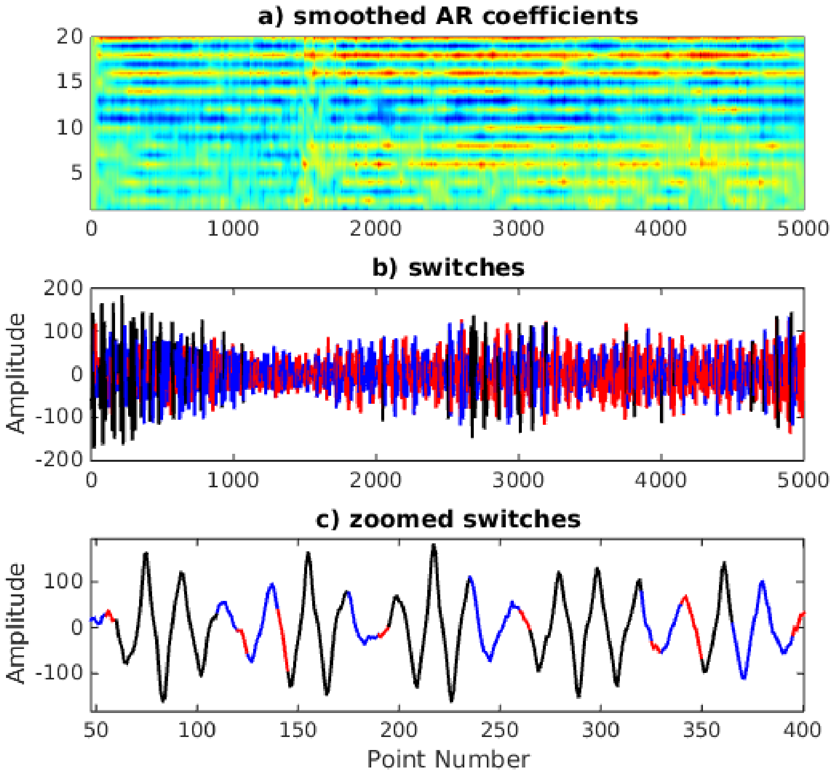

Figure 5.7: Illustration of a switching AR process on response of 1st mass of a 3-DOF simulated system, showing a) AR coefficients estimates with a Kalman smoother, b) time series response grouped by model component and c) zoom-in to grouped time series response

the one required for filtering with the addition of K unobserved components:

$$p(\mathbf{z}_{t+1}, \mathbf{y}_{t+1}) = \sum_{\mathbf{z}_t} \int_{x_t} p(\mathbf{z}_{t+1}, \mathbf{y}_{t+1} | \mathbf{z}_t, \mathbf{x}_t, \mathbf{y}_t) p(\mathbf{z}_t, \mathbf{x}_t | \mathbf{y}_{1:t}) d\mathbf{x}_t \tag{5.25}$$

This is the filtering distribution for estimating $\mathbf{x}$ and $\mathbf{z}$, and the K^{t-1} complexity is introduced due to the summation over the states required at every time step. Several approximate filtering solutions exist, generally based on particle filters, and collapsing the K Gaussians generated at every time step to one. The Gaussian sum filter is one popular example of the former [83], while the Rao-Blackwelised particle filter [116] is a good example of the sampling approach. Murphy provides a relatively recent review of various approximate inference techniques in his PhD thesis [111]. While it would be possible to use approximate inference in a switching model, and to use the likelihood functions in the same fashion as they have been presented here in order to perform damage detection, a full comparison of all the available filtering techniques would be a huge digression. Instead, a discussion is provided in the next section on using models that are computationally tractable, yet still capture the effects of EOVs in the dynamics being monitored.

5.4.1 Inference with factorial switching linear systems

A constraint on the transition probability is not always necessary in the context of accounting for environmental variability, and inference is made computationally much easier if a mixture of models framework is used on top of the linear dynamical system, as opposed to an HMM. There are various ways in which one may adopt a mixture structure over linear dynamical systems; Ghahramani [117] discusses a factorial switching linear dynamical model structure which defines K different observation models and K independent state-space models, and the switching variable dictates which model is observed. The Bayesian network for this model is depicted in Figure 5.8. Note the dashed arrow that describes the Markov chain, which indicates that one may or may not want to model the switching as a Markov chain. If the switching variable is not connected through time, the inference algorithm is computationally very efficient, as it involves running K independent Kalman filters, and switching between them only according to the posterior probability of $\mathbf{z}_t$. The computational complexity is low, because each component has a separate state vector which is only conditioned on the expectations of its past values.

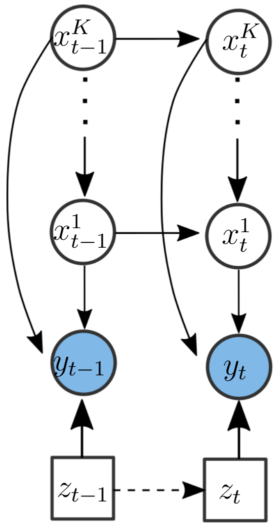

Figure 5.8: Dynamic Bayesian network representation of factorial switching observation model. Note the dashed arrow on the temporal link between $\mathbf{z}_t$ and $\mathbf{z}_{t-1}$ to indicate the possibility of removing this link

EM can be used in order to obtain the model parameters [118, 119, 117], which as in the case of the Gaussian mixture model, provides model segmentation as well as parameter optimisation. Care should be taken though, to constrain this model to physically significant model switches; in particular, given that EM does not guarantee arriving at a global optimum solution for the parameters and segmentation. After all, this type of model now assumes two sets of unobserved variables (the state, $\mathbf{x}_t$ and the switching variable $\mathbf{z}_t$), so the more physically meaningful constraints one can add to the learning scheme, the better. In particular, when considering operational and environmental changes on does not want the model to switch between models every two or three points, if the environmental trend takes hours, days or months to evolve. If there is some knowledge about the dynamics of the system, placing a "no-switching" constraint every N points would be a good solution. In this case, N could be chosen to allow a reasonable number of cycles to occur before switching. While this may not necessarily be the case, if prior knowledge exists over the segmentation of the different operational or environmental conditions for the training data set, then an EM (or other) learning scheme can be applied to each data segment individually, and the factorial model assembled from this.

The question of model order in this case has been left out so far, but it is still an important question nevertheless, and model order will need to be estimated from the data (unless of course, prior knowledge on this exists!). AIC and BIC should offer a reasonable solution, unless one adopts a Bayesian parameter inference approach by specifying a prior over the parameter space; this is a much more complex matter and outside the focus of this work. Even taking the pragmatic solution, and using BIC to estimate model order is not very practical; the order of each linear dynamical system must be estimated, for each K (number of components). Given that EM convergence for a full factorial switching linear model takes a long time using average computing power, this seems in-practical, especially seeing as sampling methods have been avoided here due to their computational inefficiency. For this one could turn to the static data methods of the previous chapter, together with BIC, for the purposes of selecting a reasonable K, and then fitting individual linear dynamical models using EM to now already segmented data. Inference for damage detection, can then be carried out as described above. Note that the model order issue raises interesting questions. For instance, should the models of different components be of the same order? And if not, then what is a good learning strategy for identifying the number of components, with individually optimal "sub-model" orders? This is now outside the scope of this thesis, and falls into the regime of Bayesian model comparison, where some interesting work has been done over the last decade [26].

Once again, the 3-DOF mass-spring-damper system is used to illustrate the usage of the likelihood derived from the factorial switching dynamical model, for SHM purposes. In this case, to add environmental variability, the stiffness of the training set consists of data generated from three overall stiffness settings at 100%, 80% and 60%, representing an undamaged condition with a temperature induced global stiffness change. Damage is introduced as a 20% reduction in stiffness between mass 1 and 2. In this case, finding the number of components and data segmentation was carried out with a GMM with the frequency spectrum as a feature vector; the individual models were then trained with EM. The resulting likelihoods, shown in Figure 5.8, illustrate the ability of this modelling approach for taking into account this kind of environmental variability while still highlighting adverse change in the dynamics.

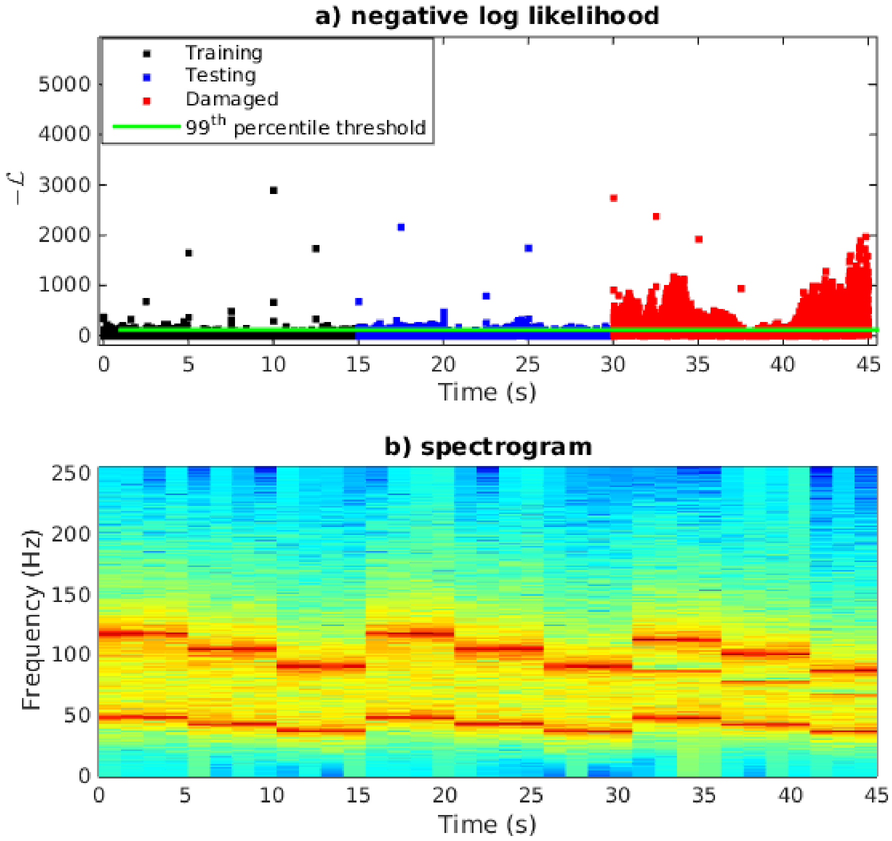

Figure 5.9: Factorial switching observation model applied to a 3-DOF simulated system with stiffness variability a) Negative log likelihood for training, test and damaged set, consisting of 20% stiffness reduction, and b) spectrogram of raw time series of second mass for reference

Chapter 6

Case study: Damage detection in a simulated structure

This chapter demonstrates the application of the PCA, Kalman filtering and mixture models described in Chapters 4 to the problem of detecting damage using vibration measurements on a simulated structure. The motivation for doing this on a computationally simulated structure is the control over the model, the model parameters, and variability in loading conditions and the environment. The motivation for using vibration as a (simulated) measurement is that it is industrially relevant, easy to interpret, and straightforward to simulate. Detecting damage in a structure with linear dynamics could be considered a solved problem. To make this challenging and relevant, stiffness nonlinearities and environmental variability were added to the structure, in order to demonstrate the ability of the generative modelling framework to deal with these changes and perform damage detection using a likelihood function.

Two main types of variabilities relevant to SHM are explored in this chapter. The first is a stiffness variability, designed to demonstrate how one may handle such cases which arise typically as a result of temperature fluctuations. Variabilities introduced due to nonlinearities and the resulting changes in dynamics due to changing loading conditions is also discussed. For each of these cases, the use of inference using static and dynamic Bayesian networks is discussed. The focus throughout the chapter is on the application of Bayesian network inference as a tool for damage detection.

To make this problem representative of an operational system, the input excitation will be treated as an unknown. Furthermore, regardless of whether the system in question is nonlinear or not, changes in operational conditions will still cause a simple linear model (such as linear Kalman filter) to fail to capture this variability effectively, thus motivating the mixture-type extensions available through Bayesian networks. The general philosophy for the treatment of the problem is the same as has been presented in Chapters 4 and 5. The idea is to capture the regimes associated with different operational/environmental conditions using mixture distributions, and to use the overall model likelihood function as a novelty/damage index.

6.1 Simulated Structure

The structure being investigated here is a 3-DOF lumped parameter system with mass, stiffness and damping. The system was simulated using a fourth order Runge-Kutta numerical integration scheme with a sampling rate of 512 Hz. The Figure shown in 6.1 is described by the following equations of motion

$$\begin{aligned} F_1 &= m_1\ddot{y}_1 + c\dot{y}_1 + y_1(k_1+k_2) - k_2y_2 + g_1y_1^3 \\ F_2 &= m_2\ddot{y}_2 + c\dot{y}_2 + y_2(k_2+k_3) - k_2y_1 - k_3y_3 \\ F_3 &= m_3\ddot{y}_3 + c\dot{y}_3 + y_3(k_3+k_4) - k_3y_2 \end{aligned} \tag{6.1}$$

where m_i, c_i, k_i are the i_{th} mass damping and stiffnesses respectively. The system has a cubic stiffness nonlinearity introduced between mass m_1 and ground. This is represented by a parameter g_1.

The parameters used for the simulations were $m_{1:3} = 1kg$, $k_{1:4} = 1 \times 10^4 N/m$, $c = 0.2Ns/m$ and $g_1 = 1 \times 10^9 N/m^3$ when demonstrating nonlinearity effects, otherwise $g_1 = 0N/m^3$. A Gaussian random excitation was applied on the first mass, with different levels of energy; these will be highlighted in the next sections. Also, the data recorded from the simulations was the resulting accelerations at the three masses. This constitutes the measurement of three DOFs for a structure with three modes. All the data was corrupted with white noise with 1% of the standard deviation of the resulting accelerations.

While it could be argued that measuring all the degrees of freedom of this system may not be representative of a real structure (which is not possible in practice),

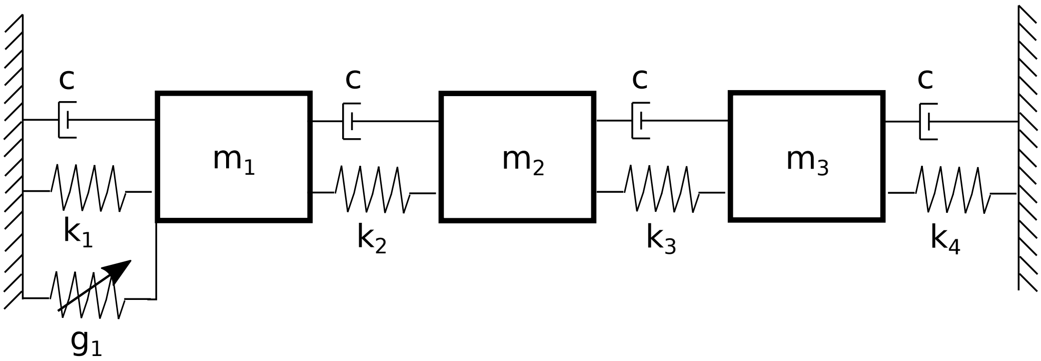

Figure 6.1: 3-Degree of Freedom (DOF) mass-spring-damper system used for numerical simulations. Note the spring connecting ground and first mass represents a cubic nonlinearity. Damage is introduced as a reduction of k_2, whilst changes in stiffness due to the environment are modelled as a global reduction on all stiffnesses.

in practice this translates to measuring sufficient DOFs to capture the relevant modes of the system. In this relatively small system, three degrees of freedom are necessary to capture the first three modes of the system, and hence all of the DOFs are measured. In a larger structure this would imply measuring enough DOFs to capture the modes where damage is most likely to manifest itself.

The stiffness was varied in two different ways: to simulate damage, the stiffness term connecting masses m_1 and m_2 was reduced; while to simulate environmental variability all of the stiffness terms were reduced. Throughout this chapter, $k_{1:4}$ is referred to as the global stiffness.

When dealing with linear models, the data was decimated to a sample rate of 128Hz, as this is more appropriate given the system natural frequencies, which lie between 10Hz to 30Hz, while when dealing with the nonlinear system a sample rate of 512Hz was used.

6.1.1 Operational and Environmental Variations

The excitation of real engineering structures can be quite complex and varied, therefore the simulations were carefully chosen to give a representative “environment” for the purpose of assessing and demonstrating the application of an SHM system.

A bridge, for example may be excited by traffic loading as well as wind and possibly water waves, all of which have their own difficulties in characterising, but will tend to be stochastic in nature and may be reasonably approximated by a Gaussian distribution to satisfy the objective of assessing an SHM system.

On the other hand, aerospace structures may have different types of excitation, perhaps slightly more predictable in nature. This may include flow induced vibration. A rotary wing aircraft will be primarily excited by sinusoids at the fundamental frequency and the harmonics of the main and tail rotors, together with the meshing frequencies of the gearboxes. This has the effect of exacerbating the effect of structural nonlinearities, which tend to be excited when more energy is placed in a narrow frequency range. EOVs in the case of aerospace structures can be introduced by effects such as icing and temperature changes.

The objective of this work is not to investigate in detail the effects of structural nonlinearities or environmental variations. The idea here is to design a representative experiment whereby the application of the graphical model interpretation of state space systems can be demonstrated in an SHM context.

To this end, three cases of variation are examined. Section 6.2 presents the damage detection problem where no variations are present. Section 6.3 discusses damage detection with a changing global stiffness, as a demonstration for changing temperature. Finally, Section 6.4 discusses the case when the system has a nonlinearity, and the undamaged condition contains excitations at different energies.

6.2 Damage detection with no environmental variation

As a baseline, the damage detection problem on a linear system without environmental variation will be discussed first. This problem is much easier than with the presence of variability, so it will help solidify the Bayesian network novelty detection approach, and discuss some of the practical points of computing and interpreting likelihood functions, in a simpler model. Both static and dynamic Bayesian networks will be used, in order to compare and contrast the approaches. In the case of a static Bayesian network, a feature vector first needs to be derived before performing inference. A dynamic Bayesian network, on the other hand, performs inference directly on the measured data $\mathbf{y}_t$.

The system response is illustrated in Figure 6.2 in the time domain. In these trials,

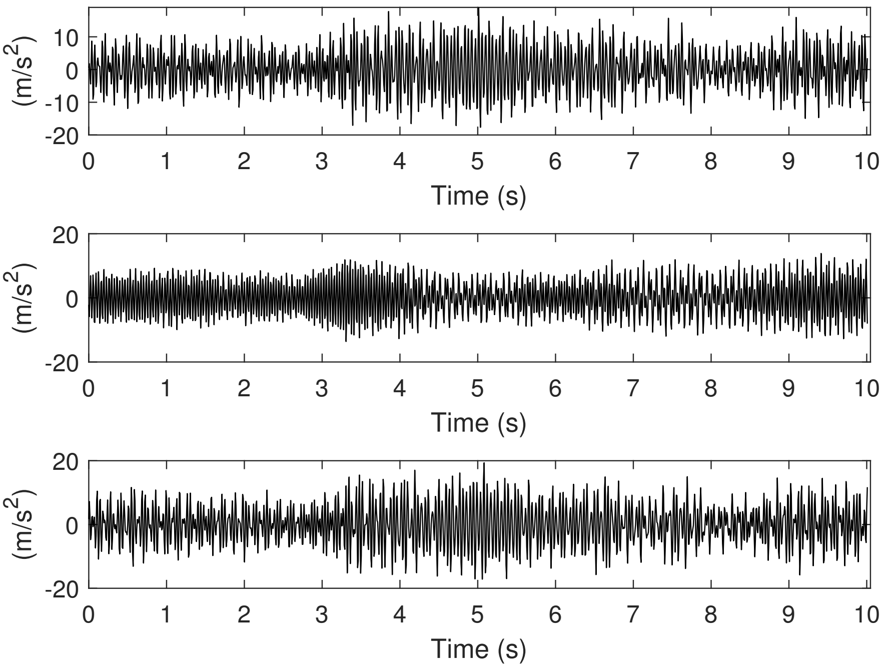

Figure 6.2: Illustration of time domain response to Gaussian random excitation of the three masses for the simulated structure shown in Figure 6.1.

the system was excited with white Gaussian noise and as before, damage is introduced by reducing the stiffness between the first and second mass. Note that the Gaussian excitation was not the same for all instances of the simulations (a similar random seed was not used).

6.2.1 Static Bayesian network inference

There is a wide variety of features to choose from for this problem. The feature vector needs to capture the periodicity in the dataset; one knows that the natural frequencies will change with damage, so a transformation of the non-stationary time domain data into a time-frequency domain makes sense, and this is easily achieved with a Short Time Fourier Transform (STFT). The natural frequencies of the system will be well represented by the Fourier coefficients although this is a rather sparse representation of the information required to detect damage; in order to get the required frequency resolution one must use a large enough STFT window, but damage will induce change in only a handful of coefficients amongst hundreds or thousands. The choice of using a STFT to build feature vectors is not necessarily optimal in this case; it has some clear downsides:

- A choice of window size needs to be taken; a large window implies poor time resolution but good frequency resolution, and vice-versa.
- The Fourier coefficients associated with the system natural frequencies will generally be sparse, as a large window will imply good frequency resolution.
- Fourier coefficients are not invariant to changes in excitation amplitude, and frequency content.

Nonetheless, the sparsity of the Fourier coefficients can be dealt with well using PCA, and any potential changes in loading could be dealt with using a mixture model, though this point is demonstrated later. An STFT represents a relatively simple feature, from which engineering information can be readily extracted; a spectrogram or autospectrum (a time average of a spectrogram) can be easily read and interpreted. Much more processing would be required to perform, for example, an automated operational modal analysis process (such as the one used in the Z-24 bridge dataset). The time-frequency resolution trade-off could be eliminated through the use of a wavelet transform, but this adds more processing steps and modelling choices, and may not be as readily interpretable as a frequency spectrum. The Bayesian network applications, however, are applicable to any feature vector used. The objective is to illustrate the use a static Bayesian network for inference.

A PCA model was formed from the training set using Fourier coefficients for time windows, extracted using an STFT. The sample rate used in the simulated system was 128 Hz, and a window size of 256 points was used.

Because the feature vector is relatively high-dimensional and sparse, PCA is more efficient in terms of learning a density model, as it effectively represents the data in lower dimensions, so fewer parameters are required in the learning step. This is even more so in the case of mixture models, used later when considering the effect of environmental variability. One could always fit a Gaussian distribution to the first few principal components and use these as features for novelty detection. This could even be extended to fitting a Gaussian mixture to principal components, when EOVs are present. This approach was illustrated in Section 4.3.3. To do this, an estimate of the latent variables, $\mathbf{x}$ is first required, using

$$\mathbf{x} = (\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}\mathbf{y} \tag{6.2}$$

from which a density estimate can be formed over $\mathbf{x}$ with a (multivariate) Gaussian. This approach may be reasonable, but it does not define a proper probability model. The idea here is to use inference in the PCA model for novelty detection. Because PCA defines a covariance matrix over the observed data, $\mathbf{y}_t$, novelty detection can be done by evaluating the likelihood function directly on the observations, rather than having to actually compute the latent variables/principal components. Recall from Chapter 4 that PCA defines a covariance matrix over the observations of the form:

$$\mathbf{R} = \sigma^2\mathbf{I} + \mathbf{C}\mathbf{C}' \tag{6.3}$$

which can be used to evaluate a Gaussian likelihood. To illustrate the damage process with this likelihood function, a PCA model was fitted to the STFT of the acceleration response of the second mass of the 3-DOF system, on an undamaged condition, and with no environmental variations. A testing set was also created, by producing two more 10-second simulated trials in an undamaged case. The damage set, consists of trials with 10% and 20% reductions in the k_2 stiffness. The negative log-likelihood of the PCA model, evaluated for the training, testing and damaged cases is illustrated in Figure 6.3. In this case, five principal components were used in this model. It is clear from this that the model is capable of detecting damage, while minimising false positives within the testing set.

6.2.2 Dynamic Bayesian network inference

If a structure is linear and its excitation can be approximated as white Gaussian noise, then a Kalman filter is arguably the best tool there currently is for making predictions of the system response. It has already been discussed in Section 5.3, that SHM could be performed in two ways with a dynamic Bayesian network; by either making predictions of the measurements, and therefore first running a system identification step, or by using the dynamic bayesian network to identify time-varying system parameters, and to then use these to decide whether damage is present. In this example, system identification is performed, via the EM algorithm in order to

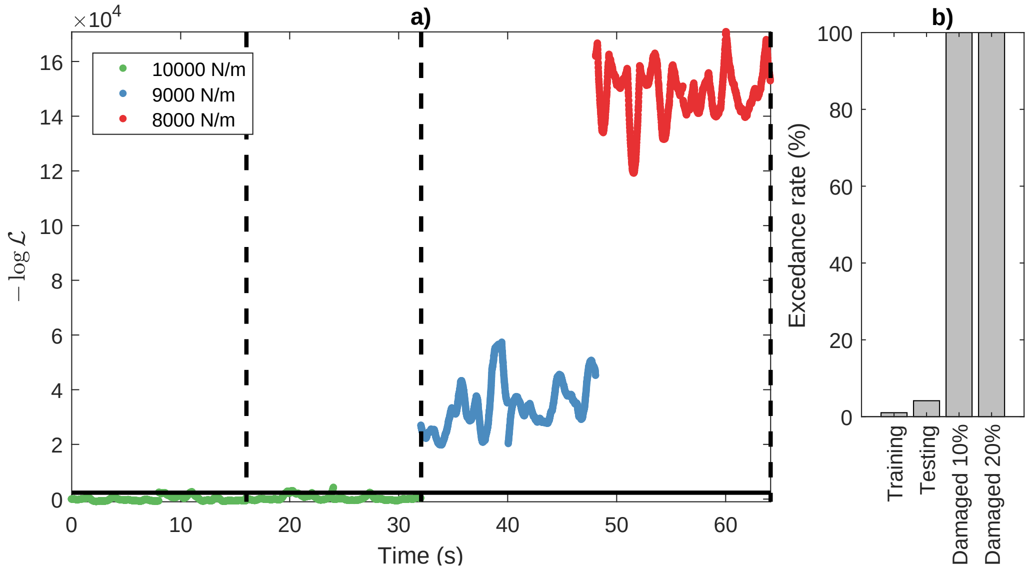

Figure 6.3: a) Negative log-likelihood of PCA model evaluated on training, testing and damaged data for a system with no EOVs. Data is grouped by stiffness of k_2. Horizontal line shows 99th percentile threshold over training data. b) shows the exceedance rates of the threshold.

extract a state space model directly from observed data and to use Kalman filter inference to perform predictions, and to compare measured data against those predictions. The interest is on the use and interpretation of the likelihood function as a novelty index. Recall from Section 5.3.1 that the complete data likelihood is the sum over the density given by the filtering recursion $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$. This complete-data likelihood is given in equation (5.19) for the linear Kalman filter, while the individual point-by-point likelihood estimates are given by equation (5.21), which (in its log form) is effectively a Mahalanobis distance measure on the residuals, where the Kalman filter also recursively updates the estimate for the covariance over those residuals.

The time domain response of the 3-DOF system, consisting of three accelerations is shown in Figure 6.2.

In this case there are three dimensions for the observation vector $\mathbf{y}_t$ and the choice of dimensions for the latent variable $\mathbf{x}_t$ specifies the model order. In certain circumstances one may want to choose a model order that is of a lower dimension than the observation vector, namely if the dynamics contain fewer degrees of freedom than the observations. In these simulations, however, the dimension of a vector that fully defines the dynamics is twice the size of the observations; the equations of motion

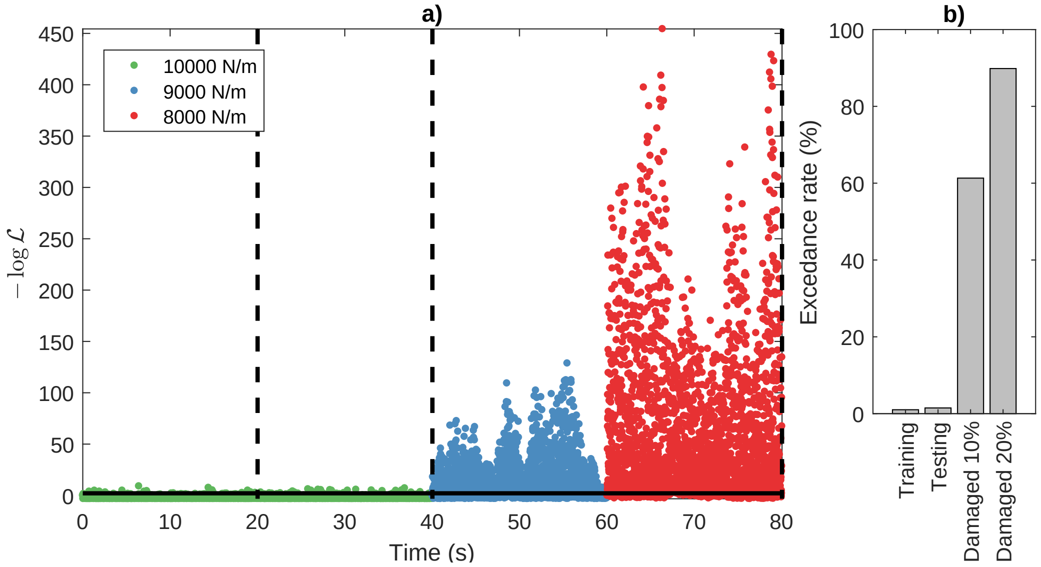

Figure 6.4: Negative log likelihood point estimates derived using a Kalman filter for the acceleration response of the 3-DOF simulated system for training, testing and damaged sets, consisting of 10% and 20% stiffness reductions. Data points are grouped by k_2 stiffness (damage). Horizontal line shows 99th percentile threshold over training data. b) shows the exceedance rates of the threshold.

are second order differential equations. If one derivative is being measured (in this case acceleration/2nd derivative) then the other two need to be *estimated* to be able to project the physics forward in time. Another way to look at this problem is by noting that the state space has an Auto-Regressive Moving Average (ARMA) representation, and a model order can be selected by using AIC, BIC, or alternatively by checking the spectral content of the ARMA process, and seeing that it contains the frequencies of interest in the process. In these simulated experiments, a model order of six has been selected based on prior knowledge of the system, though it could be shown that the same result could be arrived at using BIC. Two 10-second simulated trials were used each for model training, testing and for damage detection cases. The resulting negative log likelihoods are shown in Figure 6.4. Two damage conditions are shown, for 10% and 20% stiffness reduction.

It is no surprise that the likelihoods derived from this dynamic Bayesian network are clearly able to generalise well and identify damage . The key aspect to notice is that the filtering (negative log) conditional densities $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ define precisely that: a density. On average they are able to capture the change in the system well, but expecting all points in that probability estimate to individually novelty would be a misinterpretation of probability. More "consistent" results can be derived if

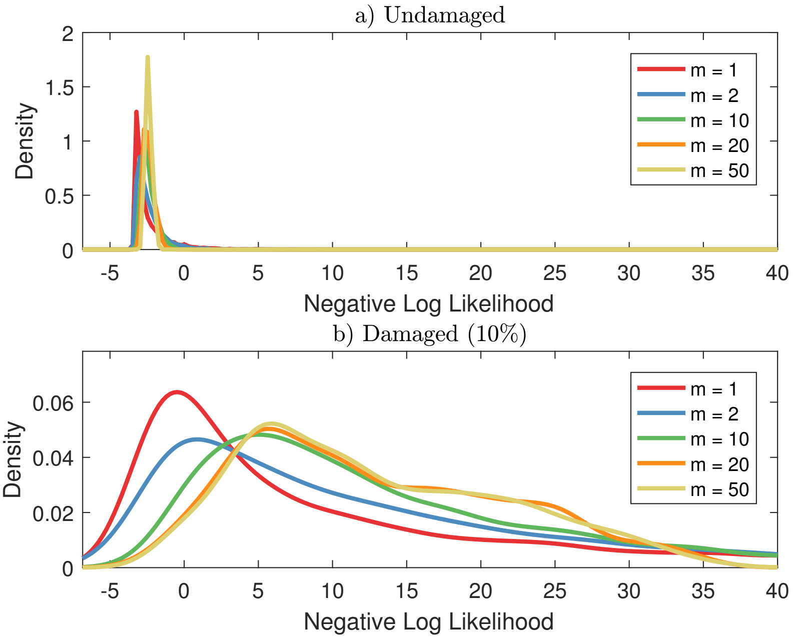

Figure 6.5: Kernel density estimate for Kalman filter negative log-likelihoods of acceleration response of a) undamaged and b) 10% damaged systems. The densities are shown for moving means of the likelihoods with increasing window sizes m.

one looks at the statistics of $-\log p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$.

Figure 6.5 shows the density estimates of the point-by-point negative log likelihoods (from Figure 6.4) where a moving mean is taken over the likelihoods, with an increasing window size. Note from figure 6.5a, that as discussed in section 3.2, taking a moving expectation with increasing window size m, the baseline undamaged density looks more like a Gaussian distribution. On the other hand, the presence of damage (Figure 6.5b) introduces long tails to the distribution of negative log likelihoods when $m = 1$. When m is increased, the density converges to a multi-modal distribution with much higher $-\log \mathcal{L}$. Note the multi-modality comes from the fact that this density was taken over two damage states. This all highlights the probabilistic interpretation of the likelihood; it is being derived from a residual which will grow on average if the system generating the data changes. The interpretation of doing a moving average is that of using the expectation $E[-\log p(\mathbf{y}_t|\mathbf{y}_{1:t-1})]$ as a novelty index, and noting that there is a limiting distribution of that expectation, as the number of observations (of each window), m is increased. As a conclusion here, it could be stated that the if the dynamic Bayesian network correctly captures the physics of the process, the number of false positives could be reduced to zero while bringing the true positive rate to 100% by using the expectation of the negative log

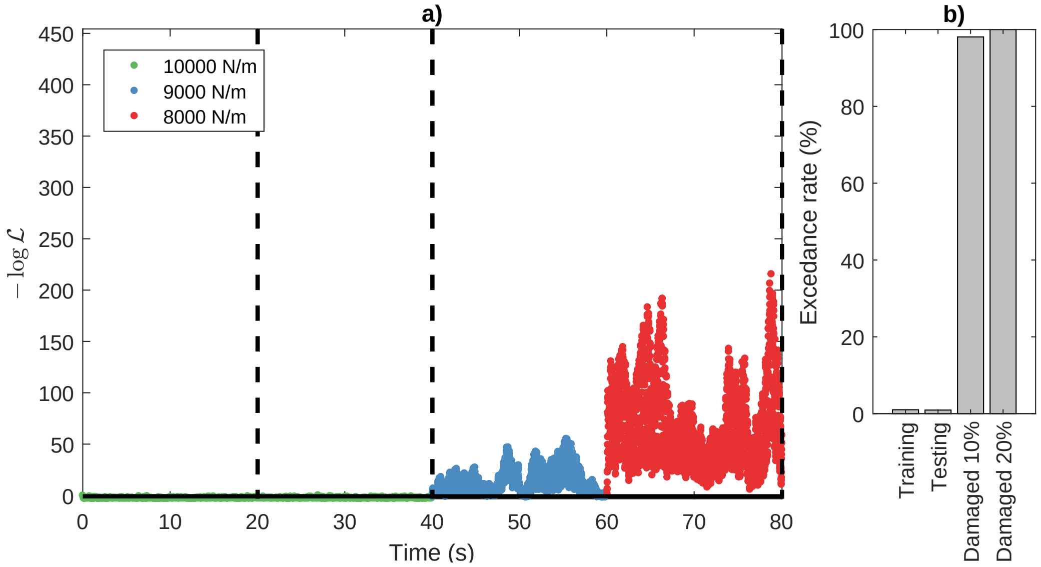

Figure 6.6: Expectation of negative log-likelihood derived using a Kalman filter, with a 50-point moving average window, for the acceleration response of the 3-DOF simulated system for training, testing and damaged sets, consisting of 10% and 20% stiffness reductions. Data points are grouped by k_2 stiffness. Horizontal line shows 99^{th} percentile threshold over training data. b) shows the exceedance rates of the threshold.

likelihood as a novelty index. This is illustrated in Figure 6.6.

6.3 Damage detection under varying global stiffness

This section discusses and shows the use of static and dynamic Bayesian network inference, for the case where there are stiffness variabilities present in the structure. Note that this is till a linear system ($g_1 = 0$). The treatments is similar to that of the case with no environmental variability, but in this case, both the mixture extensions of PCA and Kalman filters are used to perform learning and inference. As before, inference is referred to as the estimation of the data likelihood against the parameters learned using healthy condition data.

A global reduction in stiffness is used to simulate stiffness variability. More specifically, the stiffness reduction is applied by a reduction on all $k_1, .., k_4$. Damage is still considered as a reduction on k_2 only. In the healthy condition, the structure is assumed to operate at 100% and 90% *global* stiffness. Damage is shown here as 10% and 20% reductions in k_2. The reader may note that EOVs may sometimes manifest themselves as continuous changes to a system, but often as discrete changes. This discrete change in stiffness clearly represents a discrete change to the dynamics of the system (for example, due to freezing temperatures, or sudden changes of mass). However, continuously changing EOVs could still be modelled with this approach given enough number of mixture components.

6.3.1 Static Bayesian network inference

The STFT is used again to extract the time-varying Fourier coefficients of the time series, but in this case, a mixture of PCA models is fitted to the Fourier coefficient data. Two ten second cases for both global stiffness conditions are considered for learning purposes, while three ten second cases are use for validation. In this case, a two component mixture, with two principal components each, was learned using EM.

Figure 6.7 shows the averaged Fourier coefficients computed using an STFT, where these are grouped by the global stiffness on the top subplot (Figure 6.7a) for the undamaged condition, and by k_2 for all conditions on the bottom subplot (Figure 6.7b). The main variations due to damage are introduced in the second and third natural frequencies; there is a clear shift to the left when the global stiffness is reduced, and the second natural frequency is more prominent when only k_2 is reduced.

The negative log-likelihoods for the PCA mixture model are shown in Figure 6.8 evaluated on the training, testing and damage datasets. A 99^{th} percentile threshold is shown on the horizontal axis. It is clear that the PCA mixture model easily captures this type of stiffness variability and highlights damage, under this variability. Note that, because damage is introduced as a 10% and 20% stiffness reduction on all the different global stiffness values (two in this case), this yields a total of four different levels for the damaged cases. The negative log likelihood values for the 20% damage points are greater than the 10% damaged points, which makes sense given that the feature vector is further away from any of the regular conditions. This is the same as with the case with no variation. Note also that in the damaged condition, the negative log likelihood has some excursions below the threshold, especially at the low damage levels. Note that k_2 takes a value of 9000 both under the

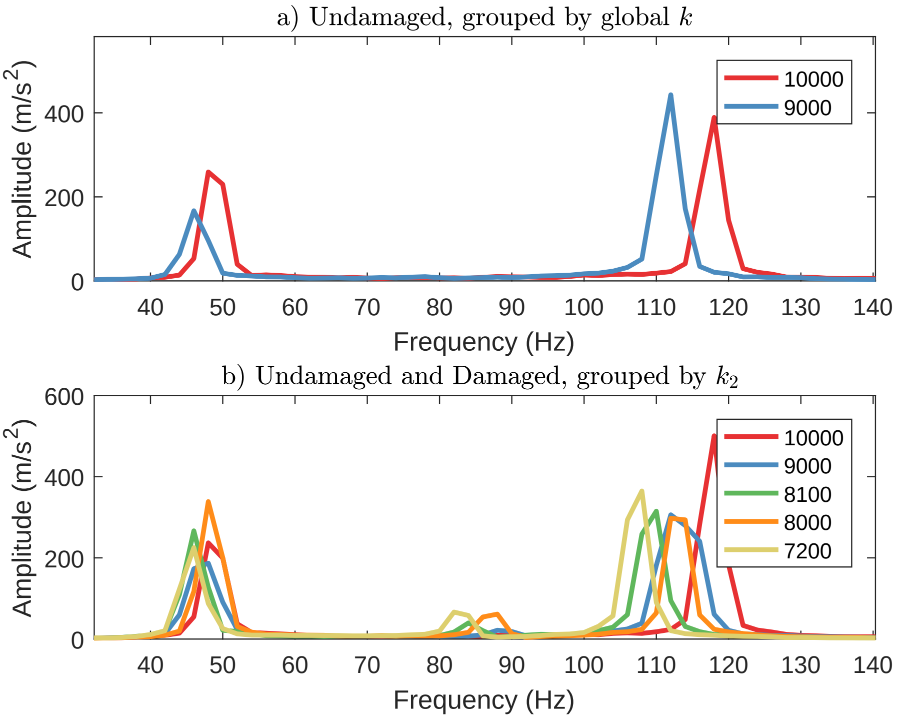

Figure 6.7: Averaged autospectrum for the 3-DOF response on the second mass, showing the variability in frequency content caused by the two different stiffnesses

undamaged condition with 10% global stiffness reduction, and on the 10% damaged condition with 100% global stiffness. This contributes to the false positives close to 30-32 seconds in Figure 6.8. It also makes this a good result, implying that this modelling framework is capable of distinguishing these two cases clearly.

6.3.2 Dynamic Bayesian network inference

The dynamic Bayesian network used in Section 6.2.2 is extended here using the mixture framework described in Chapter 5, which is effectively a switching Kalman filter. It was discussed how different dynamic Bayeian network structures can lead to different representations for a switching model, and the implications of some representations for inference. Networks of the type shown in Figure 5.6b, which represent a latent linear dynamical system, where the switching variable has transition probabilities over time, are avoided here. As discussed in Section 5.4, inference for this class of models scales exponentially with the number of time points, making it impractical for SHM usage, where even low frequency vibration data may require

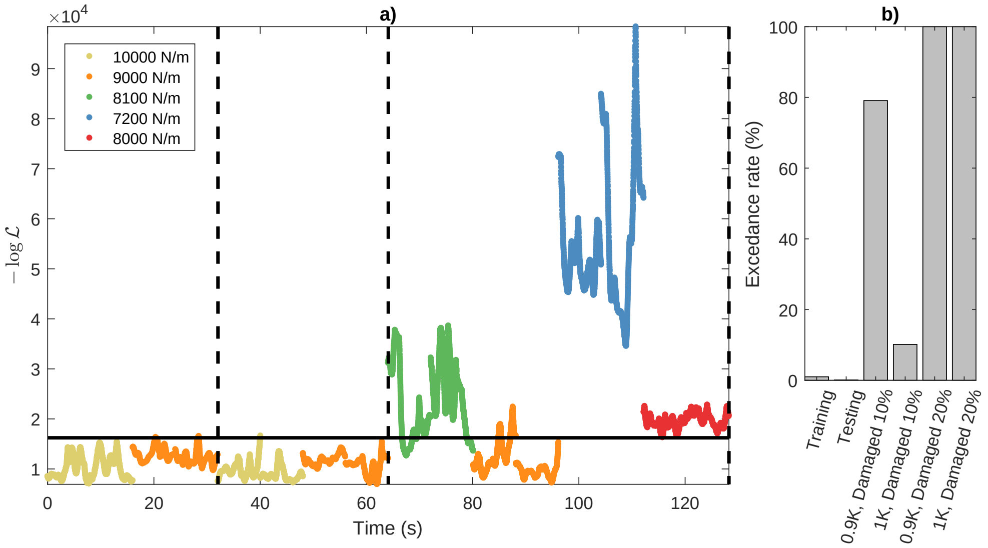

Figure 6.8: Negative log-likelihood for a PCA mixture model, fit to the 3-DOF STFT data for training, testing and damaged sets. Data points are grouped by stiffness. The training and testing sets consist of two different global stiffnesses, while the damaged sets consist of 10% and 20% reduction in k_2. Horizontal line denotes the 99^{th} percentile threshold. a) shows the exceedance rate of the threshold for training, testing and damage sets.

computation over thousands of points. The factorial Switching Linear Dynamical System (SLDS), discussed in Section 5.4.1 is used here instead, and without transition probabilities between the switching variable, $\mathbf{z}_t$. The factorial SLDS thus reduces the inference problem to K individual Kalman filters, which compete using their likelihood functions, and Bayes' rule, to determine which model better explains the observations. The damage detection hypothesis in this case is that if none of the models explains the observation, then the underlying system, together with the known variabilities, has changed.

An example has already been provided in Section 5.4.1 on how one may use SLDS inference with vibration data for novelty detection under changing stiffness, but this section expands the previous example with more details.

The stiffness variabilities introduced are the same as for the results presented above using a mixture of PCA models with an STFT feature vector. The key difference is that learning and inference is performed directly on the measured accelerations, and less consideration is taken as to which measurement channels are taken into account, as recalling from Chapter 5, the Kalman filter projects the observation vectors $\mathbf{y}_t$ into a state space vector, $\mathbf{x}_t$, that can describe the dynamics.

From Chapter 5.4.1, recall that the factorial SLDS is described by K parameter vectors, each for the k^{th} model component: $\boldsymbol{\theta}_k = \{\boldsymbol{\pi}_k, \mathbf{A}_k, \mathbf{C}_k, \mathbf{Q}_k, \mathbf{R}_{,k}, \mathbf{x}_0, \mathbf{V}_0\}$. The initial state and covariance ($\mathbf{x}_0$ and $\mathbf{V}_0$) are part of the model parameters, although the same ones are assigned to all models.

Recall from the discussion in the last chapter that instead of the actual data log likelihood, the conditional density $p(\mathbf{y}_t|\mathbf{y}_{1:t-1})$ is evaluated as a novelty measure, as this effectively represents the point-by-point data likelihood.

The learning step in this case can involve the use of EM to both segment the data and learn the model parameters at the same time. However, this does not tend to segment the data very well, as EM is prone to local minima and therefore a good initialisation is crucial. If one is prepared to put the effort into pre-segmenting the data to initialise the model, it is then a better idea to perform EM for each individual model, on the respective segmented data. In the interest of pragmatism, the segmentation for the factorial SLDS models presented here was carried out with PCA mixtures on Fourier coefficients and the number of segments was selected based on the BIC for this model. This is much cheaper to compute than the factorial SLDS full EM, and converges to a similar result (in this case) in a few seconds compared to hours of the factorial SLDS EM optimisation.

The negative log-likelihoods for the training, testing and damage sets are shown in Figures 6.9 and 6.10. In the first figure, a ten-point moving average of the negative log likelihood is shown, whereas Figure 6.10 shows the actual point by point likelihood. In both plots, the thresholds shown correspond to the 99^{th} percentile of the training Negative Log-Likelihood (NLL).

The effect of the "smooth" NLL separating the effects of damage, while keeping the testing NLL under the threshold is expected; this is a probabilistic model, so the Kalman filter residuals will grow *on average*, if the data does not correspond to the model. Hence, setting thresholds on averages or even maxima of the likelihood is bound to give a better indication of damage, with less false negatives, than setting thresholds based on point-by-point likelihoods.

The factorial SLDS does do a slightly better job than the PCA mixture. This can be seen from the fact that the negative log likelihoods of the undamaged testing set of the PCA mixture cross the 99^{th} percentile threshold more markedly than those for the factorial SLDS. Furthermore, the factorial SLDS operates on all measurements

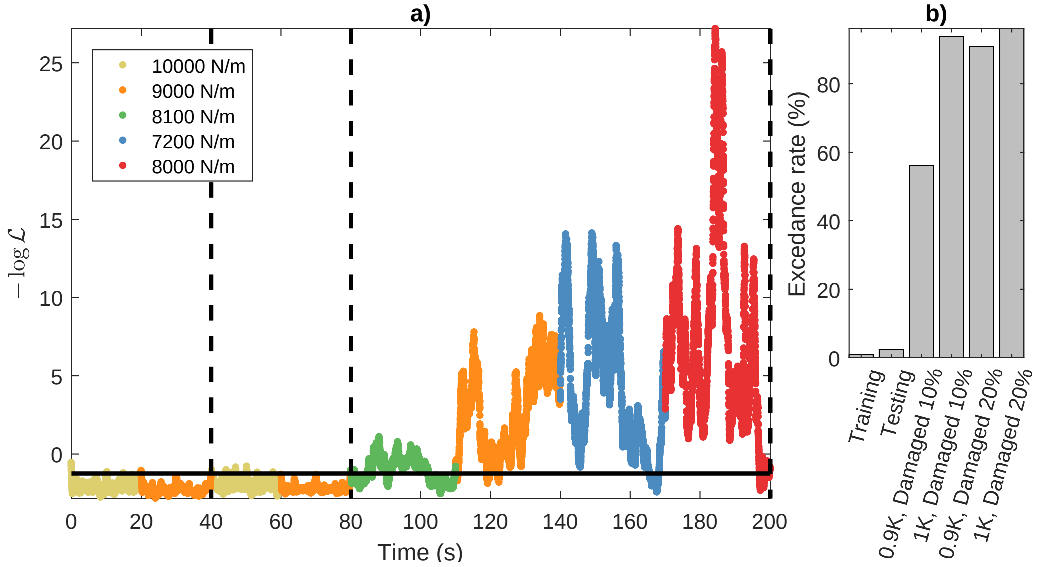

Figure 6.9: a) Ten-point average of negative log-likelihood for factorial SLDS model, fit to the 3-DOF system acceleration data for training, testing and damage sets. Data points are grouped by k_2 stiffnesses. The training and testing sets consist of two different global stiffness, while the damage sets consist of 10% and 20% reduction in k_2. Horizontal threshold shows 99^{th} percentile of training set. b) shows the exceedance rates of the threshold for training, testing, and damage sets.

channels, as well as directly on raw data, so it does not rely on the choice of a feature vector. Note however that the detection for the 80% damaged case is not as prominent, given that the dynamics that generate this data are "closer" to the 10% global stiffness reduction, which forms part of the undamaged model.

6.4 Damage detection with nonlinearities under changing loading

Modelling the behaviour of a nonlinear system is not the central idea of this thesis. What the author wishes to achieve here, is to demonstrate the applicability as well as limitations, of a Bayesian network framework for SHM-oriented inference, where the effects of a nonlinearity play a role in the changing dynamics of the system being monitored. Modal analysis plays a mayor role in the modelling and understanding of structural dynamics. It builds on the principle of linear superposition, which allows one to describe multi degree-of-freedom dynamics as a superposition of single degree of freedom oscillators. This greatly simplifies the treatment as the system

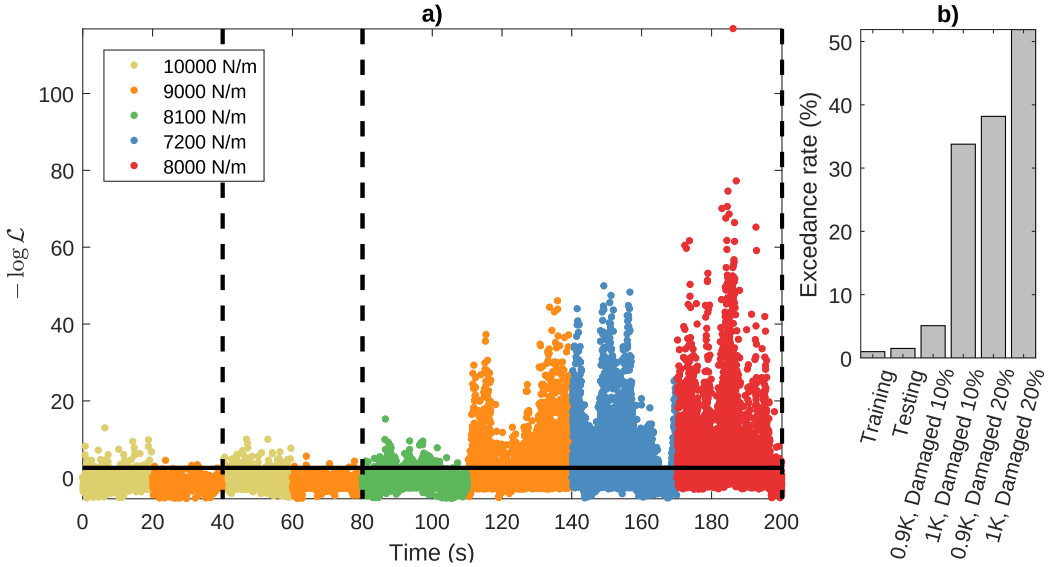

Figure 6.10: Negative log likelihood for factorial SLDS model, fit to the 3-DOF acceleration data, showing testing and damaged sets. Data points are grouped by k_2 stiffness. The training and testing sets consist of two different global stiffness, while the damaged sets consist of 10% and 20% reduction in k_2. Horizontal threshold shows 99th percentile of training set. b) shows the exceedance rate of the threshold, for training, testing and damage sets.

can be fully described by a set of natural frequencies, damping factors and mode shapes. One of the stumbling blocks when modelling nonlinear systems is that the principle of superposition does not hold any more, and decomposing the system into a superposition of SDOF oscillators is not straightforward if even possible, though many attempts have been made. One notable example (and relevant to the system being investigated here), is Shaw and Pierre's approximate decomposition of a 2-DOF oscillator into a series of SDOF oscillators [120], using Taylor series approximation. This approach is semi-successful, there are several approximation steps and extending this to more DOFs involves an unreasonable amount of algebra, that does not scale well to real-world engineering structures.

Nonlinear systems are also subject to bifurcations, involving step changes in the frequency response of the system across different energy levels. They can be chaotic, making the prediction of their response incredibly sensitive to initial conditions. A linear system will respond at the frequency of excitation, whereas a nonlinear system may also respond at harmonics of the natural frequencies if excited at high enough energy. So, avoiding a great and potentially very interesting digression, treatment here is limited to questioning the extent to which one could one make use of Bayesian network inference, as presented thus far, to monitoring a nonlinear system. No attempt is made here to fully explore the dynamics of the system or to even explore all the possible cases in which inference and learning for the Bayesian networks in question may or may not work for the SHM problem. This would be a monumentuous task.

To this end, the use of both static and dynamic Bayesian networks is explored for damage detection on the 3-DOF simulated system used so far, except that now the cubic stiffness g_1, between masses 1 and 2 is nonzero (this has been zero on all the previous examples). In this simulation, the cubic stiffness was given a value of $g_1 = 1 \times 10^9 N/m^3$. This is set two orders of magnitude higher than the linear spring stiffnesses, so that the nonlinear effects are clear while being excited by Gaussian white noise. The effects of operational variation are thus investigated by exciting the system at various energy levels, and performing novelty detection on this. In this simulated experiment the global stiffness remains the same and it is the introduction of nonlinearity and excitations at different energy levels that cause variability.

An example of the response of this system to Gaussian noise is shown in Figure 6.11 in the form of the autospectrum response of the second mass, grouped by excitation level. The system was excited with white noise at $\sigma = \{1, 4, 16\}N$ in order to highlight the sudden change in dynamics. Note that the difference between $\sigma = 1N$ and $\sigma = 4N$ involves a slight shift increase in natural frequencies, as the energy increases. The change seen at $\sigma = 16N$ is however completely different, characterised by shifting as well as harmonics of the natural frequencies. The 3-DOF "simple" example is not so simple any more. It is clear that the response at higher energy levels is more complex than at lower ones. As before, inference using the likelihood functions from static and dynamic Bayesian networks, as performed above will be carried out.

6.4.1 Static Bayesian network inference

Once again, the STFT is used to derive a feature vector to feed into a latent variable mixture model. In this case, however, because the excitation changes, each instance of the feature vector $\mathbf{y}_t$ was normalised by its maximum amplitude. Large scale dissimilarities can cause numerical overflow and underflow issues when computing log-likelihood functions. These scale differences cause particular difficulty

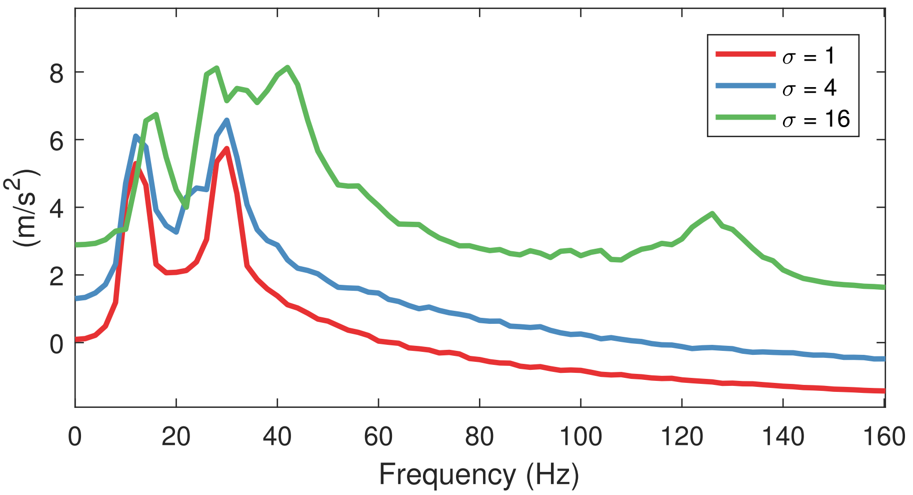

Figure 6.11: Autospectrum for acceleration response on the second mass of the nonlinear 3-DOF system, grouped by excitation level (standard deviation). Note that the response of the system shifts significantly to the right, and a harmonic appears at 126 Hz. These responses represent the undamaged condition of the system.

when initialising models, as they can lead to covariance matrices with very small determinants. Furthermore, when segmenting data, the resulting component likelihoods could be biased if the segmentation is not perfect, which it will not be given that this is a probabilistic model. Normalising against a feature variance or maxima allows one to compare the relative shape of the features, although it does throw away some information about relative amplitudes.

Recall that the covariance of the PCA model assumes that the observation noise is isotropic. The noise variance captures whatever the model cannot explain, so the higher the variance, the more the model cannot explain. The isotropic noise assumption of PCA may not be a good thing if some dimensions can be captured better than others. In general, it can be observed that the response of a nonlinear system may vary more at some frequencies than at others. This could be because, depending on the type of nonlinearity, the response of the system could be given by an underlying linear model plus the response of the nonlinear component. Naturally, the dimensions associated with the nonlinear response will tend to be more“volatile”.

This author suggests that Factor Analysis (FA) is better suited to this problem; it is a similar model to PCA, except that it relaxes the isotropic noise assumption, and allows each variable to adopt its own variance. It is still a latent variable model that

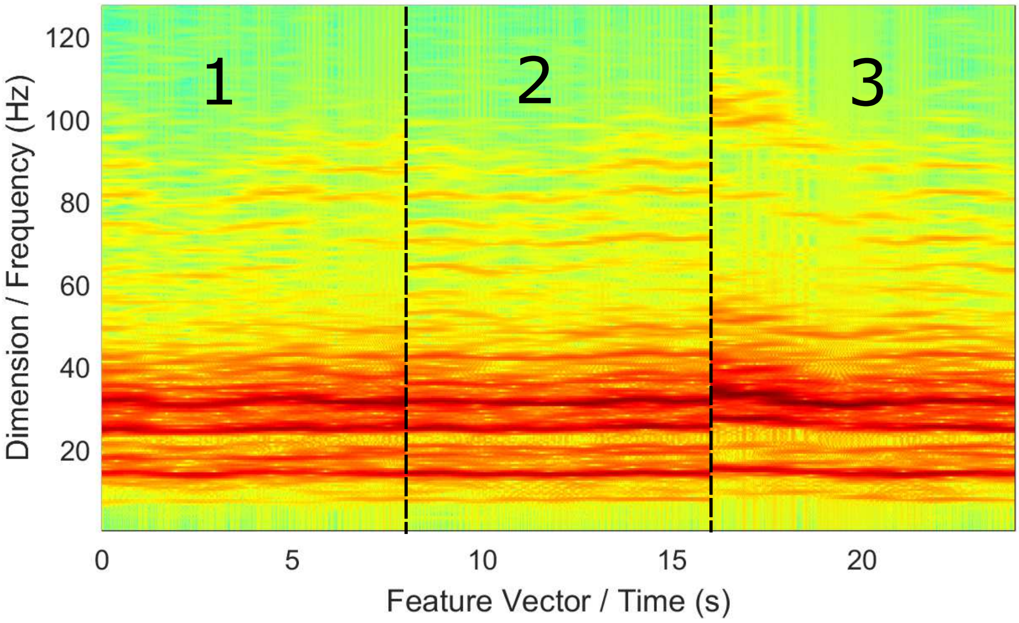

Figure 6.12: Spectrogram of the response of the nonlinear 3-DOF system at the second mass. The responses of three different trials are shown, which constitute excerpts from the training data for the static data models.

embeds the observed data $\mathbf{y}_t$ into a lower dimensional vector $\mathbf{x}_t$ through a linear map $\mathbf{C}$. The link between PCA and FA is well studied [117], and both models have been cast as effectively the same, except for the way they treat the observation noise. Intuitively, their mixture model extension will be similar. In fact the EM algorithm for FA mixtures was written before that for PCA mixtures [121]. Formally, the covariance structure of $p(\mathbf{y})$ for the factor analysis model is given as:

$$\mathbf{R} = \mathbf{C}\mathbf{C}' + \text{diag}(\boldsymbol{\sigma}) \tag{6.4}$$

where $\boldsymbol{\sigma} = \{\sigma_1, ..., \sigma_d\}$ is a vector of the variances for each dimension, which goes along the diagonal of the observation noise covariance. To contrast this with PCA, consider first the response of the nonlinear 3-DOF system at the highest energy (considered here). Figure 6.12 shows the spectrogram of the response of the second mass for an undamaged system, with all three excitation levels. Three different trials are shown in this figure, each of ten seconds duration. The key aspect to notice is that the response contains harmonic components beyond the natural frequencies, which vary in amplitude and frequency due to a combination of the random excitation and the nonlinear nature of the system. It can be observed though, that the lower frequency components do remain somehow more constant. As an illustration, mixtures of PCA and FA were fitted to the dataset from Figure 6.12 using two components. The resulting means and variances are shown in Figure 6.13. Note that each component identifies a slightly different mean, corresponding to the cluster centre, but most importantly, FA assigns a much lower variance to certain dimensions, such as the frequencies close to 15, 25 and 31 Hz. Because the variance is tight around these frequencies, the model is more sensitive to changes in these dimensions. Furthermore, the fact that this is a mixture model allows such tight variance. If a single (FA) component were to be fitted, the variance in these dimensions would be much higher. To show how this affects damage detection, a three-component mixture model was fitted using both PCA and FA, to the spectral response of the system, still using the highest energy level. The resulting negative log-likelihood for the training testing and damage sets is shown in Figure 6.14. Here, only a 20% stiffness reduction was used to represent damage. Note that to achieve this separation, three cluster components were required, and the model order for the individual local FA/PCA models was 25. A smaller model order would completely fail to generalise and/or discriminate damage. This highlights that the response of nonlinear systems is complex; three components and a relatively high model order were required to capture this problem (from an SHM inference perspective). This is in contrast with the linear case, where one component (per operational case) and an isotropic covariance sufficed.

PCA mixtures are, in general, more sensitive to change than their FA counterpart. In the case of the PCA mixture a better result could be achieved by fitting more mixture components. Because the noise model is isotropic, this would result in the model segmenting the data by the variance of different dimensions. On the other hand, because FA assigns a variance to each dimension, it is more prone to segment data according to the shape of the feature vector. In this example, PCA mixtures discriminate damage well, but don't generalise well on a normal condition: the testing set contains more false positives. FA mixtures are slightly worse at discriminating damage, but contain much less false positives.

Finally, a loading variability of the nonlinear system is considered by taking the first two energy levels of excitation ($\sigma = 1, 4N$) and fitting a five-component FA mixture model to the Fourier coefficient vector (of the second mass, again). In this case, the change due to damage is more subtle, and a relatively large number of components were required to fit only two operational changes; this highlights once again that the problem is made much more complex with the introduction of a nonlinearity.

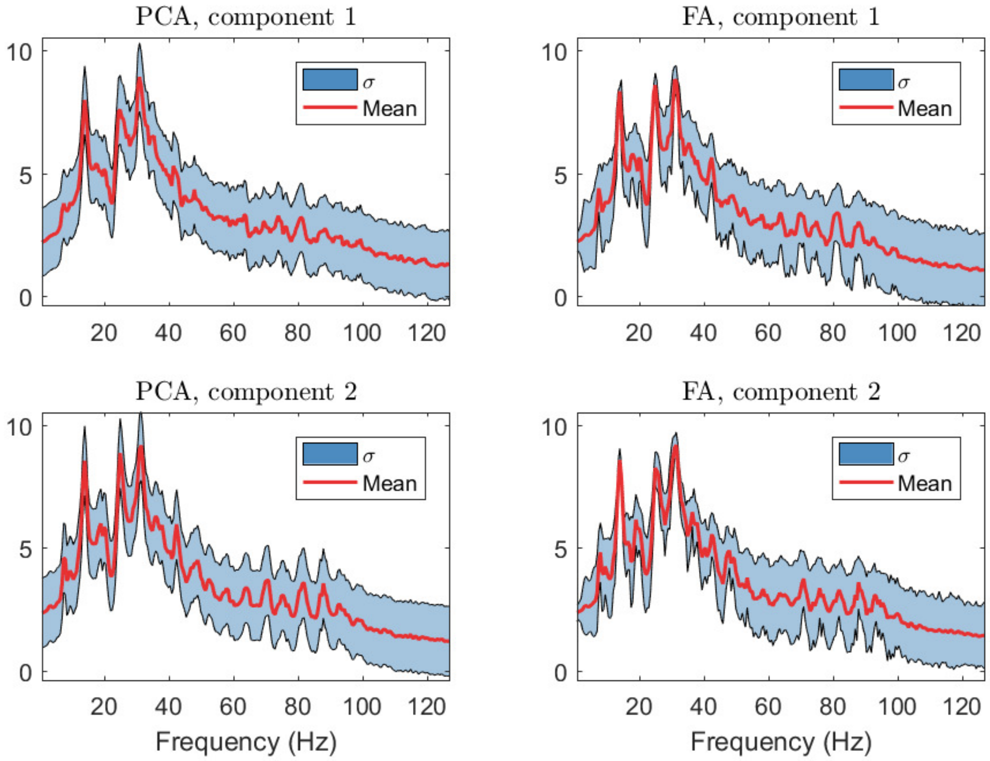

Figure 6.13: Mean and variance identified through a 2-component mixture model of PCA and FA to the 3-DOF nonlinear response at excitation $\sigma = 16N$. Note the variances assigned by FA are different across dimensions. The underlying linear natural frequencies are assigned less variance.

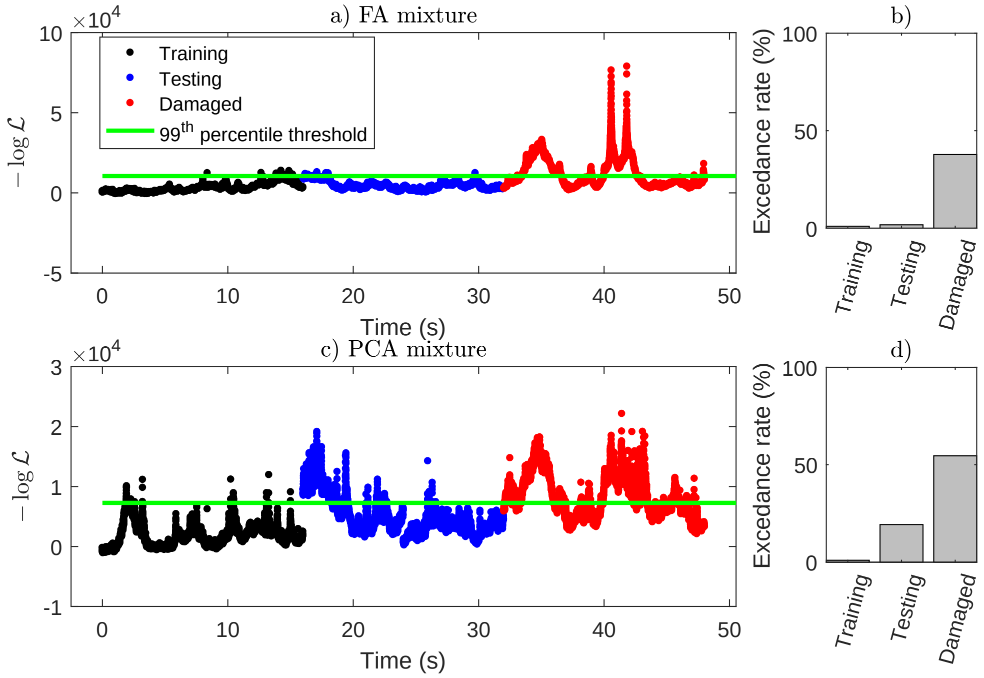

Figure 6.14: Negative log-likelihood evaluated on feature vectors from nonlinear 3-DOF system response at high excitation level, comparing mixtures of a) Factor Analysis and c) PCA. The data are grouped by training, testing and damaged sets condition. The horizontal line shows a 99^{th} percentile threshold on the training set. b) and d) show the threshold exceedance rate for FA and PCA models respectively.

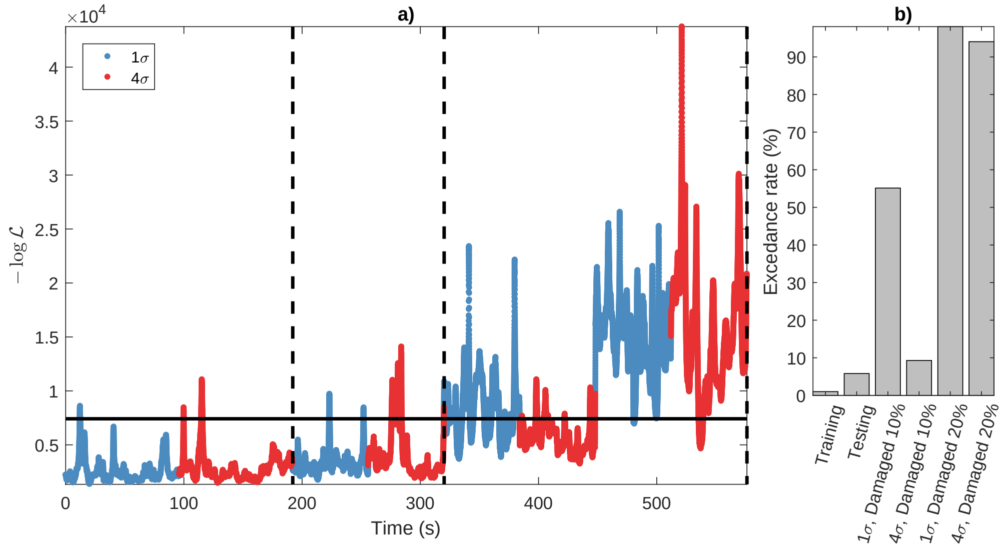

Figure 6.15: a) Negative log-likelihood evaluated on feature vectors from the nonlinear 3-DOF system response at different excitations conditions, using a Factor Analysis model, for training testing and damage sets. Data is grouped by energy level of excitation. Horizontal line shows 99^{th} percentile threshold. b) shows exceedance rates for the threshold for training, testing and damage sets.

The negative log-likelihoods for the training, testing and damage sets are shown in Figure 6.15.

While detection is more prominent in the 20% damage scenario, the more subtle 10% damage case suffers from more false negatives which is undesirable. Generalisation is also not as great as with the linear cases. Performance in terms of true positive and false negative rate generally decreases with a higher excitation level. Performance could be improved by refining the model and increasing the number of training observations; however, this nicely highlights the issues at hand. Also, in this numerical simulations the excitation and nonlinearity levels have been artificially increased to exacerbate the nonlinear effects.

6.4.2 Dynamic Bayesian network inference

Finally, this section shows the application of the factorial Switching Linear Dynamical System (SLDS) to the problem of detecting damage in a nonlinear system with changing loading conditions. The loading conditions used to demonstrate this will be the same as those just used to demonstrate the application of the FA mixture model, namely excitation under white Gaussian noise, under two different energy levels with $\sigma = 1, 4N$. The complexity of the model in this case has to be much higher than that for distinguishing between two different linear systems, as the nonlinear response has more frequencies of interest and these vary so both the number of models K, as well as their individual order is much greater than for the counterpart linear problem. In this case it was found that a three-component model was appropriate as it yielded good generalisation together with discrimination of damage. Figure 6.16 shows the expected negative log likelihoods evaluated using the 3-component factorial SLDS for the training, testing and damage sets. Note that damage discrimination in this case is much better for the higher energy cases, with no false negatives. This is in contrast with the detection using the FA mixture model. The difference could be explained by the fact that a feature vector is required, with some level of preprocessing, by the FA mixture model. A dynamical system, operating directly in the time domain is invariant to changes in amplitude of the signal, which is possibly why it also generalises better than the static Bayesian networks explored here.

As was also the case with the stiffness variability, the factorial SLDS is generally better at discriminating the lower levels of damage, even though at the low excitation levels this is more subtle. This shows that the mixture modelling framework is quite feasible even when dealing with the difficult problem of detecting damage in a nonlinear system, and a factorial SLDS provides an efficient way of doing this, while retaining interpretability, by effectively linearising the dynamics through segmentation of the data using mixture components. This is not the only way one could go about this problem using a dynamic Bayesian network; a particle filter, or other nonlinear model could be used instead. However, here it is clearly demonstrated that the filtering density, or rather its expectation, which represent the data likelihood under this model, is a viable measure of novelty.

6.5 Chapter conclusions

An application of the Bayesian network approach to damage detection has been demonstrated in this chapter, to a simulated mass-spring-damper system representative of real engineering EOVs. Three different scenarios for damage detection have been considered, where damage was introduced as a stiffness reduction on k_2 (see Figure 6.1):

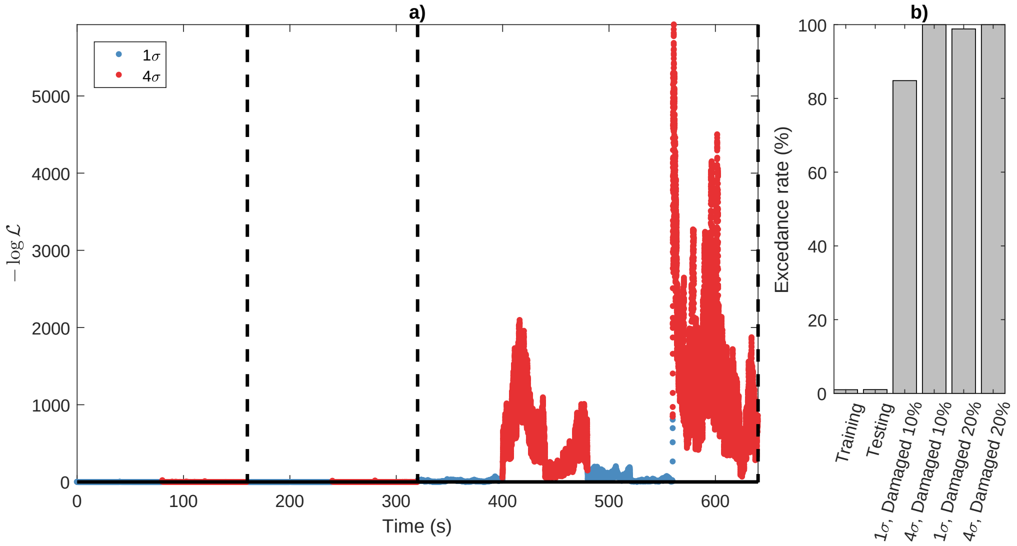

Figure 6.16: a) negative log-likelihood derived from the 3-component Switching Linear Dynamical System (SLDS) on the training, testing and damaged data set of the 3-DOF nonlinear system excited at different energy levels. Data are grouped by different loading conditions. Note the vertical axis has been truncated to show the data more clearly. Horizontal line shows 99^{th} percentile threshold. b) shows exceedance rates for the threshold for training, testing and damage sets.

1. No environmental variation, to establish a baseline EOV-free case
2. A system with variation in global stiffness, emulating temperature effects
3. A nonlinear system, subject to varying excitation levels.

In all three cases, the use of both static and dynamic Bayesian networks was demonstrated. The purpose of this was to illustrate the applicability of more than one approach to modelling the system with a Bayesian network. The main point here is the consistent use of the likelihood function as a novelty measure, capable of detecting damage accross a wide range of models. On the baseline, EOV-free scenario, it was shown that a linear Kalman filter can be used as a dynamic Bayesian network directly on the raw data. PCA can be used on features such as Fourier coefficients extracted through an STFT. The likelihood function of both of these models was successful at detecting damage. However, the advantage of the Kalman filter was highlighted due to the ability to deal effectively with raw measurements, making it a suitable technique for real-time damage detection applications.

In the second scenario, with varying global stiffness in an undamaged condition, a mixture extension of both the above models was used to treat the problem of environmental variability. In the static data model case, this was a mixture of PCA models, while in the dynamic data framework a factorial SLDS was used due to its computational efficiency. Some of the computational issues of switching linear dynamical systems were discussed. In both cases, it was shown how the model likelihood was able to discern between an environmental variation and damage.

In the third scenario, with structural nonlinearity the mixture PCA and factorial SLDS were used again. However, (and as expected) this case proved much more difficult to tackle, requiring models of higher order and greater number of mixture. To illustrate this, the models (mixture PCA/FA and factorial SLDS) were fitted first to the response of the system at a single high excitation level of $\sigma = 6N$. This system on its own proved to have enough complexity and variability, requiring a three-component mixture with individual orders of at least 25 to just semi-successfully capture damage. Variability in loading was then demonstrated on two different excitation levels of $\boldsymbol{\sigma} = \{1, 4\}N$ where both FA mixtures and SLDS performed well in terms of likelihood based damage detection.

Chapter 7

Case Study: Detecting Damage on the Z24 Bridge

The purpose of this chapter is to provide an application case study of the use of Bayesian networks for damage detection, using a well-studied dataset. It also provides an opportunity for further discussion about some of the practical implications of the usage of the methods described in the previous two chapters, when applying these methods in real world datasets. The Z24 bridge, has been the subject of numerous SHM studies [11, 8, 18, 27, 16, 13, 81], and this is why it was decided to dedicate a chapter to it. It is a very useful data set as the bridge was monitored for an entire year, during which the natural frequencies revealed a stiffness dependence with temperature. Furthermore, damage to the bridge was introduced towards the end of the monitoring period. The damage was introduced progressively, and consisted of the following changes to the bridge [122] (in this order):

- Pier settlement
- Tilt of foundation followed by settlement removal
- Concrete spalling
- Landsliding
- Concrete hinge failure
- Anchor head failure

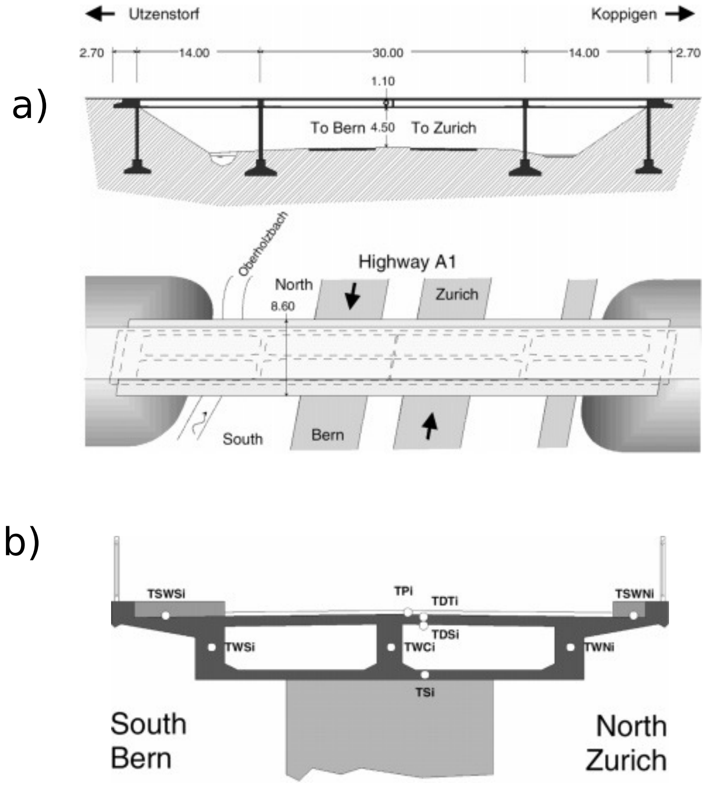

Figure 7.1: Z24 Bridge, showing a) a longitudinal and top view, and b) a cross sectional view [12].

- Tendon rupture

This type of data set is rare, given that one of the struggles of SHM research is the difficulty in acquiring data in a damaged condition, in high value engineering structures operating in their regular environment (outside the lab). The Z24 bridge was located in Zurich, Switzerland and a schematic of it is shown in Figure 7.1.

The Z24 dataset used in this work did not consist of raw vibration waveforms, but instead of natural frequencies extracted through an automated modal analysis procedure, where the modal parameters were identified using Stochastic Subspace Identification (SSI), with an automated pole-picking procedure described in [64]. The SSI method provides a solution to the state space model described in Section 5.3, so it is suitable as an output-only, or operational modal analysis method. Therefore, the system identification procedure used to extract the natural frequencies shown in Figure 7.3 makes use of the natural excitation sources of the bridge, such as wind and traffic loading. The mode shapes corresponding to the four natural frequencies

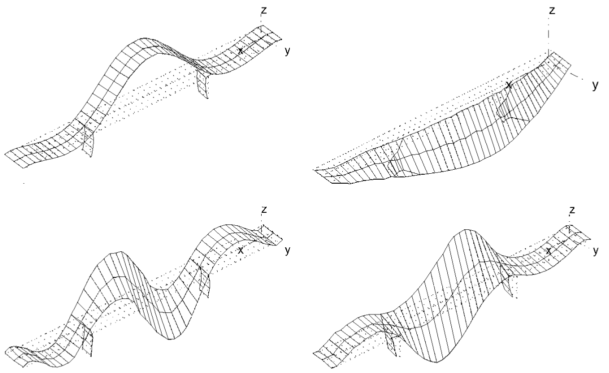

Figure 7.2: Mode shapes corresponding to the first four natural frequencies of the Z24 bridge [64].

being studied in this chapter are illustrated in Figure 7.2. The automated modal analysis procedure did not succeed at finding a stable solution for an eigenvector all of the time, and hence the data set contains a lot of missing data. The lowest success rate was in the fourth mode, with a rate of 77%. The rest of the modes had success rates above 97%. In this study, the data has been sanitised in such a way that only instances where all four modes are available are considered for analysis.

This dataset is no longer a “research challenge”, but it does provide an excellent opportunity to investigate new algorithms. However, before presenting the Bayesian network approach to this data set, it is worth reviewing some of the meritable work carried out on the removal of environmental trends from this dataset. This is provided in Section 7.1. The application of mixture models to characterise the different environmental regimes of this bridge is presented in Section 7.2.

7.1 Previous SHM work carried out on the Z24 bridge

The Z24 bridge has been the subject of numerous studies. The dataset was first introduced in [12]. The first four natural frequencies of the bridge, against temperature are plotted in Figure 7.3. It is clear from them that a bilinear stiffness-temperature relationship exists. When temperatures drop under $0°C$, stiffening of the asphalt layer on the deck is thought to cause an overall increase in stiffness, inversely proportionately with temperature. The SHM problem here is to be able to determine

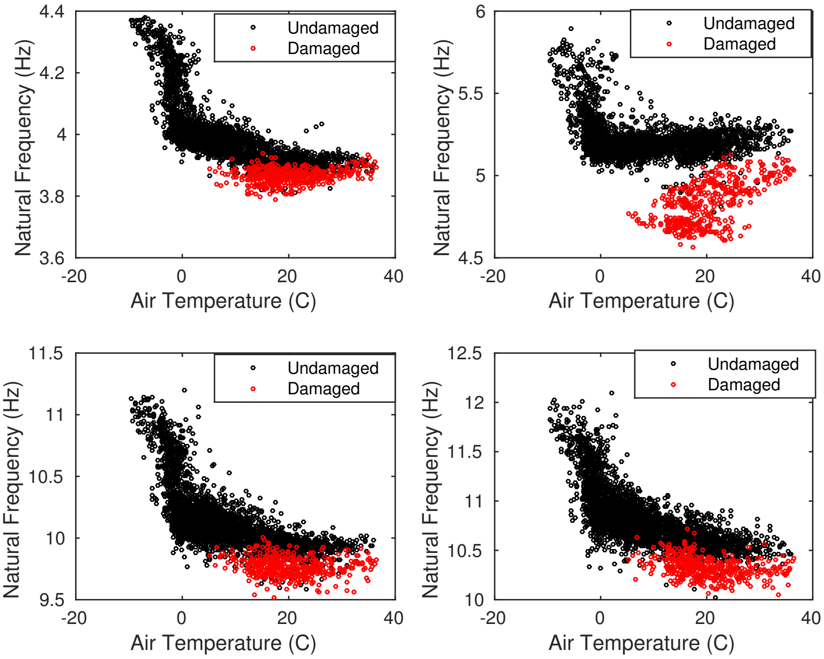

Figure 7.3: First four natural frequencies of Z24 against ambient temperature for the entire monitoring period. Data is separated here in known damage and undamaged conditions [12]

when the change in the natural frequencies is due to temperature or due to damage. Peeters proposed a solution to the problem of separating changes to the natural frequency using linear regression, and Autoregressive with eXogenous input (ARX) models [64]. This approach yields reasonable results, as the ARX model is clearly able to identify the bridge damage. However, this makes use of temperature as an input into the model, and this is the main factor driving the change in stiffness, and hence variation of the features within the undamaged condition. More recently, robust regression analysis has been proposed as a method for separating the influences of the variations due to environmental effects and those from damage. One could use regression, as suggested in [64], to remove the effect of the temperature trend from the dataset, and monitor the residuals of the regression. The issue is that complex environmental relationships may bias the estimate of the regression parameters, yielding a sub-optimal monitoring algorithm. Robust statistics have been proposed to deal with the problem of outliers in SHM data [123], where the argument is made that certain environmental variations could be modelled as an outlier. Dervilis, et al. [123] proposes the use of a robust multivariate statistical distance, the MCD, described in [124]. If a data set contains outliers, the maximum likelihood computation of the mean ($\mu = 1/N \sum_i^N x_i$) will be biased, and fail to estimate the true mean of the data. In this situation, a median is a better estimate of the average of the data. The MCD is a procedure for finding a multivariate Gaussian distribution that ignores outliers when estimating the mean and covariance of the distribution. Dervilis, et al. also suggests the use of a robust regression techinque known as Least Trimmed Squares (LTS), to remove the trend between temperature and the bridge natural frequencies [13]. It is shown that when comparing the Mahalanobis squared distance, from a Gaussian found through MCD, against the residual of the robust regression process, the outliers due to environmental variation appear as "vertical" outliers, while the outliers due to damage appear as "horizontal" outliers; for illustrative examples refer to [13].

Another approach for the removal of environmental trends from SHM features is the use of cointegration [22, 8]. This method is rooted in financial time series analysis, where it is used for the identification and removal of common trends in data. The idea behind cointegration is to reduce a nonstationary series of observations $\mathbf{Y} = \{\mathbf{y}_1, ...\mathbf{y}_n\}$ to a stationary one. It effectively uses regression in order to find common trends within the different dimensions of a multivariate data set. Thus, cointegration seeks to find a linear combination of itself, $\boldsymbol{\beta}'\mathbf{y}_t$ that is cointegrated:

Cross [22] makes use of the cointegrating residuals as feature vectors in a damage detection process for both the Tamar and Z24 bridges. It was found however that this (linear) cointegration only worked well to remove a single trend from the Z24 natural frequencies; it worked best only when data corresponding to temperature above freezing conditions was considered, thus motivating the use of nonlinear cointegration methods.

Finally, some studies have already looked specifically at using GMMs to try to characterise environmental variability, and some have looked specifically into the Z24 bridge. The key difference between these investigations, and the approach presented here, is the systematic use of a likelihood function. For example, [125, 126] investigate the use of a local linear models to represent the Z24 bridge data, but both use ad-hoc damage indexes based on errors from local model. A similar approach is taken in [27]. The contrast is that these studies have not used the GMM as a flexible density estimator, to its full extent. As discussed in the introduction, Kullaa [11] has previously applied a GMM to the Z24 data set for damage detection in a similar fashion to the work presented here, but using a distance metric that first selects the closest cluster member, and then computes a distance weighed by the variance of the local cluster, akin to a MSD measure.

In contrast with all the work described above, the point being made in this work is that the likelihood function derived from a Bayesian network is useful for novelty detection. The idea being that Bayesian network modelling does not solve problems on its own, but rather gives the user the flexibility to build a model that provides an accurate representation of the mechanism generating the data. The added bonus for damage detection is that computation of likelihoods is inherent in the Bayesian network framework. In this context, the mixture framework explains the data generating process for Z24, as physically different dynamics have generated the two main different trends in the natural frequencies. So, the GMM is being used partly as a clustering algorithm, and partly as a density estimator. The next section reviews the application of a Gaussian mixture, to the Z24 dataset in the novelty detection context.

7.2 Modelling bridge natural frequencies with Gaussian mixtures

This section presents the application of the mixture modelling framework to extract a likelihood function for damage detection purposes on the natural frequencies of Z24. Recall from Chapter 4 that the likelihood of a Gaussian mixture model is simply a weighted sum of mixtures. This is re-written here:

$$p(\mathbf{y}) = \sum_{k=1}^{M} \pi_k \mathcal{N}(\mathbf{y}|\mu_k, \mathbf{S}_k) \tag{7.1}$$

where k is the component number, π_k denotes the k_{th} prior mixing proportion, and μ_k and $\mathbf{S}_k$ are the k_{th} mean and covariance respectively. The first step would be to split the undamaged data set into a training set and a validation, or testing set. After sanitation of the data (removal of all failed modes), there are 3500 points that are "undamaged", the majority of which represent data above freezing condition. For the mixture modelling approach to work well, the training data must capture as much of the environmental variability as possible. For this reason the first 2250 points were assigned to the training set, 1250 as validation, and the rest belong to a damaged condition, as damaged was introduced in point 3500 (of this sanitised set).

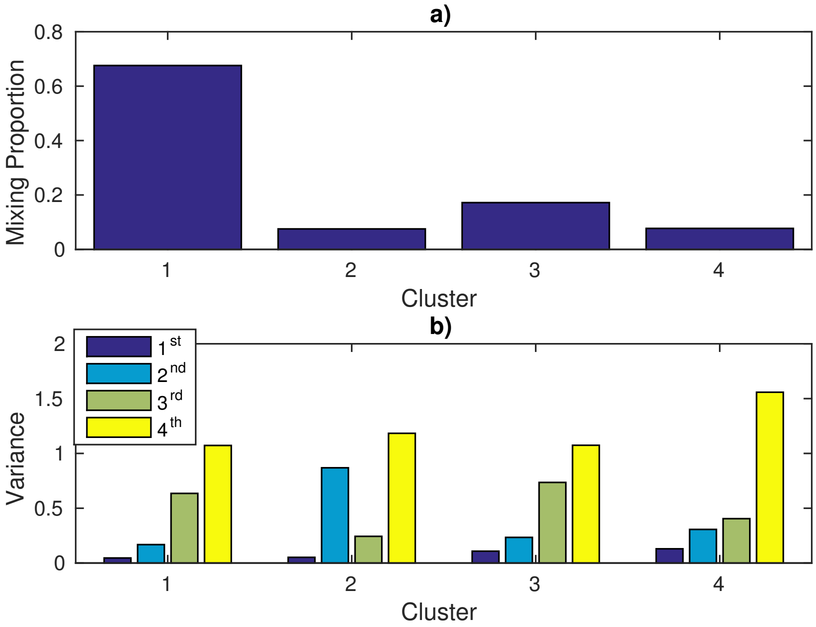

Figure 7.4: 1) Prior probability/mixing proportion for each cluster fit to the Z24 natural frequencies, and b) variance along the diagonal, for each natural frequency, on each cluster

The Bayesian network interpretation of the mixture modelling framework is not fully Bayesian, in the sense that no priors are specified over the parameters of the model $p(\theta)$ and so in this case, the model order was selected according to a minimum BIC index. For this data set, this results in four clusters, with the following prior probabilities ($p(z_k = 1)$), and diagonal elements of the variance illustrated in Figure 7.4.

Although it is hard to visualise, since there are four dimensions in this problem, the resulting model is illustrated in Figures 7.5 and 7.6 in terms of cluster assignments.The clustering is illustrated as hard clusters, assigning the data point to the cluster with the highest posterior responsibility (equation (4.13)). Notice how the GMM partitions the data effectively in regions below freezing, above freezing, and at the transition. The fourth cluster, with lowest mixing proportion tends to take explain all the points that are not very clearly explained by clusters one, two and three. Figure 7.6 shows the clustering, on the trend of the second natural frequency against air temperature. It is interesting to see that the training set did not include temperatures above $18.6°C$, so in effect the GMM is extrapolating its density predictions outside of the regime covered in the training set. This could be an issue if the dynamics had a step change to another regime at this higher temperature. If

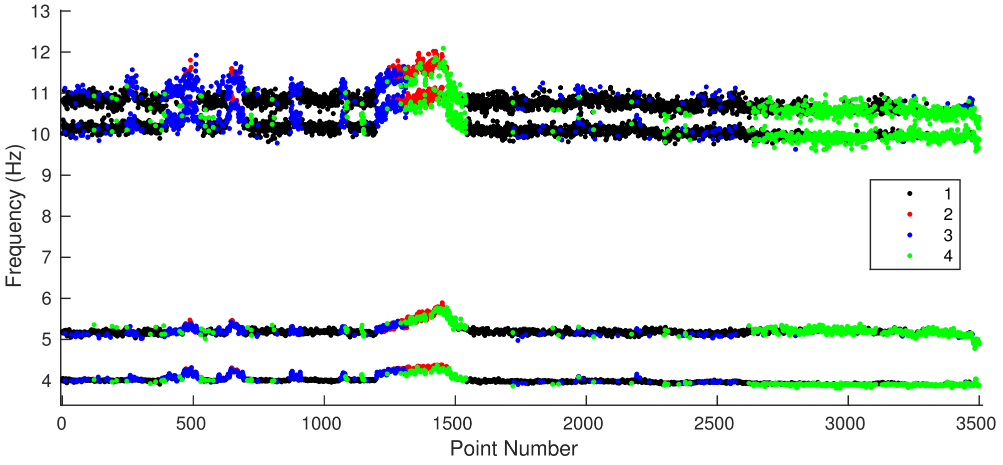

Figure 7.5: Natural frequencies of training and test set, hard clustered by the mixture model

that was the case, this GMM would not be a good novelty detector. Because the dynamics do not change regime at this higher temperatures, the model still explains most of these points above 18.6°C as normal, putting the responsibility on the 4th cluster, even though the center of the fourth cluster is further away than the first cluster.

The first cluster explains the majority of the data, as it has a tight variance, while the fourth cluster has a low mixing proportion, but large variance, so when one evaluates its posterior responsibility on points that may be outliers, they are more likely to be assigned to cluster four. This is also noticeable in the double modality of the density of cluster four, in Figure 7.6; this component is explaining data that does not belong to it. Note that this would not be the case had all the available undamaged condition data been used for training the model, but this would defeat the point of this exercise.

Finally, the negative log-likelihood of the the entire dataset, against the model trained on points 1:2250 is evaluated and shown in Figure 7.7a where it is labelled by training, testing and damaged cases and Figure 7.7b, where the trend of each natural frequency against temperature is shown, and coloured by its negative log likelihood. In this case, a 99th percentile threshold is defined, based on the training data set. It is clear in this case, that the model is successful at capturing environmental variabilities and identifying damage.

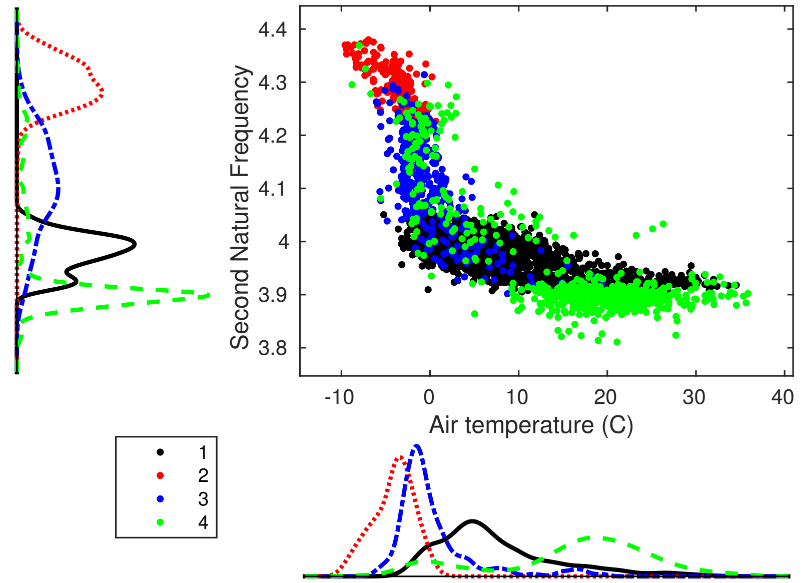

Figure 7.6: Air temperature vs second natural frequency, grouped by cluster, and showing marginal densities. This data comprises the training and test sets. Marginal kernel density estimates are shown for each cluster on the bottom and side of each axis.

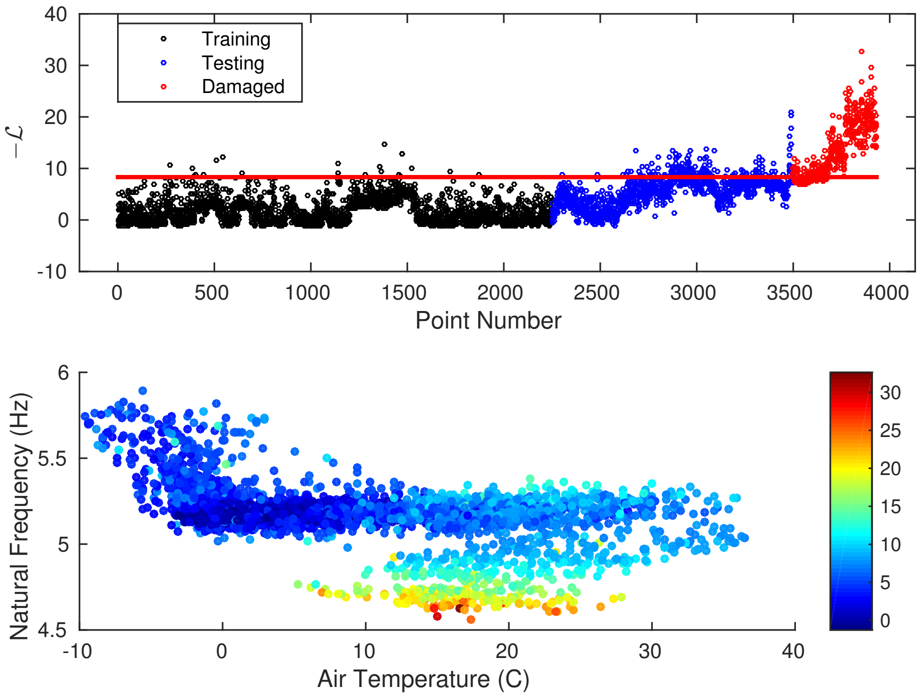

Figure 7.7: Negative log likelihood of model, partitioned by training, test and damaged data, with a 99th percentile threshold

7.3 Chapter Conclusions

This chapter presented a case study on the application of Gaussian mixture models to the well known Z24 dataset; the idea was to contrast the results of other studies on Z24 in terms of the approach to outliers and novelty detection methodology. Here, the mixture model is used a density estimator; the negative log-likelihood is effectively the negative log-density. The threshold that defines the class separation between damaged and undamaged was defined for this exercise as the 99th percentile over the negative log-likelihood of the training set. This may not necessarily be optimal in terms of probability of detection, but it represents true measure of the actual observations against this density. A percentile threshold in this case is conservative; it is close to the mass of the data and thus will have increased sensitivity to damage. A less conservative threshold could be defined using Monte Carlo trials, or by modelling the tails of the distribution of the likelihood function with EVS. Furthermore, there are various interesting extensions to the Gaussian mixture model, that would be able to better estimate the number of components. Two notable approaches are the Dirchlet process mixture [127], and the variational Bayesian Gaussian mixture model [128], which more than anything tackle the problem of model order selection and thus may be able to segment the data better. However, these methods still define a Bayesian network and so their likelihood functions are thus equally suitable for damage detection algorithms.

Finally, the example provided in this chapter has made use of pre-processed features, and the author had no involvement in this pre-processing, only on the machine learning/novelty detection aspect of it which is the subject of this thesis. Chapter 6 offers a better discussion on the use of a mixture model directly on the raw data (through a dynamic Bayesian network) or on pre-processed and different features.

Chapter 8

Wind Turbine Bearing Failure Detection

This chapter presents a case study on the application of the concepts presented so far, to the problem of detecting damage on a wind turbine bearing. The study presented here was part of a larger investigative project carried out by the author, in partnership with Ricardo ltd, and Kongsberg Maritime. Throughout this project, an experimental investigation was carried out into the detection of bearing damage, using an experimental rig with seeded faults, instrumented with AE and vibration sensors. The rig was designed to represent the operational conditions of a planet bearing inside a planetary gearbox. The reason this is of interest is due to their propensity for failure, owing to the fact that the loading is applied at a constant point on the bearing, as depicted in Figure 8.1. As part of this project, a field investigation was carried out also on an operational turbine. However, for the purposes of this work it was decided to include only the experimental rig investigation results for the following reasons:

- Compared with previous studies, the defects introduced by the author were of a realistic nature, therefore novel and worthy of discussion.
- In the experimental rig, the operational variations introduced cover a greater operational spectrum in terms of bearing speed, load and temperature.
- The most important point of all is that the condition of the bearing in the experimental rig is known exactly which means one can know with greater

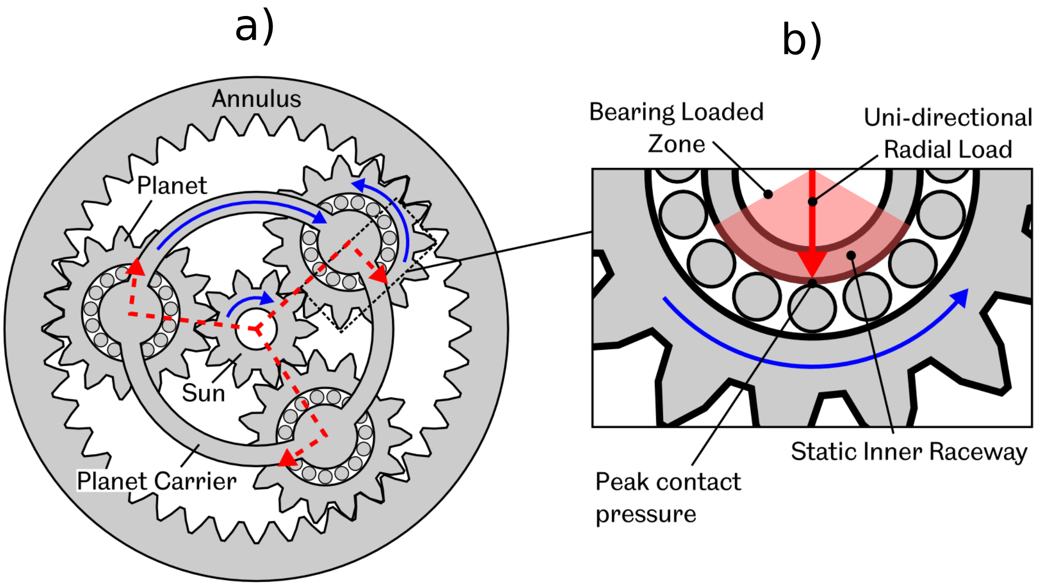

Figure 8.1: a) Diagram showing gearing setup for a planetary gearbox. Note the planet bearings are constantly loaded in the torque direction, indicated by the red arrows. b) zoom-in to one of the planet bearings, highlighting the loaded zone [129]

> certainty whether a damage detection algorithm is providing a false positive or false negative. The condition of the entire gearbox in the turbine is highly uncertain, and hence the exploratory data analysis can yield a lot of speculation.

This chapter will thus focus on the application of Bayesian networks for novelty detection on the experimental AE dataset. Section 8.2 will provide some background motivation for bearing monitoring, and especially for using AE. Sections 8.3 and 8.4 discuss the experimental setup, and the feature extraction methods used on AE data, respectively. Understanding the features being extracted is paramount to understanding the results, so some emphasis will be placed on this, hoping it won't distract the reader too much from the main point in this chapter, which is using a Bayesian network likelihood to infer the presence of damage. This is presented in Section 8.7.

8.1 Acoustic Emission Testing

Acoustic Emissions (AE), when used within a Structural Health Monitoring (SHM) or Non-Destructive Testing (NDT) context, are high frequency stress waves that propagate through a material. These waves can be generated by a number of different mechanisms including stress, plastic deformation, friction and corrosion. A change in the internal structure of a solid will tend to generate AE waves, and monitoring these waves has been shown to be a successful method for detecting the early onset of cracks in various applications [41, 42, 38]. AE testing is a passive method, in the sense that one is listening to the acoustic response of the material when mechanical stress is applied to it. Most materials will have a certain level of AE activity even in an undamaged state when stress is applied to them, the technical term for this is the Kaiser effect [36]. However, when defects such as cracks or spalling are present in the material, the AE response when stress is applied will tend to be more frequent, of higher amplitude, and may have different spectral characteristic depending on the wave modes that get excited. In general, the frequencies that get excited by a defect are inversely proportional to its size. This is the main reason AE testing is often able to capture the onset of damage before it propagates to a large visible crack. There is a well established relationship between the AE response of a metal and fatigue crack growth [38].

Although there is no clear definition of the frequency range for AE, as this depends on material properties and excitation mechanisms, since this depends on the physics of the particular defect, a typical AE stress wave generated from the initiation of a crack in steel can range from 50kHz to 2MHz (from this author's experience). A typical AE measurement system consists of:

1. A piezoelectric transducer, that converts the stress wave into voltage.

2. An amplifier, which amplifies the signal from the transducer; these can be separate or built into the transducer to reduce noise.

3. An Analogue-to-Digital (ADC) converter, to sample the continuous output from the amplifier into discrete, digital waveforms.

The arrangement is illustrated in Figure 8.2.

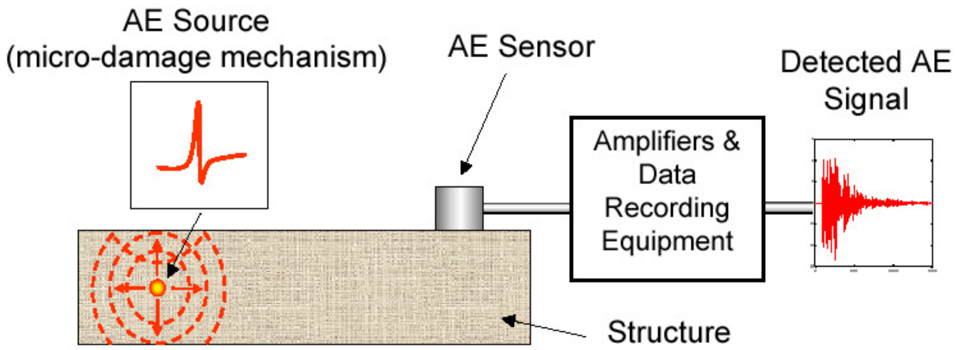

Figure 8.2: Illustration of normal setup required for AE data acquisition

8.2 Background, and comparison against other relevant work on AE monitoring

While AE monitoring is now a well established field, with commercial software being available for its use in engineering structures, not a lot of studies have been published on its use in rotating machinery, especially in wind turbine bearings. In this context, the use of AE monitoring is still considered a research topic. More importantly, the literature in this community has not adopted the use of machine learning methods for damage detection.

The current processing for defect detection in rotating machinery tends to make use of the periodicity within the AE signal to identify and classify damage. This relies on the assumption that a defect will cause a strong component in the frequency spectrum of the raw signal or its envelope. The fact that the component rotates at a known speed makes identification, not only of the defect, but its location in the gearbox relatively straightforward, since the defect will emit energy periodically, at the frequency of the component in question. Identifying the type of defect thus becomes a task of finding the component associated with a particular frequency (relative to the main rotational speed) in a lookup table.

Two paradigms in signal processing have made this task easier. One is the realisation that taking a frequency spectrum of the signal envelope, rather than the raw signal, leads to much clearer identification of spectral lines related to damage [130]. This is because the rolling of a rotating component over a crack tends to send a wave of much higher frequency than the roller passing frequency.

The second paradigm is performing signal processing in the rotational, or angle domain. If one records tachometer pulses from the rotating component, it is possible to convert the raw data into an angular domain. This effectively normalises the data in the frequency domain, so that any further spectral analysis yields frequencies relative to the main rotating component.

Originally, angle domain processing and spectral analysis of signal envelopes started out as methods applied to vibration signals, but soon these methods were applied to AE data as well [130, 131]. Enveloping, is in fact, arguably much better suited to AE data processing than vibration, because AE is recorded at much higher sampling rates, and AE sensors are more sensitive at higher frequencies than vibration sensors.

The problem examined in this study is the detection of subsurface and early surface damage on wind turbine planetary bearings. This is a rather specific but very relevant problem. As the size of wind turbines has grown, to provide more power, the size of the gearboxes has grown as well. Epicyclic gearboxes remain one of the most efficient ways of transferring the power delivered by wind turbine blades, at very low rotational speeds, to the much faster speeds required by generators. This has introduced a fatigue problem with planetary bearings, which observe a constant load in the torque direction of the gearbox as depicted in Figure 8.1, where an epicyclic gearbox and the load direction on planetary bearings is depicted. The problem currently facing the wind industry is that planetary bearings are failing much before their prescribed fatigue life. The reason for this is somewhat uncertain at this point, but it is clear that there is a lack of understanding of the true loading spectrum of these bearings. The fact that the load direction is constant is problematic, as only one area within the circumference of the bearing is fatigued.

It is a standard result from Hertzian contact mechanics, that when a bearing is loaded in compression, pressed by the rolling elements, its stress field reaches a maximum under the surface of the bearing. This results in fatigue cracks that start under the surface and slowly propagate out as the bearing is loaded. Once the crack reaches the surface, progression of damage is quick due to spalling and failure of the entire gearbox is imminent (caused by the introduction of debris into the oil system). One of the current hypotheses for fatigue life in planetary bearings being so unexpectedly low is that gearboxes often observe high impact loads from transient events such as high gusts. This works to start the crack under the surface and this gets further propagated through the regular cyclic loading.

Regardless of the actual explanation for low fatigue life, a monitoring strategy is one way to avoid such situations. For it to be effective, it should be capable of detecting damage at the subsurface stage.

The study presented here focuses on the use of Baysian network inference to attempt a solution to this problem. The experimental investigation used to carry this out is laid out in the next section, while the application of Bayesian networks follows towards the end of the chapter.

8.3 Experimental Setup

This section describes the test rig used to validate the detection of defects using AE. The rig used for this study was originally designed to perform accelerated life tests on bearings. The test rig was designed in order to represent the operational conditions of the planetary raceway in a wind turbine gearbox. The experimental rig is representative of the real-life setup, and contains an inner bearing housed inside an outer support bearing. The key point from a reliability perspective is that, in operation, the inner raceway of a planet bearing has a constant radial load (see Figure 8.1), which means there is a relatively high stress concentration at this point, leading to poor reliability.

The rig comprises of an inner “test bearing, housed by an outer “rig bearing, illustrated in Figure 8.3. A compressive load is applied to load the inner raceway of the test bearing. This is achieved using a hydraulic press, shown in Figure 8.3b (the yellow component under the bearing). The load of the hydraulic press is transmitted to the bearing through two lugs which pull from the left and right side of a shaft. Because the load is split into two components, they are referred to here as left and right-side loads. It has been determined in previous work [132], that a large surface defect (approximately 200 μm width) is detectable by measuring AE directly at the inner raceway. The objective of this study is to determine whether smaller defects and more realistic defects are detectable from measurement positions located on the outer casing. In practice, it would be hard to place a sensor inside the housing of a wind turbine gearbox bearing, and doing so might compromise the integrity of the bearing itself. The measurement locations used in this study are illustrated in 8.3, while the seeded defects are discussed in section 8.3.2.

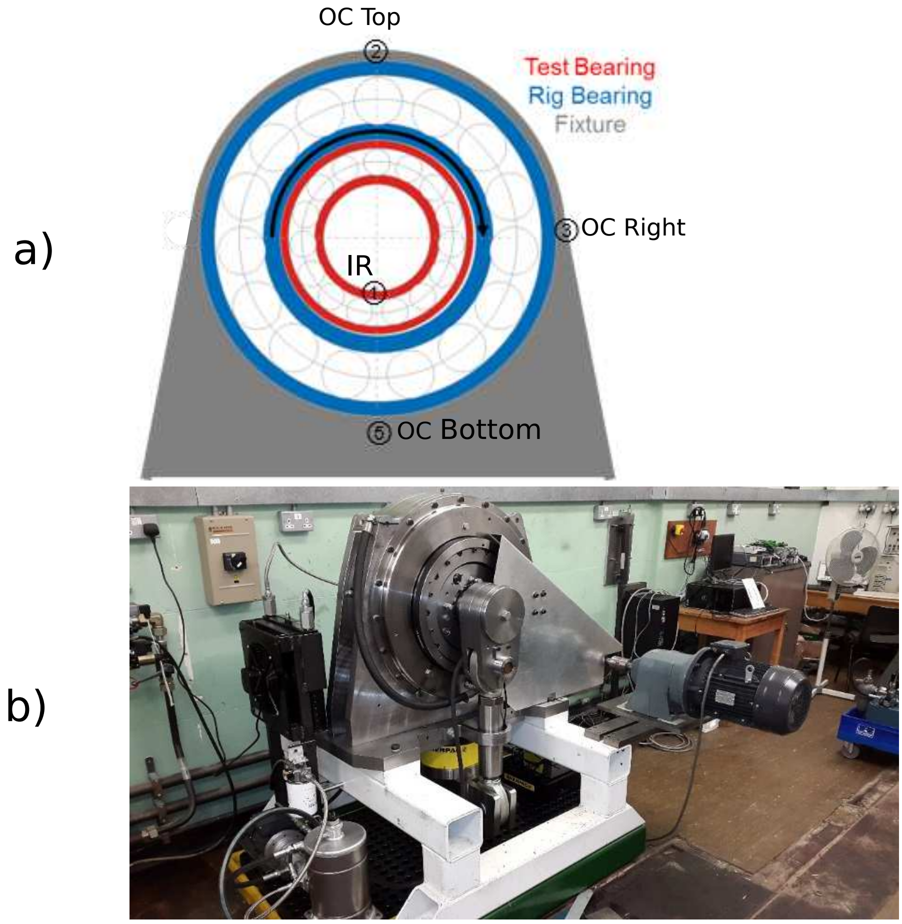

Figure 8.3: a) Diagram showing fixture, rig bearing and test bearing, outlining the location of AE measurement channels. b) Photograph of rig in the lab.

There were a total of five raceway conditions:

1. Undamaged
2. 50 μm surface scratch
3. 20 μm surface scratch
4. 5 μm surface scratch
5. Subsurface damage

which constitute the different seeded defects examined in this study. The objective of the tests is to capture, for each of these raceway conditions, the effects of varying load, speed and temperature. Preliminary tests were conducted, stepping the compressive load at 100kN steps from 0 to 1200kN. This pointed to three major regimes of AE activity, around the low load (0-400kN), medium load (400kN-800kN) and high loads (800kN-1200kN). For this reason three loads were selected for a test schedule: 200kN, 600kN and 1000kN. The temperature of the rig proved difficult to control precisely. The factors that affect the rig and oil temperatures are the operating load, bearing condition (a failed bearing introduces debris into the system and drives the temperature up through friction), ambient temperature, accrued usage time, whether the heat exchanger is present and whether the fan is engaged. Therefore the only means of controlling the temperature directly are via the heat exchanger, and the operating load. In general it is easier to warm up the rig, than to cool it down, as once it runs and a load is applied, it will quickly warm up and reach a stable temperature. Therefore, it was decided to split the tests into low and high temperatures. This split is reasonable, given that the main effect that temperature introduces (to the AE activity) is an increase in friction at higher temperatures from a reduction of viscosity [133, 134]. In order to keep the temperature down, the low temperature runs were performed:

- Early in the morning.
- Testing low loads first.
- Keeping cooling fan on.

Table 8.1: Bearing test schedule. Each row was performed sequentially, and this schedule was used for every bearing condition.

Speed (RPM)	**Load (kN)**	**Temperature**
20	200	Low (Fan ON)
60		
100		
20	600	
40		
60		
80		
100		
100	1000	
100	1000	High (Fan OFF)
20	600	
40		
60		
80		
100		
20	1000	
60		
100		

Table 8.1 shows the complete schedule of tests carried out. This is also a realistic scenario given that the temperature in a wind turbine gearbox will vary in a similar fashion. When not operational, or at low wind conditions, modern gearboxes will keep circulating the oil through a heat exchanger, to keep it from getting too cold (and thus highly viscous). When the gearbox is engaged the oil temperature will quickly tend to reach a stable value.

8.3.1 Instrumentation

The acoustic path between the inner raceway (the source of the damage) was from its bottom to the bottom of the outer casing, and then both clockwise and anti-clockwise through the outer casing. Due to symmetry, it was deemed reasonable to not include a measurement position on the Outer Casing (OC) left side. A National Instruments (NI) C-DAQ chasis was used for all data acquisition. This comprised of several modules:

- NI-9223, four-channel analogue with 1MHz max sampling rate.

- NI-9234, three-channel Integrated Electronic Piezo-electric (IEPE) with 51.2kHz max sampling rate.
- NI-9213, thermocouple, for temperature measurements.

A total of four AE and two vibration channels were acquired, at the positions indicated in Figure 8.3. Several operational parameters were also acquired in order to assess the influence of each one on the AE response. These parameters were:

- Test Bearing Speed (RPM).
- Left-side and Right-side Load (kN).
- Oil Temperature.
- Casing Temperature.

The rig was controlled using bespoke software (coded with the contribution of this author at the University of Sheffield) using NI software which also transmitted these operational variables through a local area network.

The AE sensors used in this study were Mistras 3MICRO-30D sensors with a differential cable for noise reductions and one NANO-30 sensor. The main objective of this study was to determine the detectability at the OC locations. The NANO-30 sensor was used in the Inner Raceway (IR), as it had been fitted there from previous installations and although it was not the focus of this study (it being mounted directly on the inner raceway) it was left there and data were captured for reference purposes, as it constitutes a measurement location where damage should be clear to see in the AE response. The sensitivities of both sensor models are illustrated in Figure 8.4. Note that the Micro30D has a marked resonance at approximately 350kHz, while the Nano30 has a flatter response on the range 200kHz-500kHz. This is relevant as the sensor frequency response shapes the acquired signals significantly. Compared with vibration sensors, the frequencies of interest are much broader, in the range of 50kHz - 2MHz. It is therefore much harder to achieve a flat frequency response across all frequencies of interest, and so one must accept the significant filtering that the sensor applies to the "true" underlying signal. It must also be noted that sensor-to-sensor variability is usually much more significant in AE sensors, compared with vibration instrumentation. The sensitivity curves provided in Figure 8.4 are a generic ones from the manufacturer to illustrate the general trend, they are not sensor-specific.

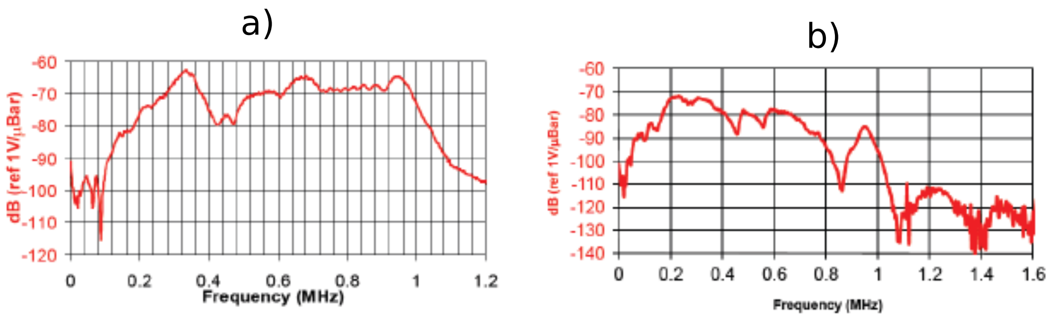

Figure 8.4: AE sensor sensitivities for a) Mistras Micro-30D and b) Mistras Nano 30. Source: Mistras UK

8.3.2 Seeded Defects

This section details the defects seeded into the inner raceway. For the purposes of this study, two types of defects were seeded, surface and subsurface. The idea behind using these two types of defects is to investigate the detectability of faults at different stages of development. In a typical wind turbine bearing, a fault will start below the surface before it propagates out.

Surface defects

In previous work [132], a spark erosion technique was used to etch the surface of the bearing to emulate a surface crack, which generated surface defects of approximately 200μm width. In order to achieve a smaller defect, more representative of the early stages of a surface crack, a Cubic Boron Nitrite (CBN) grit was used to scratch the surface. This was performed at three different pressures, with each one at a different angular position on the raceway. The aim of using three different angular locations along the raceway was to be able to perform a test with three different defect sizes by simply positioning the different defects on the loaded zone of the bearing. The angular position of the defects in the raceway is shown in Figure 8.5. The target sizes for the seeded defects were 5μm, 20μm and 50μm. The etches were performed with the help of the Advanced Manufacturing Research Centre (AMRC) with Boeing, who facilitated the use of a 6 axis CNC machine that could scratch a straight line, at various levels of pressure, with very high precision across the outer surface of the bearing raceway.

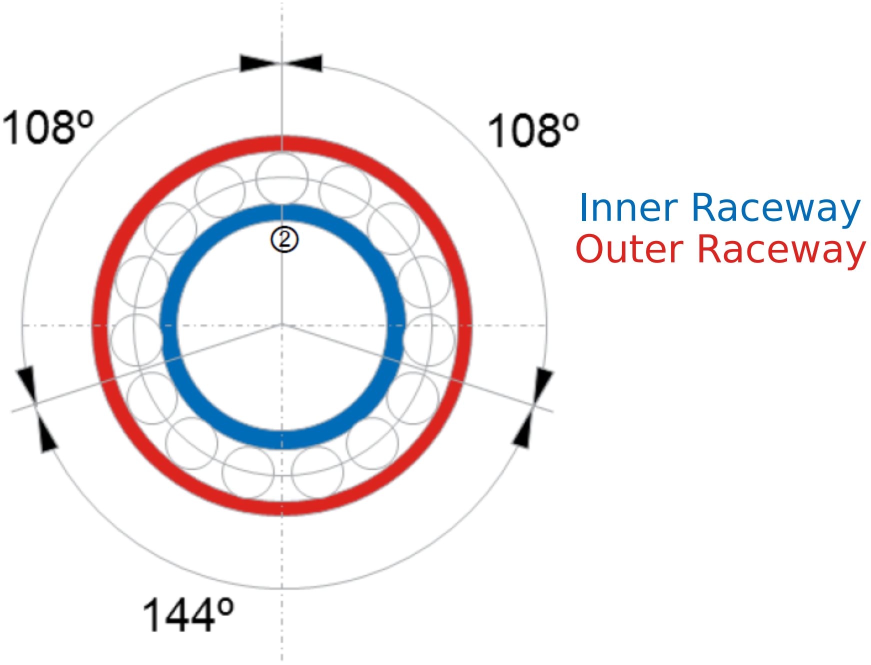

Figure 8.5: Angular positions of defects along raceway. Note that this angular separation ensures that when one roller passes over one of the defects, none of the rest of the rollers will pass over the other two defects at the same time.

Subsurface defects

The subsurface defect was seeded by means of compressing the outer surface of the bearing raceway with a rolling element. The compression was applied using a hydraulic press capable of applying up to 2000kN. Subsurface yield was estimated to occur at 1000kN for this bearing, using Hertzian contact mechanics relationships. To ensure that subsurface yield occured, while also preventing the damage to propagate to the surface, the yield process was monitored using AE. Some of the observations on AE from this damage seeding are further discussed in [37].

During the tests, a large increase in AE energy was observed in the 800kN to 1200kN range. Visible surface damage was only found when the bearing was loaded beyond 1700kN. Although several tests were carried out on numerous raceways, on the final raceway, three compressions were applied with maximum loads of 800kN, 1000kN and 1200kN. These were applied on the same circumferential indexes as for the surface damage. In this case, however, contact between the damage and the rollers in the unloaded zones is not a concern since the damage is under the surface.

To summarize, two bearing raceways were damaged. One raceway contained three surface etches with increasing sizes, to emulate increasing levels of damage. The second raceway contained three seeded subsurface cracks, with increasing levels of maximum compressive load. From the subsurface damaged raceway only one damage site was used in this study; the one with where the maximum compressive load was applied.

8.4 AE Signal Processing and Feature Extraction

Custom signal processing methods had to be developed for this study. The objective of this is to process the data from raw AE signals to yield information regarding the state of the system, in this case the bearing. The approach to fault identification using AE is fundamentally different to that used in structural damage detection using vibration, and therefore the signal processing required is also different. When performing damage diagnosis in non-rotating structural systems using vibration, one would typically look for changes in resonance frequencies of the structure, which can be inferred from frequency-domain processing, modal analysis and time-domain models. In the case of AE monitoring, one is listening for the release of stress waves being emitted at the damage source. The key point, from a signal processing point of view is that the bursts captured by the AE acquisition system are very short in comparison to the large amount of time that needs to be spent monitoring. Because of the high sample rates required to capture these high frequency waves, this means that a lot of noise is recorded, in comparison to the amount of useful AE bursts. To put this in context, the bursts recorded from a yielding steel specimen may last on the order of 2000 μs. If one were to monitor at 1MHz for 1 second, and expect 15 bursts (which is roughly how many bursts are expected in this rig at 100 RPM), this would mean approximately 3% of the data points are informative and the rest is noise. Given the high sample rate, data storage and handling becomes an issue if one wishes to monitor for long periods of time.

This has led the AE community to develop hit-extraction strategies, where an AE hit is defined as a burst large enough for it to be likely to be caused by material fracture. These hits are then useful features to perform further processing and statistical pattern recognition. It was found during this investigation that simple features, extracted from hits were sufficient to identify the presence of a fault. In non-rotating systems, a simple threshold is often sufficient to identify a hit within the background noise. In rotating systems, however, there are more sources of acoustic noise, and bursts that are not related to damage but generated by friction, roller impact, etc; this is one of the primary reasons that it is hard to find comprehensive studies on the use of AE in rotating machinery. In the investigations presented here, misalignment of the bearing rig was also a significant source of noise. The contribution to the noise from the misalignment arises due to edge loading on the bearing; the compressive loads being applied by the lugs are not even. On a "quiet" structure, it is normally straightforward to identify hits by setting a threshold on the overall AE signal; the value of the threshold would mostly be determined by the background noise level of the environment and the electrical noise. Having a constantly changing noise level introduced by periodic friction complicates matters. In this case an adaptive threshold was clearly required to generate a hit extraction procedure, robust to changing noise levels. This is described in the subsection below.

The overall fault identification procedure using AE hits is to first detect the hits, derive features from them, and then characterise these using statistical pattern recognition. The last step in this case consists of using Bayesian network inference.

8.4.1 Adaptive threshold methodology

While the adaptive threshold is not the main focus of this work, it plays a central part in computing the features used to detect damage, so it is important to devote a small section to this. The main objective of thresholding is to identify the presence of a hit, in any given channel. This does not mean identifying the *onset*. A hit onset is defined (here) as the precise time of arrival of the first wave mode. Before computing this value it is useful to first define whether a hit is present. This task is referred to here as"‘hit extraction", and this is where the adaptive threshold plays an important role. The AE response is characterised by the presence of numerous *spurious* hits. The source of spurious hits can be electrical noise, external sound, and any other process alien to the rotating system. When selecting a thresholding strategy for hit extraction, there are two approaches:

- attempt to weed out spurious hits in order to make the pattern recognition step easier;
- collect as many hits as possible and let the pattern recognition algorithm separate them.

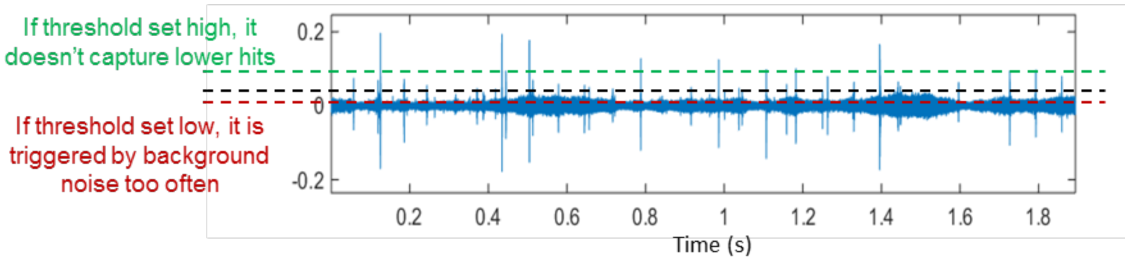

Figure 8.6: Illustration of effects of setting a threshold either too low or too high, showing a sample of raw AE data with periodically varying noise level

In this study, the latter approach is taken, in the spirit of examining the capability of Bayesian network inference upstream of the process, and not risking information being lost through assumptions in the thresholding strategy.

It is the author's observation, that in this rig, and in wind turbine gearboxes generally, there are various sources of AE hits that are unrelated to the damage process. The impact of rollers with the bearing casing, if there is a sudden change in speed and their centrifugal force is reduced, and friction, in both the inner and outer casing tend to generate large amplitude, high energy hits. Debris in the oil system also causes a great deal of additional AE hits, but these tend to be of lower amplitude.

Furthermore, misalignment in the bearing or load being applied can easily lead to a periodically changing noise floor level; this is a problem that is not encountered in non-rotating systems. The consequence here is that applying a threshold over the overall noise floor, results in either identifying noise as hits, if the threshold is set too low, or missing out lower energy hits, if the threshold is set too high. This dilemma is illustrated in Figure 8.6, showing an example output from an AE stream collected on the bearing rig (in an undamaged state).

A thresholding strategy is required that identifies the presence of a hit, within a constantly changing noise floor. The methodology developed here makes use of a simple thresholding function, which computes the difference between the local signal energy E_t and some lagged version of itself E_{t-a}. The difference is then normalised against the local noise level at $t-a$. The local signal energy is easily computed with a moving Root Mean Square (RMS), with a window size within the order of lag a. This thresholding function can be computed using:

$$T(t) = \frac{E(t) - E(t-a)}{E(t-a)} \tag{8.1}$$

where $E(t)$ is the local signal energy computed using a moving RMS statistic. Because this function normalises the difference of the local signal energy against the background local noise, the value of the threshold is also normalised. The selected threshold value over the adaptive thresholding function of equation (8.1) is denoted here as $\mathcal{T}$. Choosing a value of $\mathcal{T}$ over the thresholding function $T(t)$ means selecting those bursts that rise $\mathcal{T}$ times above the local noise floor. The value of a determines the lag used in comparing current local energy against previous values. In this case a was tuned empirically, to be approximately equal to the average rising time of an AE hit.

AE data streams comprise millions of points, and the feature extraction process being described here is applied to hundreds of data files. Efficient computation of features is therefore required. To reduce the number of time points that moving RMS, and thresholding functions need to be computed, for a multi-level wavelet decomposition is used as a preprocessing step. The AE sensors are only sensitive within a relatively narrow frequency band of the actual bandwidth of the sampling process. The wavelet decomposition can make use of this fact and "throw away" any wavelet coefficients that do not belong to a frequency band where the AE sensor is sensitive. If using a broad-band AE sensor and data loss is a concern, this step could be skipped. and the adaptive threshold computed directly on the raw signal.

In summary, the steps taken for detecting the presence of a hit, for every AE channel are:

- Decimate signal with a wavelet decomposition, adjusted to filter out any frequencies where the sensor is not sensitive
- Take an envelope $E(t)$, of the wavelet coefficients, to capture the amplitude modulation of the process. The envelope could consist of a Hilbert transform or a moving RMS to compute the local signal energy.
- Compute the thresholding function given by equation (8.1), and derive T.
- Set a meaningful threshold over T, using any of the methods discussed in Section 3.2.2. In this case, empirical CDF percentiles were used.

These steps are illustrated in figure 8.7. Once an appropriate threshold is obtained, that normalises out background noise, one can proceed to extract hits based on this threshold.

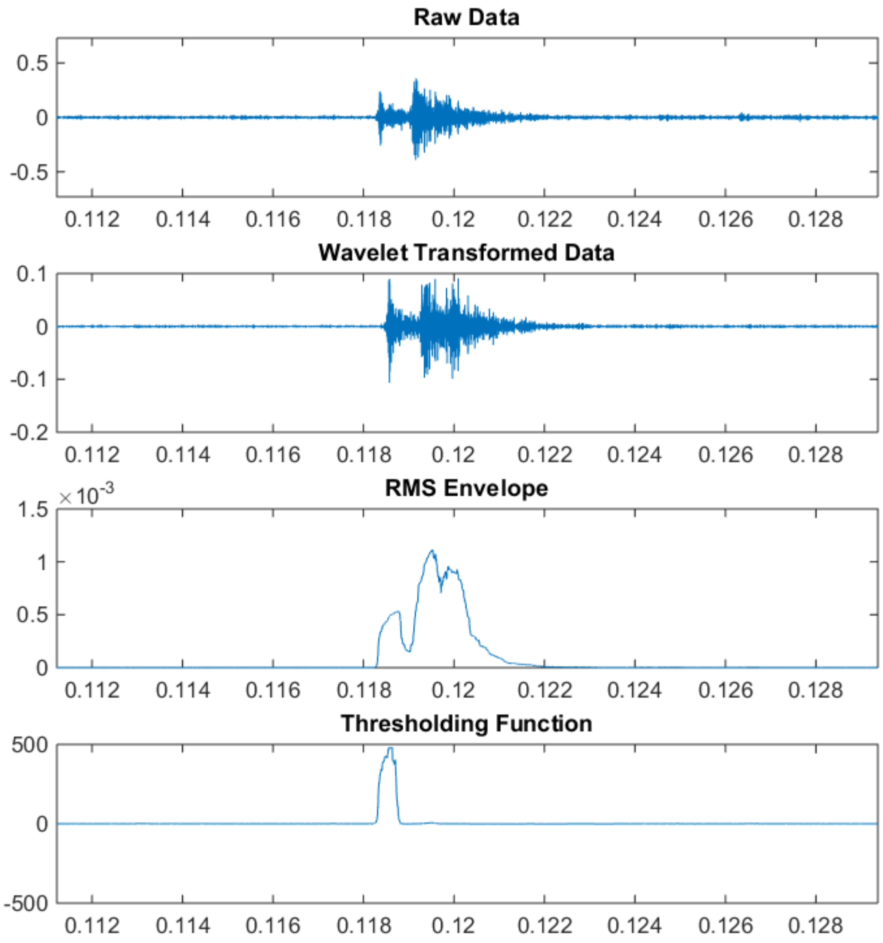

Figure 8.7: Illustration of steps taken for the computation of the thresholding function, from top to bottom

8.4.2 Hit extraction procedure

Once the presence of a hit has been identified by the thresholding strategy (adaptive or non-adaptive) some further steps are required to identify when the burst roughly starts and finishes. The start and end time indexes are merely required to store data, so as not to record the entire AE data stream, and to try to retain as much information that relates to the damage process. Note clearly that in the method described here, features are computed as a post process of the data recorded

Hit identification is not trivial, but a lengthy discussion of the process and its alternatives would distract from the main point. The procedure is therefore only outlined here for reference, as it is still important to understand the assumptions made while extracting the AE hits. The general process is described in Figure 8.8, which includes the steps outlined above for computing the adaptive threshold.

The key output of this process is the start and end times of the hit, using the trigger from the thresholding strategy. A rough start time is already given by the exceedance of $\mathcal{T}$ on the adaptive threshold $T(t)$. Recording the stop time, requires first scanning forward to find the maximum amplitude of the burst, within a specified forward scan length. Once the time index for the maximum amplitude is found, the algorithms scans forward again, to find the point where $E(t)$ reaches within $n\%$ of its value before the burst started; the local noise floor. In this study, a maximum hit length is also defined, as it is possible for a hit to to last unrealistically long times if the local noise floor rises suddenly towards the end of the burst.

Lastly, in a rotating system of this kind it may be possible to encounter hits that are very closely spaced, and where the second one comes before the first one decays back to the noise floor level. For this reason, the algorithm implemented here (Figure 8.8) checks whether another threshold exceedance has occurred before defining the hit end time. If one does occur the new hit is defined then, with a small number of buffer of points before it, and the first hit is ended at that time index.

8.4.3 AE hit features

Once a table of start and stop positions has been extracted from the AE data for every channel, it is relatively straightforward to go back to the signal and save only the waveforms at those time instances. This is the strategy that has been adopted; it significantly reduces the amount of data stored, and focuses all the post-processing on the data points corresponding to AE hits only, which as discussed before, comprise only a small percentage of the data points in the signal. There are numerous features that can be extracted once the waveform has been captured. Because an AE waveform is a transient event, there are some simple key features that can characterise it in general, but simple, terms. Note however that an accurate computation of some key features depends on one thing: an accurate estimation of the onset of the AE waveform. The method used for onset estimation will be discussed later in Section 8.4.4.

Possibly the most informative feature is the energy contained in the waveform.

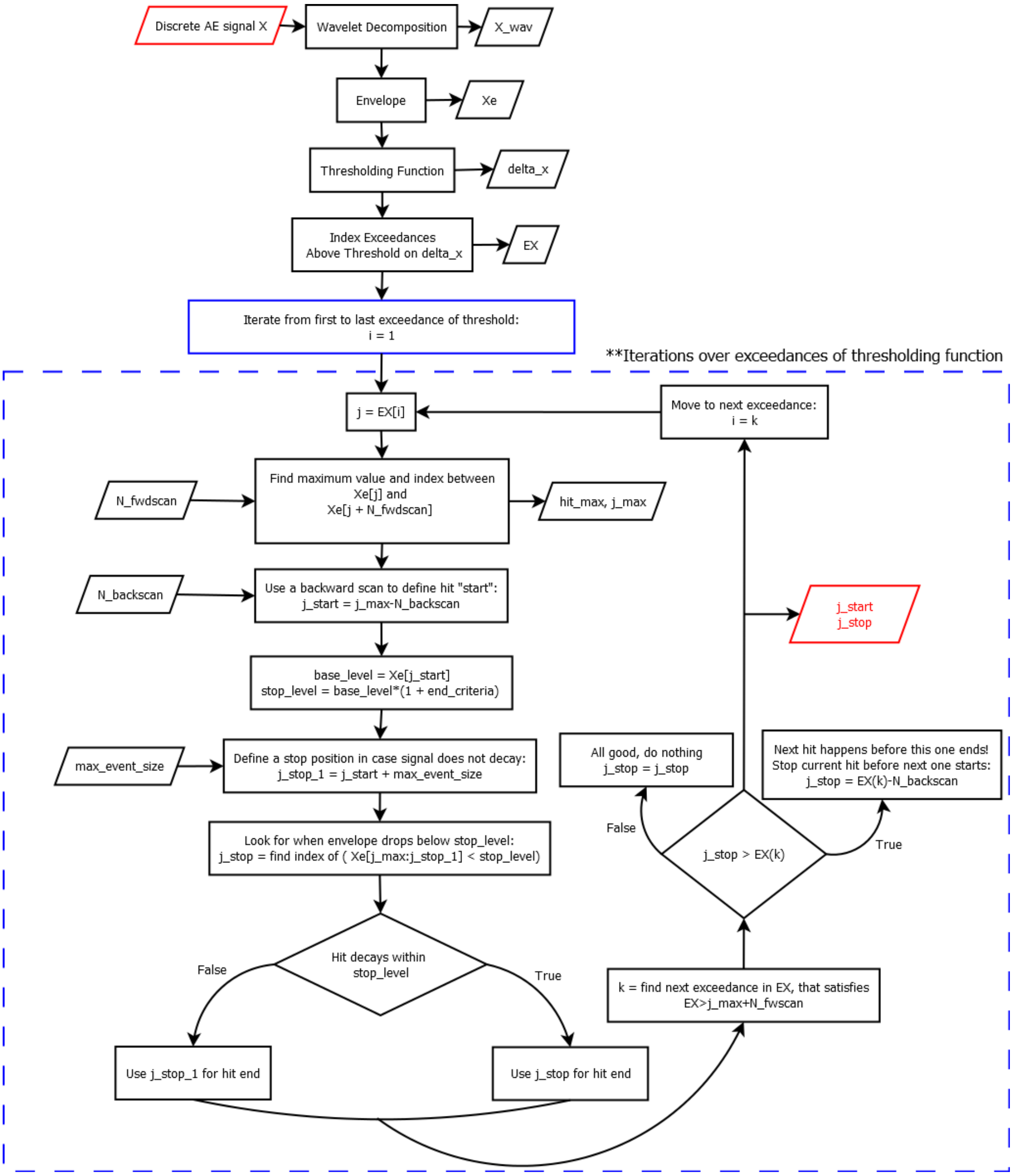
Discrete AE signal X
Wavelet Decomposition
X_wav
Envelope
Xe
Thresholding Function
delta_x
Index Exceedances
Above Threshold on delta_x
EX
Iterate from first to last exceedance of threshold:
i = 1
**Iterations over exceedances of thresholding function
j = EX[i]
Move to next exceedance:
i = k
N_fwdscan
Find maximum value and index between
Xe[j] and
Xe[j + N_fwdscan]
hit_max, j_max
N_backscan
Use a backward scan to define hit "start":
j_start = j_max-N_backscan
j_start
j_stop
base_level = Xe[j_start]
stop_level = base_level*(1 + end_criteria)
max_event_size
Define a stop position in case signal does not decay:
j_stop_1 = j_start + max_event_size
All good, do nothing
j_stop = j_stop
Next hit happens before this one ends!
Stop current hit before next one starts:
j_stop = EX(k)-N_backscan
False
j_stop > EX(k)
True
Look for when envelope drops below stop_level:
j_stop = find index of (Xe[j_max:j_stop_1] < stop_level)
False
Hit decays within
stop_level
True
k = find next exceedance in EX, that satisfies
EX>j_max+N_fwscan
Use j_stop_1 for hit end
Use j_stop for hit end

Figure 8.8: AE hit extraction algorithm flow chart

Different sources of AE will release stress waves at widely different energy levels. It is outside the scope of this investigation to characterise the energy levels of different processes within a wind turbine (or other) gearbox; the approach in this study uses a Bayesian network to characterise the energy of "normal" AE hits, against those of abnormal AE hits. Energy, power and RMS are all very closely related. The energy is easily computed as the sum of squares of the data points. The power normalises the energy by the duration of the signal, and the RMS is simply its square root. In the case of a transient waveform, such as that of an AE hit, energy, power and RMS will all be related to each other since the duration is a function of the total energy, because of the exponential decay in amplitude. For this reason, only the power has been used here as a feature, since all three convey similar information.

The risetime, is defined as the time difference between the waveform onset, and its maximum amplitude. The information this carries is valuable because due to the difference in speeds of different wave modes, some will arrive first and some later, thus giving a rough indication of how far the source is from the sensor. In practice, in a steel structure, waves will propagate as longitudinal, transversal, surface, and possibly Lamb waves. The Lamb wave modes may or may not be excited, as their existence requires that the wavelength of the AE be of the same order of magnitude to the thickness of the material it is travelling through. An investigation of Lamb waves is outside the scope of this report, but their use should not be discarded and is marked as future work. In steel, longitudinal, shear and surface waves arrive in that order. The amount of energy they carry is also given in that order. Therefore the first arrival will always be from a longitudinal wave, and the maximum amplitude will tend to be recorded at the arrival of a surface wave. The usefulness of this is that the risetime of an AE hit is a useful feature as it gives an indication of how far the wave has traveled. Waves that come from far away will have high risetime (separation between longitudinal and surface waves) while the opposite is true for short rise times.

Other features that are collected are the peak amplitude of the signal the total duration and the decay time. The duration is defined, during the hit extraction process, as a decay after the peak amplitude to a level within a specified tolerance of the baseline noise, immediately before the hit. The duration will tend to be a function of the energy in the waveform, but also of the physical mechanism exciting the wave, and it is therefore a useful feature. Once all of these features for each hit are computed, they are assembled so that inference with a Bayesian network

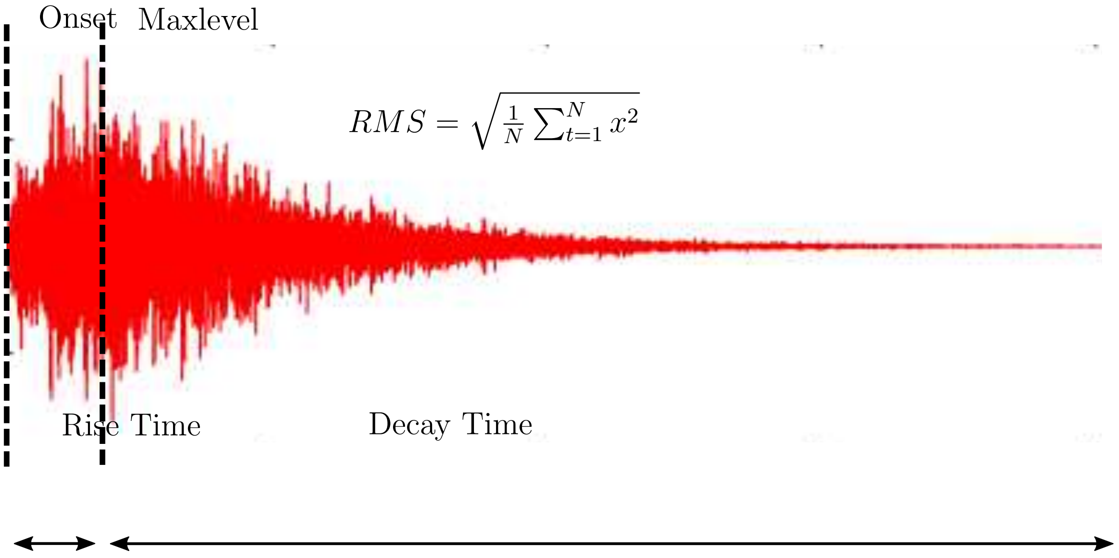

Figure 8.9: Illustration of AE burst and some of the features derived from it.

can be performed with them. An illustration of the features extracted from an AE waveform is given in Figure 8.9.

8.4.4 Onset estimation

The hit extraction process goes as far as establishing that a hit exists, capturing its maximum amplitude, and establishing when it has finished, based either on the decay to a level close to baseline noise, or the presence of a subsequent, closely space hit before such decay. The early days of AE testing use threshold crossings to establish the onset of the waveform. This is, however, not a reliable way of computing onsets given that the longitudinal wave, which arrives first, will have orders of magnitude less energy than the peak amplitude of the AE hit. In this investigation the methodology proposed by Kurz [135], based on the Akaike Information Criterion (AIC) is used. This method computes a cumulative variance forwards and backwards and creates an AIC function as the superposition of these two. Where this function reaches a minimum indicates the highest change of information (or variance) in the signal and thus the onset of the AE wave can be established by looking for a minimum of this function. The AIC function can be computed by [135]

$$AIC(t_w) = t_w \log\left(var(R_w(t_w, 1))\right) + (T_w - t_w - 1) \log\left(var(R_w(1 + t_w, T_w))\right) \quad (8.2)$$

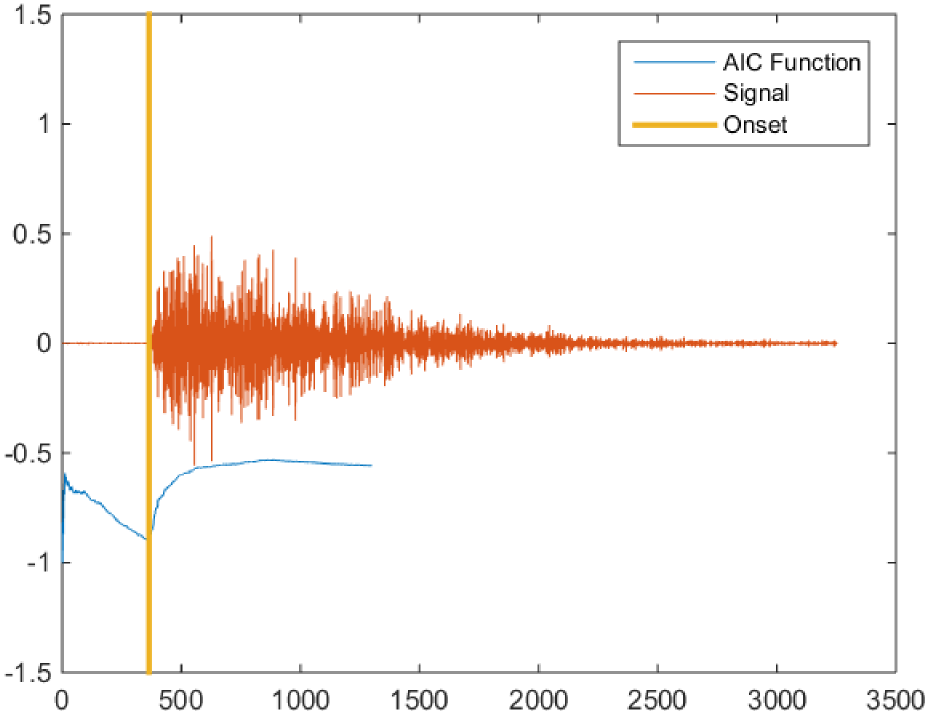

Figure 8.10: Illustration of AIC onset function on a sample AE hit. The onset is defined where the function reaches its minimum, indicating the greatest change in variance within the signal window.

where t_w denotes the time at index w, relative to the current window, while T_w denotes the total time of the window being analysed. $R_w(t_w, 1)$ denotes the time history of the extracted hit with the range given by the time indices inside the parenthesis. An illustration of the AIC function indicating the minimum, where the onset is defined is shown in Figure 8.10.

8.5 Measurement and analysis channels

Recall that four AE channels were collected, as shown in Figure 8.3. While AE is well known for being able to accurately locate the spatial location of a defect based on time-of-flight differences across sensors, the objective here is to investigate the detectability of a realistic defect at different gearbox locations. The sensor placed in the IR presents the easiest of cases, as this is relatively close to the source. Any sound source sent from the inside of the raceway to the OC, will be significantly attenuated, as the sound waves will have to travel through two layers of steel casing, steel rollers and oil. Once the sound wave does make it to the OC, it then travels around it, as only the bottom of the inner raceway is in contact with its casing, leaving a large air gap on the top. The author's working assumption is thus that out of the three sensors placed in the OC, the bottom one is closest to the source, followed by the right and then the top position.

Although detection is easier from measurements in the IR positions, this is not a practical measurement location in a wind turbine; it would require machining of a hole that fits the AE sensor in either the shaft or the bearing raceway itself. The OC measurement locations on the other hand are relatively practical to install, but the sound is attenuated by the time it gets there. There is a particular interest thus to investigate whether detection is possible from the OC locations. For this reason, the features from AE hits of different channels are treated separately, and attention is paid to the OC top.

Sensor failures

Even the laboratory environment can be harsh for the data acquisition equipment, in particular the AE sensors and power amplifiers, which are located close to, on, and/or inside the rig. During these trials, rig temperatures were observed on a range between 10° to $75°C$. The temperature can cycle close to this range almost daily when trials are underway, which can have a harmful effect on the amplifiers as well as the cyanoacrylate bond between the sensors and the rig. The rig is also subject to low frequency vibration and impact loads (from rolling elements), which although not quantified here, contributed numerous times to internal breakage of cables, amplifiers and de-bonding of sensors. This resulted in two sensor failures. The Inner Raceway (IR) location, intended as a baseline, was exposed to oil flow and precession of the raceway. The sensor failed intermittently, producing inconsistent readings throughout the test campaign. Also, the power amplifier for the Outer Casing (OC) bottom location suffered mechanical failure of the amplification switch setting during the data collection for the healthy bearing. The effect of this was a lower amplification applied to this channel, which also resulted in a different analogue filter being applied to the electrical signal. For this reason, it is not possible to establish a baseline model for AE data on this channel, and to perform inference on further AE data with a different amplifier setting. Due to this, the analysis presented here focuses around the OC right and OC top sensor locations, with an emphasis on OC top, given that it is the most difficult location to sense from, given that the attenuation (from the damage source) is highest.

8.6 Operational and environmental effects

Temperature and applied load are the two principal factors where operational variability makes novelty detection challenging. The effect of applied load on total AE hit-count and hit energy is illustrated in Figure 8.12. Bearing speed also affects the AE response, but in a more predictable way; it has been observed to increase the number of hits in an almost linear fashion. This is illustrated in Figure 8.11, for results grouped by ten-second tests; the effect of speed is clear if one looks at trends across similar temperatures.

Temperature plays a role principally by changing oil viscosity, which in turn affects the overall friction levels in the entire gearbox assembly. Oil viscosity is inversely proportional to temperature; high temperatures mean reduced viscosities which tends to drive asperity contact within gearbox components up. This has an enormous effect on the background, baseline AE activity. Higher friction can exacerbate the effect of misalignment and lead to a periodically changing noise floor, which can be dealt with effectively using an adaptive threshold for AE hit extraction (as discussed above). Lower viscosity also means higher contact between bearing components. This impact source can have a similar “sound” to that of a fatigue crack under stress, albeit with different features.

The presentation of the results will omit the effects of bearing RPM and focus instead on the effects of load and temperature. This is well justified given that a modern wind turbine gearbox tends to spend most of its time at a near constant RPM, for optimum transmission of power into the generator and into the grid; blade pitch is normally adjusted to keep the gearbox at the correct speed under varying wind speed. Low RPM regimes will thus only tend to be observed during start-up and shut-down of the turbine.

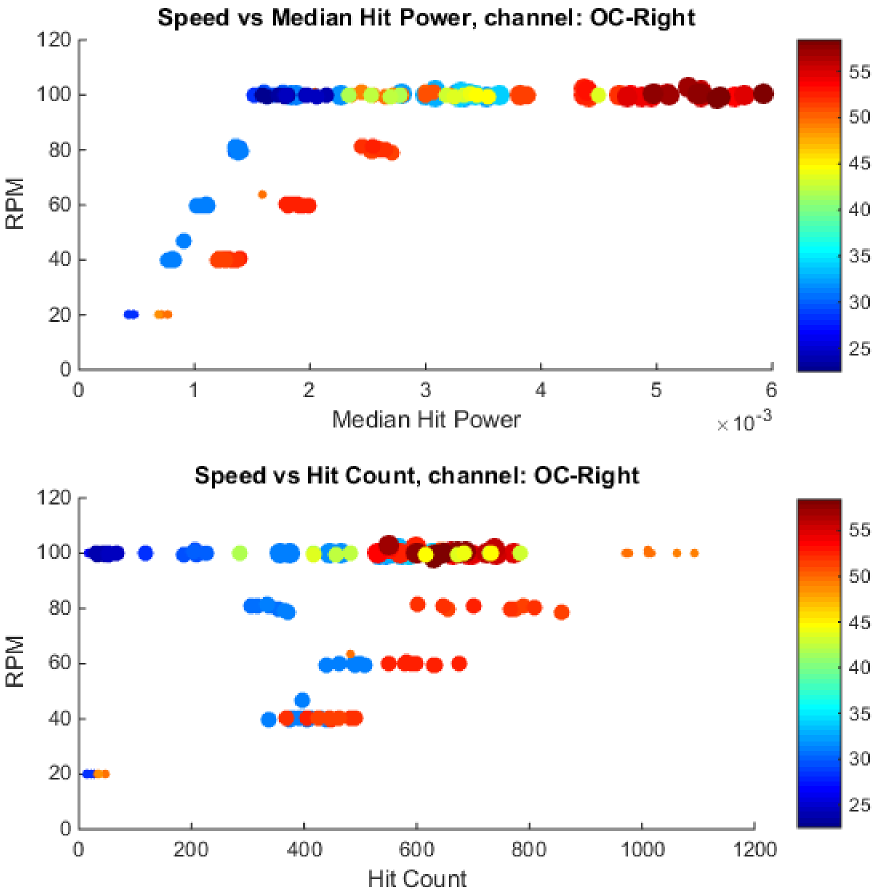

Figure 8.11: Effect of bearing speed on AE hit count and median hit power, where each point represents a ten-second recording. The colour band indicates oil temperature in $^{\circ}C$.

8.7 Damage detection with a Gaussian mixture model

So far, this chapter has described the experiment, the data collection and feature extraction process, as well as some of the data acquisition issues. This section now discusses the use of Bayesian network inference for novelty detection, based on these extracted AE features. From here, the problem could be approached in a number of different ways; two different procedures are examined here.

First, novelty detection is explored using averaged features from the hits of ten-second trials. While this simplification could throw away certain information, it

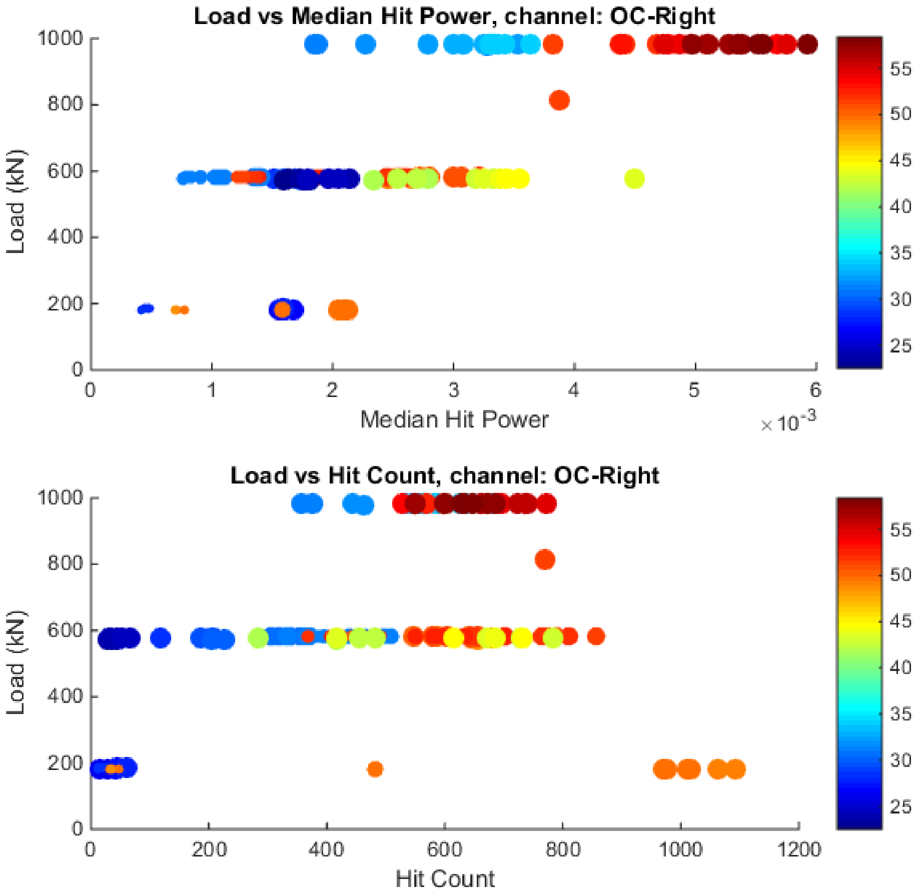

Figure 8.12: Effect of bearing load on AE hit count and median hit power, where each point represents a 10 second recording. The colour band indicates oil temperature in $°C$.

greatly simplifies treatment, and also provides one extra feature: the hit count for a specific period of time. This is particularly useful, as damage will not only present itself as AE hits with different features, it will also increase the number of hits generated within a time window. This is particularly true for a rotating system. In the specific case of this bearing rig, at full speed (100 RPM) the roller rolls over the damaged zone at a rate of 15 Hz (assuming all of the roller passings over damaged zone generate an AE burst). So, one may expect at least 15 more AE hits per second if damage is present, however stress waves could be generated not only when the roller is in direct contact over the damaged zone but also when it is getting close, as the stress field will smoothly increase and decrease as the roller passes the loaded zone of the bearing. If the load is high, the stress field around the contact zone may be enough to generate more stress waves and thus, one could expect more AE hits

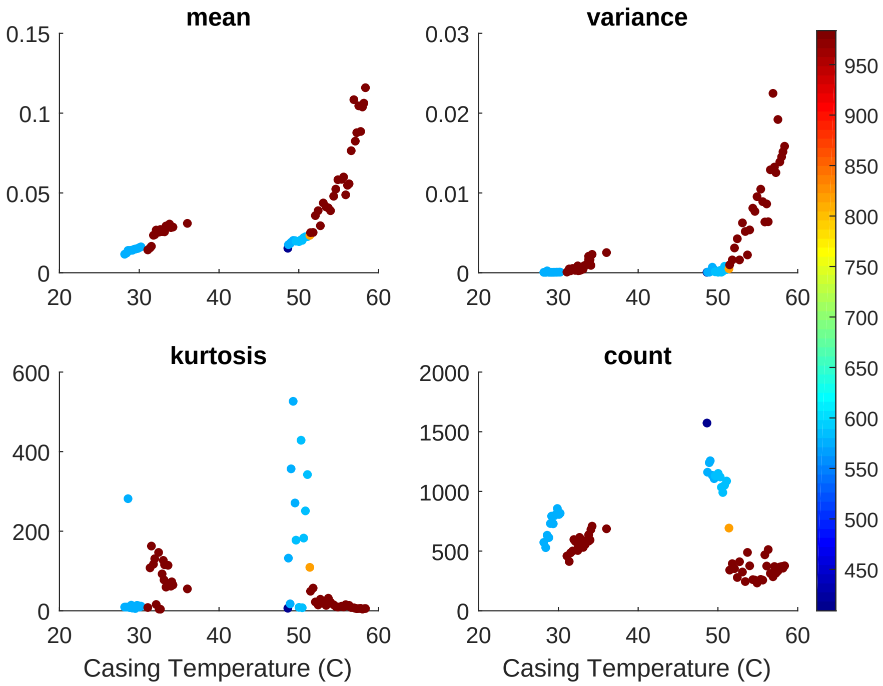

Figure 8.13: Summary statistics for AE hits on undamaged condition data, plotted against casing temperature. Each point represents a ten-second test, and its colour represents the load applied to the rig in kN.

in a time window than that given by the ball-pass frequency alone.

The second approach is to consider the features of individual AE hits, and fit a Bayesian network to this data. This retains the information given by the relative differences in the features of each hit, but throws away the information given by the rate of generation of AE hits. This is also a different problem from detecting damage from a loss of stiffness. Under a damaged state, a rotating system of this kind will still generate the AE response from regular operation (friction, roller impact, etc), with the *possible* addition of more AE hits generated from the damage process, if it hasn't been attenuated at the sensor location. Hence, one should not only look for a permanent increase in the novelty index (negative log-likelihood), but for an increase in the rate of generation of low likelihood AE hits. The expectation (or other moments) of the likelihood function for AE hits over a reasonable time window needs to be assessed, in order to capture this problem well. Ultimately, the performance of the SHM algorithm, in terms of false positive and negative rates can only be assessed by testing the hypothesis that a given recording belongs to a damaged or undamaged state. One cannot (at this stage) assess whether an individual hit was generated from a damaging process or from a different source, but it is possible to assert that a T-second recording was performed in a damaged, or undamaged condition. This point will be returned to towards the end of this section; the first approach will be discussed next. Note that in both cases applied load and temperature were not used as features inside the GMM, in order to model the more realistic scenario where these are unobserved. In general it would be relatively difficult to measure applied load in an operational wind turbine, whilst oil temperature is more accessible and monitored regularly.

8.7.1 Modelling summary statistics of AE hits

To illustrate the first approach, consider the mean, variance and kurtosis of the AE hit power, as well as the AE hit count (for a ten-second test) from an undamaged condition, given in Figure 8.13. This figure illustrates the variability with respect to temperature and load. Recall that in this analysis, a high (1000kN) and a low (600kN) load are both considered, and this leads to a wide range of temperatures. Higher loads lead to higher temperatures due to the increase in friction in the bearing. A Gaussian mixture model was fitted to this data set, consisting of the summary statistics of each trial. A “trial” here refers to a ten-second test, so that each mean, variance and hit count represent a ten-second time window. In the undamaged condition, a total of 118 trials were recorded. This was split into a training and testing set, sampling uniformly to select 70 trials at random to assemble a training set.

The covariance matrices for the Gaussian mixture were restricted to be diagonal, otherwise there is a tendency for one Gaussian to fit the majority of the data, assigning a tiny mixture component to the rest. This could be the result of the relatively low quantity of training data, which is a result of using summary statistics for time windows. If the required segmentation of data is high, due to a high number of operating regimes, this could result in not enough data points to define a cluster. Constraining the covariances restricts the number of free parameters that define each cluster, which is more appropriate in this case since given the low number of overall training points, some clusters may be assigned only one or two points.

Using Bayesian Information Criterion (BIC), it was determined that five cluster components provided a good fit for a GMM with a constrained covariance, so this was fitted to the training data. The resulting negative log-likelihood for the training, testing and damage data set are shown in Figure 8.14, where it can be seen that this approach is, in general, very successful. The model is able to correctly characterise the process across two different operating loads, and a wide temperature range, evidenced by the fact that the majority of the testing points remain within the 99$^{\text{th}}$ percentile threshold of the training set. The most notable result here is the mostly correct identification of subsurface damage, at a far away measurement location. The reason that this is notable is that this damage scenario is not only subtle but realistic. Most studies limit the use of seeded defects to surface damage, which, using AE, would be relatively easy to detect, given that the roller will impact the defect, thus generating high energy stress waves. The subsurface damage introduced here is not visible at the surface, so any “novel” AE bursts generated from this subsurface-damaged bearing are likely to come from stress waves generated by the high stress concentrations around the yield crack. An alternative explanation for these exceedances of the novelty index for the subsurface damaged bearing is that while applying the compressive load (seeding the defect), the geometry of the bearing was changed, which could lead to higher friction and thus higher AE levels. Although the bearing geometry has not actually been measured after subsurface damage was introduced, a large geometrical change is unlikely for two reasons. The stress applied was enough to yield the material only locally, under the contact point, and so a global plastic deformation is unlikely. A large change in geometry is bound to generate a significantly higher change in the AE response, much higher at least than the 50 μm surface etch; the change in AE response observed in Figure 8.14 is consistent as a higher change is seen with increasing levels of damage.

While this approach, based on features derived from summary statistics, is generally successful, the simplification of diagonality of the covariance matrices for the clusters had to be made, due to the low number of training data points available. Regardless of this, the features derived were highly informative, as they contain the rate of generation of AE bursts, and the first and second order moments (mean and variance) for three key features: hit power, rise time and duration. The second approach, based on fitting mixture models to individual features, instead of their summary statistics is discussed next.

8.7.2 Modelling individual AE hit data

A training set was assembled, by first sampling uniformly over the undamaged tests. Recall that each trial here corresponds to a ten-second recording of AE. As before, a

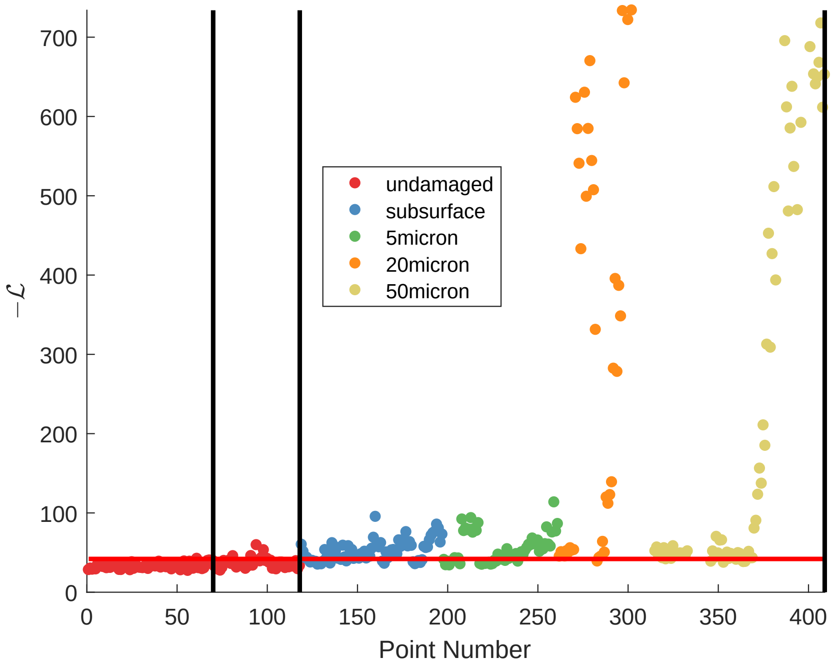

Figure 8.14: Negative log-likelihood for training, testing and damage sets for a four-component Gaussian mixture fit to the summary statistics of the bearing AE hit data. Note that the damage is arranged in increasing levels of severity. Threshold shown denotes 99th percentile of training set.

total of 70 trials were used to assemble a training set, except that now the individual hit features from each of these 70 tests were used for the training set. This approach was favored instead of sampling from the complete pool of individual hits, which would result in some trials containing a mixture of training and testing sets. In total, this yielded 22135 training points, 15404 testing points, and 116397 points from damaged conditions.

A Gaussian mixture model was now trained using the individual hit data as features (instead of their summary statistics). The advantage this provides is that the information retained in the relationship between the features of individual hits; however this ignores the rate of generation of AE hits, which has a relationship with residual fatigue life. Figure 8.15a shows the negative log-likelihood of the AE feature data, evaluated under a Gaussian mixture model with eight components. Note that a much higher number of components was required in this case to get a good fit of the model to the data. The AE features used for this analysis were maximum amplitude, power, duration, and risetime. It is clear from observation of this negative log-likelihood that the the effect of damage is to increase the number of abnormal AE hits, while retaining the same number of “normal” hits. It would be beneficial to see this information in terms of the average likelihood of a specific time window. Because data was collected in ten-second windows, it is convenient to look at descriptive statistics of the ten-second windows. Figure 8.15b shows the mean of the negative log-likelihood for each ten-second window, together with a 99^{th} percentile threshold on the training data. The likelihood means, grouped by test, are slightly easier to interpret than the raw likelihoods. In particular, at the higher damage levels, the excursions from the threshold are very clear. The lower damage levels, however, remain largely within the threshold, even though a significant amount of the points in the subsurface and 5 μm classes exceed the threshold of the raw negative log-likelihood (Figure 8.15b). There is a much bigger portion of “normal” hits in these conditions, and because the damage is subtle, the low novelty index is not enough to push the test-mean upwards. It is much more informative to turn to the exceedance rate of the negative log-likelihood threshold; this is shown in Figure 8.15c. The exceedance rate is defined as the number of crossings of the threshold defined over the negative log-likelihood on the training set. It highlights only those hits that are potentially abnormal. This exceedance rate could be potentially treated as a novelty index. The threshold shown in Figure 8.15c is defined as the 99^{th} percentile of exeedances on the training set. In summary, this could be described as putting a threshold on the number of threshold crossings, and it may seem like an odd approach, but it enables the use of a likelihood function from a Bayesian network to highlight those tests (time windows), with more abnormal hits than the nominal condition.

The features based on summary statistics are simpler and, in this case, perform better at discriminating damage. Fewer steps were required to assemble the feature vectors, and a less complex model could be used. When using the individual features, an eight-component Gaussian mixture had to be used, with full covariance matrices. Only a four-component Gaussian mixture with diagonal covariance matrices sufficed to provide a good fit to the data density. The usefulness of obtaining summary statistics, of either the features, or the likelihoods, is that it is possible to observe the false positive and negatives. It would not be possible to do this on individual AE hits, as the true origin of each hit is unknown, whereas at least in the case of this experimental investigation, the true state of a ten-second window is given by is a known state.

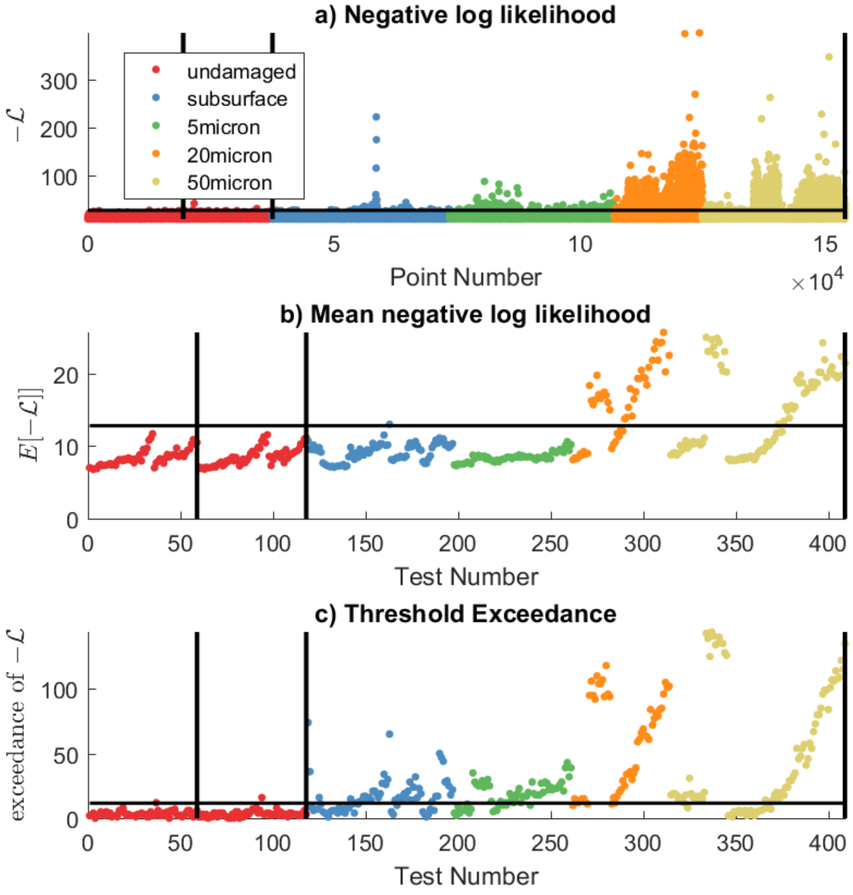

Figure 8.15: Negative log-likelihood for OC top AE sensor features, derived from a Gaussian mixture fitted to individual AE hits. Vertical lines divide training, testing and damage sets. The horizontal lines indicate 99^{th} percentile thresholds on the training set. The three rows show: a) $-\log \mathcal{L}$ for each individual hit, b) the mean $-\log \mathcal{L}$ for each ten-second trial and c) the exceedance rate of a 99^{th} percentile threshold of $-\log \mathcal{L}$, for each ten-second trial.

8.7.3 Taking periodicity into account

One of the key aspects that the overall approach presented here has not taken into account is the periodicity of AE hits. In the sections above, it was argued that a hit-based analysis could be more suitable than a perdiodicity-based one. The principal reason for this being that even in a normal operating condition, strong periodicities may be already present at the relevant frequencies of the rotating component. This effect has been observed by the author on both the rig described here as well as on data from an operational turbine. The current industry state of the art in periodicity-based damage detection is to use the Fourier transform of a signal envelope [136, 137]. This extracts the frequency information of the modulated signal. There is nothing preventing the use of this envelope of frequency spectra as features for a Bayesian network. This would yield a similar analysis to that of Chapter 5, where a mixture of PCA or Factor Analysis may be more suitable due to the high dimensionality of the resulting feature vector. It is, however, outside the scope of this work to make a comparison of both features. For example, the work by [132] makes use of a feature with a similar concept, where the cepstrum of one of the levels of a wavelet packet decomposition is used as a feature. The feature is only characterised by a Mahalanobis distance measure; a single multivariate Gaussian. This type of feature extraction approach is starkly different from the hit extraction methodology. One of the key advantages of using hits as features is the level of data compression. In this type of application, only about 1% to 3% of the data points are informative, so retaining only those points makes sense. The problem is that the way it has been done here has thrown away all the information regarding the timing of these hits. This timing may be relevant to the diagnostic, so it is worth discussing how one would be able to extract this information, while using hit data as the main features.

One possibility would be to use a tachometer to record the angular position of the bearing as it rotates, so that it can be recorded for every hit. A reasonable amount of data pre-processing would need to be carried out; the entire raw waveform does not need to be resampled into the revolution domain, only the angular position of each extracted hit needs to be recorded. This does require a tachometer with high resolution. The tachometer mounted on the rig used in this rig ticked every revolution, where the period of the revolution is around 1 second. In this second, the speed of the bearing can vary significantly, which would have yielded inaccurate readings for angular position.

Another possibility is to use a dynamic Bayesian network. A Hidden Markov Model (HMM) is particularly well suited to this problem; it could be thought of as a mixture model, where the mixture responsibility is also dependent on its previous value. The likelihood of a data point would thus also consider the value (generating cluster) of previous hits. This is, however, left as motivation for future work here.

8.7.4 Uncertainty over damage detection performance

Before concluding this chapter, it is useful to examine some of the elements that may have introduced uncertainty into the results presented here. The question that will be discussed here is mainly that of whether the successful failure detection presented in the previous sections is a result of correct performance of the algorithm or inadvertently introduced by the testing methodology. This point deserves some discussion, because changing the bearing condition involves a partial disassembly of the rig. Therefore, a disassembly and reassembly of the rig was required in the following cases:

- In order to swap the undamaged inner raceway with the damaged one.
- In order to rotate the damaged inner raceway to align the loaded path to the different seeded defects.
- To swap the first damaged bearing with surface defects, with the second damaged bearing containing the sub-surface seeded defect.

While most of the rig is left untouched in this process, there are several steps where a change could be potentially introduced, listed as follows:

- Compressive loading lugs have to be removed and put back on; this could potentially change the loading alignment.
- The driving belt has to be removed during disassembly, and the same tension during re-assembly is not guaranteed.
- The oil filter is replaced, and any debris in the oil system is removed.

- For the surface damage cases, the inner bearing raceway is rotated so the damaged condition lies on the loaded zone. In the subsurface damage and undamaged cases, the raceway is replaced. Thus, a total of three raceways were used, and geometrical variability could exists within them.

From these steps, the potential shift in load alignment is likely to cause the most change in-between tests as the rig is very sensitive to the positioning of the loading lugs, which is set manually during each re-assembly of the rig.

During operation, the controller of the hydraulic pump that applies the compressive load can only keep a constant load to within 15kN. This is reasonable, as it is only 1.5% of the maximum load applied. However, while examining the relationship between temperature and load, during the post-processing of the results, it transpired that most of the damaged condition tests were carried out at a slightly higher load than the undamaged condition tests. This is evident in Figure 8.16, where the casing temperature is shown against the total test load. It is not the same story for the temperature range; the undamaged set covers the same temperature range as the damage sets. This leads to the question of whether the higher novelty indexes observed in these discussions are a result of damage, or the only thing that is being detected is a higher load. Note that Figure 8.16 splits the data into high and medium load ranges. Recall that the novelty detector has been trained using data from both of these conditions. It is unlikely that this small load variation, of 15kN, would cause such a step change in the novelty index, especially given that it was trained in two load conditions 400kN apart. So, in conclusion, even though the tests carried out left a small amount of uncertainty about whether the effect detected by the signal processing and Gaussian mixture relates to damage or to a small change in load, it is unlikely that such a small load change would have caused the change in likelihood observed here. The fact that the novelty index increases with the level of damage progression could be considered as evidence that the main source of change is the damage process, and not the variations in loading.

Another key question, largely left unanswered is whether the disassembly process introduced any changes in the overall condition of the test rig. The answer is most likely yes, but any such changes clearly did not create a large change in the AE response. It also helped that the undamaged condition, used to train the model consist of two different instances of bearing assembly.

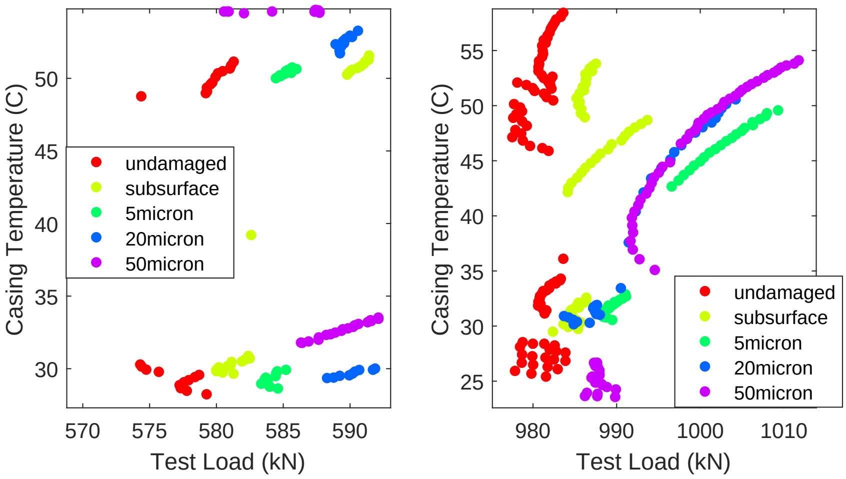

Figure 8.16: Relationship between rig casing temperature and measured compressive load for a) low load, and b) high load. Data is grouped by bearing condition

8.8 Chapter Conclusions

This chapter presented an investigation into the use of Bayesian networks, to model features derived from AE data on a wind turbine bearing rig. Various realistic damage scenarios were investigated, with the lowest level of damage being a subsurface crack, which was carefully seeded into the bearing by yielding it in compression with a steel roller. A Gaussian mixture model, a simple instance of a Bayesian network, has been shown to highlight all of the damage levels, through the use of the likelihood function as a novelty index.

The feature extraction process described in this chapter is perhaps lengthier than for other application domains, but this is seen as necessary here because the useful information pertaining to the damage process is very sparse (in the time domain). The particular route to feature extraction taken here is not a requirement for the algorithm to work. Other features could also potentially work well, an investigation and comparison of different features is not the point of this work. The objective of this chapter was to demonstrate the use of Bayesian network inference for a relevant SHM problem, with the use of the likelihood function as a novelty index. In this case the Bayesian network interpretation, and the use of the negative log-likelihood function as a damage index proved to be a successful approach for this difficult problem.

Beyond the Bayesian network application, the results presented here are significant.

Detection of subsurface damage in a realistic environment is a new result, not found in the literature. This is a very relevant problem in industry, as once damage propagates to the surface of a bearing, progression is typically very fast through spalling. Any effort to prevent or quantify this at the subsurface state is desired and has a direct impact on bearing life assessment In this study, detection is also performed under a changing operational environment involving changes in load, which affect the stress concentration and overall gearbox friction, as well as temperature which has an inversely proportional relationship with oil viscosity, thus affecting friction too. Characterising the AE response under this environment is not only challenging but novel. The method developed for seeding the subsurface damage is also novel, and is described in more detail in [37].

Chapter 9

Conclusions and Future Work

9.1 Thesis summary

This thesis has presented an approach to the statistical pattern recognition step of SHM based on Bayesian networks, with an emphasis on the use of likelihood functions derived from them as damage indices.

The particular class of Bayesian networks of interest in this work is that of generative models: PCA, Factor Analysis (FA), Kalman filtering, Hidden Markov Models (HMMs), and Gaussian mixture models all belong to this class. While these individual models have seen applications in SHM before [2], their Bayesian network interpretation has not previously been emphasised in this context. The presentation of Bayesian networks here has strongly focused around the novelty detection problem. One of the useful features of this approach is that, in principle, the extension of these generative models to mixtures is straightforward. It has been shown that this has powerful implications for SHM applications, namely that modelling damage sensitive features with a mixture model can account for certain Environmental and Operational Variabilities (EOVs).

The Bayesian network generative models have been been split into two classes:

1. Static data models: those that do not model temporal relationships between variables, and are thus useful for modelling the density of damage sensitive features that do not vary with time. While in SHM, most problems are of a

dynamic nature, the dynamics are often taken into account by a pre-processing or feature extraction stage that will embed temporal relationships into different dimensions of a feature vector. Such models include modal parameter extraction, Fourier and wavelet analyses.

2. Dynamic data models: those that do account for temporal relationships between variables, and are thus suitable for modelling damage sensitive features. These are particularly useful when modelling raw data where the dynamics are implicit in the temporal dependencies.

The thesis starts with an overview os SHM and its current challenges in Chapter 1, where Environmental and Operational Variabilities (EOVs) are highlighted as the current challenge in SHM research. The overall approach to the use of Bayesian networks and likelihood functions for damage detection was outlined and motivated. Chapter 2 provided background on SHM, and discussed aspects such as the data types used and damage sensitive features; it also provided a brief introduction to some machine learning aspects of importance to SHM such as regression and density estimation. This led to Chapter 3, which introduces the approach of using likelihood functions for damage detection. One of the key aspects discussed in this chapter is the different thresholding strategies that can be adopted; in particular, three are discussed: percentiles based on empirical cumulative distribution functions, Monte Carlo sampling, and Extreme Value Statistics (EVS). These thresholding methodologies were applied to a simple toy problem, for which the negative log likelihood of a Gaussian mixture model was used. One of the key points from this discussion, is that provided enough training data is available from an undamaged state to define the tails of the CDF, a percentile provides a simple and effective way of establishing a threshold. If less training data is available, a parametric model based on EVS is probably more suitable.

The use of both static and dynamic Bayesian networks, in the context of structural damage detection has been discussed in Chapters 4 and 5 respectively. Chapter 4 included an introduction to Bayesian networks in general; it introduced the idea of representing conditional probabilities through graphical models and motivated the use of these as a tool for deriving likelihood functions for models. This chapter also introduced static Bayesian networks, with a focus on the derivation of likelihoods and also with an emphasis on the mixture extensions. Two specific cases were drawn: Gaussian and PCA mixtures. As an illustration, the chapter briefly put these models into context through two damage detection examples: the Z24 bridge data set, and a nonlinear 3-DOF mass-spring-damper system.

Chapter 5 extended this discussion to the use of dynamic Bayesian networks, again with a focus on the use of likelihood functions and their use in novelty detection. Special attention was paid to autoregressive models and state space models such as the Kalman filter; this is due to a current lack of application of model likelihoods for such models in a damage detection context. One of the main points, applicable to damage detection on a linear dynamical system with no EOVs, was that the form of the negative log likelihood function of a Kalman filter has the same form as that of the squared Mahalanobis distance on its residuals, with one key difference: the Kalman filter updates the covariance of the residuals recursively. The use of mixture extensions to state space models was introduced as a means of dealing with EOVs; this effectively yields Switching Linear Dynamical Systems (SLDS). It was highlighted that the problem is not as straightforward as with the static data models, as likelihood inference in an SLDS requires evaluation of K^T Gaussian components, where T is the number of time points in the time series. This is a clear computational impediment. To address this, the factorial SLDS of Ghahramani [117] is suggested as it scales well with the number of time points. The use of switching autoregressive models based on Hidden Markov Models is also discussed, perhaps as a slight digression. However, this illustrates the mixture/switching interpretation of a model commonly used in SHM, but where this mixtures approach is rarely used. The chapter ends by illustrating the factorial SLDS, and its likelihood function, to a linear 3-DOF numerical mass-spring-damper system with variations in global stiffness in its undamaged condition.

The identification of structural damage was demonstrated in three systems of engineering interest in Chapters 6, 7 and 8:

1. A numerical 3-DOF mass-spring-damper system, with stiffness variability and variations in loading with a structural nonlinearity.
2. The Z-24 bridge data set, consisting of four natural frequencies which vary with temperature
3. Damage detection on a wind turbine bearing using Acoustic Emissions (AE).

The numerical simulations of a mass-spring-damper system presented an opportunity to demonstrate some of the capability and advantages of the Bayesian network view of damage detection. One of the key aspects of this chapter is that it makes use of both static and dynamic Bayesian networks discussed previously in Chapters 4 and 5. The use of both models is an important point here as it highlights the fact that a general approach is being presented, which involves the modelling of SHM with Bayesian networks. The particular model depends on the damage sensitive feature being chosen. The comparison between PCA and Kalman filtering in Chapter 5 was thus important as PCA works well on a feature that encodes the dynamics of the process such as Fourier coefficients, whereas a Kalman filter removes that requirement on its features since it could be seen as an extension of PCA that models temporal relationships. The negative log-likelihoods for both models are shown to work well on a simple linear system with damage. Their mixture extensions are then demonstrated on two scenarios representing environmental and operational variability: varying stiffness, and varying loading under structural nonlinearity. It is shown how the negative log likeklihood of the mixture PCA and factorial SLDS are well suited for dealing with this kind of EOV, and successful at detecting damage. It should be stressed that, not only is the application of these individual models new to SHM, this general view has not been applied in this context.

Chapter 7 demonstrated the application of the likelihood of a simpler Bayesian network on a well-known problem in the SHM literature: the Z24 bridge. The Bayesian network used in this case was a Gaussian mixture model, which is well suited to the problem, given that the Z-24 dataset consists of four natural frequencies, so it is relatively low dimensional. The Gaussian mixture model, with the negative log-likelihood as a damage index, was shown to work well for the problem of damage detection in this data set.

Finally, the third application, presented in Chapter 8, applies the Bayesian network, likelihood inference framework to AE data from an experimental investigation of damage detection on a wind turbine bearing. This investigation focused on detecting small defects in large planetary bearings. The smallest type of defect studied was subsurface damage, which presents the lowest level that one may possibly still consider “damage”. Detecting defects at this stage is of great interest in wind turbine gearbox monitoring, as their propagation to the surface accelerates failure of the bearing, and gearbox, significantly. The Chapter presents some of the necessary signal processing developed by the author, required in order to derive damage sensitive features from the raw AE data that one can then treat with a Bayesian network. This included the detection and extraction of hits, an adaptive threshold methodology, and the computation of features from AE hits. Because the AE hits yield a low dimensional damage sensitive feature vector, a Gaussian mixture is a suitable modelling choice, so it was demonstrated on two scenarios:

1. Features consisting of summary statistics of AE hits collected on ten-second trials.
2. Features of individual AE hits were passed to the Gaussian mixture.

One of the key difference between the first and second approach is that in the first, one of the summary statistics is the total hit count, which is expected to rise in the presence of damage. Because the latter approach does not encode this information, the exceedance rate of a negative log-likelihood threshold is also used. The conclusion of this Chapter is not only that Bayesian networks and the log likelihood function are a viable technique for dealing with this type of subtle damage detection problem, but also the detection of subsurface damage on a bearing of this scale, from a practical measurement location (outside the bearing casing) is a novel result. The merit of this work is not only the successful detection of damage, but the ability to carry this out under changing loading conditions, and temperature both of which affect the dynamic response of the system in the frequency range of interest of AE.

Overall, this thesis has presented an approach to dealing with SHM features derived from probability computations based on Bayesian networks. It has been shown how the systematic use of a likelihood function on a Bayesian network appropriate for the type of damage sensitive feature can detect damage under a range of operational and environmental changes.

9.2 Future Work

This thesis presents only a small subset of the possibilities that arise when treating SHM problems as Bayesian networks. Here, the treatment was focused on the use of likelihood functions as novelty measures as the author felt that the links established across models in the machine learning literature had not been applied consistently to SHM, and general engineering problems. The major limitation of all the models presented in this work is the linear-Gaussian assumption. In some instances, it was shown that locally linear models can be used when the complexity of the data warrants it. In the Bayesian network viewpoint this was provided as mixtures of Gaussians, PCA, factor analysis and Kalman filter models. However, this approach is rather pragmatic. In instances, where multiple operating regimes exist, this approach is arguably well suited. This was the case in the stiffness variability examples of Chapter 6, as well as the Z24 and the wind turbine AE monitoring case studies of Chapters 7 and 8. However, in instances where the data complexity arises from a fundamental system nonlinearity, the locally linear approach to SHM is simply just an approximation, and it is not principled in any way. Such was the case of the novelty detection examples on the nonlinear 3-DOF system of Chapter 6. In this case, even though the models performed well at detecting a change in the system (the dynamic network having better performance than the static one), the approach would not be successful if the models were asked, for instance, to predict at previously unseen excitation levels. This is because the mixture models do not actually capture the physics of the process, they simply model data density. Capturing the right physics would require a nonlinear model, and this complicates matters significantly. A Kalman filter, for instance, is simply the result of Gaussian assumptions when considering the propagation of uncertainty in a linear dynamical system. Simple linearisations of the Kalman filter that deal with system nonlinearities exist, such as the Extended Kalman Filter (EKF) and the Unscented Kalman Filter (UKF). However, these two could both be considered as ad-hoc approximations to the propagation of uncertainty. The author's interest lies in investigating methods that involve a principled approach to the uncertainty propagation in dynamic Bayesian networks for nonlinear systems. This includes the various flavours of particle filters as well as Gaussian Process-based state space models.

Furthermore, an area of interest to the author is to tackle the problem of prognosis (of mechanical systems) through a combination of data-based and model-based approaches. Prognosis is an interesting problem that has not been paid as much attention in the SHM research community as the identification and localisation problems have. The author feels that there is a strong research gap in this area.

Appendices

Appendix A

Appendix A: Kalman filter

The Kalman algorithm involves two steps, a time update and a measurement update. In the first, the Gaussian mean and variance of the state vector is propagated forward using the physical model to generate a state prediction:

$$\mathbf{x}_t^{t-1} = \mathbf{A}\mathbf{x}_{t-1}^{t-1} \tag{A.1}$$

$$\mathbf{V}_t^{t-1} = \mathbf{A}\mathbf{V}_{t-1}^{t-1}\mathbf{A}' + \mathbf{Q} \tag{A.2}$$

where $\mathbf{x}_t$ and $\mathbf{V}_t$ are the state mean and covariance respectively.

The subscripts denote the time index while the superscripts in this case indicate up to what time index the value includes information from; it denotes whether the value is a prediction, a filtered or a smoothed estimate. The time update equations are effectively a prediction filter, and are a simply just making use of the identity of a Gaussian random variable undergoing a linear transformation. Note that at every time-update step, the uncertainty grows by at least $\mathbf{Q}$. The measurement-update takes in new measurements and compute the filtered state vector, $\mathbf{x}_t^t$ in the light of the measurements gathered at time t. They effectively use Bayes' rule to shrink the uncertainty of the state, given the measurements, via the Kalman gain (matrix) $\mathbf{K}$:

$$\mathbf{K}_t = \mathbf{V}_t^{t-1}\mathbf{C}'(\mathbf{C}\mathbf{V}_t^{t-1} + \mathbf{R})^{-1} \tag{A.3}$$

the gain effectively represents the confidence in the measurement, a small gain denoting high confidence and vice-versa. It is used to weight the observed measurements at t using

$$\mathbf{x}_t^t = \mathbf{x}_t^{t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{C}\mathbf{x}_t^{t-1}) \tag{A.4}$$

$$\mathbf{V}_t^t = \mathbf{V}_t^{t-1} - \mathbf{K}_t\mathbf{C}\mathbf{V}_t^{t-1} \tag{A.5}$$

in order to generate the filtered estimates of the state vector, which since it is a Gaussian distribution, is entirely defined by its mean $\mathbf{x}_t^t$ and covariance $\mathbf{V}_t^t$. The next two sections will show how one may use a Kalman filter in the two contexts discussed so far: as a parameter learning scheme, or to perform prediction and filtering with the aim of computing data probability.

Bibliography

[1] S. Roweis and Z. Ghahramani, “A unifying review of linear gaussian models.,” *Neural computation*, vol. 11, no. 2, pp. 305–345, 1999.

[2] K. Worden and C. R. Farrar, *Structural health monitoring: a machine learning perspective.* John Wiley & Sons.

[3] A. Rytter, *Vibration Based Inspection of Civil Engineering Structures.* Phd thesis, Aalborg University, 1993.

[4] K. Worden, C. R. Farrar, G. Manson, and G. Park, “The fundamental axioms of structural health monitoring,” *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 463, pp. 1639–1664, 2007.

[5] C. Grappasonni, N. Ameri, G. Coppotelli, D. J. Ewins, a. Colombo, E. Bianchi, and V. Barraco, “Dynamic identification of helicopter structures using operational modal analysis methods in presence of harmonic loading,” *Proceedings of the International Conference on Noise and Vibration Engineering ISMA 2012*, pp. 2017–2038, 2012.

[6] L. Hermans and H. Van Der Auweraer, “Modal testing and analysis of structures under operational conditions: industrial applications,” *Mechanical Systems and Signal Processing*, vol. 13, no. 2, pp. 193–216, 1998.

[7] M. Abdelghani, M. Goursat, and T. Biolchini, “On-line modal monitoring of aircraft structures under unknown excitation,” *Mechanical Systems and Signal Processing*, vol. 13, no. 6, pp. 839–853, 1999.

[8] E. J. Cross, *On structural health monitoring in changing environmental and operational conditions.* Phd thesis, The University of Sheffield, 2012.

[9] Y. Xia, B. Chen, S. Weng, Y. Q. Ni, and Y. L. Xu, “Temperature effect on vibration properties of civil structures: A literature review and case studies,” *Journal of Civil Structural Health Monitoring*, vol. 2, no. 1, pp. 29–46, 2012.

[10] P. C. Chang, A. Flatau, and S. C. Liu, “Review paper: health monitoring of civil infrastructure,” *Structural Health Monitoring*, vol. 2, no. 3, pp. 257–267, 2003.

[11] J. Kullaa, “Structural health monitoring under nonlinear environmental or operational influences,” *Shock and Vibration*, vol. 2014, 2014.

[12] C. Kramer, C. de Smet, and G. de Roeck, “Z24 bridge damage detection tests,” *Proceedings of the International Modal Analysis Conference - IMAC*, vol. 1, pp. 1023–1029, 1999.

[13] N. Dervilis, K. Worden, and E. Cross, “On robust regression analysis as a means of exploring environmental and operational conditions for SHM data,” *Journal of Sound and Vibration*, vol. 347, pp. 279–296, 2015.

[14] Z. Wang and K. C. G. Ong, “Autoregressive coefficients based Hotelling ' s T2 control chart for structural health monitoring,” vol. 86, pp. 1918–1935, 2008.

[15] Z. Wang and K. C. G. Ong, “Structural damage detection using autoregressive-model-incorporating multivariate exponentially weighted moving average control chart,” *Engineering Structures*, vol. 31, no. 5, pp. 1265–1275, 2009.

[16] J. Maeck, B. Peeters, and G. D. Roeck, “Damage identification on the Z24 bridge using vibration monitoring,” *Smart Materials and Structures*, vol. 10, pp. 512–517, 2001.

[17] A. M. Yan, P. De Boe, and J.-C. Golinval, “Structural damage diagnosis by Kalman model based on stochastic subspace identification,” *Structural Health Monitoring*, vol. 3, pp. 103–119, jun 2004.

[18] A. Santos, R. Santos, E. Figueiredo, C. Sales, and W. A. Costa, “A structural damage detection technique based on agglomerative clustering applied to the Z-24 Bridge,” in *Proceedings of the European Workshop in Structural Health Monitoring*, no. July 2016, p. 24, 2016.

[19] EWEA, “The European offshore wind industry key 2015 trends and statistics,” Tech. Rep. January, 2015.

[20] N. Dervilis, M. Choi, S. G. Taylor, R. J. Barthorpe, G. Park, C. R. Farrar, and K. Worden, "On damage diagnosis for a wind turbine blade using pattern recognition," *Journal of Sound and Vibration*, vol. 333, no. 6, pp. 1833–1850, 2014.

[21] G. Manson, "Identifying damage sensitive, environment insensitive features for damage detection," in *3rd International Conference on Identification in Engineering Systems*, (Swansea, UK), 2002.

[22] E. J. Cross, K. Worden, and Q. Chen, "Cointegration: a novel approach for the removal of environmental trends in structural health monitoring data," 2011.

[23] S. Johansen, *Likelihood-based inference in cointegrated vector autoregressive models.* Oxford University Press, 1995.

[24] R. F. Engle and C. W. J. Granger, "Co-integration and error correction: representation, estimation, and testing," *Econometrica*, vol. 55, no. 2, pp. 251–276, 1987.

[25] H. Shi, K. Worden, and E. J. Cross, "A nonlinear cointegration approach with applications to structural health monitoring," *Journal of Physics: Conference Series*, vol. 744, p. 012025, 2016.

[26] C. M. Bishop, *Pattern recognition and machine learning.* Springer-Verlag New York, 2006.

[27] E. Figueiredo, L. Radu, K. Worden, and C. R. Farrar, "A Bayesian approach based on a Markov-chain Monte Carlo method for damage detection under unknown sources of variability," *Engineering Structures*, vol. 80, pp. 1–10, 2014.

[28] D. a. Clifton, S. Hugueny, L. Tarassenko, and R. Drive, "Novelty detection with multivariate extreme value theory , part I : a numerical approach To multimodal estimation," *Proceedings of IEEE Machine Learning in Signal Processing*, no. x, pp. 1–6, 2009.

[29] K. Worden, G. Manson, and N. Fieller, "Damage detection using outlier analysis," *Journal of Sound and Vibration*, vol. 229, no. 3, pp. 647–667, 2009.

[30] A. Santos, E. Figueiredo, M. F. M. Silva, C. S. Sales, and J. C. W. A. Costa, "Machine learning algorithms for damage detection: kernel-based approaches," *Journal of Sound and Vibration*, vol. 363, pp. 584–599, 2016.

[31] E. Figueiredo, G. Park, C. R. Farrar, K. Worden, and J. Figueiras, "Machine learning algorithms for damage detection under operational and environmental variability," *Structural Health Monitoring*, vol. 10, no. 6, pp. 559–572, 2011.

[32] J. N. Yang, S. Lin, H. Huang, and L. Zhou, "An adaptive extended Kalman filter for structural damage identification," *Structural Control and Health Monitoring*, vol. 13, no. 4, pp. 849–867, 2006.

[33] Y. Chen and M. Q. Feng, "Structural health monitoring by recursive Bayesian filtering," *Journal of Engineering Mechanics*, vol. 135, no. April, pp. 231–242, 2009.

[34] A. Deraemaeker, A. Preumont, and J. Kullaa, "Modeling and removal of environmental effects for vibration based SHM using spatial filtering and factor analysis," in *Proceedings of the International Modal Analysis Conference - IMAC*, 2006.

[35] S. Doebling, C. R. Farrar, B. Prime, M, and D. Shevitz, "Damage identification and health monitoring of structural and mechanical systems from shanges in their vibration characteristics: A literature review," 1996.

[36] J. Kaiser, *Untersuchungen über das auftreten von geräuschen beim zugversuch.* PhD thesis, Technical University of Munich (TUM), 1950.

[37] R. Fuentes, T. P. Howard, M. B. Marshall, E. J. Cross, and R. S. Dwyer-Joyce, "Observations on Acoustic emissions from a line contact compressed into the plastic region," *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, vol. 230, no. 11, pp. 1371–1376, 2016.

[38] J. Baram and M. Rosen, "Fatigue life prediction by distribution analysis of acoustic emission signals," *Materials Science and Engineering*, vol. 41, no. 1, pp. 25–30, 1979.

[39] J. J. Hensman, *Novel techniques for acoustic emission monitoring of fatigue fractures in landing gear.* PhD thesis, University of Sheffield, 2009.

[40] M. J. Eaton, R. Pullin, J. J. Hensman, K. M. Holford, K. Worden, and S. L. Evans, “Principal component analysis of acoustic emission signals from landing gear components: An aid to fatigue fracture detection,” *Strain*, vol. 47, no. SUPPL. 1, 2011.

[41] K. M. Holford, R. Pullin, S. L. Evans, M. J. Eaton, J. Hensman, and K. Worden, “Acoustic emission for monitoring aircraft structures,” *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, vol. 223, pp. 525–532, aug 2009.

[42] J. Hensman, K. Worden, M. Eaton, R. Pullin, K. Holford, and S. Evans, “Spatial scanning for anomaly detection in acoustic emission testing of an aerospace structure,” *Mechanical Systems and Signal Processing*, vol. 25, no. 7, pp. 2462–2474, 2011.

[43] M. G. Baxter, R. Pullin, K. M. Holford, and S. L. Evans, “Delta T source location for acoustic emission,” *Mechanical Systems and Signal Processing*, vol. 21, no. 3, pp. 1512–1520, 2007.

[44] J. Hensman, R. Mills, S. G. Pierce, K. Worden, and M. Eaton, “Locating acoustic emission sources in complex structures using Gaussian processes,” *Mechanical Systems and Signal Processing*, vol. 24, no. 1, pp. 211–223, 2010.

[45] K. K. Nair, A. S. Kiremidjian, and K. H. Law, “Time series-based damage detection and localization algorithm with application to the ASCE benchmark structure,” *Journal of Sound and Vibration*, vol. 291, pp. 349–368, 2006.

[46] K. Krishnan Nair and A. S. Kiremidjian, “Time series based structural damage detection algorithm using Gaussian mixtures modeling,” *Journal of Dynamic Systems, Measurement, and Control*, vol. 129, no. 3, p. 285, 2007.

[47] R. Yao and S. N. Pakzad, “Autoregressive statistical pattern recognition algorithms for damage detection in civil structures,” *Mechanical Systems and Signal Processing*, vol. 31, pp. 355–368, 2012.

[48] Y. Lu and F. Gao, “A novel time-domain auto-regressive model for structural damage diagnosis,” *Journal of Sound and Vibration*, vol. 283, no. 3-5, pp. 1031–1049, 2005.

[49] R. H. Shumway and D. S. Stoffer, *Time series analysis and its applications.* Springer New York, 2011.

[50] S. Chen, S. a. Billings, C. F. N. Cowan, and P. M. Grant, “Practical identification of NARMAX models using radial basis functions,” *International Journal of Control*, vol. 52, no. 769892610, pp. 1327–1350, 1990.

[51] S. Billings, *Nonlinear system identification NARMAX methods in the time, frequency, and spatio-temporal domains*, vol. 21. 2013.

[52] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning.* Cambridge, Massachusetts: The MIT Press, 2006.

[53] K. Worden, G. Manson, and E. J. Cross, “On gaussian process NARX models and their higher-order frequency response functions,” in *Springer Proceedings in Mathematics and Statistics*, vol. 97, pp. 315–335, 2014.

[54] L. Bornn, C. R. Farrar, G. Park, and K. Farinholt, “Structural health monitoring with autoregressive support vector machines,” 2009.

[55] R. Fuentes, E. J. Cross, A. Halfpenny, K. Worden, and R. J. Barthorpe, “Aircraft parametric structural load monitoring using Gaussian process regression,” in *7th European Workshop on Structural Health Monitoring*, 2014.

[56] J. Mottershead and M. Friswell, “Model updating in structural dynamics: a survey,” 1993.

[57] A. Pandey, M. Biswas, and M. Samman, “Damage detection from changes in curvature mode shapes,” *Journal of Sound and Vibration*, vol. 145, no. 2, pp. 321–332, 1991.

[58] G. Hearn and R. B. Testa, “Modal analysis for damage detection in structures,” *Journal of Structural Engineering*, vol. 117, no. 10, pp. 3042–3063, 1991.

[59] B. Jaishi and W.-X. Ren, “Damage detection by finite element model updating using modal flexibility residual,” *Journal of Sound and Vibration*, vol. 290, no. 1-2, pp. 369–387, 2006.

[60] B. Peeters and G. De Roeck, “Reference-based stochastic subspace identification for output-only modal analysis,” *Mechanical Systems and Signal Processing*, vol. 13, no. 6, pp. 855–878, 1999.

[61] M. W. Vanik, J. L. Beck, and S. K. Au, “Bayesian probabilistic approach to structural health monitoring,” *Journal of Engineering Mechanics*, vol. 126, no. 7, pp. 738–745, 2000.

[62] H.-J. Lee and S. Roberts, “On-line novelty detection using the Kalman filter and extreme value theory,” *2008 19th International Conference on Pattern Recognition*, no. June, 2008.

[63] P. Hayton, S. Utete, D. King, S. King, P. Anuzis, and L. Tarassenko, “Static and dynamic novelty detection methods for jet engine health monitoring,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1851, pp. 493–514, 2007.

[64] B. Peeters and G. De Roeck, “One-year monitoring of the Z 24-Bridge: environmental effects versus damage events,” *Earthquake engineering & structural dynamics*, vol. 30, no. January 2000, pp. 149–171, 2001.

[65] F. Magalhães, A. Cunha, and E. Caetano, “Vibration based structural health monitoring of an arch bridge: From automated OMA to damage detection,” *Mechanical Systems and Signal Processing*, vol. 28, pp. 212–228, 2012.

[66] Y. Lu and F. Gao, “A novel time-domain auto-regressive model for structural damage diagnosis,” *Journal of Sound and Vibration*, vol. 283, pp. 1031–1049, 2005.

[67] C. Bishop, *Neural networks for pattern recognition*, vol. 1995. 1995.

[68] C. M. Bishop, “Bayesian neural networks,” *Journal of the Brazilian Computer Society*, vol. 4, jul 1997.

[69] B. a. Warner and R. M. Neal, “Bayesian learning for neural networks,” *Journal of the American Statistical Association*, vol. 92, p. 791, 1997.

[70] S. F. Masri, M. Nakamura, a. G. Chassiakos, and T. K. Caughey, “Neural network approach to detection of changes in structural parameters,” *Journal of Engineering Mechanics*, vol. 122, no. 4, pp. 350–360, 1996.

[71] V. Lopes, G. Park, H. H. Cudney, and D. J. Inman, “Impedance-based structural health monitoring with artificial neural networks,” *Journal of Intelligent Material Systems and Structures*, vol. 11, no. March 2000, pp. 206–214, 2000.

[72] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[73] B. Silverman, "Density estimation for statistics and data analysis," *Chapman and Hall*, vol. 37, no. 1, pp. 1–22, 1986.

[74] S. Gill, B. Stephen, and S. Galloway, "Wind turbine condition assessment through power curve copula modeling," *IEEE Transactions on Sustainable Energy*, vol. 3, no. 1, pp. 94–101, 2012.

[75] C. M. Bishop, "Bayesian PCA," *Advances in neural information processing systems*, vol. 11, pp. 382–388, 1999.

[76] H. Hoffmann, "Kernel PCA for novelty detection," *Pattern Recognition*, vol. 40, no. 3, pp. 863–874, 2007.

[77] V. H. Nguyen and J.-C. Golinval, "Fault detection based on kernel Principal Component Analysis," *Engineering Structures*, vol. 32, no. 11, pp. 3683–3691, 2010.

[78] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[79] M. Titsias, "Variational learning of inducing variables in sparse Gaussian processes," *Aistats*, vol. 5, pp. 567–574, 2009.

[80] R. Fuentes, T. Howard, E. J. Cross, R. Harald-Hestmo, T. Huntley, B. Marshall, Mathew, and R. Dwyer-Joyce, "Detecting damage in wind turbine bearings using acoustic emissions and Gaussian process latent variable models," in *Proceedings of the 10th International Worklshop in Structural Health Monitoring*, (Stanford University, Palo Alto, CA), 2015.

[81] N. Dervilis, *A machine learning approach to Structural Health Monitoring with a view towards wind turbines.* Phd thesis, The University of Sheffield, 2013.

[82] H. Nickisch and C. E. Rasmussen, "Gaussian mixture modeling with Gaussian process latent variable models," *Arxiv preprint arXiv:1006.3640*, 2010.

[83] D. Barber, *Bayesian reasoning and machine learning.* Cambridge University Press, 2012.

[84] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, dec 1974.

[85] R. E. Kass and L. Wasserman, “A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion,” *Journal of the American Statistical Association*, vol. 90, p. 928, sep 1995.

[86] A. Dempster, N. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society Series B Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[87] J. R. Beniger, V. Barnett, and T. Lewis, “Outliers in statistical data.,” *Contemporary Sociology*, vol. 9, no. 4, p. 560, 1980.

[88] P. J. Rousseeuw and B. C. van Zomeren, “Unmasking multivariate outliers and leverage points.,” *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 633–639, 1990.

[89] S. S. Wilks, “The large-sample distribution of the likelihood ratio for testing composite hypotheses,” *Annals of Mathematical Statistics*, vol. 9, 1938.

[90] D. A. Clifton, L. Clifton, S. Hugueny, D. Wong, and L. Tarassenko, “An extreme function theory for novelty detection,” *IEEE Journal on Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 28–37, 2013.

[91] R. A. Fisher and L. H. C. Tippett, “Limiting forms of the frequency distribution of the largest or smallest member of a sample,” *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, no. 02, pp. 180–190, 1928.

[92] W. Weibull, *A statistical theory of the strength of materials.* 1939.

[93] G. S. Mudholkar, D. K. Srivastava, and G. D. Kollia, “A generalization of the Weibull distribution with application to the analysis of survival data,” *Journal of the American Statistical Association*, vol. 91, p. 1575, dec 1996.

[94] P. Embrechts and T. Mikosch, *Modelling extremal events for insurance and finance.* Berlin: Springer-Verlag, 4th editio ed., 2008.

[95] E. Castillo, *Extreme value theory in engineering.* San Diego, CA: Science, Ademic Press Series in Statistical Modelling and Decision, 1988.

[96] S. J. Roberts, "Novelty detection using extreme value statistics," *IEE Proceedings - Vision, Image, and Signal Processing*, vol. 146, no. 3, p. 124, 1999.

[97] K. Worden, D. W. Allen, H. Sohn, D. W. Stinemates, and C. R. Farrar, "Extreme value statistics for damage detection in mechanical structures," tech. rep., Los Alamos National Laboratory Report LA-13903-MS, 2002.

[98] A. F. Jenkinson, "The frequency distribution of the annual maximum (or minimum) values of meteorological elements," *Quarterly Journal of the Royal Meteorological Society*, vol. 81, no. 348, pp. 158–171, 1955.

[99] J. R. Hosking, J. R. Wallis, and E. F. Wood, "Estimation of the generalized extreme-value distribution by the method of probability-weighted moments," *Technometrics*, vol. 27, no. 3, pp. 251–261, 1985.

[100] H. W. Park and H. Sohn, "Parameter estimation of the generalized extreme value distribution for structural health monitoring," *Probabilistic Engineering Mechanics*, vol. 21, no. 4, pp. 366–376, 2006.

[101] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of palusibble inference*. San Francisco: Morgan Kaufmann, 1998.

[102] S. Roweis, "EM Algorithms for PCA and SPCA," *Computing*, vol. 10, no. 13, pp. 626–632, 1997.

[103] S. Mohamed, Z. Ghahramani, and K. Heller, "Bayesian exponential family PCA," *Advances in neural information processing systems*, pp. 1089–1096, 2009.

[104] M. N. Nounou, B. R. Bakshi, P. K. Goel, and X. Shen, "Bayesian principal component analysis," *Journal of Chemometrics*, vol. 16, no. 11, pp. 576–595, 2002.

[105] O. Cappe, K. Mengersen, M. Titterington, and C. P. Robert, "Online Expectation-Maximisation," in *Mixtures: Estimation and Applications*, pp. 1–53, Wiley, 2011.

[106] J. M. Marin, K. Mengersen, and C. P. Robert, "Bayesian modelling and inference on mixtures of distributions," *Handbook of Statistics*, vol. 25, pp. 459–507, 2005.

[107] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, “Markov Chain Monte Carlo in practice,” 1996.

[108] M. Tipping and C. Bishop, “Mixtures of probabilistic principal component analyzers.,” *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999.

[109] L. A. McGee and S. F. Schmidt, “Discovery of the Kalman filter as a practical tool for aerospace and industry,” no. November, p. 21, 1985.

[110] Z. Ghahramani and G. H. Hinton, “Parameter estimation for linear dynamical systems, Technical Report CRG-TR-96-2,” tech. rep., Department of Computer Science, University of Toronto, Toronto, 1996.

[111] K. P. Murphy, *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, UC Berkeley, 2002.

[112] M. Verhaegen and P. Van Dooren, “Numerical aspects of different Kalman filter implementations,” *IEEE Transactions on Automatic Control*, vol. AC-31, no. 10, 1986.

[113] P. Caravani, M. L. Watson, and W. T. Thomson, “Recursive least-squares time domain identification of structural parameters,” *Journal of applied mechanics*, vol. 44, no. March 1977, p. 135, 1977.

[114] C. Paleologu, J. Benesty, and S. Ciochia, “A robust variable forgetting factor recursive least-squares algorithm for system identification,” *IEEE Signal Processing Letters*, vol. 15, no. 3, pp. 597–600, 2008.

[115] F. P. Kopsaftopoulos and S. D. Fassois, “Vibration based health monitoring for a lightweight truss structure : experimental assessment of several statistical time series methods,” in *Proceedings of the International Conference on Noise and Vibration Engineering ISMA*, vol. 24, pp. 867–892, 2010.

[116] A. Doucet, N. D. Freitas, K. Murphy, and S. Russell, “Rao-blackwellised particle filtering for dynamic Bayesian networks,” *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pp. 176–183, 2000.

[117] Z. Ghahramani and G. E. Hinton, “Variational learning for switching state-space models.,” *Neural computation*, vol. 12, no. 4, pp. 831–864, 2000.

[118] K. Murphy, “Switching kalman filters,” Tech. Rep. August, 1998.

[119] R. H. Shumway and D. S. Stoffer, “Dynamic linear models with switching,” *Journal of the American Statistical Association*, vol. 86, no. 415, pp. 763–769, 1991.

[120] S. W. Shaw and C. Pierre, “Normal modes for nonlinear vibratory systems,” *Journal of Sound and Vibration*, vol. 164, no. 1, pp. 85–124, 1993.

[121] Z. Ghahramani and G. E. Hinton, “The EM algorithm for mixtures of factor analyzers,” *Compute*, pp. 1–8, 1997.

[122] A.-M. Yan, G. Kerschen, P. De Boe, and J.-C. Golinval, “Structural damage diagnosis under varying environmental conditionsPart I: A linear analysis,” *Mechanical Systems and Signal Processing*, vol. 19, no. 4, pp. 847–864, 2005.

[123] I. Antoniadou, N. Dervilis, E. Papatheou, A. E. Maguire, and K. Worden, “Aspects of structural health and condition monitoring of offshore wind turbines,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 373, pp. 20140075–20140075, jan 2015.

[124] P. J. Rousseeuw and K. V. Driessen, “A fast algorithm for the minimum covariance determinant estimator,” *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[125] E. Figueiredo and E. J. Cross, “Linear approaches to modeling nonlinearities in long-term monitoring of bridges,” *Journal of Civil Structural Health Monitoring*, vol. 3, no. 3, pp. 187–194, 2013.

[126] A. M. Yan, G. Kerschen, P. De Boe, and J. C. Golinval, “Structural damage diagnosis under varying environmental conditions - Part II: Local PCA for non-linear cases,” *Mechanical Systems and Signal Processing*, vol. 19, no. 4, pp. 865–880, 2005.

[127] R. M. Neal, “Markov Chain sampling methods for Dirichlet process mixture models,” *Journal of Computational and Graphical Statistics*, vol. 9, pp. 249–265, jun 2000.

[128] D. M. Blei and M. I. Jordan, “Variational inference for Dirichlet process mixtures,” *Bayesian Analysis*, vol. 1, pp. 121–143, mar 2006.

[129] T. Howard, *Development of a novel bearing concept for improved wind turbine gearbox reliability.* PhD thesis, The Univerisity of Sheffield, 2015.

[130] R. B. Randall, *Vibration-based Condition Monitoring.* Chichester, UK: John Wiley & Sons, Ltd, jan 2011.

[131] J. Shiroishi, Y. Li, S. Liang, T. Kurfess, and S. Danyluk, "Bearing condition diagnostics via vibration and Acoustic Emission measurements," *Mechanical Systems and Signal Processing*, vol. 11, pp. 693–705, sep 1997.

[132] J. R. Naumann, *Acoustic emission monitoring of wind turbine bearings.* PhD thesis, The University of Sheffield, 2015.

[133] O. Reynolds, "On the Theory of Lubrication and Its Application to Mr. Beauchamp Tower's Experiments, Including an Experimental Determination of the Viscosity of Olive Oil," *Proceedings of the Royal Society, London*, vol. 40, pp. 191–203, 1886.

[134] D. P. Hess and A. Soom, "Friction at a Lubricated Line Contact Operating at Oscillating Sliding Velocities," *Journal of Tribology*, vol. 112, no. 1, p. 147, 1990.

[135] J. H. Kurz, C. U. Grosse, and H. W. Reinhardt, "Strategies for reliable automatic onset time picking of acoustic emissions and of ultrasound signals in concrete," *Ultrasonics*, vol. 43, no. 7, pp. 538–546, 2005.

[136] R. B. Randall, "State of the art in monitoring rotating machinery Part 1," *Journal of Sound and Vibration*, vol. 38, no. 3, pp. 14–20, 2004.

[137] R. B. Randall, "State of the art in monitoring rotating machinery Part 2," *Journal of Sound and Vibration*, vol. 38, no. 5, pp. 10–16, 2004.