THE UNIVERSITY OF SHEFFIELD

PhD Mechanical Engineering

Advances in Data-Based Modelling for Structural Health Monitoring Systems

by

Christopher A. Lindley

July 2024

K. Worden, N. Dervilis, R.S. Dwyer-Joyce

Thesis submitted to the Department of Mechanical Engineering, University of Sheffield for the degree of Doctor of Philosophy

Acknowledgements

First and foremost, I would like to give my profound thanks to Prof. Keith Worden, who always was – and continues to be – an incredibly supportive, generous, and encouraging supervisor and friend. I would also like to give my thanks to my secondary supervisors, Prof. Nikolaos Dervilis and Prof. Rob Dwyer-Joyce, whose expertise and advice have been invaluable towards my growth as a researcher. I would like to extend my thanks to Dr. Tim Rogers, for getting out of his way to provide insights that helped concretise many of the ideas presented in this work.

I would also like to acknowledge the work of Robin Mills and Mat Hall, whose expertise made possible the realisation of some of the experiments presented in this thesis. This work was also made possible with the support of Esben, Jesper, Claus, and the rest of the team from Siemens Gamesa. My thanks extend to them and the technicians who helped me during my stay in Brande.

I feel most fortunate to have worked in the Dynamic Research Group. During my time here I have met and worked alongside an amazing group of people; I thank you all for your kindness, inspiration, and shared laughs. I would particularly like to thank Matty, for being a great friend and an excellent colleague to collaborate with. I would also like to thank Marcus, Max, Brandon, Aidan, Tina, Art, Aris, George, Chandy, Dan P., Dan B., and Matt T. In addition, I sincerely thank the following people from Tribology for their support and (ongoing) friendship during my first conference: Scott, Elisha, Sam, Will, Gary, and Kieran.

I am especially lucky to have shared an office with Tristan and Jack P. during my Ph.D. I would especially like to thank Tristan for his company during the writing of this thesis.

Outside the University of Sheffield, I would like to thank people whose friendship I treasure very much, and who have made my life in the UK all the more enjoyable. My thanks go to Ricky, Rafa, Neil, Aldo, Omar, Paige, Ben, Ham, Hannah, Jack B., Blair, Emma, Leah, Dalex, Julian, Sam, Alice, Gina, and Gemma. In addition, my thanks go to "La Pandilla"; your continuous support from afar has always been greatly appreciated. I would also like to give my thanks to Liv, Michi, and Sem, for their recent encouragement.

I am deeply grateful to my family, for providing me with the opportunities to pursue my career abroad and their endless love and encouragement. My thanks go to: my parents, Dan and Erika; my siblings, Paul and Anna; my nonnas, Ximena and Arlette; my tío, Andrés; my tía, Cynthia; and my prima, Jose. Special thanks go to my Nonno René,

who probably without realising, showed me the value of investing care and diligence into one's work.

Finally, my everlasting gratitude goes to Sarah. You've kept me focused and helped me push through many overwhelming challenges along the way. I can't imagine having done this without your love and support. My thanks also go to your wonderful family: Bernd, Edith, and Anna.

It is certainly impossible to thank everyone who has supported me through the course of this Ph.D in a few paragraphs. Whether it was by sharing some advice, words of encouragement, or an interesting conversation over coffee, I thank you.

Abstract

Probability and statistical applications can be found spread across various scientific and engineering disciplines. In the field of *Structural Health Monitoring* (SHM), promising advances have been made possible with the development of statistical models. Challenges that were once too complex to solve analytically can now be addressed with the assistance of intelligent monitoring systems, which are, fundamentally, driven by statistical pattern recognition and machine learning algorithms.

Over the years, numerous data-driven approaches have emerged, collectively aiming to make SHM a standard practice. This thesis explores the use of novel statistical models to facilitate the implementation of intelligent health-monitoring systems; namely, by focussing on nonparametric Bayesian modelling and, to a lesser extent, autoencoders, to address a few prevalent challenges currently encountered in SHM.

Within this framework, the problem of model selection becomes central, determining how well a system is represented in a statistical sense. The main body of this thesis thus delves into this issue, alongside the rationale for adopting models that are both nonparametric and Bayesian in engineering applications.

A series of case studies are presented, each highlighting unique challenges in SHM. These case studies are approached with models based on either *Gaussian Processes* (GPs) or *Dirichlet Processes* (DPs).

Initially, GPs are employed to enhance localisation techniques in SHM. These methodologies demonstrate their capabilities to detect abnormal operations in a journal bearing using ultrasonic measurements. Additionally, they are used to simplify the process of localising damage sources in composite structures using *Acoustic Emission* (AE) data.

The following component of this thesis introduces an approach to identifying AE events in time-series signals. It further incorporates a DP prior into a mixture model, designed to autonomously capture changes introduced by different sources of damage. The methodology is validated using AE data collected from a fatigue test of an Airbus A320 landing gear.

While a significant portion of the SHM literature focusses on data-driven models, there is a growing interest in integrating physics into these models. Therefore, this thesis concludes with insights into the input-state-parameter estimation of journal bearings by combining GPs with the dynamics of journal response in a state-space representation.

The miscellany of methodologies presented in this thesis is but a small contribution to bridging the gap between state-of-the-art machine learning techniques and their application in SHM.

Contents

Chapter 1

Introduction

Structural Health Monitoring (SHM) is an automated monitoring practice that aims to determine whether a structure has departed from its normal operating conditions [1]. Being an ongoing research area, many statistical models have been developed in pursuit of this objective. However, certain challenges persist in real-life applications, hindering the adoption of SHM as a standard practice. This thesis offers a contribution towards addressing some of these challenges; namely, by introducing methodologies based on nonparametric Bayesian models, tailored to facilitate the practical implementation of health-monitoring systems.

1.1 A Brief Introduction to Structural Health Monitoring (SHM)

A "normal operating condition" is considered here to be a healthy operation in which no obvious form or shape of damage is known to exist. This assumption does not guarantee the structure to be free of imperfections, but it is established as a reference from which any relative deviations may be interpreted as the emergence or progression of damage. From this perspective, the most fundamental problem that SHM attempts to address is that of damage detection.

Levels of added complexity to this problem can be pursued if one wishes to learn more about the nature of the identified damage. In its most simple form, the damage detection system will simply alert the operator to deviations from normal, which may happen from factors that are unrelated to damage. Therefore, it may be necessary to rely on more elaborate monitoring strategies to robustly determine the existence of damage in a structure.

Rytter's hierarchy [2] is a clear illustration of the stages required to determine a complete diagnosis of the monitored structure. The hierarchy is the following:

- Level 1: Detection – Qualitative indication that damage might be present.
- Level 2: Localisation – Probable position of damage in the system.
- Level 3: Assessment – Estimate of the extent of damage.
- Level 4: Prediction – Estimate of residual life or other information on the safety of the system.

An adaption of this hierarchy has been proposed [3, 4], primarily via the incorporation of an additional level that precedes the assessment of damage. This additional level aims to determine information about the *type* of damage. Each level usually requires that all other lower-level information is available. Level Four is different from all other levels in that it requires an understanding of the physics of the damage propagation, to be accomplished. In contrast, Level One, and on some occasions, Level Two, can be achieved without any prior knowledge of how the system will behave when damaged.

SHM is primarily motivated by the objective of enhancing human safety. This motivation may be attributed to catastrophic events in the past that have highlighted the necessity of robust monitoring schemes capable of continuously assessing the health and integrity of structures in real-time.

Other motivators are targeted towards the development of an optimised maintenance schedule, which can only be possible under an SHM scheme. This advantage is attractive from a financial point-of-view since structures would only be intervened for maintenance when strictly necessary. This aspect of SHM differs from a more conservative preventive maintenance, which ensures maintenance is performed at regular intervals, but at the risk of a greater inventory of spare parts and additional costs by conducting more maintenance than necessary.

These ideals derive originally from a different field to SHM, but one that operates in the same principled manner; that is, *Condition Monitoring* (CM) [5], in which the monitoring schemes are primarily focussed on rotating machinery, such as bearings, gearboxes and rotors. The case studies examined in subsequent chapters involve both structures and rotating machinery, so it is important to emphasise that SHM and CM are distinct bodies of knowledge, each employing a different set of techniques. However, for the sake of narrative simplicity and consistency, the terminology used hereafter will mostly align with the context of SHM.

1.1.1 Data-Driven Paradigm

SHM problems are generally approached in one of two ways; that is, either by model-driven methods or data-driven techniques [6]. The former involves the development and implementation of analytical equations that can best describe the dynamics exhibited by the structure. The latter are also based on a modelling approach but are instead characterised by statistical representations of the system. Although both approaches are perfectly viable options, data-driven techniques offer advantages that naturally address the problems illustrated in Rytter's hierarchy.

The goal of statistical modelling is to determine the underlying process that explains a collection of observations. Achieving the development of a model that adequately represents the generating process can be accomplished with the implementation of machine-learning algorithms. The application of machine learning for SHM has been extensively researched, particularly since the introduction of *Artificial Neural Networks* (ANNs) for statistical pattern recognition in this context [1]. A review on deep-learning techniques for vibration-based SHM can be found in [7], in which it is stated that the number of papers published on this subject grew from 279 in 2012, to 661 in 2018. It is easy to see why one would choose a system that can autonomously learn from data. The analytical counterpart to model complex nonlinear systems may be too cumbersome to evaluate in many applications. The popularity of using data-driven techniques, such as ANNs, could be attributed to their relatively "simpler" implementation for modelling. In particular, by having a collection of data, an ANN can *learn* an input-output mapping of arbitrary complexity without the need to rely on in-depth knowledge about the physics of the system. The task is diverted to the development of the ANN – or any other machine-learning algorithm being employed – for the success of the statistical model.

However, this discussion should not be interpreted as a statement declaring the end of physics-based modelling in engineering. There are, indeed, several challenges associated with the implementation of data-driven models that require careful attention. Failure to do so can lead to an inaccurate representation of the system, leading to poor predictions. This consideration may arguably be of greater importance in SHM, since poor predictions can be perilous for human safety, and detrimental towards financial projections, in the event of unforeseen failures. Furthermore, physics-based models provide interpretability and extrapolation capabilities that data-driven models lack [8]. The scope of this work aims to address many of these considerations when dealing with engineering applications.

1.1.2 A Probabilistic Perspective

The numerous ways in which machine learning can be implemented for SHM can be categorised into two perspectives: deterministic and probabilistic. While deterministic approaches aim towards explicit outcomes, probabilistic approaches provide, perhaps, a broader understanding of the modelled system by quantifying the uncertainty in predictions. For example, having a probability distribution to represent the normal data, means that one can determine the probability of observing an anomaly, rather than deeming such observation as anomalous with absolute certainty.

Algorithms based on probability provide valuable insights that may help guide the decision-making process regarding the health state of a structure. In this framework, these algorithms not only provide the most likely outcome but also provide a measure of confidence, enabling a quantifiable approach to risk assessment for potential courses of action. A compelling case supporting probabilistic models in engineering is presented in [9], essentially highlighting that neglecting uncertainties translates into disregarding the full utilisation of the data. While deterministic models have shown promising results in SHM applications [10], the models employed in this thesis are exclusively probabilistic. This choice is driven by key issues outlined in this work, which are best addressed by statistical models. To be explicit, the models presented in the following chapters are based on Bayesian statistics, which is a branch of probability theory that departs from the *frequentist* interpretation of random events.

1.2 Bayesian Framework for SHM

The Bayesian approach allows the expression of knowledge in the form of a probability distribution [11]. This implication means having an initial model based on prior beliefs about the system. The ideals of Bayesian statistics thus rely on quantifying how well prior beliefs support the evidence.

Upon the introduction of observations (or evidence), deriving from the system, one can make use of the new incoming information to update the prior beliefs about the model. Formally, the action of updating one's beliefs is referred to as *inference*.

1.2.1 The Prior Distribution

In Bayesian statistics, prior beliefs are quantified with a probability distribution and introduced mathematically with Bayes' Theorem. The prior distribution has been the subject of much debate ever since the idea of *inverse probability* was formalised by

Laplace in the eighteenth century [12]. Mathematicians at the time viewed the prior distribution (or prior, for short) with suspicion, believing that probability should be determined by frequency and not by a subjective measure. In other words, there was no reason to trust a proposed prior, and thus, no guarantee of inferring the "correct" posterior.

Not until the mid-20th century did statisticians begin to revisit and improve upon the work of Laplace. The publication of *Theory of Probably* by Harold Jeffreys in 1939 [13], has been considered to be a turning point in favour of Bayesian methods [14]. Another notable theorem worth addressing here is that proposed by Bruno de Finetti [15], which implies the existence of a prior for any sequence of exchangeable random variables.

While De Finetti's theorem motivates the use of priors on parameters, finding the "right" prior, from a pragmatic point-of-view, remains a challenging task. Nevertheless, some progress can be made towards this goal by acknowledging the expert knowledge one may have about systems being modelled.

1.2.2 Engineering and Bayesian Reasoning

In addition to having a model that inherently quantifies uncertainty/error, there are a few other reasons why Bayesian analysis may be preferred in engineering contexts. The Bayesian approach to probability works on the premise of measuring the plausibility of a hypothesis, which is particularly useful in situations in which probabilities cannot be readily understood as frequencies. For example, it would be impossible to determine the probability of a bridge collapsing solely by performing many trials of this event. Instead, by stating that the probability of the bridge collapsing because of reaching the end of its lifespan, or from experiencing an unusual amount of external loading, the probabilities are naturally indicated by the degree of belief the engineer has about the bridge, given, of course, evidence to support such belief.

In fact, this form of belief can be quantified and incorporated into the prior distribution within the Bayesian framework. This idea has been pursued in the implementation of the *Gaussian Process* (GP) [16] for engineering applications, and an extensive survey on the subject can be found in [17]. In short, GPs are flexible Bayesian data-driven models that have often been employed in SHM; mostly, for regression-type problems. Although GPs learn from data, their structure allows for constraints to embed physical insights imposed by the system. Such constraints may be necessary to prevent some associated caveats encountered when adopting this reasoning. For example, when modelling mechanical systems with a set of equations of motion that describe the dynamics of a structure, selecting a Gaussian prior to modelling the damping coefficients might lead

to the possibility of sampling negative values, which are physically impossible. In such cases, a prior that best represents this constraint should be preferred to develop a more accurate model, or, if the prior is a GP, nonnegative constraints may be enforced [18].

A more concrete example of this issue has been made evident when modelling the power curves of wind turbines. This exercise is a prevalent one in SHM, since power curves are key performance indicators of wind turbines [19]. Having an unbounded prior implies that the underlying functions can extend beyond the range of power output that a wind turbine actually provides. In [9], for example, a *heteroscedastic* model, in which the noise is made input-dependent, is employed to quantify more realistic uncertainty bounds. A physically-meaningful representation of the uncertainty for power curve modelling is further explored in [20], where a non-Gaussian likelihood is chosen to ensure the confidence bounds are constrained within realistic ranges. Other areas of SHM have investigated incorporating physical knowledge directly via the prior distribution. This idea is put into practice in [21], where the prior is adapted to consider geometrical constraints in the structure. This adaptation aimed to improve the accuracy of a damage localisation model, ensuring that sensible outputs were obtained.

In light of this discussion, a compelling aspect arises regarding the Bayesian paradigm in engineering applications. That is, engineers typically possess some level of knowledge about the system under observation. From a Bayesian perspective, this prior knowledge or expert information can be seamlessly integrated into the statistical representation of a structure. Additionally, Bayesian models can provide elegant solutions to challenges such as sparse or unrepresentative data, as long as a suitable prior is employed. This perspective on modelling is important and is at the heart of all methods employed in this thesis.

1.3 Motivation, Contribution and Structure

Overall, Bayesian models offer a powerful approach to address various problems encountered in engineering applications. Their advantage lies in the ability to incorporate an understanding of physical phenomena intuitively via prior distributions. This thesis is primarily concerned with the application of nonparametric Bayesian models, a type of flexible statistical model capable of adapting to complex data while minimising assumptions about the system. Rather than comparing parametric and nonparametric models, the goal here is to introduce novel health-monitoring systems specifically tailored to certain problems in SHM. These systems yield three primary contributions:

- The first core contribution of this thesis focusses on developing intelligent monitoring systems for journal bearings. This development is partly driven by recent advancements in sensing technologies for fluid-film bearings, enabling real-time, accurate measurements of oil-film thickness. By adapting GPs to learn the complex relationship between a journal bearing's operational state and its fluid-film characteristics, a robust health-monitoring system is developed that can promptly detect potential deviations from normal operations. Additionally, in a separate study, GPs are employed to estimate driving forces and dynamic coefficients of a journal bearing while in operation. In this framework, these estimates are derived merely with measurements of the journal response and simple dynamical representations.

- The second core contribution addresses challenges linked to the practical use of *Acoustic Emission* (AE)-based technologies in SHM. The first part of this contribution outlines a strategy for efficiently generating an optimal training dataset for damage-localisation methods based on *Delta T* values. The second part presents a flexible model capable of identifying and categorising individual events in AE recordings, simplifying the manipulation and processing of AE data with minimal intervention.

- The final contribution is the outcome of an independent study where a *Variational Auto Encoder* (VAE) is used as an effective means of preprocessing data for SHM. Specifically, the VAE is demonstrated to project out the influence of environmental changes from the data. This achievement is made without the need to explicitly model long-term trends, offering an unsupervised alternative to address the influence of confounding effects in SHM.

A series of different case studies are presented throughout this thesis to demonstrate the efficacy of the developed health-monitoring systems. While some of these case studies have already been examined in one way or another in the literature, the methods proposed here offer some form of advantage that has not yet been explored. The models are thus validated in two ways: (1) by following standard machine learning practices for generalisation and (2) by highlighting how these models overcome previously encountered challenges in SHM. In parallel, special attention is given to the caveats associated with their implementation and how these may be addressed in practice.

A rich family of nonparametric Bayesian models exist and continues to be explored in the machine-learning community. The powerful ideas that emerge from these studies extend to a vast range of applications, and the current work is no exception. It is, perhaps, nowadays, necessary to have some knowledge of statistics and machine learning

to become a well-rounded engineer. The present wave of computational advancement, and almost limitless access to and exchange of digital information, means having a whole new set of tools at one's disposal for the development of more elaborate models that can help towards the understanding of physical phenomena experienced in this world all a bit better.

The structure of this thesis is as follows:

- **Chapter 2**: This chapter outlines fundamental concepts related to SHM and statistical models.
- **Chapter 3**: A summary of literature covering uses of nonparametric Bayesian models for SHM is provided.
- **Chapter 4**: A theoretical background of the main data-driven techniques employed in this thesis is offered here, distinguishing between relevant models for specific applications.
- **Chapter 5**: The use of GPs for damage localisation is explored in this chapter. In particular, this method is employed to predict the location of a journal-bearing shaft centre and detect anomalous behaviour *in situ*. Additionally, this technique is applied to localising damage within a composite helicopter blade using AE data. An enhanced method based on Bayesian optimisation is proposed for constructing a *Delta T* map efficiently in the latter case.
- **Chapter 6**: This chapter explores the use of a *Dirichlet Process* (DP) model to enhance AE-based monitoring techniques. A probabilistic model is introduced to identify and cluster individual AE-bursts in recorded time-series signals, enabling early damage detection in an Airbus A320 landing gear subjected to cyclic loading.
- **Chapter 7**: This chapter focusses on journal-bearing dynamics to construct a *Latent Force Model* (LFM). By simulating a series of states that the journal bearing might experience, it is demonstrated here how external loads can be recovered, and the importance of carrying out this exercise in practice is discussed. Finally, the model is employed to not only recover the applied loads but also the dynamic coefficients of the bearing.
- **Chapter 8**: This chapter examines two experimental case studies: damage identification in rolling-element bearings and damage identification in a composite plate subjected to environmental variations. A novel view on the use of *Variational Auto Enconders* (VAEs) is explored for extracting damage-sensitive features from the data.

- **Chapter 9**: The main ideas covered in the introduction are revisited and then complemented with the outcome of the work presented in the main body of the thesis. Challenges of these methods are reflected on, and various solutions are proposed here that can be pursued in the future.

Chapter 2

Background

This chapter serves as a concise summary of concepts related to SHM and statistical modelling, providing clarity and context for the work presented in subsequent chapters.

2.1 Structural Health Monitoring

2.1.1 Sensing Technologies

The data-driven approach to monitoring structures has been made possible with the advancement of sensing technologies that enable the acquisition of data necessary for conducting analysis based on dynamics. There are various types of sensing devices that can be utilised for the implementation of a monitoring system, each offering some advantage over the rest. One of the more commonly-chosen options for both SHM is the use of accelerometers to measure the dynamic responses. This approach, known in the literature as vibration-based SHM/CM, involves processing the recordings from accelerometers such that information about the health state of a structure can be determined. Accelerometers are essentially small piezoelectric devices that can be attached directly to the surface of the structure; they are typically favoured for their wide working frequency range (1-30kHz) compared to velocity sensors (1-10kHz) and displacement sensors (1-100Hz) [22].

Vibration-based SHM has thus been explored thoroughly by researchers, with early in-depth reviews found in [23] and [24]. The popularity of vibration-based techniques may be attributed to their interpretability and relatively-simple implementation. Even under a few assumptions, such as reducing the order of the system, or expecting a linear response when it may not necessarily be the case, the dynamics can be explained in terms of a mathematical framework. The transition to damage detection from this

framework is intuitive in the sense that the dynamic behaviour of a healthy structure will experience some change upon the presence of damage. Monitoring the evolution of key dynamical features, and determining the point at which their values depart from an established baseline, is fundamentally how vibration-based SHM operates.

Another recurring sensing technology employed in SHM is based on the detection of AE waves. Promising results in the field have demonstrated advantages over vibration-based techniques in the early detection and development of faults [25]. In short, an AE is the physical mechanism whereby strain energy within a solid is suddenly released and propagated in the form of an elastic wave. Although this notion of an AE was first defined by Kaiser in 1950 [26], it was only with the development of high sampling-frequency instrumentation and highly-sensitive transducers that the detection of AE waves was possible. Much like accelerometers, AE sensors capture and convert the response - in this case, of the elastic wave - into an electric signal using piezoelectric effects, but are capable of doing so in the ultrasonic frequency range (100kHz-2MHz). Fracture extension, for example, promotes the generation of these AE waves, and can thus be detected by the continuous monitoring of AE activity within the structure.

An extensive amount of research has been carried out to better understand the underlying correlation between the physics of AE and the mechanisms that produce them in mechanical systems [27–32]. For example, several attempts have been pursued to correlate AE features with fracture-mechanical parameters, such as the stress-intensity factor of structures subjected to cyclic loading [33–36]. These studies are based on Paris-Erdogan-type laws [37], to make estimates of the remaining fatigue life of structures, and demonstrate that count rates can be used to estimate the growth rate of cracks. Additionally, several promising results have been demonstrated when implementing AE-based monitoring techniques for damage detection in structures and rotating machines. Some examples of this exercise include rolling element bearings [38, 39], gears [40], journal bearings [41, 42], self-compacting concrete specimens [43], and aerospace structures [44, 45].

AE waves are typically processed to determine their source of origin and establish whether these correspond to damage. However, analysing AE signals can be difficult because of the large quantity of data produced from high sampling frequencies. AE-based SHM is a subject of particular interest in the present work and will be covered in more detail in Chapters 5 and 6.

Other common sensing equipment used for SHM include strain gauges, fibre optics, and ultrasonic technology [46].

It is important to note that sensors, regardless of the application or sensing technology employed, do not directly measure damage. Information regarding the structural health may be concealed or "encrypted" deep within the raw signals. In other words, incipient damage may not be imposing enough to cause a detectable deviation from readings of a healthy structure. Therefore, the essence of SHM is not merely on sensing technologies, and how they can be employed for high-quality measurements, but also in the processing of these measurements to determine the health state of the system. Further elaboration on this aspect is provided in the following section.

2.1.2 Organising Principle

Regardless of the learning approach adopted for a particular application, an organising principle is required in all cases [3, 6, 47]. It is generally necessary to follow a series of steps before developing the actual statistical model. A sensible process described in [3] suggests the following sequence of steps: (1) operational evaluation, (2) data acquisition, (3) feature selection and (4) statistical model development.

Before carrying out the actual monitoring, operational evaluation is a necessary step to determine what can be monitored and how the monitoring will be accomplished. The decision-making process involved in this part of the process is mainly targeted in evaluating how data can be measured, and to determine whether the measured data are suitable for identifying features that correlate to damage.

The subsequent step involves the assessment of sensors to be employed for data acquisition and their appropriate usage. The extent and quality of the data that can be acquired will depend on a few important considerations, such as the availability of financial resources, and the frequency with which data should be collected, once the monitoring system is active. The latter of these considerations imposes a limit to the amount of data that can be collected, defined primarily by the amount of available storage and how feasible it will be to manipulate. In this stage, inevitable sources of variability must also be identified and minimised where possible.

Data cleansing is the process of selecting which measurements to accept or reject. Cleansing can be reviewed in the light of feature selection and information provided during the development of the statistical model. Feature selection refers to the assessment of features in the data in order to distinguish the different health states and operating conditions of the structure. Large amounts of data are typically collected from structures, and qualitatively assessing these data for damage detection is not immediately obvious, even when comparing various signals from the same system over time. To address this problem, reduced representations of raw signals are required to eliminate redundancies

and extract information that is relevant for damage detection. Consequently, an inherent aspect of feature selection involves condensing or reducing the dimensionality of the raw signals. Data condensation is also important to prevent problems associated with the *curse of dimensionality* [48] and minimise the required amount of storage.

Deciding how to process raw measurements and which features to select for the development of a statistical model can be a challenging task, and at times, even prove to be the most complex part of the problem [6]. The means to achieve this end are vast and a great deal of algorithms have been exploited by researchers. A simple approach to feature extraction would involve, for example, calculating the statistics of the raw signal, such as the mean and the variance. In scenarios where the dataset is comprised of high-dimensional feature vectors, then approaches such as principal component analysis [48] can be used to transform the data into a more interpretable reduced space. Other features extracted from signals can be coefficients determined in a time-domain analysis, or coefficients of a Fourier transform when working in the frequency-domain. Naturally, feature extraction may also rely on working with a joint-time-frequency analysis, such as the *Short-Time Fourier Transform* (STFT) or *Wavelet Transforms* [49]. An alternative approach to feature extraction, explored in [50], involves monitoring nonlinear features as damage indicators, under the assumption that damage causes the structure to respond in a nonlinear manner. Overall, one should note that the approach adopted for feature extraction is application-dependent.

Nevertheless, whichever method one chooses to employ, the aim in any case should be to retain features that are only sensitive to damage. In other words, it is necessary to project out the influence of benign variations from the features. This exercise is commonly referred to as *data normalisation* in the SHM literature, and it imposes a significant challenge in the realisation of SHM as a standard practice. A review of the field can be found in [51].

The focus of this thesis favours the third and fourth steps of the organising principle. However, it should be noted that each step should be treated with equal importance for any monitoring system, as an accurate statistical model can only be achieved with a clean and comprehensive dataset.

2.2 Learning Theory

Machine-learning theory is concerned with algorithms aimed to autonomously learn computational relationships on the premise of data (or observations) and sets of rules. Within this field, learning theory has been designed to address three main problems [52]:

(1) regression, (2) classification and (3) density estimation. These areas of learning can be further categorised into either a *supervised* or *unsupervised* type of algorithm. Supervised problems correspond to cases in which data are accompanied by labels. Regression and classification problems are generally approached in a supervised framework, where the goal involves mapping a set of input data to their corresponding labels. Conversely, unsupervised type algorithms learn from datasets absent of labels; density estimation and novelty detection would therefore fall under this definition.

2.2.1 Supervised and Unsupervised Learning

In SHM, supervised learning has been used extensively to produce statistical models targeted to address Levels 1-3 in the damage identification hierarchy [6]. By having labels denoting the various damaged states of the structure, it is possible to train some classifiers that can then accurately predict the source of some anomalous observations. This type of patter-recognition is a powerful one since damage can not only be detected, but also identified in the process. The advantages of this approach are evident in determining the best course of action when faced with an emerging defect. However, the construction and implementation of a classifier can be challenging in practice, because labelling is generally a difficult and expensive task. It may be unreasonable, if not financially prohibitive, to manufacture several nominal replicas of, for example, a wind turbine blade, or an aircraft landing gear, only to simulate the potential damage states these structures may experience *in situ*.

This shortcoming means that valuable data for the development of representative statistical models are often scarce in real applications. Even if gathering data from damage states was feasible, it presents another challenge that pertains to a different aspect of learning, in which information can be transferred within a collection of structures of similar characteristics. This idea is pursued in *Population-Based SHM* (PBSHM) [53], which operates on the premise that nominally identical structures do not necessarily exhibit the same statistical characteristics and uncertainties, requiring careful consideration of their variations. In problems where replicas of the same structure are statistically analysed, the collection of structures, or *population*, is said to be *homogenous*. More specifically, a homogenous population will be one where the features of the individual members can be modelled by a common and unique distribution [53]. It is then intuitive to imagine that this challenge becomes even more pronounced when dealing with *heterogeneous* populations, in which structures feature unique geometries and/or assemblies, such as bridges or buildings [54, 55].

Although practical matters appear to suggest otherwise, these arguments should not imply that supervised learning algorithms and their application in SHM should be disregarded. In situations where damage conditions are simulated in controlled environments, supervised-based algorithms serve as an ideal benchmark to aspire to when dealing with real structures. Additionally, experimental setups and computational models designed specifically for supervised learning offer valuable insights in scenarios where data availability or knowledge about the system is limited.

Nevertheless, some encouragement in the face of the aforementioned challenges is provided when the problem at hand is solely of damage detection (Level 1), because only measurements from an undamaged system are ever needed, and thus, removes the need for labels. This implication means that unsupervised learning offers an enormous advantage over supervised learning in the implementation for SHM [1]. Even if limited to the lowest level of the problem-solving hierarchy, damage detection can be incredibly beneficial in deeming a structure unsafe prior to its next scheduled maintenance, overhaul or repair. In an unsupervised setting, data are gathered from a *normal condition* state, that corresponds to measurements from a structure that is known to be undamaged. After cleaning and pre-processing the data, a statistical model can be learnt from the extracted features of the normal data. Upon new observations, deviations with respect to the model can be treated as anomalies and assumed to have been derived from the presence of damage.

Numerous approaches have been investigated that employ novelty detection for the identification of damage in structures [56, 57]. One intuitive way to construct a novelty detector is by first learning the density of the normal condition data. For example, the normal condition data may be assumed to be distributed according to a Gaussian distribution. The learning process would then involve finding the statistical moments of the Gaussian, i.e. its mean and variance, which are learnt directly from the data. Having established an underlying distribution, a measure of likelihood can then be evaluated on new data. If this measure is lower than some established threshold set to, for example, three or four standard deviations away from the mean, then the corresponding observations are flagged as anomalies. Anomalous data are not explicit indications of damage, but a sudden continuous stream of them may be a strong suggestion of such. There are, of course, limitations associated with this approach. One of the main concerns is that normal-condition data are not guaranteed to be Gaussian, and more elaborate methods may be required to infer their true underlying distribution. Some proposals to address this particular challenge are discussed more in-depth in Chapter 4.

Semi-supervised learning algorithms have also been explored as an alternative pattern-recognition approach for SHM [58]. Semi-supervised algorithms offer promising solutions to overcome prevalent challenges encountered with real systems. In particular, these challenges relate to the availability of data being limited to a subset of states for training. It is unlikely to have data representing all possible damage states or varying environmental conditions that a structure may face *a priori*. In scenarios where limited labelled data are available, semi-supervised learning attempts to construct representative models by including information from both labelled and unlabelled data.

2.2.2 Probability and Bayesian Statistics

A principled foundation for the construction of machine-learning algorithms is provided by two fundamental rules of probability [59]:

- Sum rule: $p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$
- Product rule: $p(x, y) = p(x|y)p(y)$

where x and y are outcomes of the random variables X and Y, respectively, and $\mathcal{Y}$ denotes the states of the target space corresponding to the random variable Y. In loose terms, probability can be interpreted as a degree of belief about an event. The distributions $p(x)$ and $p(y)$ correspond to *marginal distributions*, $p(x, y)$ to the *joint distribution*, and $p(x|y)$ is read as the *conditional distribution* of x given y.

In Bayesian statistics, one is often interested in making inferences of some latent variables, after having observed other random variables. An underlying latent model may be conveyed to describe a possible cause of an event, and formulated in terms of a hypothesis. In a mathematical framework, one may consider the model to be governed by a set of parameters $\mathbf{w}$, and establish a prior belief of these parameters in terms of a probability distribution $p(\mathbf{w})$, referred to as the *prior distribution*. Upon the presence of observations $\mathcal{D}$, provided by the system being modelled, some relationship $p(\mathcal{D}|\mathbf{w})$ can be established to quantify how likely it is to have observed D, given the (latent) parameters $\mathbf{w}$. The measure of interest is the updated belief after having gained knowledge from the data, i.e. the *posterior distribution* $p(\mathbf{w}|\mathcal{D})$. This inversion in the relationship between $\mathcal{D}$ and $\mathbf{w}$ is encapsulated in Bayes' Theorem,

$$\underbrace{p(\mathbf{w}|\mathcal{D})}_{posterior} = \frac{\overbrace{p(\mathcal{D}|\mathbf{w})}^{likelihood}\overbrace{p(\mathbf{w})}^{prior}}{\underbrace{p(\mathcal{D})}_{evidence}} \tag{2.1}$$

which can be derived by the product rule since $p(x, y) = p(y, x) \iff p(x|y)p(y) = p(y|x)p(x)$. Inference on $\mathbf{w}$ can thus be achieved by calculating the posterior distribution in (2.1).

The evidence $p(\mathcal{D})$, is the result of marginalising (or integrating out), the set of parameters $\mathbf{w}$; that is,

$$p(D) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\mathrm{d}\mathbf{w} \tag{2.2}$$

Here is where one of the main areas of difficulty manifests in the application of Bayesian methods. The integral in (2.4) can be computationally infeasible (if not prohibitive) to solve when faced with a high-dimensional set of variables. Unfortunately, this inconvenience occurs very often in practice. A way to circumvent this problem is by having a prior distribution *conjugate* to the likelihood function. A prior is said to be conjugate if the functional form of the posterior is the same as of the prior [59]. Choosing a conjugate prior is mathematically convenient, since it allows calculating the posterior algebraically and by inspection. However, in scenarios where a meaningful prior is not necessarily conjugate to the likelihood, the evaluation of $p(\mathcal{D})$ becomes increasingly more difficult. In such cases, one must resort to methods that approximate the posterior distribution, rather than attempting to directly solve equation (2.4). Depending on the approximation method used, there are associated challenges involved.

2.3 Model Selection

Thus far, it has been established that many problems encountered in SHM are those of statistical-pattern recognition. Adopting this type of modelling warrants the standard practices followed in the field of machine learning. Fundamentally, this process boils down to finding a "best-fitting" model that can most appropriately represent the system. Whether the machine-learning algorithm is inherently deterministic or probabilistic, the goal of the learning process is to find some model that can not only accurately represent a set of observations, but also one that can generalise well when making predictions on unseen data.

2.3.1 Regression and Complexity

To illustrate the model selection problem, one may first consider the case in which the aim is to fit a function to a set of data. If dealing with a regression problem, then a sensible approach would be to find some weighted combination of inputs that yield values closely approximating their corresponding targets. In other words, to develop a

function f that can accurately map a D-dimensional input vector $\mathbf{x}_n \in \mathbb{R}^D$, to a target output scalar $y_n \in \mathbb{R}$, i.e. a predictor $f : \mathbf{x} \to y$. The fit on y can be thought of as an interpolation given by f.

In a statistical sense, the predictor is a function learnt from the data, albeit conditioned on certain assumptions. For example, a linear interpolant may be assumed to fit the data, which will likely result in a poor fit if the true underlying trend is strongly nonlinear. An improved fit can be achieved by increasing the order of the function; that is, a quadratic function will likely provide a better fit than the linear trend, a cubic will be better than the quadratic, and so on. At each step, the fit improves while the function becomes increasingly complex, requiring the addition of more parameters to account for the higher-order terms in the polynomial. Continuing with the inclusion of terms, however, can lead to the point where the function is said to have become too complex, and the model begins to (over)fit noise in the data. Somewhere in this process of adding terms to the predictor, an optimal amount of complexity is attained, whereby the complexity of the data is matched.

How well a model "fits" the data is generally quantified by a *loss function* $\mathcal{L}(\mathbf{y}, \hat{\mathbf{y}})$ [60], which takes the true target values $\mathbf{y}$ and the corresponding outputs of the predictor $\hat{\mathbf{y}}$, as inputs to produce some measure of the error in the predictions. For example, the loss function could be defined by the absolute difference between the truth and predictions, or alternatively, by their mean-square-error. When dealing with a probabilistic predictor, the loss is quantified by the likelihood of the observation in relation to some probability distribution. Finding the "best" parameters will be those that result in predictions that closely approximate the true observations; or equivalently, those that minimise the loss function.

A compromise, however, is desired between a minimal loss function and model complexity. To ensure good generalisation, standard practice is followed, involving the division of the entire dataset into a *training set* and a *test set*. The training set is used to search for optimal parameters, while the test set simulates unseen data, allowing the evaluation of the average loss for both sets during the learning process. In loose terms, the model overfits the data when the average loss is small for the training set, but large for the test set.

It may be important to clarify at this point that one is now faced with two learning stages. The first is the estimation of the parameters based on the training data. The second is based on the hypothesis or assumptions about the model; for example, the number of terms in the polynomial predictor, or the choice of prior distribution in a Bayesian paradigm. Assuming that the hypothesis about the model is correct, there are a few methods that can be employed to prevent overfitting. A principled solution is

to introduce a penalty term in the loss function, referred to as the regulariser. One of the simplest forms of regulariser is given by the sum-of-squares of the elements in the parameter vector [61]. Another approach is to enforce an "early-stopping" to the learning process once the generalised error begins to increase. However, this last technique has its complications when attempting to control multi-dimensions of complexity [62].

Complexity has been implicitly defined here by an increasing addition of terms - and thus parameters - in the predictor. At first, this definition seems to make sense, but in a Bayesian framework, the measure of complexity is more elaborate than merely taking the count of parameters in the model [63]. A Bayesian approach to the model selection problem is driven by an *Occam's Razor* philosophy, which proposes the idea that a model should remain simple if higher amounts of complexity are not required to explain the data. Unless favoured explicitly by the prior, Bayesian inference embodies a natural preference towards simpler models [14]. In particular, Occam's razor is embodied by the evidence, whereby its evaluation automatically incorporates a trade-off between model fit and model complexity [64]. To illustrate this concept, Bayes' Theorem (2.1) can be rewritten to include the dependencies on a particular model $\mathcal{M}_j$. That is,

$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}_j) = \frac{p(\mathcal{D}|\mathbf{w}, \mathcal{M}_j)p(\mathbf{w}|\mathcal{M}_j)}{p(\mathcal{D}|\mathcal{M}_j)} \tag{2.3}$$

with the evidence now given by,

$$p(\mathcal{D}|\mathcal{M}_j) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_j)p(\mathbf{w}|\mathcal{M}_j)\mathrm{d}\mathbf{w} \tag{2.4}$$

The evidence here can be interpreted as the probability of generating the dataset after having randomly selected parameters from a given model class $\mathcal{M}_j$. Because the evidence is a probability distribution, simpler models are unlikely to generate the dataset. On the other hand, models that are too complex are capable of generating a richer variety of datasets besides $\mathcal{D}$, making them less likely to generate this particular dataset at random. The best model will be the proposal that gives the highest evidence for a given set of observations, which will simultaneously happen to be the one that fits the data without added complexities. Therefore, this Bayesian perspective offers a principled approach to comparing models [63].

One may consider an alternative view in which it is reasonable to believe that a given dataset was unlikely to have been generated from a relatively-simple model. In the Bayesian paradigm, if the model and prior are correct, there is no provision for changing them on the basis of how much data has been collected [65]. Whether one hundred or one thousand observations are available, the complexity of the model remains the

same. This property is contrary to models based on frequentist methods, where the assumptions about the model can be modified based on the size of the observation set.

The approach to Bayesian modelling therefore allows for models unbounded by their complexity. Models that are infinitely large and tractable do in fact exist, with the GP being a prevalent example of such a model used for regression and classification problems. A clear advantage is that there is no longer a need to evaluate the evidence for models of various complexities, which is often difficult in practice. This view on "large models", however, seems to dispute Occam's razor, since it supports the idea of implementing models that are infinitely complex. As it turns out, this interpretation is not quite right, and Occam's razor is always at work discouraging overcomplex Bayesian models, regardless of the stand on the question of model complexity. In [64], it is comprehensively demonstrated that, for large models, Occam's razor manifests not in terms of the model's dimensionality, but in terms of the complexity of the functions under the priors implied by the (hyper)parameters.

The discussion covered here hopes to demonstrate that Bayes' Theorem naturally sets the stage for infinitely large, yet tractable, models. A model defined by many parameters is referred to as a *nonparametric* model. These models are highly flexible and offer a number of advantages that can be exploited in engineering applications.

2.3.2 Density Estimation

The discussion above unrolls from the perspective of models developed for regression problems. These ideas also extend to density estimation. An intuitive way to interpret complexity in this type of problem is by considering a set of observations that have been generated by an underlying multimodal distribution. Approximating the generating model with a distribution that may be too simple, such as an unimodal Gaussian, will thus result in a poor fit. Models that are too simple become a problem when normal-condition data depart from the unimodal Gaussian assumption. In this scenario, wrong assumptions about the underlying density could lead to an increased rate of *false-positives*, and an outlier approach for damage detection may fail badly.

This consideration is an important one to address in SHM, since deviations within the normal data can occur frequently. EoVs are arguably the primary driver for data departing from these assumptions [66, 67]. This issue can be more concretely illustrated when considering the operational modal analysis of a bridge. If the monitoring system is designed to track the first few natural frequencies, then possible benign factors affecting the stiffness or mass of the bridge should be carefully examined during the modelling process. It has been well-established that changes in ambient temperature will cause

variations in frequencies that may mask the presence of actual damage [68]. These benign variations can, therefore, cause the natural frequencies to depart from the Gaussian representation originally assumed. Other examples of EoVs include changes in operating speed, load variations, humidity, traffic and ice build-up, among others.

Overall, the effects that these variations can have on the system promote the existence of distributions that are naturally more complex. There are different approaches for the development of models that can cope with distributions with increased complexity. Some notable examples include the use of algorithms such as GMMs and *Auto-Associative Neural Networks* (AANNs) [69], which have recurrently been implemented for problems in SHM where distributions are characterised by shapes that are nonconvex or featuring multiple distinct regions.

AANNs - or autoencoders for short - are a type of ANN comprised of an encoder and a decoder (Figure 2.1). The reconstruction process of an autoencoder would be a trivial exercise if the hidden layers were of the same size as the inputs. Their characteristic architecture conveys a compelling aspect of how an autoencoder compresses high-dimensional inputs. Essentially, by reducing the number of nodes in the hidden layers, this architecture enables a "bottle-neck" effect that forces the autoencoder to learn underlying features in the data. Enough information is preserved in the embeddings such that the decoder can then accurately reconstruct the original data. If the activation function is chosen to be linear for a network of a single hidden layer, the data compression is equivalent to *Principal Component Analysis* (PCA) [70]. Implementing other types of activation functions, however, has the advantage of performing a transformation into a nonlinear space, offering the possibility to capture dependencies in the data that a linear transformation cannot.

The nonlinear mapping reveals information about the data that may not be immediately obvious. An example of this idea is explored in [67], where a single-node bottle-neck of a trained autoencoder is shown to reproduce the underlying mechanisms driving the dynamics of both a simulated hard drive and an experimental multi-degree-of-freedom structure. Given this flexibility to interpret complex structures in the embeddings, one is no longer limited to the assumption about the normal data being Gaussian, and the autoencoder can thus be used as a powerful novelty detector. Because of this salient aspect, autoencoders have been exploited for robust damage detection in structures. For example, in [56, 71], the autoencoder is successfully employed as an elegant solution to detect damage in structures that would have otherwise required complex FE models to achieve a similar outcome.

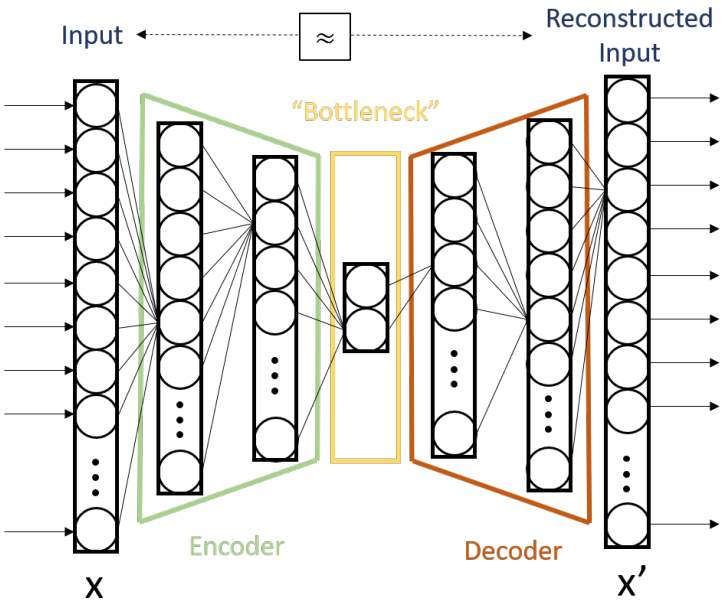


FIGURE 2.1: Diagram illustrating the structure of an autoencoder.

Conventional autoencoders are trained to reconstruct compressed representations of the input data, but are not used to explicitly estimate their true density. Several variations of the autoencoder, aimed at estimating the probability distribution of the embeddings, have been proposed [72, 73]. This notion is pursued in more detail later in Chapter 8, where the *Variational Auto-Encoder* (VAE) [72], is used for enhanced data pre-processing.

While autoencoders are perfectly valid for damage detection, a few caveats associated with them must be acknowledged – and with ANNs in general. Analogous to the model-selection problem for regression, the addition of hidden layers and nodes will improve the flexibility of the network, but at the risk of overfitting the data. Non-Bayesian neural networks are particularly susceptible to this problem, since their construction is generally based on the inclusion of numerous parameters. This problem also relates to the *curse of dimensionality*, warranting vast amounts of data to ensure a network generalises well to all possible health states of a structure. As emphasised in the previous chapter, these considerations will depend on the availability of data for a given application.

A more direct approach to density estimation can be accomplished by approximating the true distribution with a mixture of tractable distributions. The GMM, for example, makes use of K independent unimodal-Gaussian distributions to approximate distributions characterised by multiple distinct modes. Normal-condition data can thus be represented by more expressive densities. The novelty detection process would operate similarly to the case of assuming an unimodal Gaussian distribution. That is, by

assessing the likelihood of new observations to determine if they correspond to a normal-condition state. If the new observations fall within a region of high probability, they are considered normal. However, if the likelihood is below a certain threshold, set based on the inferred distribution, the observations are flagged as anomalous. One advantage of mixture models is that the number of trainable parameters is greatly reduced when compared to autoencoders. The model selection is also limited to the number of components in the mixture and the choice of elementary distributions.

In the literature, the use of GMMs for SHM is often found in applications involving the analysis of AE-data. In [74, 75], for example, a GMM is used to classify crack modes in reinforced concrete structures. In these studies, the GMM was implemented to represent two classes, shear and tensile cracks, which are displayed as two distinct clusters in a two-dimensional space span by RA (ratio of the rise time to the amplitude), vs. AF (Average Frequency) values extracted from the individual AE waves. Other uses of the GMM can also be found in [55], where the GMM is used as part of an elaborate transfer-learning exercise aimed at mapping features and labels across source and target domains in a structural dynamics context.

This subsection has been presented with the intention to demonstrate that mitigating the assumptions about the data can greatly improve the reliability of a model. It is in fact possible to remove the need for underlying parameters, and thus, keep these assumptions to a minimum. Outlined in the remainder of this chapter, this idea is shown embodied in *nonparametric* models.

2.3.3 Towards Nonparametric Modelling

A nonparametric approach operates on the premise of fewer assumptions that may help produce more accurate representations of the data [76]. In other words, nonparametric models are learnt by letting the data "speak for themselves".

One of the simplest nonparametric methods for density estimation is made possible with the implementation of a histogram over the sampling space. The functional form of a normalised histogram is given by,

$$p_i = \frac{n_i}{N\Delta_i} \tag{2.5}$$

where n_i is the number of samples in the ith bin, Δ_i denotes the width of the bins, and N is the total number of available samples. The probability measure p_i assigned to the ith bin is thus determined by the number of samples found within the bounds of the bin. Apart from Δ_i, the involvement of parameters is bypassed entirely, and the density is approximated directly from the observations. Although the number of assumptions

made about the model is reduced, there are still a few variables that determine the functional form of the estimated density. The width of the bins Δ_i and their corresponding edge locations must be carefully tuned to have the probability mass function closely approximate to the true generative distribution. These parameters are often referred to as *hyperparameters*, to make the distinction between these and the parameters $\mathbf{w}$ defined in (2.1).

In practice, the histogram may not always be the best choice for density estimation problems. It is not entirely obvious how to find the optimal width of the bins, and the number of bins required to define the histogram increases substantially with the size of the feature space. Nevertheless, histograms are good examples to illustrate how nonparametric models operate at a fundamental level. Additionally, an important concept introduced by the histogram is that of *smoothness*. Longer bin widths render "smoother" estimations, and some form of distance measure is delimited by the bins. Another way to think of the global variables is as tuners that control the functional smoothness of the estimated density.

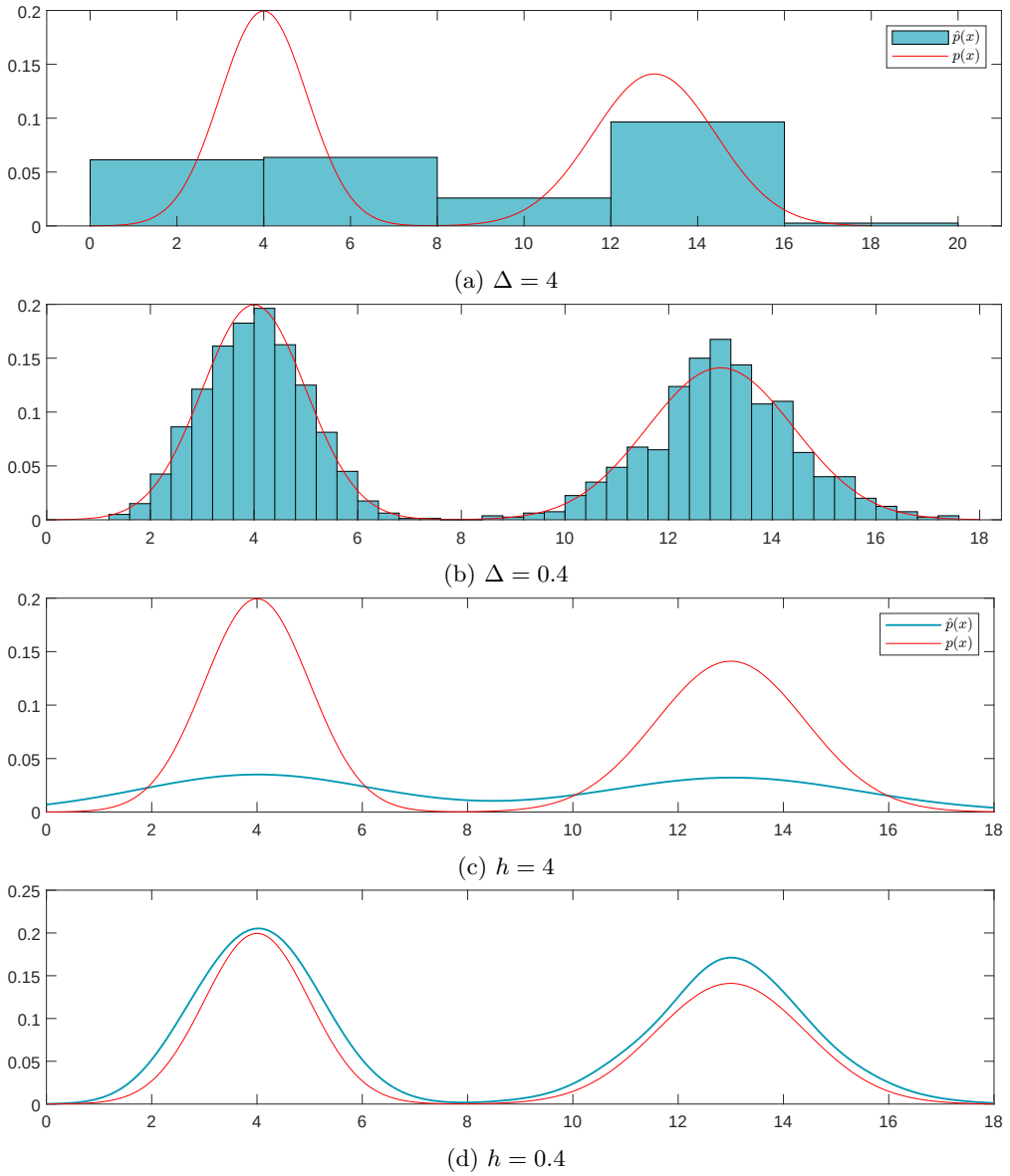
A more principled nonparametric algorithm for density estimation is the *Kernel Density Estimation* (KDE) [77], which involves a collection of probability density "atoms", each contributing to the density estimate. The atoms are defined by a *kernel function* $\mathcal{K}(\mathbf{x})$, that outputs a measure with respect to the distance of its inputs. The basic form of the estimate $\hat{p}(\mathbf{x})$, for multivariate data $\mathbf{x}$, is given by,

$$\hat{p}(\mathbf{x}) = \frac{1}{Nh} \sum_{i=1}^{N} \mathcal{K}\left(\frac{\mathbf{x} - \hat{\mathbf{x}}_i}{h}\right) \tag{2.6}$$

where $\hat{\mathbf{x}}_i$ is the centre of the ith atom and h is a smoothing parameter. A common kernel function used is the multivariate Gaussian, defined as,

$$\mathcal{K}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}||\mathbf{x}||^2\right) \tag{2.7}$$

where d is the dimensionality of $\mathbf{x}$. Choosing a Gaussian kernel means that the model is governed by the location and scale of each Gaussian atom. The atoms are centred at the locations of the data points, while the scale h, remains a tunable parameter. Finding a suitable value h is somewhat analogous to finding the correct bin-width in the histogram. Figure 2.2 shows the implementation of the histogram and KDE to estimate a multimodal Gaussian given a set of random samples drawn from the true distribution. The effects of the hyperparameters are evident when attempting to infer the underlying distribution. Finding the optimal values of these hyperparameters is a crucial step when implementing nonparametric models. This consideration is addressed more in detail in Chapter 4, when covering hyperparameter optimisation for GPs.



(a) $\Delta = 4$

(b) $\Delta = 0.4$

(c) $h = 4$

(d) $h = 0.4$

FIGURE 2.2: Density estimation exercise on a multimodal Gaussian distribution. Two histograms were constructed with windown-lengths (a) $\Delta = 4$ and (b) $\Delta = 0.4$. Additionally, the KDE was employed with parameters (c) $h = 4$ and (d) $h = 0.4$. In all cases, the same $N = 1000$ samples drawn from the true distribution were given.

The histogram and KDE are merely two examples of several nonparametric models. These examples were covered merely to illustrate the fundamental ideas governing nonparametric models; particularly, their flexibility in adapting to complex data, and the concept of smoothness, which is generally determined by a hyperparameter that requires tuning. More information about other density estimation techniques, their advantages, and limitations can be found in [77]. Additionally, comprehensive studies on the use of KDE for damage detection in structures can be found in [78–80].

In the previous chapter, the reasons for employing a Bayesian framework in engineering were covered, highlighting the possibility of seamlessly incorporating expert knowledge

about the system into the model while quantifying associated uncertainties. The inclusion of nonparametric models into this framework renders powerful algorithms capable of overcoming many problems in engineering. Two distinct models are considered in this thesis: namely, the GP and the DP. For now, these can be thought of families of nonparametric Bayesian models featuring boundless complexity in their structures. Chapter 3 presents a review of their uses in SHM, and Chapter 4 covers their theory from a machine learning perspective.

Chapter 3

Literature Review: Nonparametric Bayesian Modelling for SHM

The use of nonparametric Bayesian modelling for SHM has been a subject of increasing popularity in the literature. GP regression is perhaps the most commonly-employed nonparametric Bayesian model for SHM. The advantages that GPs have to offer seem to naturally address many problems encountered in SHM applications. The flexibility of GPs makes them a suitable choice when attempting to model relations in datasets that may be too complex to represent by means of an analytical approach. For a particular dataset, regardless of its underlying complexities, a GP considers all possible functions that fit the data and provides a predictive distribution directly over the candidate functions, rather than over a set of parameters. The inferred predictive distribution simply provides a mean prediction accompanied by confidence intervals, and this compact representation of a probabilistic model is perhaps another reason why GPs may be preferred in many engineering applications.

The DP is a different type of nonparametric Bayesian model that has also proven to be an useful tool in SHM. In particular, the DP can be advantageous when dealing with categorical-type problems. This advantage is essentially provided by their ability to estimate uncertainties regarding the number of potential classes in the dataset. Similarly, if the number of classes (or clusters) is unknown, then the DP can be used to autonomously determine the probable class distribution. In this chapter, a literature review of their respective uses and advantages in SHM is covered.

3.1 Gaussian Processes for Damage Detection, Localisation and Prognosis

GPs have been successfully adopted in many areas of SHM [81–84]. Their use, for example, has been recursively adopted when modelling the nonlinear relation between power output and wind speed in *Wind Turbine* (WT) power curves [19, 85]. Variations in the performance of a WT can be monitored when incoming new measurements depart from the power-curve model. For example, this exercise was carried out in [86], where the power curve is modelled with a GP to identify WT yaw misalignment. Another case study is presented in [87], where the GP is used to classify different health states of a WT. The results show that the GP classification model not only provides a measure of confidence in the predictions but also generally returns more accurate predictions than a *Support Vector Machine* (SVM) for the investigated case study. An alternative approach to damage detection of WT components is proposed in [88]. The study involves using the GP to learn correlations between the normalised edge frequency of each blade-pair, to then determine the presence of damage when the modelled correlations depart from normal.

Issues related to the effects of the environment and operational conditions can also be addressed in a principled manner with the implementation of GPs. This approach is based on the assumption that changes caused by mechanisms other than damage, will have a relatively longer time-scale influence. By fitting a regression model to the features of interest, predictions from this model can then be subtracted from subsequent data, leaving the remaining residuals as features with confounding influences pruned away. Regression models used in this context can be referred to as *response surface models* [89]. These residuals are then expected to be sensitive only to damage, and novelty detection techniques can be naturally incorporated in this strategy for damage identification. GPs are a convenient data-driven alternative when the relations cannot be easily modelled by simple polynomials. Several studies that have explored this strategy for SHM can be found in the literature, demonstrating the advantages and successful outcomes of using the GP for enhanced novelty detection in the presence of *Environmental and/or Operational Variations* (EoVs).

For example, in [90], a GP-based response surface model was implemented to account for benign changes experienced by the first natural frequency and deck deflection of the Tamar bridge. Another example is in [91], where a GP regression model was employed to model features extracted from the low and high-frequency data of a WT blade subjected to changes in environmental and operational conditions. The *Mahalanobis-Squared Distance* (MSD) was then used, based on the GP predictions, to determine the

presence of damage. In [92], a unified approach was proposed to distinguish influences caused by EoVs, sensor faults and structural damage. Sensor faults were detected by modelling the sensor network using a GP, and unprecedented changes were determined via a generalised likelihood ratio test. Additionally, in situations where the confounding influences are discontinuous, the *Treed Gaussian Process* (TGP) proposed in [89] can facilitate and automatically identify changes in the operational regimes of bridges.

Another technique that will be of particular interest in subsequent chapters, is the implementation of GPs for AE-based localisation techniques. In particular, the propagating behaviour of AE waves in structures exhibiting obstructions, intricate geometries, or made from anisotropic materials, may often be too complex to model analytically. This limitation hinders the reliability of AE-based techniques for structural damage detection and location. A promising solution to this matter, however, has been proposed whereby GP regression is used for interpolating the differences in time of arrival of AE-waves for source location [93, 94]. The flexibility of the GP can essentially be exploited to effectively capture the complex nature of ΔT maps. A practical example of using GPs for this task, but applied to landing gears, is presented in [95]. These studies not only demonstrate the efficacy of employing a GP for AE source location in experimental specimens, but also in complex engineering structures, aiding towards the prospects of making AE-based monitoring techniques a standard procedure.

Statistical modelling for predicting structural loading is another meaningful challenge encountered in SHM, as it is necessary for accurately estimating the remaining useful life of structures. Recent work has exploited the use of GPs to address this type of problem. In [96], a GP is employed to characterise the uncertainty of load predictions and assess the impact it has on fatigue damage. Specifically, the GP is used to model the strain of structures subjected to some latent loading. The distribution returned by the GP is then accounted for by propagating the uncertainty to the fatigue-loading predictions. The ability of the GP to provide an uncertainty measure is, perhaps, what gives them an advantage over other data-based techniques in this scenario. Indeed, information regarding how unsure one may be about the remaining life of a structure can be of value in making decisions about relevant repairs or potential decommissions. Another example of using GPs for load prediction can be found in [97]. In this study, the prediction of loads on components of landing gears, subjected to simulated landing conditions was investigated. The aim was to achieve accurate mappings using commonly-available measurements provided by the *Flight Data Recorder* as inputs. It is concluded that further improvements could lead to accurate predictions without the need to rely on strain-measuring equipment. The implications of having such a model could significantly enhance the estimation of remaining life-cycles in landing gears, and therefore, in the development of more efficient maintenance schedules.

More direct approaches for fatigue damage prognosis have also been investigated. Advancements in this area of SHM have been made possible with the introduction of GP regression for predicting the development of damage. For example, in [98], GP regression was used for the prediction of crack extension in aluminium specimens. Similarly, in [99], the GP was employed as a surrogate function for simulating fatigue crack growth, as opposed to *Finite Element Analysis* (FEA), to reduce the computational demands required to calculate the stress intensity factor of mechanical components. The choice of GPs in these studies was preferred over other surrogate modelling techniques because of their ability to return confidence intervals in the predictions. This advantage is also evident in [100], where GPs are implemented to predict the stage of damage in a composite airfoil structure. An innovative approach proposed in [101] is based on the combination of *particle filters* with GPs for fatigue crack-growth prognosis. The GP was used to directly fit a regression model to crack lengths given features deriving from vibration responses. The particle filter was then used to return a posterior estimation of the crack length, which was then fed back to the GP for updating. This updating step significantly improved the accuracy of the model in predicting the number of cycles to failure.

Overall, all the examples above rely on conventional GPs to model stationary nonlinearities, but developments in dynamic system identification have made it possible to use a variation of time-series modelling, based on GPs, for dynamic modelling [102]. In particular, a nonparametric variant of the *Nonlinear AutoRegressive model with eXogenous input* (NARX) can be made possible with the incorporation of GPs. In short, the GP-NARX is a function that (nonlinearly) relates the current value of a time series to past values of the same series and current and past values of some exogenous (i.e. external) input series. In a NARX model, the nonlinear function is usually defined by a polynomial, but in the case of a GP-NARX, a GP is used instead. The implementation of GP-NARX for SHM has been employed for modelling time-series signals featured by complex nonlinearities, such as the response to Duffing oscillators and wave loading prediction on off-shore platforms [103]. The success of GP-NARX for dynamic modelling is one prevalent example demonstrating the versatility of GPs and the added benefits they bring to the field of SHM.

This idea can be extended further by combining GPs with governing equations describing the physics of complex systems. Earlier, it was mentioned that knowing the physics of the system provides interpretability and extrapolating capabilities that data-driven models cannot have. In fact, this is one of the main limitations of data-driven models - or *black-box models* - that can be greatly alleviated with the inclusion of physics-based models - or *white-box models*. The latter may still struggle to fully capture the physics of certain mechanisms that would require solving highly-complex equations.

Introducing the GP in this context may be achieved by having the white-box part account for a simpler representation of the system while making use of data to render a more complete representation. An example of this approach for SHM can be found in [104], where the wave-loading prediction in off-shore structures is modelled according to *Morison's equation* [105], and then combined with the GP-NARX to more accurately capture nonlinearities induced by turbulence and wakes.

There are, however, some practical limitations to consider when deciding to use a GP for statistical modelling. Although relatively simple to implement, inference of the predictive posterior requires the inversion of the covariance matrix, which demands a computation on the order of $\mathcal{O}(N^3)$ [16]. Another complication when using GPs occurs when having a likelihood other than a Gaussian distribution. These issues may require the predictive posterior to be approximated with *Markov Chain Monte Carlo* (MCMC) or *Variational Inference* (VI) [48, 106], which can be laborious to implement.

3.2 Dirichlet Processes for Damage Identification

When it comes to the use of nonparametric Bayesian approaches for SHM, the DP has certainly not been explored nearly as much as the GP. A possible explanation for this preference may be because of the dominance of regression problems in SHM. The DP is better suited for density estimation problems, which may not be encountered as often and are already considered to be the hardest of all the learning-type problems [1]. Additionally, unlike GPs, the DP prior is very unlikely to be a conjugate to the likelihood employed for a particular application, and therefore, warrants the need for approximations that are generally computationally demanding or not as intuitive as inferring the predictive posterior of a Gaussian likelihood with a GP prior. Nevertheless, the development of models with DP priors for SHM has been gaining increasing interest, which could be, perhaps, for their ability to directly infer the number of components in a mixture model from the data. The importance of this last point for SHM is one of the main highlights in subsequent chapters.

One attractive application of DPs is in the construction of models comprised of an "infinite" number of independent distributions. The limiting factor in the implementation of a *Gaussian Mixture Model* (GMM) [107] for SHM is having to decide on the number of independent Gaussian distributions *a priori*. An assumption that has been made implicitly in this statement is that distinct health states will manifest as unique modes in the underlying density. Given that one cannot know for sure all the possible conditions a structure might experience, it is therefore unlikely to determine a suitable number of

components in the mixture model. If the model is made fully Bayesian, there is no theoretical reason to limit the model to a finite number of components. In fact, a tractable model exists that allows for an infinitely-large mixture model. This idea is driven by having a DP prior in the mixture model [108].

An infinite mixture of Gaussian distributions was employed for damage detection in [109]. The proposed method was implemented with data gathered from both an experimental rig and the Z24 bridge, where the premise of the model was to autonomously infer the correct number of Gaussian clusters needed to estimate the underlying density of the normal condition. The presence of damage is determined once a new cluster is generated as the result of damage causing the data to depart from normal. The use of an infinite Gaussian mixture model is also explored in [110] for AE-based damage detection in machining tools, whereby a similar reasoning is followed. A different use of DPs for SHM can be found [111], where the infinite Gaussian mixture model was developed to automatically identify real modes from spurious modes in *stabilisation diagrams*. The model was validated on a full-scale cable-stayed bridge subjected to ambient vibrations, demonstrating its ability to autonomously find mode shapes and damping values in a more consistent manner when compared to existing standard approaches used in *Operational Modal Analysis* (OMA).

Chapter 4

Theory and Applications

4.1 Gaussian Processes

GPs are powerful and flexible nonparametric models that find extensive use in machine-learning applications. The term "flexible" refers to their ability to adapt smoothly to the complexity of the data, making them suitable for a wide range of data types and patterns. Additionally, GPs are nonparametric because they do not require fixing the parameters of the model explicitly. Instead, they marginalise the underlying parameters, which allows them to learn directly from the data and adjust to its intricacies. This section aims to be as self-contained as possible, but for more in-depth information on GPs, readers are encouraged to refer to the work in [16, 48]. A gentle introduction to GPs can also be found in [112].

There are several ways in which the GP has been defined in the literature. Some commonly encountered definitions are:

- A distribution over functions.
- A multivariate Gaussian distribution with an infinite number of dimensions.
- A solution to a linear stochastic differential equation driven by white Gaussian noise.

In an attempt to clarify these interpretations, the following subsection examines the GP approach to regression.

4.1.1 Gaussian Process Regression

In order to define the GP, the following generative model is considered,

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} \tag{4.1a}$$

$$y = f(\mathbf{x}) + \epsilon \tag{4.1b}$$

where the input vector $\mathbf{x} = \{1, x_1, \ldots, x_{D-1}\}^\top \in \mathbb{R}^D$ is mapped to a scalar output (target) $y \in \mathbb{R}$, $\mathbf{w} = \{w_o, \ldots, w_{D-1}\} \in \mathbb{R}^D$ is a vector of weights, and $f(\cdot) \in \mathbb{R}$ is the modelling function. The vector of inputs is usually augmented with a unit to account for the bias (or offset) of the parameter w_o in $\mathbf{w}$. It is important to note that the model's predictions will not perfectly match the observations, and this discrepancy is captured by the additive Gaussian noise term ϵ, which follows a normal distribution $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

The inputs in (4.1) can be projected into a higher-dimensional feature-space with a (nonlinear) transformation $\boldsymbol{\phi} : \mathbb{R}^D \to \mathbb{R}^K$, thereby granting some further flexibility. For example, the following projection can be considered,

$$\phi_k(\mathbf{x}) = \exp\left(-\frac{|\mathbf{x} - \boldsymbol{\mu}_k|^2}{2s^2}\right) \tag{4.2}$$

where $\phi_k(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}$ is the *basis function* that defines the k^{th} projection of the feature vector $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}^K$. This particular basis function is called a *radial-basis function* or *Gaussian basis function*. Another recurrent example is the family of *polynomial basis functions*, where K powers of the inputs are taken so that $\phi_k(x) = x^k$ and $\boldsymbol{\phi}(x) = \{0, x, x^2, \ldots, x^{K-1}\}^\top$. By projecting inputs into higher dimensional space, the model becomes more expressive via a linear combination of the feature vectors. A variety of basis functions exist, and their choice will depend on the application at hand. Approaches for model selection are discussed later in this section. For now, if a K-dimensional feature-space is considered, the original regression model can be expressed as,

$$f(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{w} \tag{4.3}$$

where $\mathbf{w} \in \mathbb{R}^K$ now contains the linear parameters w_k.

At this stage, one might question how to handle the dimensionality of the feature vectors. As it turns out, the GP is recovered when the dimensionality of the feature space is set to infinity, in addition to defining a prior distribution over the parameters $\mathbf{w}$. This reasoning somewhat translates into marginalising the parameters in a Bayesian linear regression problem (Appendix A.1). The problem, however, is that the involved computations become prohibitively expensive as $K \to \infty$. Fortunately, an efficient solution exists that avoids explicitly computing a large matrix inversion while still allowing for a

linear combination of an infinite number of basis functions. To elucidate this solution, it is helpful to elicit the *covariance function*, also known as the *GP kernel*, denoted as $k(\mathbf{x}_p, \mathbf{x}_q)$. A common choice for the covariance function is the *Squared-Exponential* (SE) covariance function, which is defined by the following,

$$k(\mathbf{x}_p, \mathbf{x}_q) = \exp\left(-\frac{1}{2}|\mathbf{x}_p - \mathbf{x}_q|^2\right) \tag{4.4}$$

The SE covariance function is just one example of many, and its expression is the result of an infinite linear combination of radial-basis functions defined by (4.2). This operation is made possible by Mercer's Theorem, which ensures that for any positive-definite covariance function, there exists an infinite expansion in terms of basis functions [16].

In loose terms, the covariance function quantifies the similarity between any pair of random variables by means of the *kernel trick* [48]. Now, instead of having to compute the projection of the inputs into an infinite-dimensional space, the kernel trick allows for a direct comparison of variables without explicitly having to derive the elements in the feature space.

The formulation of the GP incorporates the covariance function for the construction of a distribution over the function space. Defining the prior $\mathbf{w}$ as a Gaussian with zero mean and covariance Σ_p, the GP can be fully defined as follows,

$$\mathbb{E}\left[f(\mathbf{x})\right] = \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}\left[\mathbf{w}\right] = \mathbf{0} \tag{4.5}$$

$$\mathbb{E}\left[f(\mathbf{x})f(\mathbf{x}')\right] = \boldsymbol{\phi}(\mathbf{x})^\top \mathbb{E}\left[\mathbf{w}\mathbf{w}^\top\right] \boldsymbol{\phi}(\mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})\Sigma_p\boldsymbol{\phi}(\mathbf{x}') \tag{4.6}$$

where $f(\mathbf{x})$ and $f(\mathbf{x}')$ are jointly Gaussian with mean zero and covariance $\boldsymbol{\phi}(\mathbf{x})\Sigma_p\boldsymbol{\phi}(\mathbf{x}')$. For any given input point, the values $f(x_1), \ldots, f(x_n)$ are jointly Gaussian distributed, and the GP is simply a collection of the random variables $f(\mathbf{x}_n)$. The covariance is defined by an inner product of the inputs with respect to Σ_p, and can equivalently be represented by $\psi(\mathbf{x}) = \Sigma_p^{\frac{1}{2}}\boldsymbol{\phi}(\mathbf{x})$, such that $k(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})\psi(\mathbf{x}')$. While $k(\mathbf{x}, \mathbf{x}')$ can be evaluated in this way, the same outcome can be achieved with a corresponding covariance function, enormously facilitating its calculation.

The definitions in (4.5) and (4.6) bypass the need to compute the posterior $p(\mathbf{w}|\mathbf{x}, y)$ entirely. This implication means that the GP can be completely defined by its mean and covariance function. Therefore, establishing a zero mean, and evaluating $k(\cdot, \cdot)$, implies a distribution over functions. Specifically,

$$f_* \sim \mathcal{N}(\mathbf{0}, k(\mathbf{X}_*, \mathbf{X}_*)) \tag{4.7}$$

where $\mathbf{X}_* \in \mathbb{R}^{D \times N_*}$ denotes the matrix including N_* column test inputs $\mathbf{x}_* \in \mathbb{R}^D$, and the "shape" of the realisations of f_* depends entirely on the choice of covariance function. For example, samples of f_* evaluated over a range of arbitrary test inputs, when choosing the SE covariance function, are shown in Figure 4.1(a). These samples are functions modelled by the GP before observations are introduced. In other words, it assumes a prior belief in the functional form of the data.

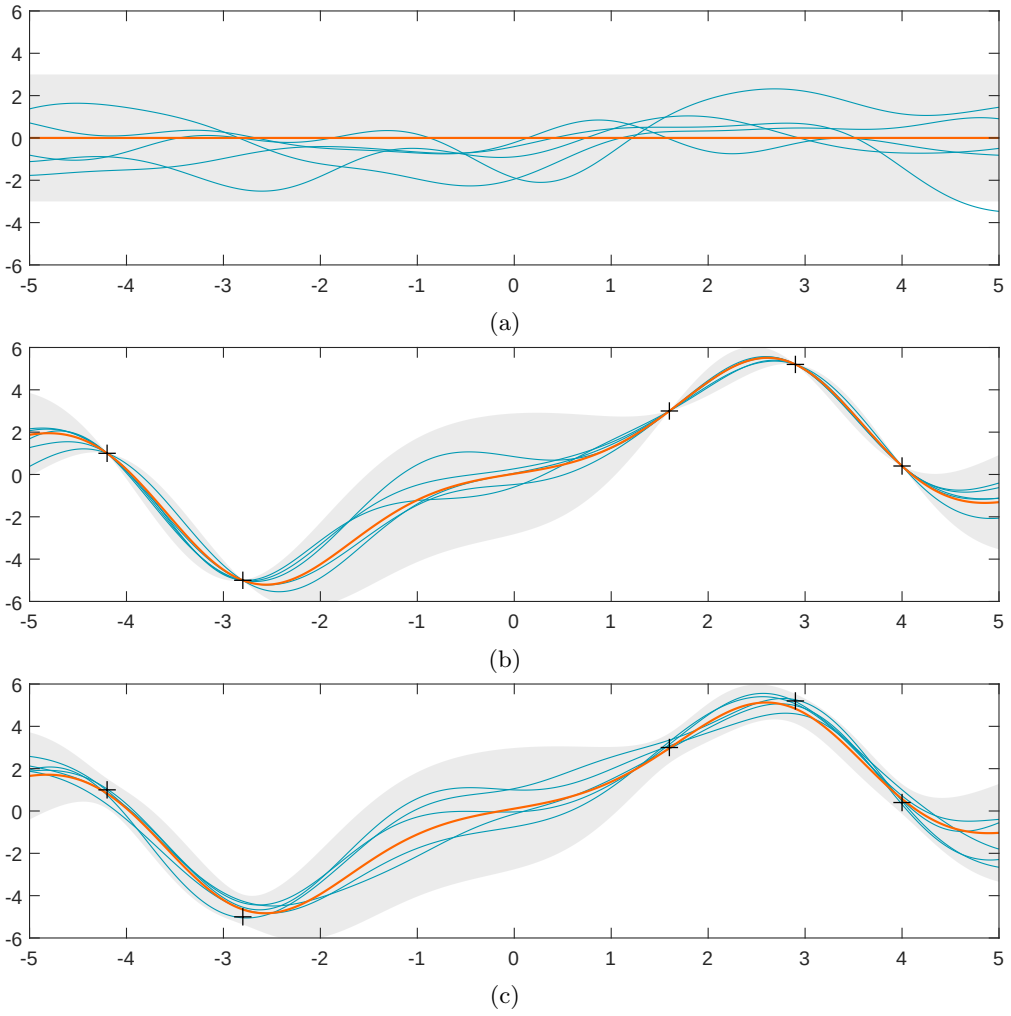


FIGURE 4.1: Demonstration of inference with a GP for regression. (a) GP prior, (b) GP conditioned on data without noise, and (c) GP conditioned on data while accounting for noise. The orange line and the grey bounds correspond to the GP mean and covariance (3σ), respectively. Seven random draws of functions shown in cyan have been included in all cases.

Presenting the GP with a training set of inputs $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{D \times N}$, alongside their corresponding targets $\mathbf{y} = \{y_1, \ldots, y_N\}^\top \in \mathbb{R}^N$, narrows down the collection of all possible functions to those that fit data. This procedure is achieved by restricting the GP to only output functions that agree with the observations. In probabilistic terms, this operation is conducted by employing the conditional properties of Gaussian distributions. To begin, the joint distribution of the observations $\mathbf{f}$, and the predictions

f_*, is defined as,

$$\begin{pmatrix} \mathbf{f} \\ f_* \end{pmatrix} \sim \mathcal{N}\left[\mathbf{0}, \begin{pmatrix} K(\mathbf{X},\mathbf{X}) & K(\mathbf{X},\mathbf{X}_*) \\ K(\mathbf{X}_*,\mathbf{X}) & K(\mathbf{X}_*,\mathbf{X}_*) \end{pmatrix}\right] \tag{4.8}$$

where $K(\mathbf{X},\mathbf{X}_*)$ corresponds to the covariance function evaluated at all pairs of training and test inputs. Since the joint distribution is a Gaussian, any subset of variables conditioned on the rest, is also Gaussian distributed. Therefore, conditioning the joint distribution on the observations gives,

$$f_*|\mathbf{X}_*,\mathbf{X},\mathbf{f} \sim \mathcal{N}(\bar{f}_*,\Sigma_*) \tag{4.9}$$

$$\bar{f}_* = K(\mathbf{X}_*,\mathbf{X})K(\mathbf{X},\mathbf{X})^{-1}\mathbf{f} \tag{4.10}$$

$$\Sigma_* = K(\mathbf{X}_*,\mathbf{X}_*) - K(\mathbf{X}_*,\mathbf{X})K(\mathbf{X},\mathbf{X})^{-1}K(\mathbf{X},\mathbf{X}_*) \tag{4.11}$$

in which the mean and covariance of the GP posterior are defined. Examples of functions sampled from a GP posterior inferred from a sequence of observations are shown in Figure 4.1(b). A useful metaphor may be to think of the prior samples as loose strands of hair, which are tied together at points where observations exist. The posterior samples are thus the result of "tying" these functions together.

There are two important notes to highlight from this outcome. First, the uncertainty in the posterior increases in areas away from the data. This behaviour is logical since the model is more reliant on the prior in regions where no information is provided by the data. Second, the GP becomes absolutely certain of the outcome at points where data exist. This outcome is clearly undesirable, as it might be unrealistic in most (if not all) applications to have error-free measurements. Fortunately, the noise in the observations can be easily accounted for in the model by assuming some corruption is provided by white noises. That is, $\mathbf{y} = f(\mathbf{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0,\sigma^2\mathbb{I})$. The same derivation for the joint distribution can be followed with,

$$\begin{pmatrix} \mathbf{y} \\ f_* \end{pmatrix} \sim \mathcal{N}\left[\mathbf{0}, \begin{pmatrix} K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbb{I} & K(\mathbf{X},\mathbf{X}_*) \\ K(\mathbf{X}_*,\mathbf{X}) & K(\mathbf{X}_*,\mathbf{X}_*) \end{pmatrix}\right] \tag{4.12}$$

where $\mathbb{I} \in \mathbb{R}^{N\times N}$ denotes the identity matrix. Conditioning this expression for the joint distribution results in the following GP posterior,

$$f_*|\mathbf{X}_*,\mathbf{X},\mathbf{y} \sim \mathcal{N}(\mathbb{E}[f_*], \mathbb{V}[f_*]) \tag{4.13}$$

$$\mathbb{E}[f_*] = K(\mathbf{X}_*,\mathbf{X})\left[K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbb{I}\right]^{-1}\mathbf{y} \tag{4.14}$$

$$\mathbb{V}[f_*] = K(\mathbf{X}_*,\mathbf{X}_*) - K(\mathbf{X}_*,\mathbf{X})\left[K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbb{I}\right]^{-1}K(\mathbf{X},\mathbf{X}_*) \tag{4.15}$$

The posterior distribution can thus be computed by adding the $\sigma^2\mathbb{I}$ along the diagonal of the covariance of the training inputs. Additionally, it is worth noting that $\sigma^2\mathbb{I}$ regularises the inversion on $K(\mathbf{X}, \mathbf{X})$. The results of this action are shown in Figure 4.1(c), where the functions are somewhat less restricted to agree with the data – the hair-ties are now a bit looser.

4.1.2 On the Choice of Covariance Function

In the previous subsection, it was stated that a variety of basis functions exist, and it was possible to define an infinite expansion in terms of basis functions by the covariance function. The covariance function dictates the expressiveness of the GP, and choosing the right covariance function for a given application is a crucial step in the implementation of GPs.

So far, the SE function (Equation (4.4)) has been demonstrated as the primary example. Although it is one of the most commonly used covariance functions in machine learning, it may not always be the best option when modelling engineering datasets. The SE function is infinitely differentiable, and a GP with this function is thus too smooth to model realistic physical processes [113]. Instead, a more suitable class of covariance functions may be the family of *Matérn* functions. The general form of the covariance functions is defined by,

$$k(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu r}}{l} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu} r}{l} \right), \qquad r = |x - x'| \tag{4.16}$$

where ν and l are positive parameters, and K_ν is a modified Bessel function [114]. A main characteristic of the Matérn class is that the parameter ν scales the smoothness of the GP. For $\nu \to \infty$ the Matérn function returns a SE. Conversely, having $\nu = 1/2$ returns the *exponential covariance function* $k(r) = \exp(-r/l)$, which is very rough. In fact, in a one-dimensional input space, the exponential function is equivalent to that of an Ornstein-Uhlenbeck (OU) process [115], which originates from modelling the velocity of a particle undergoing Brownian motion. Somewhat of a middle-ground is achieved when taking half integers of ν. That is, $\nu = p + 1/2$ with positive integers p.

The two variations of Matérn functions that will be of interest in subsequent chapters are the cases where $\nu = 3/2$ and $\nu = 5/2$. Concretely,

$$k_{3/2}(r) = \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{l} \right) \exp\left(-\frac{\sqrt{3}r}{l} \right) \tag{4.17}$$

$$k_{5/2}(r) = \sigma_f^2 \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}r}{l}\right) \tag{4.18}$$

Examples of functions drawn from GPs with covariance functions $k_{1/2}$, $k_{3/2}$ and $k_{5/2}$ are shown in Figure 4.2.

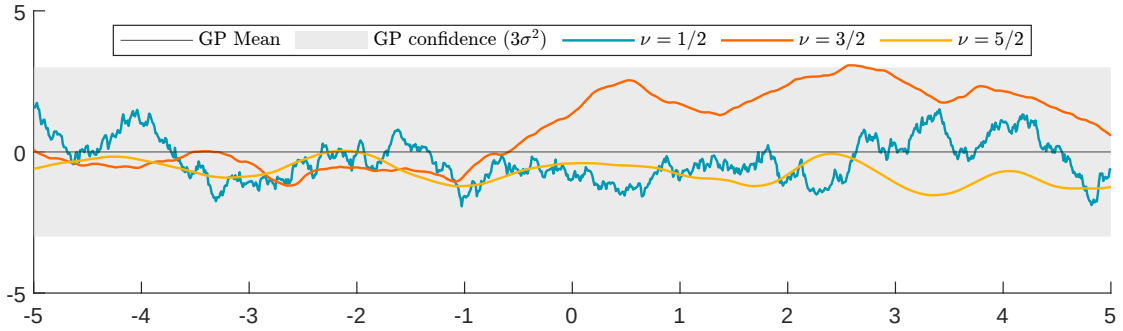


FIGURE 4.2: Random functions drawn from GP priors with the Matérn covariance function (Equation (4.16)) for different values of ν.

While the SE and Matérn functions are often used in practice, it is worth emphasising that one is not limited to these covariance functions. An important aspect not yet mentioned is that covariance functions offer some interpretability in their realisations. This note has been already demonstrated implicitly by acknowledging that the smoothness can be controlled to best represent the underlying physical nature of the data. Knowing *a priori* the physical process responsible for the observations is an essential requirement in the implementation of GPs. This implication is possible because of the Bayesian framework adopted in this type of modelling, and thus worth exploiting when possible to make better predictions.

Here is where engineering knowledge one has about the underlying mechanisms can be incorporated into this framework. In order to demonstrate the influence a covariance function has on the predictive ability of the GP, the response of a *Single-Degree-of-Freedom* (SDOF) mass-spring system (Figure 4.3) is considered.

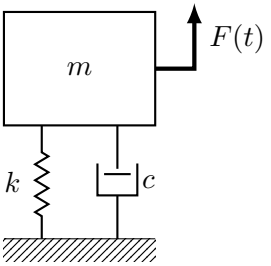


FIGURE 4.3: Mass-spring-damper SDOF system.

Here, the dynamical model can be defined by the following equation of motion,

$$m\ddot{y}(t) + c\dot{y}(t) + ky(t) = F(t) \tag{4.19}$$

where m, c and k are the dynamic coefficients corresponding to the mass, damping and stiffness of the system, $y(t)$ is the response at a given time instance t, and $F(t)$ is the input force. The dots over the variables in Equation (4.19) denote the derivatives of the variable with respect to time. Assuming that the system is linear, the impulse response function $h(t)$ of the oscillator is given by,

$$h(t) = \frac{e^{-\zeta\omega_n t}}{m\omega_d}\sin(\omega_d t) \tag{4.20}$$

where the natural frequency $\omega_n = \sqrt{k/m}$, the damping ratio $\zeta = c/2\sqrt{km}$, and the damped natural frequency $\omega_d = \omega_n\sqrt{1-\zeta^2}$. The response can now be simulated by Equation (4.20) for a given selection of coefficients. In this case, the mass, damping, and stiffness were given values of $m = 10$, $c = 3$, and $k = 100$, respectively. The resulting response is established in all three cases shown in Figure 4.4 as the true value.

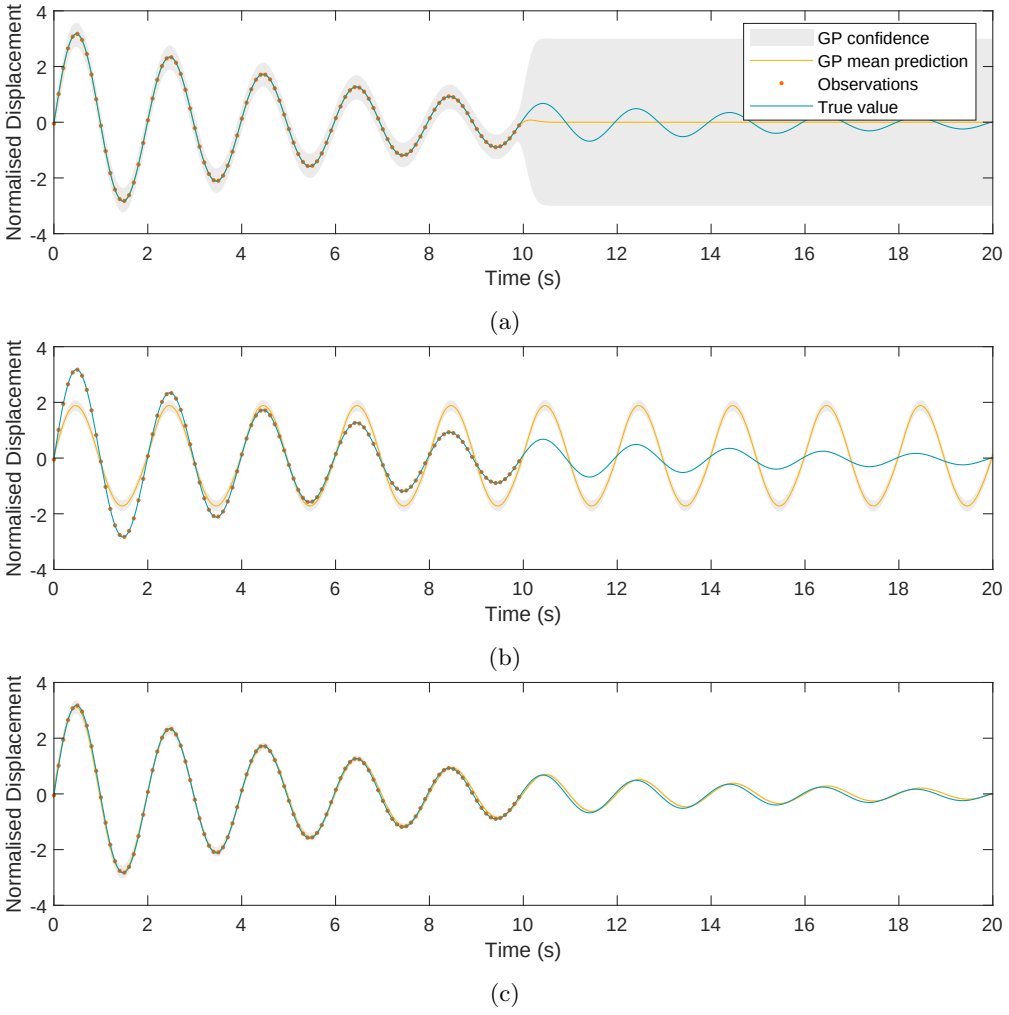


FIGURE 4.4: Fitting data from the simulated response of an SDOF system with a GP, defined by (a) a SE covariance function, (b) a strictly periodic covariance function, and (c) a combined covariance function.

In Figure 4.4(a), a SE function is first used to fit the GP over the observations. A set of observations is only included up to $t = 10$. The idea of this demonstration is to explore how the choice of covariance function affects the extrapolating capabilities of a GP. When the SE covariance is used, the GP quickly reverts to its prior upon the lack of data. Although the confidence at $t > 10$ accounts for the true response, the uncertainty is too broad to have a meaningful representation of the oscillator.

Some progress can be achieved when adopting a more informed covariance function. The harmonic nature of the response provides some information that can be used for this purpose. In the second scenario, shown in Figure 4.4(b), a *strictly periodic covariance function* [116] is employed. This function is defined as,

$$k_p(r) = \sigma_f^2 \exp\left(\frac{-2\sin^2(\pi r/p)}{l^2}\right) \tag{4.21}$$

where p represents the period of the response. The GP now learns the frequency of the oscillation from the data and can extrapolate this information beyond areas where data exist. The obvious problem is that the periodic covariance function cannot express the varying amplitude of the response. Fortunately, this limitation is remedied by combining the periodic covariance function with another that can represent the attenuation caused by the damping. The *exponential-decay covariance function* [117] is suited for this purpose, and is given by,

$$k_p(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp(\mathbf{x}l) \exp\left(\mathbf{x}'^{\top} l\right) \tag{4.22}$$

One can make use of the properties of different covariance functions by combining them accordingly. The product of two independent kernels results in another valid kernel [16], and in here, Equations (4.21) and (4.22) were combined by simply multiplying them together. Inferring the GP posterior defined with the combined covariance functions results in the prediction shown in Figure 4.4(c). The GP is now better at predicting the response of the oscillator at any instance in time, even when presented with a partial dataset. By looking at the functional form of Equation (4.20), it becomes clear why the product of k_p with k_d yields a reasonable model.

The ideas followed in the demonstration above touch upon *physics-informed* or *grey-box* modelling [8]. Although the derived model is still a *black-box* model by definition, introducing some intuition of the underlying physics greatly improved the quality of the predictions. An engineer should therefore acknowledge the dynamics of a system where possible when modelling with a data-driven algorithm like the GP.

In many cases, it may be much harder to determine which covariance function (and possible combinations) to use, and a more stringent method for model selection might be required. Nevertheless, given the Bayesian nature of the GP, it is possible to determine the choice of a suitable covariance function in a principled manner. This aspect of Bayesian modelling was raised in Section 2.3.1, highlighting that, for a given model, the evidence can be evaluated to quantify how well the model agrees with the data. In the current scenario, the model is characterised by the covariance function, and the "best" covariance function is thus the one that maximises the evidence (Equation (2.4)).

4.1.3 Hyperparameter Optimisation

Up to this point, the covariance functions have been implicitly defined by specific variables that are indirectly responsible for the expressiveness of the GP. The Matérn function, for example, is parameterised by a scaling parameter σ_f^2 and length scale l. In the expression for the periodic covariance function, the period of the response p is also included. These variables are the hyperparameters of the GP, and require careful tuning to achieve a meaningful fit to the observations. Therefore, the model-selection process for GPs generally involves two main steps. The first step, discussed above, is the comparison of different families of covariance functions, while the second step, is the optimisation of the hyperparameters for the given covariance function.

The model selection process can be viewed in a hierarchical setting. For a general Bayesian framework, at the lowest level in this hierarchy are the parameters $\mathbf{w}$. At the second level are the hyperparameters – denoted from now on by Θ – which control the distribution over the parameters $\mathbf{w}$. Finally, at the top level, there may be a discrete set of J model structures $\mathbf{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_J\}$, defined by a collection of different covariance functions.

Overall, to find the best-fitting model for a given dataset, the model-selection process involves exploring different combinations of covariance functions and their corresponding hyperparameters. The goal is to find the model that provides the best trade-off between complexity and predictive performance for the specific task at hand. Maximising the *marginal likelihood* (evidence), has the advantage of searching for the optimal hyperparameters while naturally encouraging this trade-off. The marginal likelihood for GPs can be defined by marginalising over the function values $\mathbf{f}$,

$$p(\mathbf{y}|X, \Theta) = \int p(\mathbf{y}|\mathbf{f}, X, \Theta) p(\mathbf{f}|X, \Theta) \mathrm{d}\mathbf{f} \tag{4.23}$$

resulting in the following expression for the log-marginal likelihood,

$$\log p(\mathbf{y} \mid X, \Theta) = -\frac{1}{2}\mathbf{y}^\top \left(K + \sigma_n^2 \mathbb{I}\right)^{-1} \mathbf{y} - \frac{1}{2}\log \left|K + \sigma_n^2 \mathbb{I}\right| - \frac{n}{2}\log 2\pi \tag{4.24}$$

where the dependencies on Θ are implied in the evaluation of K. The balance between complexity and predictive performance is readily interpretable by the terms of this expression. The first term – $\mathbf{y}^\top \left(K + \sigma_n^2 \mathbb{I}\right)^{-1} \mathbf{y}$ – is the only one involving the observed targets $\mathbf{y}$, while the second term – $\log \left|K + \sigma_n^2 \mathbb{I}\right|$ – is the complexity penalty that depends only on the covariance function and inputs.

One may note that this marginalisation corresponds to the second stage in the hierarchy, thereby bypassing the first stage entirely. The reason for this decision is because of the fact that the parameters $\mathbf{w}$ had already been marginalised out in Equation (4.13). By this premise, the marginalisation could, in theory, be conducted once more over the hyperparameters Θ, and thereafter again over $\mathbf{M}$. However, the former is not often executed because the integrals that result from marginalising over the hyperparameters tend to be too cumbersome to evaluate. Instead, finding the hyperparameters that maximise Equation (4.24) is achieved by some optimisation scheme. This step can be repeated over several model structures, given the same dataset, to eventually determine which one yields the highest evidence.

Finding hyperparameters that maximise the log-marginal likelihood is a problem apt for any numerical optimiser. In fact, numerous optimisation strategies exist, from which one can choose [118, 119]. It is often the case that the gradients of Equation (4.24) with respect to Θ are available, and in such circumstances, a gradient-based optimisation may be preferred. Alternatively, any gradient-free technique is equally valid, as it is likely that almost all optimisation schemes will, on average, perform similarly because of the *no-free-lunch* optimisation theorem [120]. This theorem simply states that no optimisation algorithm universally performs better than the rest across all problems. For the remainder of this thesis, the *Quantum-Behaved Particle-Swarm Optimiser* (QPSO) [121] is employed for hyperparameter optimisation of GPs. Being a population-based optimiser, the QPSO offers the advantage of being less likely to get stuck at a local minimum.

4.2 Dirichlet Processes

DPs stand out as a different family of nonparametric stochastic processes. Somewhat analogous to the definition of GPs as distributions over functions, the DP is commonly

described as a distribution over distributions. In other words, samples drawn from a DP are probability mass functions over a space of measures Θ.

The DP naturally arises in categorical problems, where the goal is to assign a given observation with the label of a discrete variable $\mathbf{y}$ that can take one of K possible mutually-exclusive states. Having to know the number of categories to model prior to observing the data is a problem that can be addressed with the DP. The expressiveness of the DP can be exploited in scenarios where the data might be more complex than anticipated.

Deriving the DP, however, is not as simple as extending a Dirichlet distribution to an infinite number of dimensions. In fact, this reasoning leads to the *Griffiths, Engen and McCloskey distribution* [122] (or GEM distribution), which is not quite strictly a DP. Before introducing the DP more formally, a small digression is pursued in the following subsections to gain some intuition on how they operate.

4.2.1 Urns

In categorical-type problems, one may ask: What is the probability of observing a particular sequence of assignments? It turns out that such a sequence, in this case, happens to be a collection of random variables that can be represented by a stochastic process. More precisely, the sequence of assignments is the result of a *Pólya urn process* [123].

The Pólya urn process is a thought experiment that works as follows: First, an urn is filled with α number of balls of one colour (e.g. red) and β number of balls of a different colour (e.g. blue). Next, a ball is drawn at random from the urn. If the ball is red, it is placed back in the urn together with an additional red ball. Conversely, if blue, the ball is placed back in the urn together with an additional blue ball. This step is then repeated for another N number of draws. At any drawing instance n, the probability of the next ball being red can be determined by the predictive distribution defined by having a Bernoulli likelihood with a Beta prior. Details of this evaluation are provided in Appendix A.2.

What makes this process interesting is that the limiting proportion of balls in the urn is a random variable distributed according to a Beta distribution. Even more remarkable is the fact that this Beta distribution is parameterised by α and β; the initial number of red and blue balls, respectively. This statement can be visualised in Figure 4.5. The proof of this outcome is made possible by de Finetti's Theorem (4.25), stating that a

sequence of random variables $(x_1, x_2, ...)$ is exchangeable if and only if,

$$p(x_1, ..., x_n) = \int \prod_{i=1}^{n} p(x_i|\theta)\mathcal{P}(\mathrm{d}\theta) \tag{4.25}$$

for all n, and for some measure $\mathcal{P}$ on θ. An exchangeable sequence of random variables is one whereby the joint probability of its elements is the same for any fixed permutation of the sequence [124]. Equation (4.25) thus demonstrates that for a sequence of exchangeable data, an underlying set of parameters must exist, and are modelled according to some measure $\mathcal{P}$ (or distribution $p(\theta)$), such that the returned data are conditionally independent. Reviews of the theorem can be found in [125].

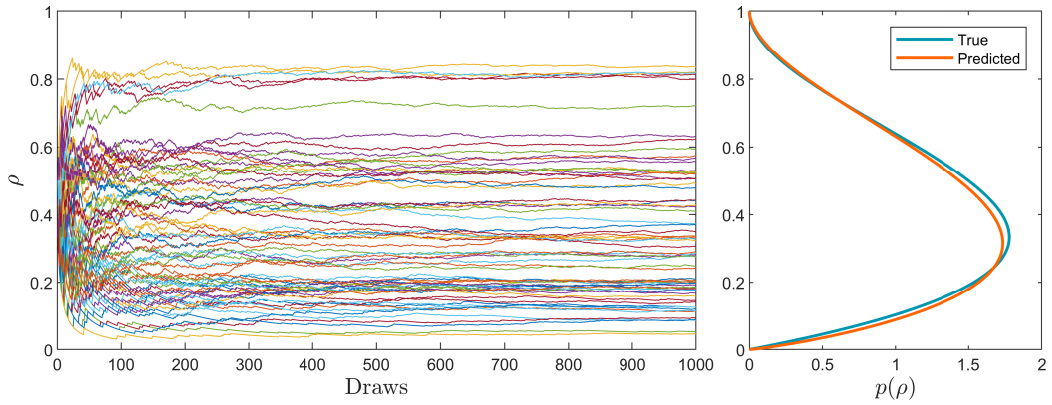


FIGURE 4.5: A series of sequences formed by 1000 samples drawn from a Pólya urn process. The limiting proportion of these sequences (left) is shown to approximate a Beta distribution parameterised by α and β (right).

In this case, the collection is the sequence of red and blue balls drawn from the urn, and the joint distribution of the sequence is the same irrespective of the order in which the balls are drawn. According to de Finetti's Theorem, this particular sequence is represented by a mixture of i.i.d. Bernoulli random variables, weighted by a Beta distribution with parameters α and β.

The notion of this process also applies when dealing with a set of multivariate random variables. In this case, the Pólya urn begins with a certain amount of K different coloured balls. The process is then constructed in the same manner, by drawing a ball and placing it back into the urn together with another ball of the same colour. The limiting proportion of balls as $N \to \infty$ is a random vector distributed according to a Dirichlet distribution, with parameters that correspond to the initial number of balls of each colour. Much like in the bivariate case, the probability of drawing a ball of a particular colour can be determined by evaluating the predictive distribution. The details of this operation are provided in Appendix A.3.

The concept of the DP can be envisioned by imagining an extensive palette containing a wide range of colours. This palette is so vast that it always allows for the creation of a ball with a brand-new colour. A stochastic process that accommodates this condition is defined as the *Hoppe urn* or the *Blackwell-McQueen urn* [126]. For the finite processes outlined above, the bivariate case has the Beta distribution as the underlying prior, while the multivariate case has the Dirichlet distribution. Similarly, the underlying prior distribution for an "unbounded" sequence, drawn from a Hoppe urn process, is a DP.

The Hoppe urn process generates such a sequence by initially placing one "neutral"-coloured ball in the urn. Whenever this neutral ball is drawn, it is returned to the urn together with an additional ball of a previously unseen colour. This process resembles that of the Pólya urn whenever an already coloured ball is drawn. Over time, as the number of draws increases, a preferential clustering effect emerges in a "rich-gets-richer" manner. Although the Hoppe urn is still somewhat of an abstract metaphor, the subsequent sections attempt to concretise the premise of this process.

4.2.2 Chinese Restaurant Process (CRP)

Much like the Hoppe urn, the *Chinese Restaurant Process* (CRP) [125] is a recurrent metaphor used to elucidate sampling from a DP. While both metaphors explain the same process, the CRP provides a scenario that may be more easily interpretable.

This imaginary scenario conveys a restaurant that can accommodate any number of customers, as it is unbounded by the number of available tables it can provide. A new customer may decide to sit at one of the occupied tables or request to be seated at a new empty table. The premise of the CRP is based on defining the probabilities of an incoming customer sitting at a particular table.

It is possible to quantify these probabilities by using k to index the tables, and assuming the first K tables to be occupied. Denoting the number of diners at the k^{th} table by c_k, the CRP defines the probabilities as,

$$p(k) = \begin{cases} \frac{c_k}{\alpha+N} & \text{for a currently occupied table} \\ \frac{\alpha}{\alpha+N} & \text{for a currently empty table} \end{cases} \tag{4.26}$$

where $N = \sum_{k=1}^{K} c_k$ is the total number of diners in the restaurant. α can be interpreted as a parameter that defines how likely it is for a dinner to choose an empty table. For example, high values of α would correspond to the case in which the restaurant is located in an area where the population is generally less sociable, whereby dinners often prefer

to eat alone, and thus more likely to request empty tables. Conversely, low values of α represent populations that enjoy socialising, and thus, diners are drawn to sitting at tables that are already occupied.

The correspondence between the CRP and the Hoppe urn becomes clear when taking α to be the mass of the "neutral"-coloured ball. If the probabilities of drawing balls from an urn are now defined by their mass, then higher values of α make it more likely to resample the "neutral" ball, leading to a sequence characterised by a wider spectrum of coloured balls. Resampling a new colour is analogous to requesting a new empty table. Similarly, the number of balls of a particular colour is analogous to the number of dinners sitting at a particular table.

As demonstrated in the following subsection, the intuition behind sampling from a DP is nicely represented by these metaphors.

4.2.3 A More Formal Definition of the DP

Thus far, some intuition has been provided in sampling categorical processes that are unbounded by the number of possible categories an observation can be assigned to. This process is precisely in line with that of drawing samples from a DP.

To formally define a DP, a continuous measurable space Θ, is first considered. The measure over Θ is quantified here by a PDF G_o, referred to as the original base distribution of the DP. Under these conditions, a measure is thus given by the probability of a value θ lying in a subset of Θ, bounded by b_1 and b_2. That is,

$$G_o(A) = p(\theta \in A) = \int_{b_1}^{b_2} G_o(\theta)\mathrm{d}\theta \tag{4.27}$$

where A is a subset of a partition over Θ, and $G_o(A)$ can be thought of as a function that assigns a probability to the set of values θ in A. The subsets of the partition cover all possible values of Θ, and do not overlap; therefore,

$$\sum_{m=1}^{M} G_o(A_m) = 1 \tag{4.28}$$

where the subscript m denotes the index of a subsection in the partition comprised of M subsets.

Now, for a subset A_m, $G_o(A_m)$ will always return the same probability, i.e. the function $G_o(A_m)$ is deterministic. If these probabilities are instead assumed to be a realisation of a stochastic process, then the corresponding distribution will be a DP, with base

distribution G_o. In other words, a random measure over G_o can be represented by the vector $G = \{G(A_1), \ldots, G(A_M)\}$, which is distributed according to a DP if G is distributed as a finite Dirichlet distribution with parameters $\{\alpha G_o(A_1), \ldots, \alpha G(A_M)\}$. That is, $G \sim DP(\alpha, G_o)$ if,

$$\{G(A_1), \ldots, G(A_M)\} \sim \text{Dir}(\alpha G_o(A_1), \ldots, \alpha G(A_M)) \tag{4.29}$$

where α corresponds to a scaling or concentration parameter. Sampling from a DP, however, is not entirely obvious by this definition.

Some clarity is provided when $\{G(A_1), \ldots, G(A_M)\}$ is thought of as a prior distribution over the probabilities of θ lying in one of the M subsections in Θ. Having observed a collection of values $\theta_1, \ldots, \theta_n$, these probabilities can be easily updated by the conjugacy between the multinomial likelihood and Dirichlet distribution. The posterior is thus,

$$\{G(A_1), \ldots, G(A_M)\} \,|\theta_1, \ldots, \theta_n \sim \text{Dir}(\alpha G_o(A_1) + c_1, \ldots, \alpha G(A_M) + c_M) \tag{4.30}$$

where c_m denotes the number of observations found in A_m. By the definition of a DP, the posterior distribution over G must also be a DP. The parameters of the posterior DP can thus be rewritten in terms of an updated base distribution $G_o^*(A)$ and an updated scaling parameter $\alpha*$. Concretely,

$$\alpha^* G_o^*(A_m) = \alpha G_o(A_m) + c_m \tag{4.31}$$

The updated scaling parameters can be obtained after taking the sum over all subsets in the partition,

$$\alpha^* \sum G_o^*(A_m) = \alpha \sum G_o(A_m) + N$$

Finally, because of Equation (4.28),

$$\alpha^* = \alpha + N \tag{4.32}$$

Plugging α^* back into Equation (4.31) yields,

$$G_o^*(A_m) = \frac{\alpha}{\alpha + N} G_o(A_m) + \frac{1}{\alpha + N} \sum_{n=1}^{N} \delta_{\theta_n}(A_m) \tag{4.33}$$

where c_m has been replaced with $\sum_{n=1}^{N} \delta_{\theta_n}(A_m)$, and δ_{θ_n} is equal to unity when $\theta_n \in A_m$ and zero otherwise. The DP posterior is thus expressed as,

$$G(A)|\theta_1, \ldots, \theta_n \sim \mathrm{DP}\left(\alpha + N, \frac{\alpha}{\alpha + N} G_o(A) + \frac{1}{\alpha + N} \sum_{n=1}^{N} \delta_{\theta_n}(A)\right) \tag{4.34}$$

The predictive distribution for θ_{n+1} can now be derived by marginalising G as follows,

$$\begin{aligned} p(\theta_{n+1} \in A_m|\theta_1, \ldots, \theta_n) &= \int p(\theta_{n+1} \in A_m|G(A_m)) p(G(A_m)|\theta_1, \ldots, \theta_n) \mathrm{d}G(A_m) \\ &= \int G(A_m) p(G(A_m)|\theta_1, \ldots, \theta_n) \mathrm{d}G(A_m) \\ &= \mathbb{E}[G(A_m)] \end{aligned} \tag{4.35}$$

which is the expected value of $G(A_m)$ with respect to the DP posterior. Since $G(A_m)$ is the m^{th} component of a Dirichlet, its expected value can be calculated by taking the corresponding Dirichlet parameter divided over the sum of the Dirichlet parameters,

$$\begin{aligned} \mathbb{E}[G(A_m)] &= \frac{\alpha^* G^*(A_m)}{\sum_{m=1}^{M} \alpha^* G^*(A_m)} \\ &= \frac{\alpha}{\alpha + N} G_o(A_m) + \frac{1}{\alpha + N} \sum_{n=1}^{N} \delta_{\theta_n}(A_m) \\ &= G_o^*(A_m) \end{aligned} \tag{4.36}$$

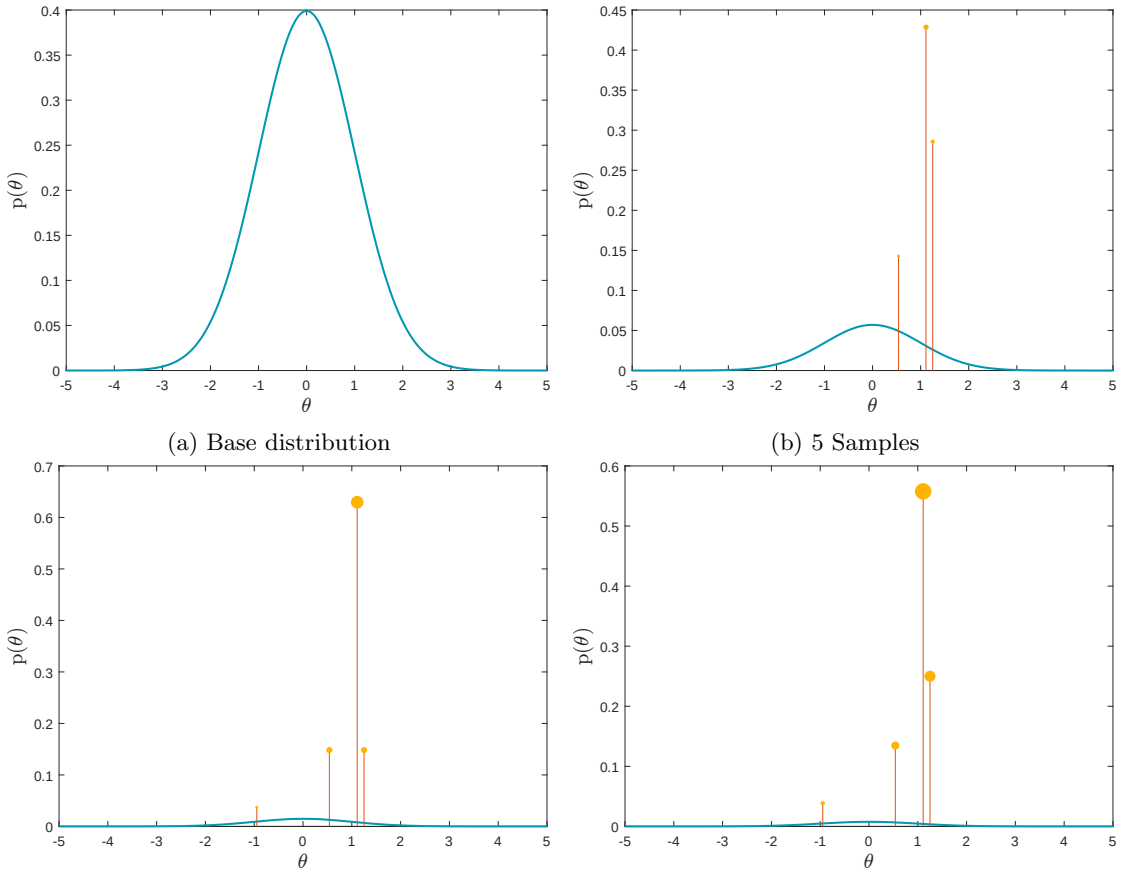
whereby the outcome of the predictive distribution equals the base distribution of the DP posterior. This derivation shows that sampling from a DP is thus carried out by sampling from its base distribution. Since the DP is a stochastic process, each sample will depend on the previous ones, corresponding to an update of the base distribution at each draw.

Sampling from $G_o(A)^*$ – and hence the DP – is now straightforward. The two components in Equation (4.33) are weighted by $\alpha/(\alpha + N)$ and $c_m/(\alpha + N)$, respectively, and one of these components is chosen proportional to their weights. If the outcome favours the first component, a sample θ is drawn from $G_o(A_m)$. Otherwise, θ is drawn from $Q(A) = \sum \delta_{\theta_n}(A)$. Because this term is a summation, it can also be thought of as a mixture of components. Choosing one of these components for sampling is done by picking any region containing an observation θ_n, proportional to the number of current observations in each of these regions. A new sample θ will be identical to θ_n.

Overall, a sample from a DP is drawn either from the base distribution with probability $\alpha/(\alpha + N)$, or by replicating one of the previous samples with probability proportional to the number of times it has been sampled already. This sampling process can be

visualised in Figure 4.6. A Gaussian distribution with zero mean and unit variance is shown as the original base distribution of the DP (Figure 4.6(a)). In the very first iteration, no samples exist, so the second term in (4.33) reduces to zero, and the only possible choice is to draw a random sample from the base distribution. As new samples are introduced, the weighting $c_m/(\alpha + N)$ becomes more prevalent in certain regions, and the chances of sampling within a region with an existing sample are greater. After drawing 50 samples, as shown in Figure 4.6(d), the second term in (4.33) eventually outweighs the first, and existing groups become imposing enough so as to grow at higher rates than any potentially-new group would.

By increasing α, the formation of new groups is promoted, as a bias favouring the first term in (4.33) is introduced. The result of sampling from a DP with different values of α is shown in Figure 4.7. It is clear that increasing α results in the formation of more distinct groups during the sampling process, as more weighting is provided towards sampling from the base distribution.



(a) Base distribution

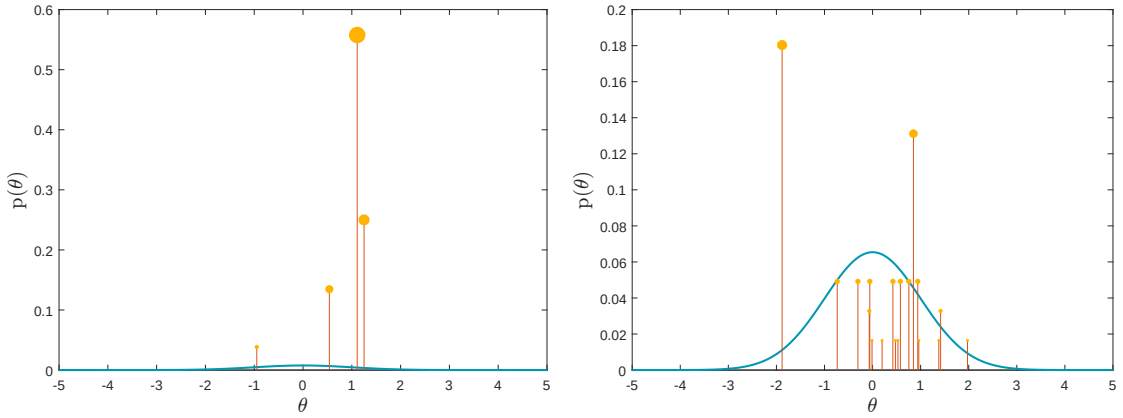
(b) 5 Samples

(c) 25 Samples

(d) 50 Samples

FIGURE 4.6: Sampling from a DP with $\alpha = 1$. The base distribution is a Gaussian with mean $\mu = 0$ and variance $\sigma^2 = 1$.

An important observation worth highlighting from this demonstration is that the DP permits sampling from the same values of θ more than once. The converging result of



(a) (b)

FIGURE 4.7: Outcome of a DP after drawing 50 samples with (a) $\alpha = 1$ and (b) $\alpha = 10$. The same base distribution is used in both cases; that is, a Gaussian distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.

this process is a vector of discrete measures, with each of its elements assigned to a unique sample drawn from the base distribution. This implication makes it possible to sample the same value more than once, even when modelled by a continuous probability distribution.

One may note parallels between this process and the CRP. The correspondence between them can be more easily envisioned by re-writting the base distribution of the DP posterior as,

$$\frac{\alpha}{\alpha + N} G_o(A) + \frac{1}{\alpha + N} \sum_{k=1}^{K} c_k \delta_{\theta_k}(A) \tag{4.37}$$

where c_k denotes the number of times a unique value of θ_k has been sampled; given that within the N current samples, there will be K unique values and $K \leq N$. Sampling from this expression is achieved by choosing whether to sample a new value with probability proportional to α, or to resample an existing value with probability proportional to the number of times it has been sampled already. These sampled values from the DP are equivalent to the tables set in the CRP, whereby a unique value θ_k would be assigned to each existing table. The value of α has the effect of promoting the generation of new values, which is precisely the same effect it had for drawing new tables in the restaurant. Similarly, the number of times a unique value is resampled is analogous to the number of dinners currently sitting at a unique table.

In a more pragmatic sense, the values of θ_k need not be arbitrary. In fact, these can be assigned to the parameters defining a collection of unique distributions in a mixture model. This notion is the premise of an infinite mixture model, as the generation of new values is boundless, with the added ability to resample from existing values more than once.

4.2.4 Stick-Breaking Process

The properties that define a DP are made explicit by the *stick-breaking* construction [127, 128], which is given as follows,

$$\begin{aligned} \rho_k &\sim \text{Beta}(1, \alpha) \\ \pi_k &= \rho_k \prod_{j=1}^{k-1}(1-\rho_j) \\ \theta &\sim G_o \\ G &= \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k} \end{aligned} \tag{4.38}$$

where G is a weighted sum of point masses distributed according to DP. The sequence of proportions π_k are samples from a GEM distribution, parameterised by α, i.e. $\boldsymbol{\pi} \sim$ GEM(α). Metaphorically, the stick-breaking construction provides some intuition for sampling from the GEM. An illustration of this process is shown in Figure 4.8.

At first, a proportion ρ_1 is broken off a unit-length "stick". This step can then be repeated indefinitely by breaking a new proportion ρ_k from the remaining length of the stick, i.e. $\prod_{j=1}^{k-1}(1-\rho_j)$. In this thought experiment, the end of the stick is never reached, resulting in an infinite sequence of proportions $\boldsymbol{\pi} = \{\pi_1, \dots\}$. The proportions broken off the stick at each iteration are randomly distributed according to a Beta distribution with parameters 1 and α. This condition is necessary for the sequence of Beta draws to converge to a GEM with parameter α. The size of the proportions depends on the value assigned to α; larger values favour smaller stick partitions, thereby promoting a higher number of small cuts, as opposed to ending with a few but larger chunks of the stick. It should be noted that a draw of proportions from a GEM is not equivalent to drawing samples from a DP. It is only by assigning these proportions to random samples drawn from a base distribution G_o, that makes G distributed according to a DP.

4.2.5 Infinite Mixture Model

The reasons for having a DP prior in the mixture model become evident when one is uncertain of the number of clusters to include. This scenario can be expected in many SHM applications, where it may be impossible to anticipate the possible number of states/conditions a structure may experience during its lifespan. Later in Chapter **??**, the use of DPs is explored in AE-based techniques used in SHM applications.

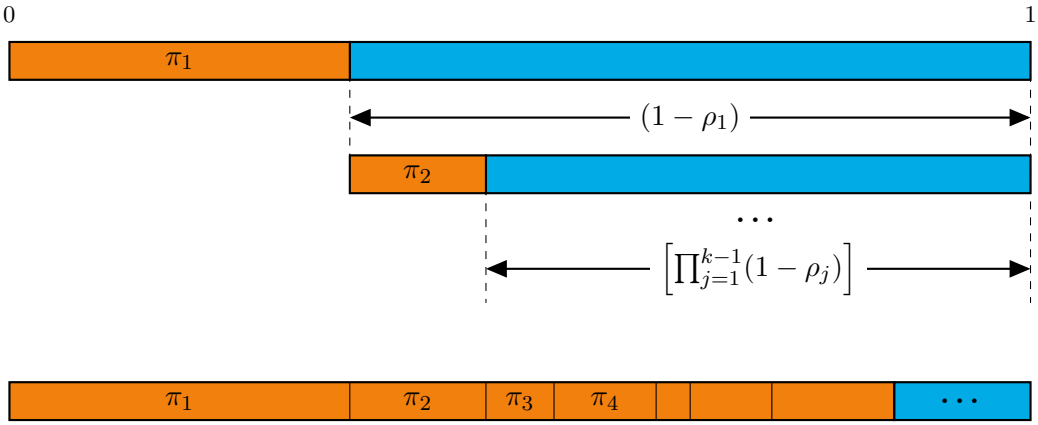


FIGURE 4.8: Stick-breaking construction.

An infinite mixture model can be constructed by allowing a set of N observations $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, assumed to derive from the following generative process,

$$\begin{aligned} G|\alpha, G_o &\sim \mathrm{DP}(\alpha, G_o) \\ \theta_n|G &\sim G \\ \mathbf{x}_n|\theta_n &\sim p(\mathbf{x}_n|\theta_n) \end{aligned} \tag{4.39}$$

where θ_n can now be interpreted as the parameters governing the independent distributions in the mixture model. By establishing a DP-prior in the mixture model, the actual number of clusters is no longer defined but is rather to be inferred directly from the data. The challenge is in inverting the generative process to infer the parameters of the mixture model after having observed the data. While this inversion is possible by employing Bayes' Theorem, the DP is often unlikely to be a conjugate prior, thereby necessitating approximation of the posterior. Methods developed specifically for this type of inference are based on either MCMC sampling [129] or VI [130].

4.3 Summary

In this chapter, a theoretical background regarding GPs and DPs was presented. The aims here were to introduce fundamental concepts underpinning these models and provide an intuitive understanding of how they operate via illustrative metaphors. In Chapters 5 and 7, the GP is employed to learn complex relationships in case studies involving damage detection and localisation. Meanwhile, Chapter 6 explores the use of DPs to develop a flexible clustering model, designed to simplify the analysis of AE data.

Chapter 5

An Exploratory Study for Enhanced Localisation Techniques for Rotational Systems and Structures

Despite only meeting the fundamental conditions governed by Rytter's hierarchy, a monitoring system merely designed to detect damage is already promising enough to prevent structural failures and provide means for an efficient maintenance schedule. Nonetheless, one may be interested, or even required, to know where damage might be present within the structure.

The reasons for this potential interest are, indeed, application-dependent, but are often motivated by defects that emerge in locations deep within the structure that cannot be easily accessed during a visual inspection. This premise also applies to systems comprised of various components, in which case one may want to know which component is faulty without having to disassemble the system in its entirety. For example, the inspection of a commercial jet engine can be planned more carefully if it is suggested by the monitoring system that the source of damage derives from the low-pressure compressor. With this information at hand, an engineer could then produce an efficient maintenance plan in which it is only required to isolate the low-pressure compressor from the rest of the engine, and thereby limit the search for the exact location of damage to this module, saving valuable time and resources in the process. It should be noted that this example grossly simplifies the attention that is required in the overall inspection of commercial jet engines, as damage in the frontal modules has the potential to propagate through to the rear of the engine, but the idea holds if monitoring systems are designed to aid the

engineers and technicians responsible for this task, with the added upside of potentially mitigating further costs.

The present chapter addresses this type of problem by means of statistical modelling. More specifically, two case studies are outlined in which localisation techniques are employed to conveniently diagnose the health-state of the system. The first of these case studies builds upon a comprehensive experimental procedure conducted in [131], where state-of-the-art ultrasonic methods are tested to accurately measure the fluid-film profile of a journal bearing *in-situ*. In the following section, details of this case study are outlined along with the proposed method for modelling the developing fluid film, which is then used to estimate the location of the bearing shaft under various operational conditions. The results offer the possibility to visualise the confidence of the predictions and allow the true location to be found within an area of high probability in the bearing's bore.

The second case study outlines a more direct approach to damage localisation by exploiting the propagation of AE waves in structures. This case study differs from the first in the sense that the system being studied is purely structural. In particular, AE-data recorded from a full-scale helicopter blade are processed to develop a probabilistic model designed to predict the location of potential sources of damage. The section covering this case study addresses the challenges related to the construction of ΔT maps and proposes a novel strategy for identifying optimal sampling points, eliminating the need for extensive data collection for training.

In both case studies, a GP regression strategy is utilised, predominantly adopting the techniques explored by Jones et al. [132]. The fundamental aim of this strategy is based on learning a functional mapping of location defined by a two-dimensional coordinate system, to a corresponding target value, i.e. $(x, y) \rightarrow z$. Given a training data set $D = \{(x_i, y_i), z_i\}_{i=1}^{N}$, the GP learns an interpolated representation of the target values over the continuous space that is covered by the coordinate system. Because the GP is employed in this regression exercise, the model returns a quantified uncertainty associated with the predictions of z_i. The location of interest can then be identified as the coordinates exhibiting the highest probability, which is achieved after evaluating the likelihood of z across the entire map.

5.1 Case Study 1: Journal Bearing Shaft-Centre Localisation

The journal bearing stands out as a specific type of fluid-film bearing extensively used in motors, pumps, turbines, and gearboxes [133]. All rotating machinery heavily depends on the smooth operation of their supporting bearings, and this holds true for the journal bearing as well. To clarify the motivation of the strategies taken in this section, it is first necessary to emphasise the contrasts between journal bearings and their rolling element counterparts in terms of developing condition monitoring systems. A simple diagram of these two types of bearings is shown in Figure 5.1.

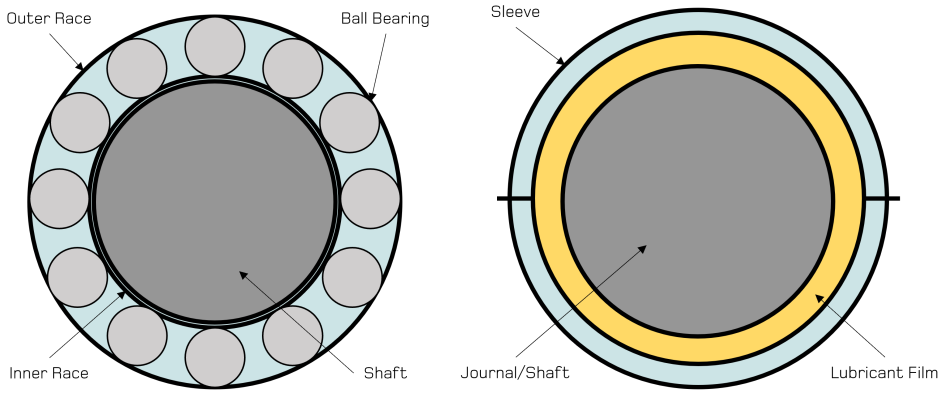


FIGURE 5.1: Diagram illustrating the cross-sections of (a) a rolling-element bearing and (b) a journal bearing.

Generally, the mechanisms governing acceleration signatures in rolling-element bearings are more comprehensible than those in journal bearings. For example, when local faults are present, the frequency spectrum of recorded signals should exhibit peaks at characteristic frequencies linked to rolling elements striking faults in the inner or outer race. The same effect may also arise when a fault in the rolling element strikes both the inner and outer race within a single basic period [134]. Given that the response is prompted by a series of periodic impulses, extracting the envelope of the carrier frequency can be accomplished by employing a Hilbert transform or a band-pass filter on the signal.

This envelope spectrum analysis aids the development of a CM system, offering insight into the frequencies affected by defects in the bearing [134]. Depending on the type of damage, the waveform might also exhibit varying amplitude modulation. For example, with inner-race damage, the peak amplitude will be highest when the damaged location aligns with the point of maximum load and gradually decreases as it rotates away from it [5]. These phenomena are distinctive to damaged bearings, to the extent that the

formulae for frequencies associated with each damage type can be defined, as provided in [135].

Since the load on journal bearings is entirely supported by the fluid film, which forms because of the lubricant being directed into a narrowing path by the rotating shaft, this process might not be directly applicable to journal bearings. In theory, local faults in a journal bearing are not expected to be captured by accelerometers in the same manner as in rolling element bearings. Instead of producing a sequence of periodic impulses, the fluid film in journal bearings might just flow over a local fault with minimal impact on the recorded waveform. Consequently, traditional vibration monitoring methods might not be well-suited for early fault detection in journal bearings.

Because of this consideration, and the intricate dynamics arising from a load-supporting film, the CM of journal bearings becomes a compelling subject, warranting investigation. In fact, extensive research on this topic can be found in the literature, with many studies proposing innovative AE-based monitoring techniques for detecting, and in some cases, identifying various failure modes that journal bearings might undergo, such as wear [136, 137], contamination [138, 139], seizure [140, 141], starvation [142], and vapour cavitation [41].

In the present work, however, an alternative approach is explored, where monitoring systems are developed based on ultrasonic measurements of the fluid-film thickness. Monitoring the fluid-film thickness as a primary feature is of interest because of its high dependency on the operating parameters of the bearing. Changes in the operating conditions cause the bearing shaft to adjust to a new equilibrium position, where internal hydrodynamic forces are in balance with applied loads [143]. Therefore, by monitoring the fluid-film thickness, it might be possible to ascertain the operating state of the bearing and assess whether such operation is within normal conditions.

Various methods have been proposed for measuring the fluid-film thickness, with some examples demonstrated to be reliable when based on electrical techinques [144, 145] or optical techniques [146, 147]. Although the predictions of these methods tend to be in good accordance with simulations and numerical solutions, they can be complex or rely on impractical requirements (e.g. invasive sensor placement or transparent surfaces for optical-based measurements). On the other hand, ultrasonic-based techniques have the potential to provide accurate measurements without the aforementioned disadvantages. This method of measuring thin fluid films has been successfully implemented by evaluating changes that ultrasonic pulses experience as they traverse a liquid medium between solid surfaces [148, 149]. The observed differences primarily stem from changes in amplitude and phase of the reflected pulses, encompassing the *spring model* for film measurements, as these are governed by the stiffness of the medium. The spring model

has limitations in the range of thicknesses it can measure, and an additional model, the *resonant dip technique*, is necessary to account for thicker layers [150]. The application of ultrasonic-based techniques has been extended to journal bearings and extensively explored for online fluid-film thickness measurements [131, 149, 151, 152].

While the results from these studies are promising, some limitations become evident by the constraints imposed by the amplitude-change method, phase-change method and resonant-dip technique. Fortunately, these limitations can be partially alleviated by complementing the measurements with a statistical model capable of estimating the thickness whenever these three methods fail to do so accurately. Before delving into how this enhancement can be achieved, a brief background in ultrasonic-based techniques is covered.

5.1.1 Ultrasonic-Based Techniques for Measuring Oil Film Thickness

During operation, pulses of ultrasonic waves traverse the oil film as the shaft spins, with some of its energy bouncing off the film and adjacent surfaces. The reflected waves are then picked up by the same emitting sensors, and their signals are processed to identify differences from the originals. One of these changes may be noted as a reduction in amplitude, provided by the energy loss that is transferred to the lubricant. The idea is to somehow relate this measured change in amplitude to the actual thickness of the film. This relation is achieved by defining the reflection coefficient R, which is given by the ratio of the reflected signal amplitude to the incident signal amplitude, together with the acoustic properties of the bearing and lubricant film. The result is an expression,

$$h = \frac{\rho c^2}{\Omega z_1 z_2} \sqrt{\frac{R^2(z_2 + z_1)^2 - (z_2 - z_1)^2}{1 - R^2}} \tag{5.1}$$

where h is the fluid-film thickness, ρ is the lubricant density, c the velocity of sound in the lubricant, Ω the angular frequency of the ultrasound wave and z_1 and z_2 are the acoustic impedances - or resistance to the wave's propagation - of the materials on each side of the lubricant film.

The second feature that can also be monitored is the phase change of the reflected signal. Following a similar procedure as before, but this time relating the phase of the reflected wave ϕ_R to the acoustic properties, yields another equation for the fluid-film thickness,

$$h = \frac{\rho c^2 (\tan \phi_R)(z_2^2 - z_1^2)}{\Omega z_1 z_2^2 \pm \sqrt{(\Omega z_1 z_2^2)^2 - (\tan \phi_R)^2 (z_2^2 - z_1^2)(\Omega z_1 z_2^2)^2}} \tag{5.2}$$

Both methods are limited by the range of boundary lengths they can cover. As the thickness increases, the system is said to be no longer stiffness dominated [131], and the reflection coefficient R tends to unity for the amplitude-change Equation (5.1), while ϕ_R tends to zero for the phase change Equation (5.2). The calculated values of h are clearly indefinite at such limits, thereby making these methods valid only for small boundary lengths.

The measurable range of fluid-film thickness can be extended by incorporating multiple sensors of varying frequencies but at the expense of higher hardware complexity in the ultrasonic system. An alternative to this problem is by means of a third method; the resonant dip method, which essentially involves tracking amplitude "dips" that manifest in the frequency domain when the lubricant film resonates under the excitation of the incident wave [153]. This method, however, is also limited by its measuring range, as higher-order dips often blend with background noise.

5.1.2 Experimental setup

The data examined in this case study correspond to the measurements obtained from the experimental procedure in [131], which was originally conducted by Dr. S. Beamish from the University of Sheffield. The methodology employed involved the design of a bespoke test rig, designed specifically to measure the fluid-film thickness of an operating journal bearing. A schematic is shown in Figure 5.2.

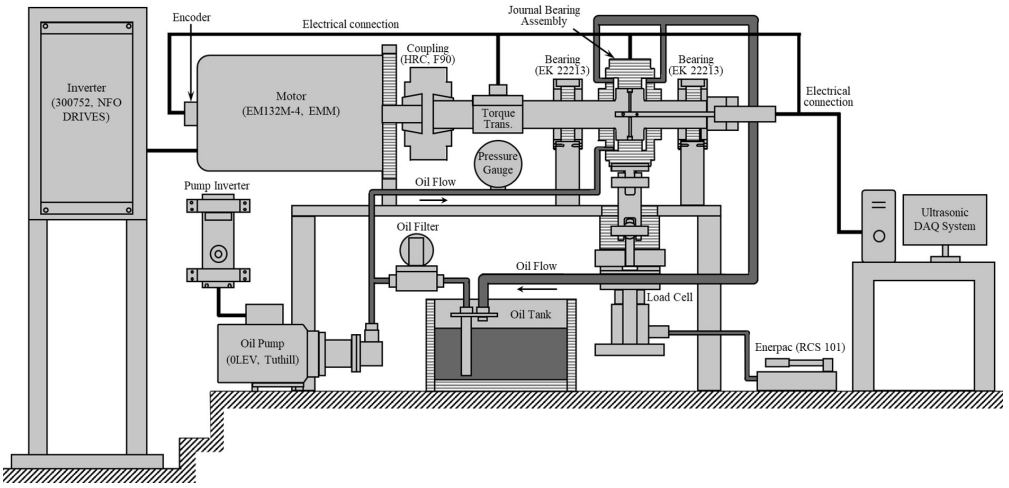
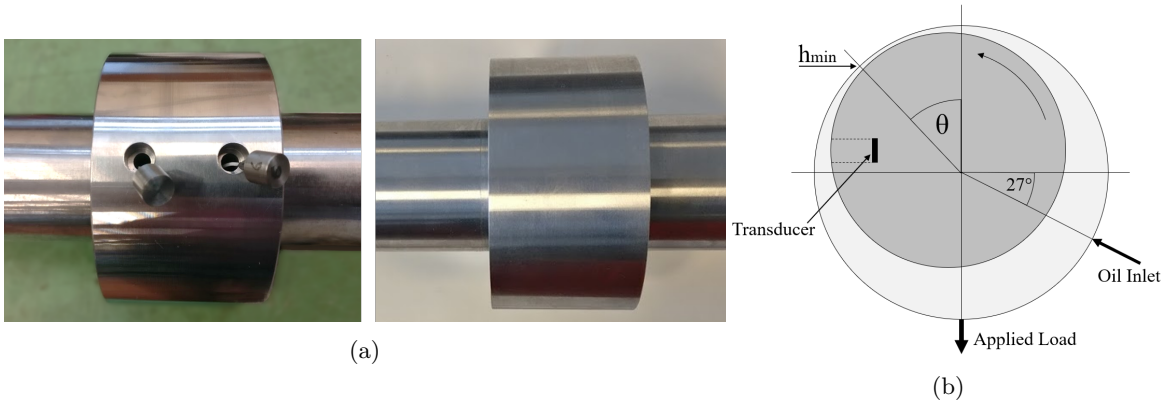


FIGURE 5.2: Schematic of journal-bearing test rig. From Beamish et al. [131].

In this configuration, rolling-element bearings were included in the rotor and placed on either side of the journal bearing to adjust the alignment of the shaft. A flexible linkage allowed the rest of the bearing assembly to move freely and adopt its equilibrium position. Controlled variations of the rotational speed, static load and inlet oil temperature

were possible in this setting. An important consideration to highlight at this point is that the static load is applied downwards via a hydraulic actuator. The lubricant was stored in a heated bath as part of a continuous circulating system of a fully-flooded bearing assembly. Six longitudinal ultrasonic transducers were embedded within the shaft, feeding to the data acquisition hardware at a rate of 80kHz via a multi-channel slip ring. Other measuring equipment were also installed, such as thermocouples to monitor oil temperature, load cells to measure the applied load, encoders to keep a record of the rotation angle, and four eddy-current gap sensors for indirect fluid-film thickness measurements.

The novelty of this rig lies in the set of ultrasonic emitters and receivers being embedded in the shaft, allowing for measuring the fluid-film thickness all around the bearing, rather than at a fixed point. This concept is shown in Figure 5.3.



(a) (b)

FIGURE 5.3: (a) Photographs of shaft during pin installation and after grinding. (b) Schematic of bearing cross-section including position of oil inlet. From Beamish et al. [131]

The interested reader is encouraged to refer to the original source ([131]) for the full details regarding the experimental test rig.

5.1.3 Learning Strategy

The fluid-film measurements of the bearing operating at 400rpm and subjected to a load of 20kN are shown in Figure 5.4. At first glance, the aforementioned limitations of the methods are evidenced by gaps in the measurements; the ultrasonic transducer is incapable of capturing enough information in these regions to derive meaningful results. Although the given results qualitatively appear to agree well with the theoretical solution (Raimondi-Boyd method [154]), these "dead zones" might restrict the derivation of other physical effects, such as the film pressure distribution. These gaps motivate the use of a GP regressor to sensibly interpolate the available data.

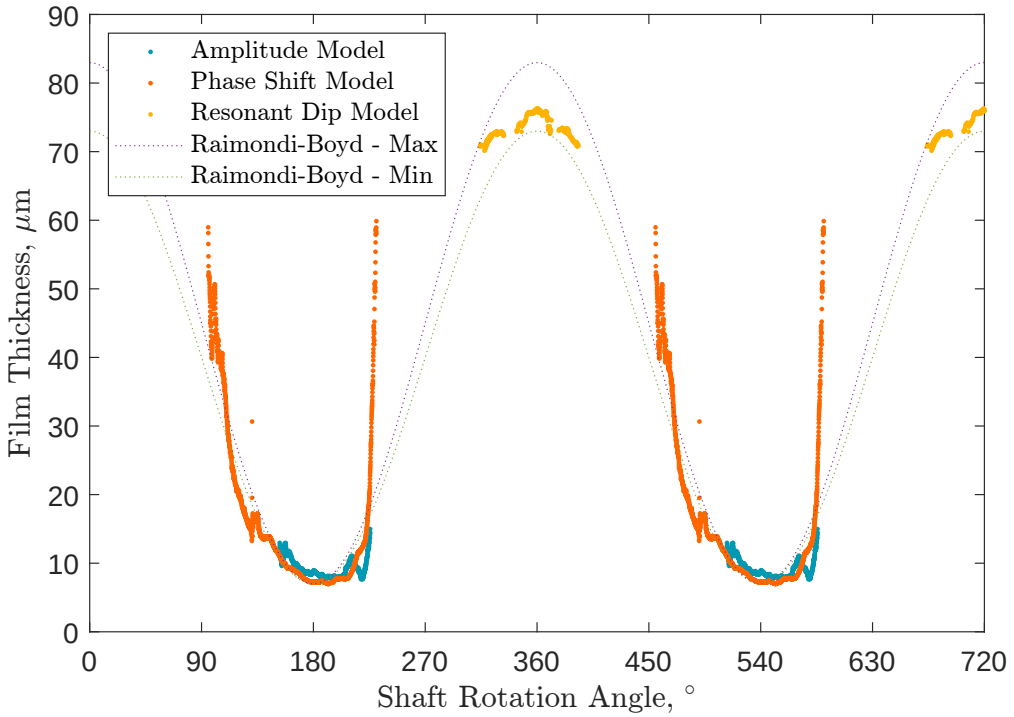


FIGURE 5.4: Experimental results where the measurements obtained from the amplitude-change model, phase-change model and resonant-dip technique are displayed. (Reproduced from [131])

Having the collection of measurements at hand, it is possible to estimate what the fluid-film thickness may be at all points around the bearing. One could, for example, assume a deterministic framework and assign the minimum value data point to be the true minimum fluid-film thickness h_{min}. The remaining values of h can then be obtained geometrically via,

$$h = c(1 + \epsilon \cos(\Theta)) \tag{5.3}$$

where c is the bearing radial clearance, Θ is a polar coordinate about the shaft-centre, and ϵ is the eccentricity ratio defined as $\epsilon = \frac{e}{c} = \frac{r - h_{min}}{c}$, with e being the off-set distance of the shaft-centre from the bearing centre. However, adopting a deterministic framework implies absolute certainty that the bearing is perfectly circular, and that the observed experimental results are true, which is unlikely to be the case given the inevitable presence of noise and uncertainties in the measurements. The GP is thus employed to map a set of observations (shaft rotation angle) to a corresponding vector of targets of the fluid-film thickness measurements. The implementation of a Bayesian approach allows for the prediction of the fluid-film thickness based on the observed data, its noise, uncertainty and any informed beliefs of the underlying physics.

For this particular application, it is convenient to introduce a covariance function that operates in polar domains [155]. Stationary covariance functions are generally computed with respect to the Euclidean distance $||x - x'||$, which underestimates the influence

of points located on the same concentric circles. A covariance function that satisfies the conditions of semi-positive definiteness over the polar domain is the C^2-*Wendland function* [155],

$$W_c(t) = \left(1 + \tau\frac{t}{c}\right)\left(1 - \frac{t}{c}\right)_+^{\tau}, \qquad c \in]\,0, \pi\,], \tau \geq 4 \tag{5.4}$$

where t is the angular distance between points x and x', τ is a steepening parameter, and $c = \pi$ if the measured distances are geodesic. The conditions in Equation 5.4 are necessary for the covariance to be zero at $t = \pi$ and also strictly positive when $t > \pi$.

The influence of the polar kernel can be combined with an Euclidean kernel to achieve a more accurate representation of the bearing. This operation is conducted using an ANOVA combination,

$$k_{2D}(x, x') = s^2(1 + \alpha_1^2 k_{rad}(\rho, \rho'))(1 + \alpha_2^2 k_{ang}(\theta, \theta')) \tag{5.5}$$

where the radial and angular coordinates of x are now represented by ρ and θ, respectively. The corresponding radial and polar covariance functions are denoted by k_{rad} and k_{ang}, where $k_{ang}(\theta, \theta') = W_\pi(d(\theta, \theta'))$.

Recalling the mathematical framework for GPs presented in Chapter 4, the following subsections outline the strategies conducted in the construction and implementation of the models for fluid-film prediction and shaft-centre localisation.

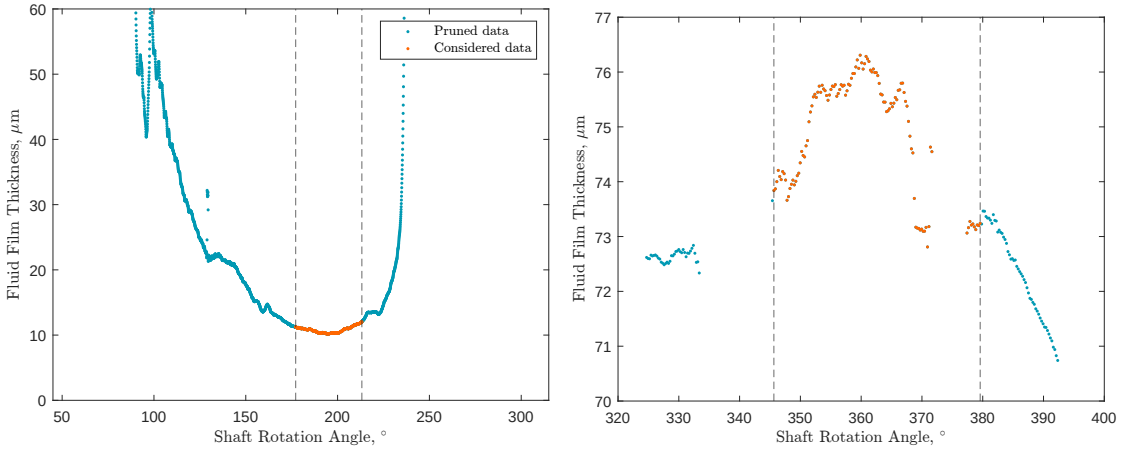
5.1.4 Modelling Journal Bearing Fluid-Film Thickness

To illustrate, measurements obtained under a specific operational condition are initially taken into consideration. These measurements correspond to the ultrasound measurements gathered while the bearing was operating at a constant speed of 400rpm and subjected to an applied static load of 20kN. The outcomes are depicted in Figure 5.4, illustrating the estimates of fluid-film thickness provided by the amplitude model, phase model, and resonant dip method.

While directly constructing a GP to fit the observations is feasible, preprocessing the observations is necessary to eliminate potentially misleading measurements. This step is warranted as the ultrasound technique tends to be more dependable in converging regions near the minimum point. Specifically, in diverging regions, cavitation causes the film to rupture, distorting the measurements collected from these areas [156]. This phenomenon is evident in Figure 5.4, where the phase model predictions converge to

unrealistic observations as the sensor readings move away from the point of minimum thickness. Similarly, the amplitude model yields an additional dip that appears to challenge the cylindrical geometry of the bearing. Neglecting these inconsistent observations could adversely affect the predictive capabilities of the GP model.

Therefore, only data contained within a small region about the minimum fluid-film thickness are taken into account. The selected data region for this analysis encompasses a 10% span of a full cycle; that is, 18° on each side of the minimum point (Figure 5.5). The Raimondi-Boyd (min) solution is employed as a reference to estimate the angular position of the minimum fluid-film thickness. The same procedure is applied to the observations obtained via the resonant dip technique.



(a) (b)

FIGURE 5.5: Data-selection from (a) the phase change method and (b) the resonant dip method.

The resulting data set comprises a combination of the observations from the phase change method with those from the resonant dip method, which are then partitioned into a training and testing set. Indeed, the data selection process is based on a predefined range, potentially delimiting the most consistent observations. This factor unavoidably influences the learning process, and a form of validation is essential to ensure that the considered portion optimises the performance of the model. Although such validation steps are not expounded upon in the current work, they remain a topic for future exploration. Nevertheless, even with the reduced data set, meaningful outcomes are achieved, effectively showcasing the intended concepts of this section.

A variety of different covariance functions is worth evaluating to find the one that is better suited for the current application. In this case, three covariance functions are used to construct each corresponding GP: (a) a squared-exponential function, (b) a Matérn 3/2 function and (c) a strictly-periodic function. The definition of these functions are recalled from Chapter 4 as,

$$k_{sqe}(x, x') = \sigma_f^2 \exp\left(\frac{r^2}{2l^2}\right) \tag{5.6}$$

$$k_{3/2}(r) = \sigma_f^2 \left(1 + \frac{\sqrt{3}r}{l}\right) \exp\left(-\frac{\sqrt{3}r}{l}\right) \tag{5.7}$$

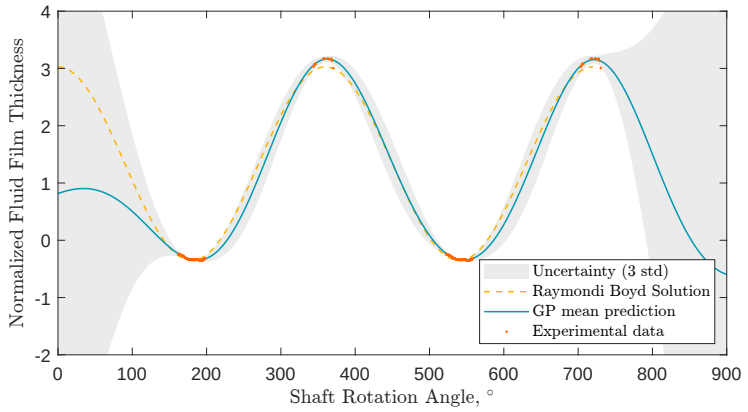
and,

$$k_p(x, x') = \sigma_f^2 \exp\left(-2\frac{(\pi|x - x'|/p)}{l^2}\right) \tag{5.8}$$

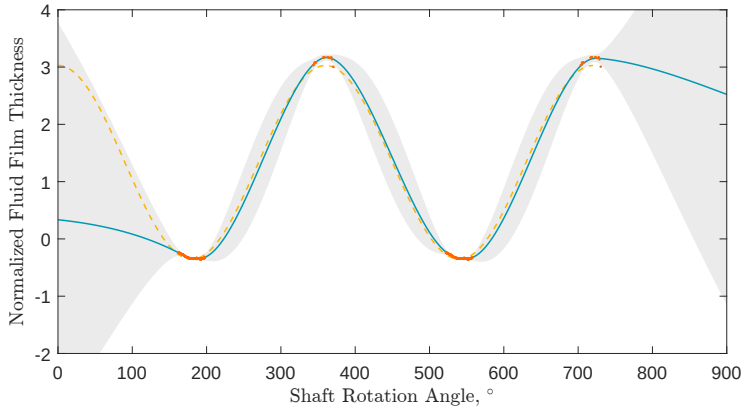
respectively, where $r = ||x - x'||$. The (hyper)parameters σ_f^2 and l represent the signal variance and length scale, respectively, and the new parameter p in equation (5.8) simply determines the period of the function. The squared-exponential function and the Matérn function are common choices when one is uncertain about the characteristics of the underlying regressor. If enough data are provided, the fit provided by these functions should approximate closely to the best-fitting regressor. Since the available data in this experiment were limited, it was perhaps necessary to examine a covariance function that could embody the periodic nature of the data, and for this reason, the strictly-periodic function was included in the model selection process.

The resultant models for each data set are shown in Figure 5.6. Unlike the first two covariance functions, the periodic function maintains the periodicity of the GP beyond the range of observations. This outcome aligns with the physical attributes of the bearing operating at a constant speed. Additionally, after evaluating the log marginal likelihood of the testing sets in relation to each GP, it becomes evident that the periodic prior appears to be the one that best supports the new evidence. Given these advantages, the GPs constructed for all other operational conditions can be defined by the periodic covariance function.

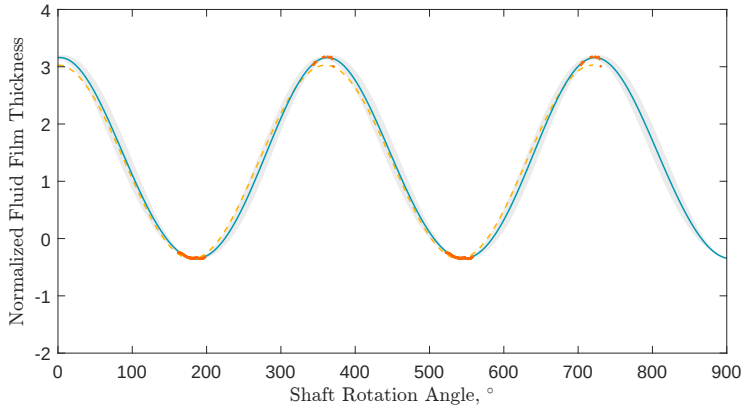
It is worth noting that the GP interpolates the film measurements over the full range of the bearing. All three GPs appear to encompass the Raimondi-Boyd solution within their delimited confidence bounds. This interpolation is not based on the observations of the minimum fluid-film thickness only, but it also takes into consideration all the other observations in the training set. By leveraging the available measurements, the GP can probabilistically predict the actual film thickness for all shaft rotation angles. Another noteworthy observation is the confidence the model has in making predictions, which decreases in areas where observations are unavailable. This outcome comes as no surprise, as the model provides estimates in areas lacking empirical evidence. Nevertheless, the GP posterior mean indicates where observations are most likely to exist, allowing one to



(a)



(b)



(c)

FIGURE 5.6: GPs over phase change fluid-film thickness observations defined by the (a) squared-exponential function, (b) Matérn 3/2 function and (c) strictly periodic function. Log marginal likelihood evaluated on test data resulted in (a) 592.46, (b) 578.14 and (c) 605.16 for each model, respectively.

use the Maximum-a-Posteriori (MAP) estimate to model the fluid-film thickness. Alternatively, the uncertainty in the predictions can be propagated through further analysis to estimate some bearing dynamics with a quantified measure of confidence.

In a CM context, for example, training a GP to learn from measurements corresponding to a healthy bearing could help discern whether new observations derive from a normal

operating condition. If wear or misalignment were to emerge during operation, the ensuing observations would somewhat deviate from normal, consequently diminishing their likelihood in relation to the previously-inferred model. This notion is pursued in the subsequent subsection with the integration of the C^2-Wendland function to predict the shaft-centre location in the bore.

5.1.5 On the Use of GPs for Shaft-Centre Localisation

The shaft-centre location can be defined by two coordinates: its eccentricity ratio, ϵ, and attitude angle, Φ. These coordinates represent the radial and angular components of the shaft-centre location with respect to the absolute bearing centre and load vector, respectively. It is possible to extract these coordinates directly from the fluid-film thickness observations. This approach involves finding the minimum value of the observations and its corresponding polar coordinate with respect to the bearing's *Top Dead Centre* (TDC). However, one would assume a deterministic framework by deeming the minimum point observation to be true. As mentioned above, assuming absolute certainty in the observations would imply perfection and thus lead to inaccurate predictions of the shaft-centre location. Therefore, the probabilistic framework described in the previous subsection was preferred to establish the most likely location of the shaft when given a specific operational condition.

The series of experimental runs here involved either changing the applied load W or rotational speed ω in controlled steady-state environments (e.g. constant oil temperature and supply). Each of the data sets was then modelled with a GP defined by the periodic covariance function, followed by the extraction of the MAP estimates of h_{min}, and their corresponding angular positions in terms of the attitude angle. Figure 5.7 shows the coordinates of the shaft centre inferred from each speed-load combination. Although unusual, the shaft found itself at the top half of the bearing throughout the experimental procedure. This was expected as the bearing assembly adjusts to the build-up of pressure produced in the film by reacting to the load being applied downwards on the housing (while maintaining the shaft fixed). This setup is equivalent to having the bearing assembly fixed and the loaded shaft free to adjust to its stable position.

At this point, it was necessary to assign labels to each location in order to construct a data set that could be used to map the shaft-centre location from its respective operational parameters. A convenient combination of these parameters exists via the Sommerfeld number [157], defined by,

$$S_o = \left(\frac{R}{c}\right)^2 \frac{\mu\omega LR}{2W} \tag{5.9}$$

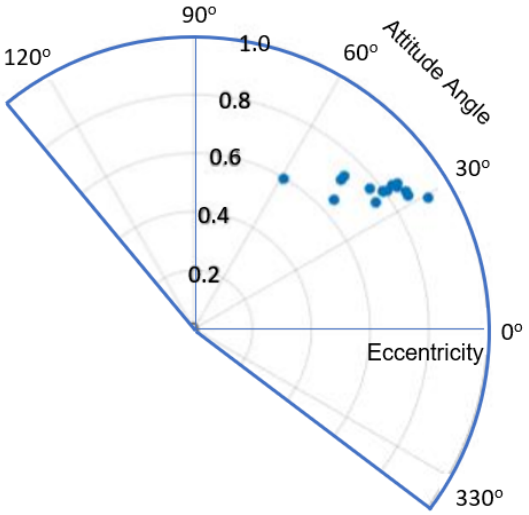


FIGURE 5.7: Training data set including all shaft-centre locations.

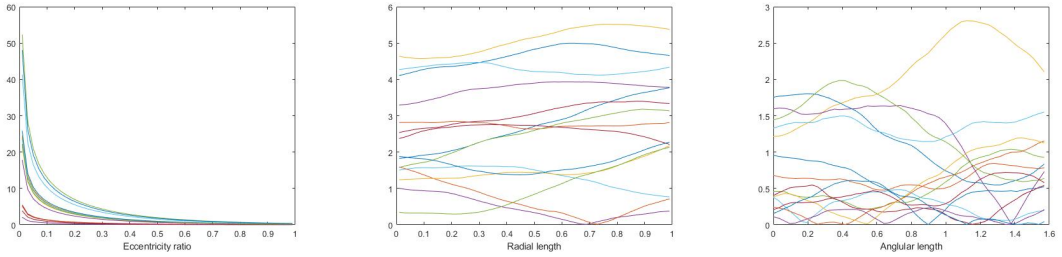
where R and L are the internal radius and length of the bearing, respectively, and μ is the lubricant dynamic viscosity. The Sommerfeld number has been commonly used to compare the performance of journal bearings [143]. It should be noted that the rotational speed and applied load variables in Equation (5.9) are combined by the ratio ω/W. Given that the experimental procedure was performed in a controlled steady-state environment, the viscosity term is assumed constant. Hence, each location can be labelled by a simplified Sommerfeld number defined only as a proportion of their corresponding speed-load ratios. This assumption allows one to have a training set $D = \{X_i, y_i\}_{i=1}^{N}$ of N observations where X_i is the coordinate (ρ_i, θ_i) of the target $\mathrm{t}_i = \omega_i/W_i$. Since the bearing is analysed in the polar domain, the coordinates are expressed in terms of a radial distance ρ_i and angular position θ_i from the bearing's centre and TDC, respectively.

Having established a data set that relates the shaft-centre position to the corresponding operating parameters, it is possible to develop a new GP regression model aiming to capture this relationship at all locations that the shaft might adopt. This procedure involves placing a grid over the bearing's cross-section. The nodes in the grid correspond to the predicted values of the model based on the training set D and any informed beliefs about the function being modelled. These considerations are managed with a suitable choice of covariance functions that best describe the underlying behaviour of the journal bearing.

Indeed, the polar nature of the bearing analysis means that a unique Euclidean-based covariance function would underestimate the influence of changes in the angular direction, and disregard the circular geometry of the bearing. A more prudent approach was thus followed with the inclusion of a polar covariance function k_{ang} defined by the

C^2-Wendland function (Equation (5.4)) for geodesic distances. Two other covariance functions were also considered for the radial space and then combined with k_{ang} via an ANOVA operation; these are the polynomial exponential-decay function (k_{pol}) and the Matérn 3/2 function ($k_{3/2}$); each combination will be referred to as *Model A* and *Model B*, respectively.

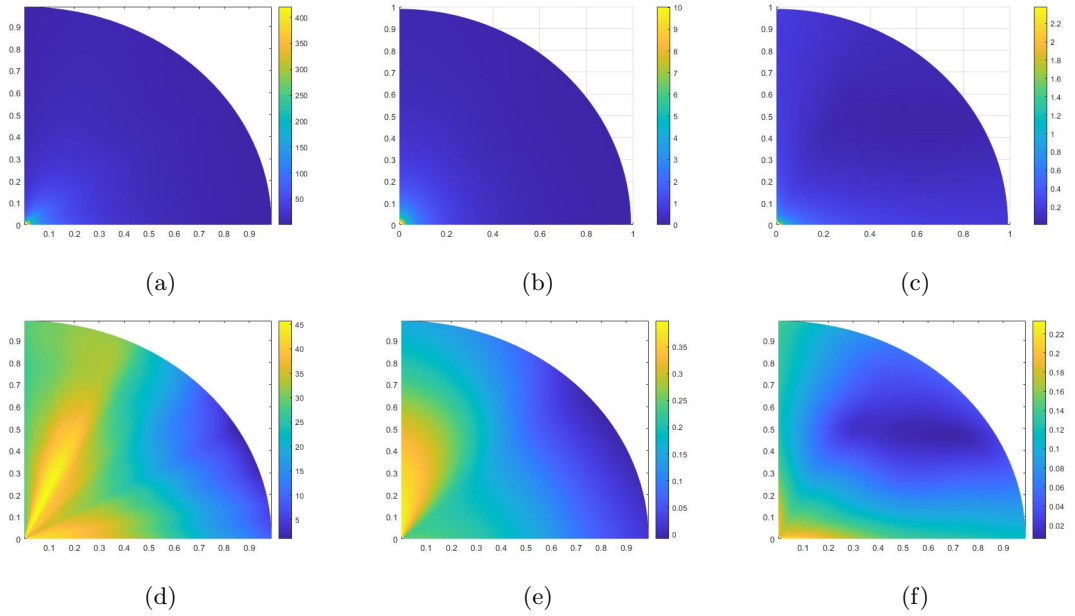
The former radial covariance function (k_{pol}) is justified when considering the expected behaviour of the journal bearing under varying operational conditions. Increasing the rotational speed results in the journal adjusting towards the centre of the bearing. In theory, given a steady-state fixed load operation, the journal will be concentric to the bearing when $\mathrm{t} \propto \omega \to \infty$ as $\omega \to \infty$. Conversely, increasing the load while maintaining a constant speed will have an opposite effect and $\mathrm{t} \propto 1/W \to 0$ as $W \to \infty$. An exponential decay function will somewhat model these characteristics and is thereby expected to provide a more realistic predictive framework. Meanwhile, the Matérn function is elicited to compare the outcome of a somewhat informed covariance function (i.e. the radial covariance function), to another that one would employ if unaware of the physics involved. Figure 5.8 shows some random samples from GP priors defined by each of the individual covariance functions.



(a) (b) (c)

FIGURE 5.8: Samples drawn from zero-mean GP priors defined by: (a) a polynomial exponential-decay kernel, (b) a Matérn 3/2 kernel, and (c) C^2-Wendland kernel. Each line represents a different random function sampled from the GP prior. The vertical axis represents the output of these functions when presented with values from the input space.

Having defined the priors, the mean and variance of the GP posterior can be inferred with the introduction of the training set D. The results are shown in Figure 5.9, showing both the mean and variance of GP prior and posterior corresponding to each model. Only the first quadrant of the bearing's cross-section was modelled since the shaft was found to have positioned itself only within this area in all experiments. This choice was also convenient to minimise computational expenses. Much like in the previous section, the models (more evidently displayed by Model B), are less confident in regions away from the observations.



(a) (b) (c)

(d) (e) (f)

FIGURE 5.9: (a)-(c) GP prior sample, GP posterior mean and variance defined by Model A. (d)-(f) GP prior sample, GP posterior mean and variance defined by Model B.

One may note that the predictive distributions do not explicitly estimate the location of the journal centre. Fortunately, having assigned speed-load ratio predictions to each grid point (ρ, θ), a coordinate X that maximises the probability of observing a new speed-load ratio t_{new} can be found. Inverting the problem in this way means assessing the log-likelihood of a new observation, t_{new}, as,

$$\log p(\mathrm{t}_{new}|D, X_*) = -\frac{1}{2}\log \mathbb{V}[\mathbf{f}_*] - \frac{(\mathrm{t}_{new} - \mathbb{E}[\mathbf{f}_*])^2}{2\mathbb{V}[\mathbf{f}_*]} - \frac{1}{2}\log 2\pi \tag{5.10}$$

where X_* is an array comprised of all the candidate coordinates where the log-likelihood of t_{new} is evaluated, and $\mathbf{f}_*$ denotes the GP predictions in relation to X_*. The results are shown in Figures 5.10 and 5.11, displaying the likelihood map of t_{new} at all grid points across the bearing section. One can thus infer the journal position by locating the point of maximum likelihood in the grid, which is provided with an associated uncertainty in the prediction.

It is important to note that the model simply attempts to predict the mean position of the shaft centre, given the initially-presented observations. In this case study, there is no reason to believe that any form of elliptical orbiting occurred. However, if some form of elliptical motion exists, and enough observations are provided, then the GP model should account for the added uncertainty given by the variation of the shaft displacement about its centre. If this were to be the case, then the estimated location should indicate

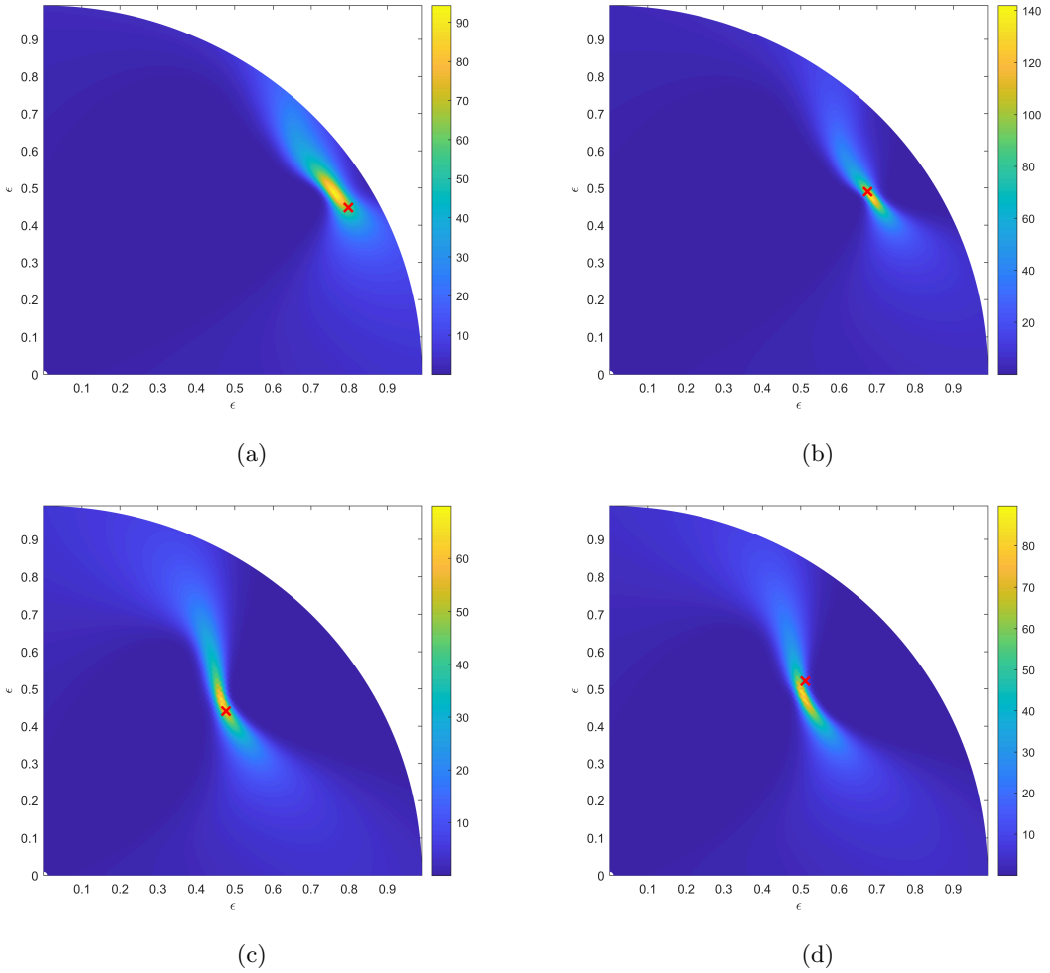


FIGURE 5.10: **Model A** likelihood maps of bearing shaft-centre location. Each figure corresponds to an arbitrary test-point: (a) 100rpm/10kN, (b) 400rpm/14kN, (c) 400rpm/4kN and (d) 800rpm/10kN. The actual measurements are indicated by a red cross.

the area covered by the orbital motion of the shaft centre, but predicting the orbital path may require a different approach in the learning process.

Another point that should be mentioned, is the limitation of having the rotational speed and static load as the only two input parameters of the model. The journal bearing was studied in a controlled environment, and it would be unrealistic to expect an identical behaviour in a practical application. In order to improve the robustness of the GP model, one would need to consider other important parameters, such as dynamic loads, lubricant temperature variations, elastic deformations and surface roughness, among others. This work leaves room for these considerations and will be accounted for in future experiments.

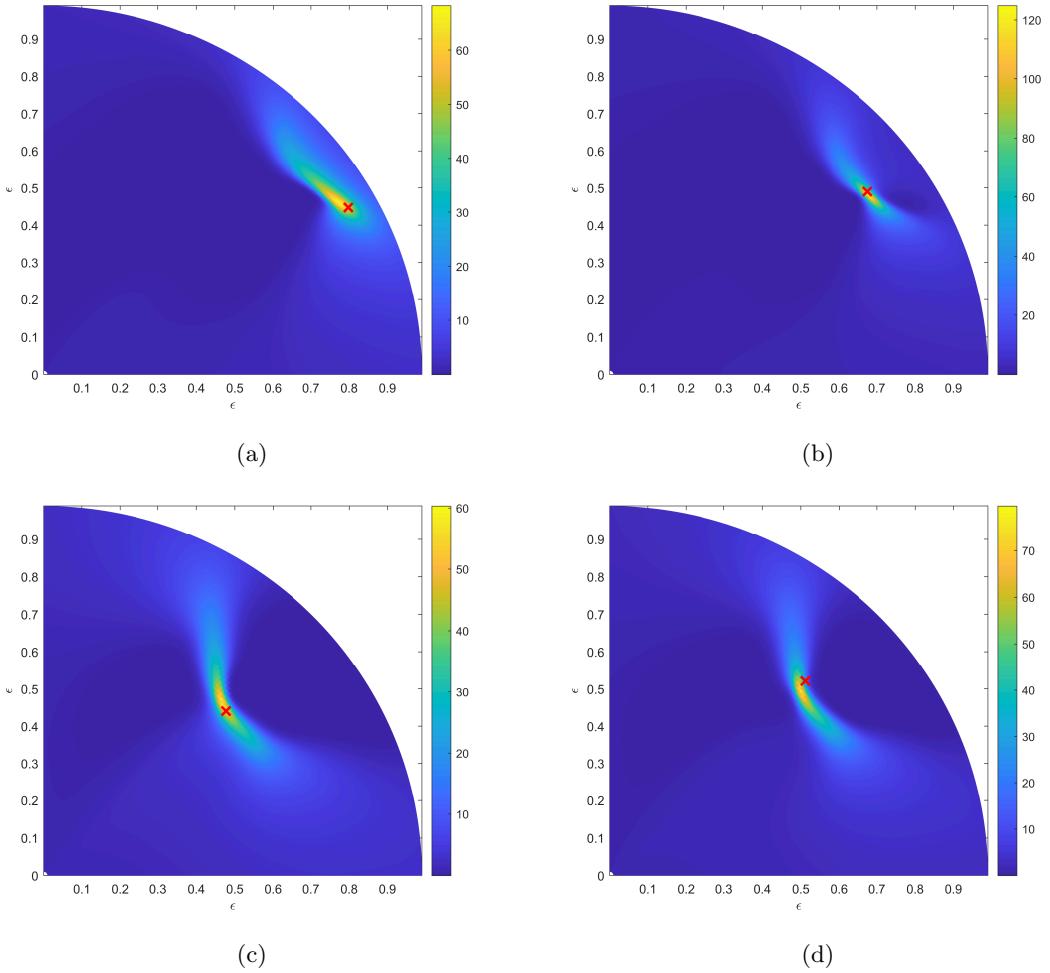


FIGURE 5.11: **Model B** likelihood maps of bearing shaft-centre location. Each figure corresponds to an arbitrary test-point: (a) 100rpm/10kN, (b) 400rpm/14kN, (c) 400rpm/4kN and (d) 800rpm/10kN. The actual measurements are indicated by a red cross.

5.1.6 Performance Evaluation

Because of the small available data set, splitting the observations into a training and test set could have limited the learning process. Nevertheless, alternative validation techniques exist to account for small data sets, such as *k-fold cross-validation* and *leave-one-out cross-validation* (LOOCV) [158]. Given that the number of observations was particularly small in this work ($N = 15$), the latter approach was adopted.

The LOOCV approach involves splitting the whole data set into K parts equal to the total number of observations (i.e. $K = N$). The model is then trained on $K - 1$ parts and tested with the remaining one. This process is repeated with a newly-assigned test point until all K parts have been a test set. The calculated errors in each iteration are then averaged to give a final performance evaluation. Although training the model

repetitively can be computationally expensive, a small data set compensates for this inconvenience.

The localisation algorithms developed in the previous section were based on different prior beliefs. The small number of observations makes the choice of a suitable GP prior even more important. One can make use of the marginal likelihood to compare how likely unobserved data (t_{new}), are given the choice of prior belief. The errors in the validation process were hence assessed by calculating the likelihood of each test set with Equation (5.10). The results shown in Table 5.1 demonstrate that the more informed prior leads to a higher overall likelihood of the test data.

TABLE 5.1: Performance evaluation results of GP models indicating the averaged marginal likelihood and root-mean-square-error (RMSE). The results are averages of the error metrics derived from the test sets assigned at each iteration in the cross-validation process.

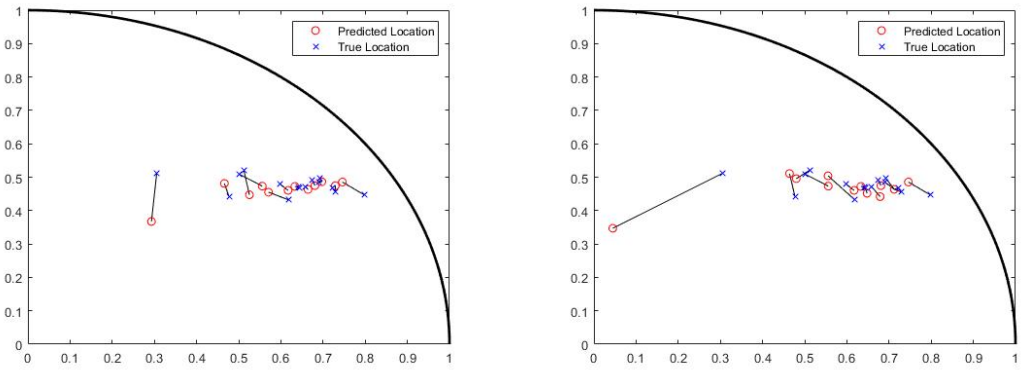
Model	Covariance function	Averaged marginal likelihood	Averaged RMSE
A	$k_{pol}(\rho, \rho')$ & $k_{ang}(\theta, \theta')$	70.8083	0.0442
B	$k_{3/2}(\rho, \rho')$ & $k_{ang}(\theta, \theta')$	50.4647	0.0685

Another performance assessment method was also considered, where the most likely predicted location was compared to the true value via the *Root-Mean-Square-Error* (RMSE). The RMSE can be expressed as,

$$RMSE = \sqrt{\frac{\sum_{\rho,\theta}((\rho,\theta)_{pred} - (\rho,\theta)_{true})^2}{\mathrm{N}_{test}}} \tag{5.11}$$

where N_{test} is the number of samples in the test set. As before, the RMSE was evaluated at each iteration test set of the LOOCV process and then averaged to attain the overall performance. The results have also been included in Table 5.1 and a depiction of these differences is illustrated in Figure 5.12.

Similarly, the averaged RMSE indicates that Model A predicted locations that were generally closer to the true values than Model B. By looking at the predicted locations in Figure 5.12, one can note that a major contribution to this error is given by the point lying nearest to the bearing centre. This point corresponds to the journal bearing operating at a speed of 400rpm and with an applied static load of 2kN. Under these conditions, the shaft is located away from all the other locations recorded in the data set. As demonstrated thus far, in areas of sparse data, the GP becomes less certain, thereby making poorer predictions, like the one observed for this data point. However, when considering the confidence bounds around the shaft-centre location, the true location is said to exist somewhere within the spread of the prediction. That is, the GP predictions are not limited to a single location, and account for the probabilities of the true values



(a) $k_{pol}(\rho, \rho')$ & $k_{ang}(\theta, \theta')$ (b) $k_{3/2}(\rho, \rho')$ & $k_{ang}(\theta, \theta')$

FIGURE 5.12: Distance of the predicted locations to their corresponding true values. (a) Model A and (b) Model B.

existing anywhere in the bearing. The practical advantage of this probabilistic framework is then provided by reduced areas within the bearing where the shaft-centre true location is more likely to be found. These areas only become narrower (i.e. more certain) under the presence of data, and more measurements would be needed to improve the performance of the proposed models.

5.1.7 Damage Detection on Simulated Data

Before concluding this section, a scenario is considered where a journal bearing undergoes gradual oil starvation during operation. This simulation is conducted with the aim of showcasing how the proposed model could be applied to detect deviations from normal operations. To begin, the response of the journal was simulated assuming it to be a *short-width* bearing [159], and by adopting the operational parameters outlined in Table 5.2. These parameters align with the actual dimensions and operating conditions of the journal bearing studied so far.

TABLE 5.2: Journal bearing operating parameters.

Parameter	Symbol	Value	Units
Bearing radial clearance	c	55	μm
Journal diameter	D	110	mm
Rotational speed	ω	1-2000	rpm
Static load	W	2000	N
Oil kinematic viscosity	μ	0.0604	Pa s

The equilibrium position was calculated for speeds ranging from 1rpm to 2000rpm, while keeping the remaining parameters constant. The obtained coordinates were then corrupted with artificial noise modelled by i.i.d. samples from a Gaussian distribution with zero mean and standard deviation equivalent to 2% of the standard deviation of

the noise-free signals. The corrupted data were subsequently used to train a GP to learn the functional mapping between the coordinates and speed-load ratio.

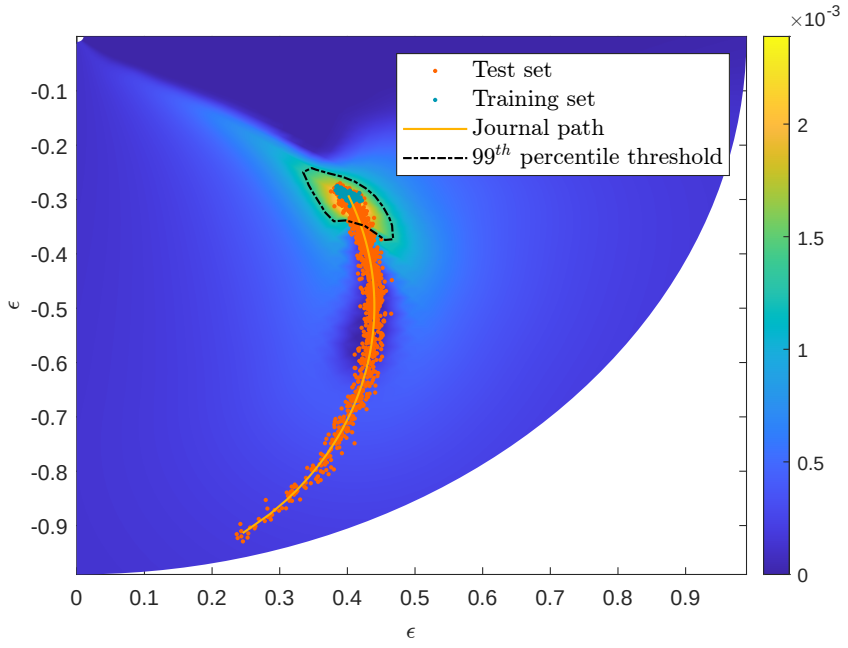
Let the nominal operation of the bearing be such that the equilibrium position is at $X = \{0.5, 54^\circ\}$. In this particular case, according to the short-width bearing model, this location corresponds to the bearing rotating at 115rpm and subjected to a static load of 2kN.

In practical scenarios, a collapsing film might cause the shaft to rub against the bore at the point where the thinning film is unable to support the load of the rotor. Such an event is undesirable and can be damaging to the rotor, especially when dealing with high-speed turbo-machinery systems. A monitoring system capable of promptly detecting the onset of this phenomenon could provide operators with sufficient time to take necessary actions.

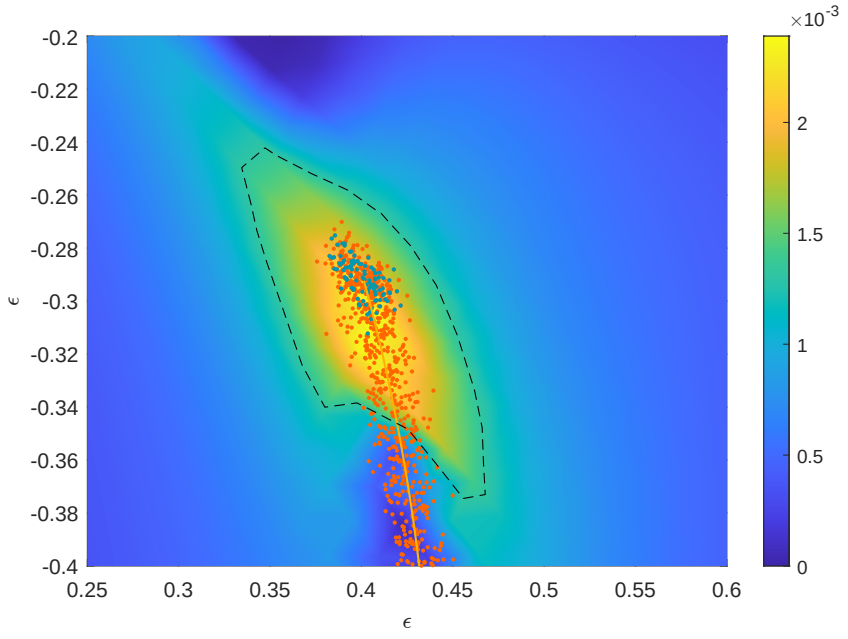
The strategy employed here to detect this abnormal behaviour is based on the concept of using the negative log-likelihood as a novelty metric. Concretely, the GP likelihood map favours the journal sitting near the equilibrium position in terms of mass distribution in the inferred likelihood map. In other words, if the location of the shaft were to change from factors other than speed and load - such as a reduced oil supply - the computed likelihood would decrease as the shaft departs from its equilibrium position. Given the probabilistic framework of the model, a threshold for flagging abnormal operations can be empirically defined using a sufficiently high percentile.

The outcome of this exercise is illustrated in Figure 5.13, where a 99^{th}-percentile threshold is employed. The standard bearing configuration is adopted, focusing now on the lower quadrant. The reduction in oil supply is simulated by gradually reducing the thickness of the fluid film at these conditions, and the resulting downward trajectory followed by the journal is depicted in red, accompanied by simulated measurements artificially corrupted by noise. The negative log-likelihood of these measurements is calculated chronologically as shown in Figure 5.14.

Although this simulation may offer a simplified representation of oil starvation, it serves as an illustration to demonstrate that any form of deviation from the equilibrium position can be easily detected by means of the proposed model. A prevalent advantage of this strategy is its adaptability to operational variations. The model is somewhat impervious to changes in rotational speed and/or static loads since the GP is trained to estimate the equilibrium position in relation to these parameters. However, as discussed earlier, other benign variations are expected to affect this relation, potentially leading to an increased rate of false positives. Addressing these variations in the relationship learned by the GP



(a)



(b)

FIGURE 5.13: Illustration of oil starvation showing (a) the trajectory of the path followed by the shaft centre, and (b) a detailed view of the threshold defined in terms of the likelihood map. The colour bars indicate the likelihood given by Equation (5.10).

could enhance the model to accurately distinguish genuine abnormal responses. While beyond the scope of this chapter, this issue remains a topic for future research.

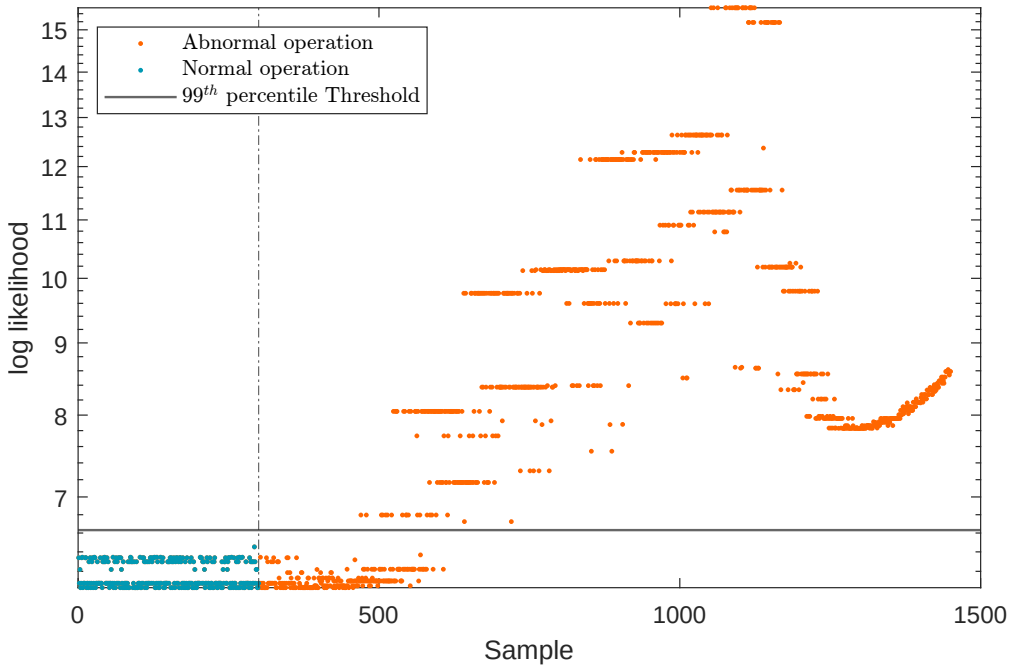


FIGURE 5.14: Novelty indices given by the negative log-likelihood, as a function of the operational parameters $\omega = 115$rpm and $W = 2$kN.

5.2 Case study 2: Acoustic Emission Source Location Using Bayesian Optimisation for a Composite Helicopter Blade.

The second case study outlined in the present chapter is concerned with the implementation of AE-based techniques to localise sources of damage. This method is made possible by having multiple AE-sensors acting simultaneously. The location of an emerging defect can effectively be determined by measuring the *Time of Arrival* (ToA) of a wave to each sensor.

A traditional approach to AE source localisation is based on triangulation, where the source is determined to be at the intersection of the contours resulting from calculating the differences in ToA (ΔT) from multiple sensor-pairs [160, 161]. Localisation may also be approached as an optimisation problem, where the origin of a source is estimated by minimising the difference between the recorded ΔT and values calculated at candidate locations.

Although these methods have been demonstrated to work successfully, there are a few associated challenges that can be arduous to overcome. For example, the propagating path of an emitted wave is an important factor to consider, as localisation becomes increasingly more difficult when dealing with composites, or when obstructions are present in the propagating path of the wave. Some progress has been achieved with anisotropic

materials [162, 163], but it may still be challenging to analytically determine how the wave travels through the medium.

Data-based techniques have been developed for this type of application by instead learning statistical models that are uniquely concerned with the ΔT measurements [164]. The idea behind this approach is to artificially construct a map of ΔT values at various locations, to then match the ΔT from a real AE to those on the map. A probabilistic framework for this technique has also been explored and implemented successfully in [93, 165].

The problem is that ΔT maps can be laborious to produce. A vast collection of data may be required to cover the entire surface where damage might be expected. The ΔT maps must also be fine enough for accurate predictions, considerably increasing the size of the data set needed for the learning process.

While attention is given to localising sources of AE, the efficient construction of a ΔT map is also explored, where a Bayesian optimisation approach is used for an optimal selection of sampling points. The subsequent sections detail the experimental procedure undertaken for data collection, outline the employed modelling strategies and subsequently discuss the obtained results from the proposed methods.

5.2.1 Experimental Procedure

The experimental set-up studied here was designed for a fatigue test of a helicopter blade. An eccentrically-loaded motor attached to the blade was tuned to have the blade resonate in its second mode of vibration, to promote the extension of fracture somewhere near the root.

The detection of the early onset of fracture in fatigue is possible with the use of AE-based monitoring systems [45]. However, the effectiveness of monitoring the progression of fracture during the fatigue test relies on correctly characterising the detected AE waves. By having the monitoring system also locate the (possible) various sources of AE, one can identify and isolate those deriving from damage. This consideration is important for correlating the features extracted from AE waves with the potential failure modes of the structure, thus highlighting the need for a robust localisation system. In [44, 95, 166], for example, localisation techniques are exploited to directly characterise AE activity emerging near a developing crack.

The configuration of AE sensors used here is shown in Figure 5.15, where six sensors can be seen attached near the root of the blade. This set of sensors comprised three *WD differential* and three *Micro* 30*D differential Mistras* AE sensors. All sensors were

attached to the blade surface using adhesive glue, and small contact gaps were filled with grease to improve the acoustic coupling. The sensors were spaced as evenly as possible near the root of the blade; roughly covering an area of 20cm × 60cm. A *2/4/6 Switch Selectable Gain* was used to pre-amplify the sensors to 20dB. The data acquisition system used was a *Mistras Micro-II compact PCI AE Chassis*.

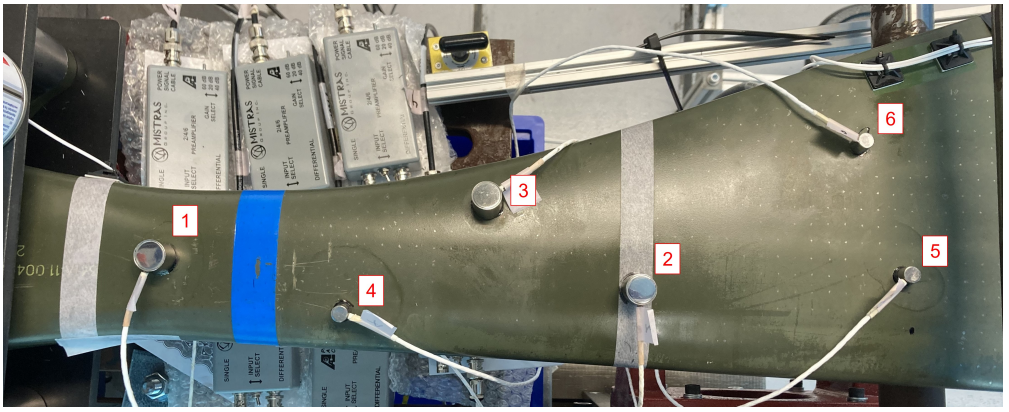


FIGURE 5.15: Helicopter blade root and AE sensor configuration.

In order to develop a localisation monitoring system, it was first necessary to collect enough data for training the model. The training set was therefore gathered by measuring the ΔT values of waves propagating through the blade. By knowing the location of the source beforehand, one can record the time it takes for the wave to trigger each sensor from its origin. The procedure was then followed with the construction of the ΔT maps for each sensor-pair combination. In this case, 15 unique sensor-pair combinations were possible.

A standard way to simulate realistic broad-band AE waves is by breaking pencil leads on the surface of the blade. In this way, one can choose the location of the source using a *Hsu-Nielsen* device [167]. A collection of AE sources can then be artificially generated at different locations and the respective ToA recorded to gather a training set. The challenge with this approach is deciding where to sample on the blade. Intuitively, a grid could be placed over the blade with uniformly-spaced sample-points. The spacing between sampling points will determine the precision of the localisation algorithm, and thus a finer grid might be preferred in most applications. Unfortunately, there is the practical expense of having to sample from an inconveniently large number of points. The trade-off between accuracy and cost of the implementation will be application-dependent, warranting the need for an autonomous solution that can provide a balance between both.

Given that these samples had to be carried out by hand, a compromise was made in the fineness of the grid by having a sample-points spaced by $10mm$. Nevertheless, additional sample-points were included in between layers nearest to the root, since the fatigue test

was designed to promote the extension of fracture in this area. The grid was designed on graph paper and was then wrapped over the blade to mark the locations of the sampling points on its surface. These points can be seen faintly in Figure 5.15.

Having established a sampling grid, the Hsu-Nielsen device was used to generate an AE-wave at each location. The waves were recorded and stored to determine their true ToA at each sensor. This last step was necessary, given that the data-acquisition system only places a time-stamp at the instance in which a threshold crossing occurs, which can happen even after the actual onset of the wave is picked up by the sensor(s). To account for this consideration, the true ToA of the recorded AE-waves can be determined based on the *Akaike Information Criterion* (AIC) [168]. The AIC-picker technique can be defined by the following equation,

$$AIC(t) = t\log(\mathbb{V}(R_w[1:t])) + (T-t-1)\log(\mathbb{V}(R_w[t:T])) \tag{5.12}$$

where at each time step t, the entropy is calculated before and after each time step. The entropy of the windowed signal, R_w, will return a minimum when the proportion of the signal up to the time step t is uniquely composed of uncorrelated noise, happening to also correspond to the onset of the AE-wave. The indexing in R_w indicates the range of values used to calculate the variance.

The results of the difference in time of arrival are shown in Figure 5.16 (top) for the sensor-pair 3-4. Some areas of the blade could not be sampled because of the sensors themselves occupying the space, or the limited accessibility because of the test-rig configuration, resulting in some sparsity that can be seen in the results. A somewhat gradual transition can be noted along the blade, which is to be expected in an ideal scenario, but a few inevitable exceptions can also be observed that appear in dispute with this trend. These "anomalous" data could be the result of noise in the measurements, mainly attributed to the complexity of the propagating paths in the blade.

A schematic of the internal structure of the blade was not available, but it is known to be made of a carbon composite with a complex internal honeycomb structure. The propagating paths could have therefore resulted in complex wave dispersion and/or damping attenuation that made it difficult to discern background noise from actual AE activity. In particular, waves originating furthest from a sensor were the most cumbersome to capture. After curation, the data set comprised a total of 595 sampled points, and then was split into a training and test set, where 50 random sample points were allocated to the test set.

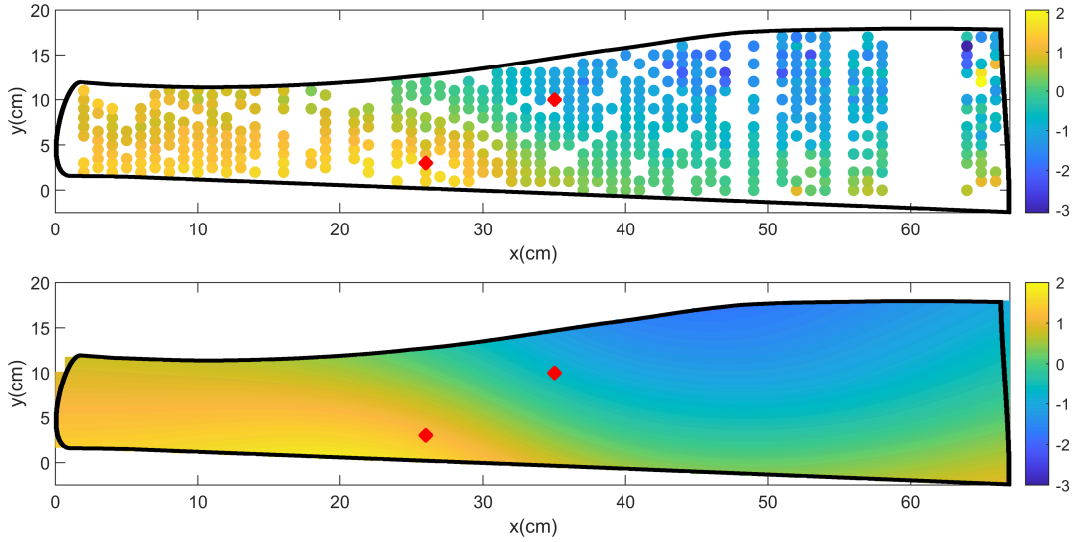


FIGURE 5.16: (Top) ΔT measurements recorded from sensor-pair 3-4. (Bottom) Inferred GP posterior mean. Red dots indicate the locations of sensors 3 (right) and 4 (left). Colour bars indicate the standardised ΔT.

5.2.2 Localisation Strategy

Much like in the previous case study, this approach is based on the same regression model, but the mapping is instead based on a location on the blade to a corresponding difference in ToA, i.e. $(x, y) \rightarrow \Delta T$. Given the data gathered from the lead breaks, it is possible to construct a model capable of returning the probability of observing a ΔT value at a new location (x_*, y_*). A (GP) regression model is again employed to predict ΔT values over the continuous space covered by the grid.

Implementing the GP with the results presented in the previous subsection, returns an interpolated mapping over the space covered by the grid, as shown in Figure 5.16 (bottom). By repeating the implementation of the GP on all 15 sets of results (corresponding to each sensor-pair), one can make use of these models to pinpoint the most likely location from which an AE-wave may originate.

The location can be determined by calculating the likelihood of observing the set of ΔT values for each sensor-pair, given the functional mapping $(x, y) \rightarrow \Delta T$ learned from the training set. For the set of $J = 15$ models, the log-likelihood for each model m_j can be assessed as,

$$\log p(t_{*,j}|D, (x_*, y_*), m_j) = -\frac{1}{2}\log \mathbb{V}[\mathbf{f}_{*,j}] - \frac{(t_{*,j} - \mathbb{E}[\mathbf{f}_{*,j}])^2}{2\mathbb{V}[\mathbf{f}_{*,j}]} - \frac{1}{2}\log 2\pi \tag{5.13}$$

where $D = (\mathbf{x}, \mathbf{y})$ denotes the location of all datapoints in the training set, and $t_{*,j}$ is the predicted ΔT at the new location (x_*, y_*), given by the j^{th} sensor-pair. The log-likelihood maps produced by assessing Equation (5.13) will display contours of the most probable locations of $t_{*,j}$. A unique solution is generally found at the intersection of the contours evaluated for each sensor-pair model. Mathematically, this outcome can be derived by marginalising each model. That is,

$$p(\boldsymbol{t}_*|D, (x_*, y_*)) = \sum_{j=1}^{J} p(t_{*,j}|D, (x_*, y_*), m_j)p(m_j) \tag{5.14}$$

If each model is assumed to have equal importance, the marginalised probability distribution is given by,

$$p(\boldsymbol{t}_*|D, (x_*, y_*)) = \frac{1}{J}\sum_{j=1}^{J} p(t_{*,j}|D, (x_*, y_*), m_j) \tag{5.15}$$

which is simply the average of the likelihood maps for each sensor-pair combination. The most likely location will therefore be given by the point (x_*, y_*) that maximises Equation (5.15). The outcome of this approach for an arbitrary test point is shown in Figure 5.17.

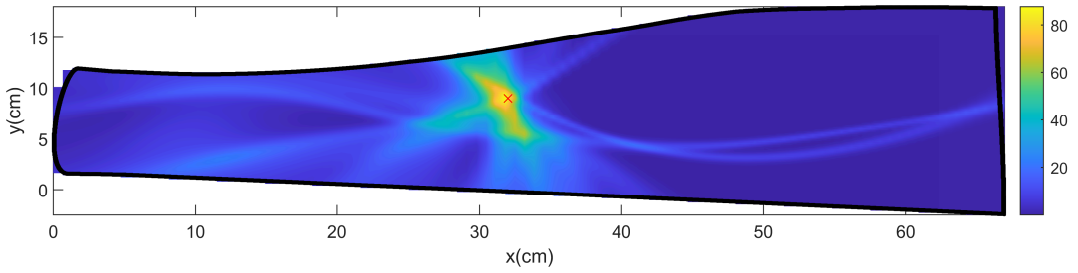


FIGURE 5.17: Prediction of the GP model for an arbitrary test point. The red x marks the true location of the AE source. The colour bar represents the combined likelihood value.

5.2.3 Bayesian Optimisation for Feature Selection

As mentioned above, the sampling grid is subject to some practical limitations. In this subsection, the idea of developing an optimum grid is explored, employing *Bayesian Optimisation* (BO) [169]. In a BO scheme, an objective function is optimised by finding its local maximum (or minimum) with as few parameter proposals as possible. This method does not rely on gradients and can be an attractive alternative in applications where the objective function cannot be easily determined or is expensive to compute.

The objective function is approximated with a surrogate function, which is usually chosen to be a GP. A new proposal reveals the corresponding output of the objective function, followed by conditioning the surrogate GP with the revealed observations. Deciding on the proposals is determined by an *acquisition function*. A variety of acquisition functions exist, but the commonly used *Expected Improvement* (EI) [169] is adopted here. The EI is defined as,

$$EI(x) = \begin{cases} (\mu(x) - f(x^+) - \eta)\Phi(Z) + \sigma(x)\phi(Z) & \text{if } \sigma(x) > 0 \\ 0 & \text{if } \sigma(x) = 0 \end{cases} \tag{5.16}$$

where $f(x^+)$ is the value of the current best sample x^+, and $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation of the surrogate GP posterior, respectively. $\Phi(Z)$ and $\phi(Z)$ correspond to the *Cumulative Density Fucntion* (CDF) and *Probability Density Function* (PDF) of the standard normal distribution, evaluated with respect to $Z = (\mu(x) - f(x^+) - \eta)/\sigma(x)$, respectively.

Acquisition functions suggest new proposals by balancing the exploitation and exploration of the search space. In short, exploitation refers to sampling more from areas that improve the output of the objective function, while exploration promotes sampling from areas of high uncertainty.

These actions are readily interpretable in Equation (5.16), whereby the expected improvement is high either when the difference $(\mu(x) - f(x^+))$ is high – promoting exploitation – or when the uncertainty $\sigma(x)$ is high – promoting exploration. Already sampled values reduce the expected improvement to zero, to prevent sampling the same values more than once. The parameter η is an adjustable parameter that controls the amount of exploration.

Although used for finding optimal parameters, the BO can be adopted to select the best points to sample from in the ΔT grid. The idea here is to minimise the number of samples one would need to construct a grid simple enough to determine the location of new AE signals. Simultaneously, this approach would also avoid having to decide on the density of the grid, since new samples are proposed sequentially in an active setting. This idea has been explored in [170], where a BO scheme is used to promote sampling from damaged areas in specimens for an autonomous ultrasonic inspection.

What remains is to decide on the objective function for this scenario. An ideal objective function may be one that promotes sampling in areas that yield the highest gain in accuracy while allowing for some exploration in areas of high uncertainty in the blade. The objective function used here is one constructed by calculating the reduction in the

RMSE, $r(\Delta T)$, that a single sample-point in the training set provides to the localisation of the AE-waves in the test set. Overall, the map-construction strategy is described in Algorithm 1.

Algorithm 1 ΔT map construction

Define uniform grid on surface (x, y)
Sample $r(\Delta T_o)$ at random point $\mathcal{D}_o = (x_o, y_o)$ ▷ Starting point for the BO
for $i = 1, 2, \ldots$ **do**
  $(x_i, y_i) \leftarrow \text{argmax}_{(x,y)}\, EI((x, y); \mathcal{D}_{1:i-1})$ ▷ Maximise Expected Improvement
  Sample $r(\Delta T_i)$ at (x_i, y_i) ▷ Update surrogate function
  $\mathcal{D}_i \leftarrow \{\mathcal{D}_{1:i-1}, (x_i, y_i)\}$ ▷ Add new sampled observation to dataset
end for
Predict t_* using Equation (5.15)

This algorithm is active in the sense that it proposes to the operator where to sample from in real-time. In other words, once a ΔT measurement is taken, the optimiser determines the next sampling point that will likely provide the greatest improvement in accuracy. After having conducted the entire sampling process, the gathered data are used to train the GP for damage localisation by following the steps described in Subection 5.2.2.

5.2.4 Results and Discussion

To demonstrate the implementation of the BO strategy for an efficient ΔT map construction, a scenario is considered in which limiting resources only allow for 100 samples to be taken. These samples would follow on the test set already collected at 50 random locations within the borders of the grid. The results of the sequential sampling are shown in Figure 5.18. A series of random-sampling strategies were also performed to demonstrate the benefits of choosing an optimal (under the BO) collection of sampling points. The RMSE calculated after having sampled all training datapoints has also been included as a reference.

A few observations are worth addressing from these results. Firstly, the BO strategy appeared to have found a sequence of samples that returns a reduction in error at a greater rate than the random sequence. It may be possible to find a better combination of points at random but at the high risk of having to reiterate the test if otherwise. The results from the random test also appear to show that some datapoints can negatively affect the performance of the model. This outcome could be explained by the "anomalous" measurements described above. It may be necessary to repeat this study but with a "clean" data set gathered from a simpler structure, such as an aluminium plate.

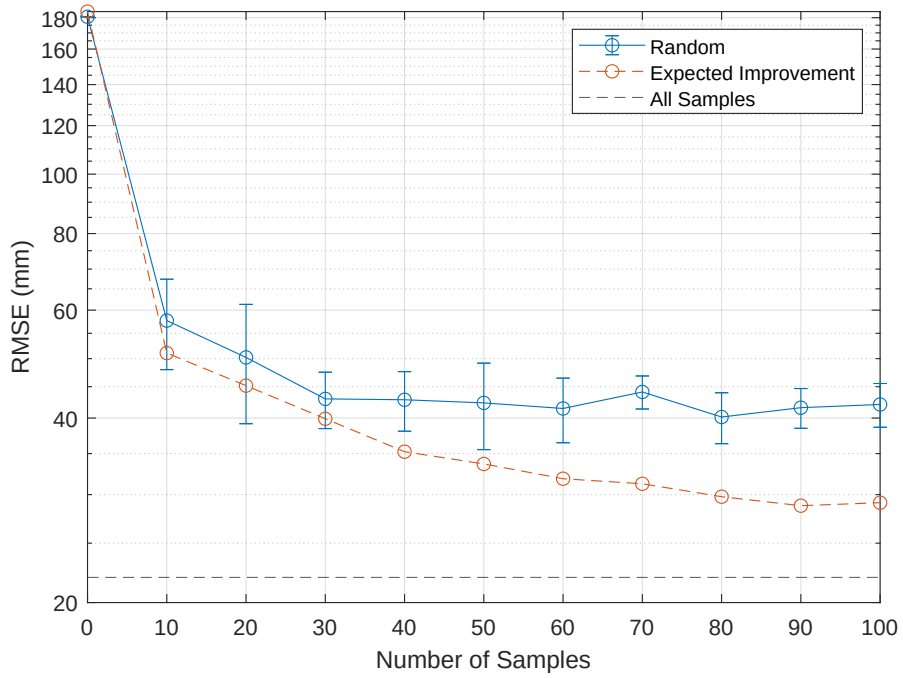


FIGURE 5.18: RMSE calculated with data sampled from a random process (blue) and by using the EI acquisition function (red). The RMSE calculated from a complete data set has been included as a reference (black dashed line).

The objective function was approximated as samples were sequentially introduced. The resulting mean of the surrogate objective function, after assessing each point on the grid, is shown in Figure 5.19 (top). Similarly, as shown in Figure 5.19 (bottom), one may note a higher concentration of samples found near the global maximum of the objective function. It is important to note, however, that the aim here deviates from the purpose of a conventional Bayesian optimiser in finding a unique solution that maximises the objective function. Although this was the intention of the proposed strategy, some substantial exploration had to be enforced by increasing the parameter η. This last step is important since the accuracy of the localisation model improves with a comprehensive representation of the search space, which is achieved by gathering data covering the entirety of the blade. A balance of these aspects is nicely encapsulated by the acquisition function.

Finally, it is worth acknowledging that the proposed objective function is by no means the only option one could use in practice. In [170], for example, the objective function is based on novelty detection techniques used in SHM. The success of this approach may rely on the correct choice of the objective function, acquisition function, and correct tuning of the parameters involved. Deciding on these parameters will naturally depend on the structure and application at hand.

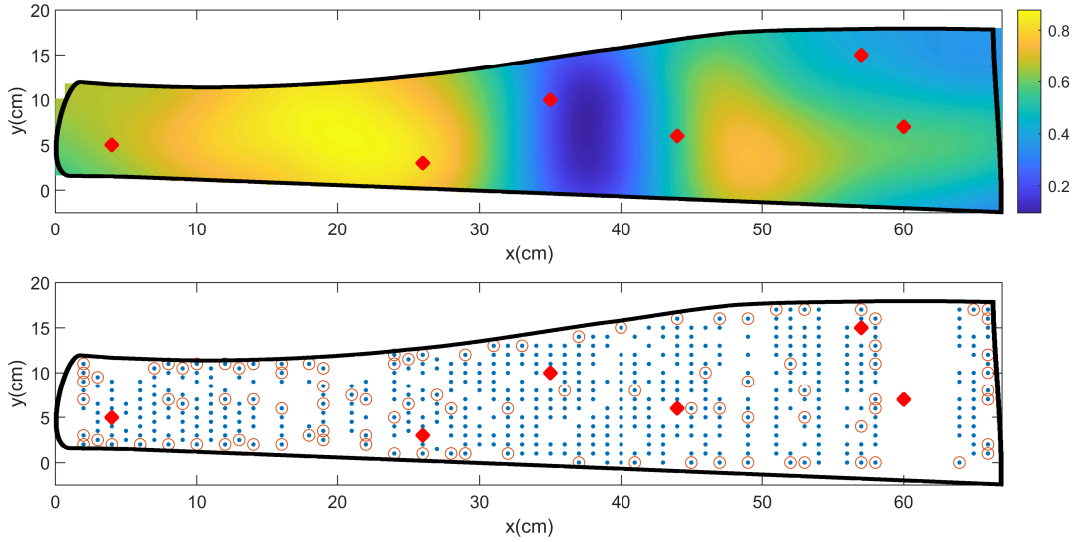


FIGURE 5.19: (Top) Objective function over the search-space. (Bottom) Red circles indicate the sampled locations proposed by the EI acquisition function. Red dots indicate all sensor locations. The colour bar represents the normalised values of the objective function.

5.3 Conclusions

This chapter delves into the application of GPs for localisation-oriented problems in SHM. Two distinct case studies are presented, demonstrating the utility of data-driven models for localisation purposes. In the first case study, the GP is used to predict the probable equilibrium position that a journal-bearing shaft might adopt under a specific operational state. In contrast, the second case study explores a more direct approach to damage localisation by employing a GP to model ΔT maps of a composite helicopter blade.

The motivation for using GPs in these case studies is driven by the complex functional mappings that may otherwise be too intricate to derive analytically. In the case of the journal bearing, strong nonlinear dependencies exist in the relationship between the equilibrium position and operational parameters. GPs can effectively capture these relationships from fluid-film thickness measurements alone. While the selection of a suitable covariance function is essential to attain a good representation of this relationship, considering a simplified notion of the influence of rotational speed and static load on the equilibrium position facilitates the selection process. Similarly, the complex propagating paths of AE-waves in composites make the GP apt for modelling time-of-arrival differences in relation to their corresponding sources.

Indeed, a set of challenges were encountered in both case studies, warranting acknowledgement to enhance the proposed methods. Notably, the likelihood maps for shaft-centre localisation were generated from rather small data sets ($N = 15$), which might not comprehensively represent the journal response to operational variations. In light of more data, the GP becomes less reliant on biases imposed by the covariance function, potentially offering predictions that agree better with the true journal response.

Further validation with additional experimental work might be desirable, including the evaluation of the model with other journal-bearing configurations, such as non-conformal bearings or tilting-pad bearings. Conversely, an alternative approach to limited data sets could involve incorporating journal-bearing physics within a *grey-box* framework. This approach may not only improve estimates within interpolated areas, but also provide more accurate predictions beyond the confines of the available data.

Regarding the case study involving the helicopter blade, the strategies proposed for an efficient ΔT map construction require further investigation. While using a Bayesian optimiser to determine the sampling points demonstrated promising outcomes, the validity of the method might need testing on other structures. Essentially, conducting this exercise with a composite structure might not be enough to determine its success. The additional complexities introduced by composite structures in the propagation of AE waves likely contributed to variables affecting the outcome that were not considered. Therefore, it might be valuable to implement the BO approach for constructing ΔT maps on simpler structures first, such as an aluminium plate. This benchmarking comparison could help reveal any potential discrepancies. Furthermore, selecting an objective function remains an open issue; conducting a comprehensive exploration of various objective functions and comparing their outcomes across systems could enhance the sampling process. Finally, the proposal of incorporating a *Multi-Objective Bayesian Optimisation* (MOBO) scheme for this purpose is left for future investigation.

Chapter 6

A Novel Probabilistic Approach for Acoustic Emission-Based Monitoring Techniques

In Chapter 2, it was briefly mentioned that, compared to other forms of monitoring techniques, AE measurements can provide a more sensitive means of detecting the onset of damage, facilitating the earlier diagnosis of an unhealthy structure. Much like in the case of implementing AE-based techniques for damage localisation - as discussed in Chapter 5 - this aspect can be of great value for the development of an optimised maintenance schedule. For example, considering an offshore wind turbine, it is more convenient to determine remotely whether maintenance is required, given the associated high cost of having the wind turbine put out of service for a potentially unnecessary repair. Conversely, if the detected damage in the wind turbine suggests that it will fail prior to its next scheduled maintenance, then action can be taken to prevent a potentially-catastrophic outcome. The acquisition of a more sensitive damage-detection system therefore means having more time to plan a suitable course of action.

Unlike the approach followed in the previous chapter, it is important to note that the concern here is not on localising sources of damage, but rather on their early detection. Specifically, the aim of this chapter is constrained to the first level in Rytter's hierarchy. Although the approach presented in this context might be considered "simpler" within this hierarchy, there are complexities associated with AE-based techniques for damage detection that require careful attention [171].

One of these limitations pertains to the high sampling frequencies required to record AE signals. In just a matter of a few seconds, a recording can generate millions of

data points, leading to challenges in storage and manipulation. Consequently, relying on feature-extraction methods becomes indispensable to reduce the overall size of the measured data. The features of primary interest are those that characterise the AE, and a few examples of these features are shown in Figure 6.1, illustrating a typical AE waveform.

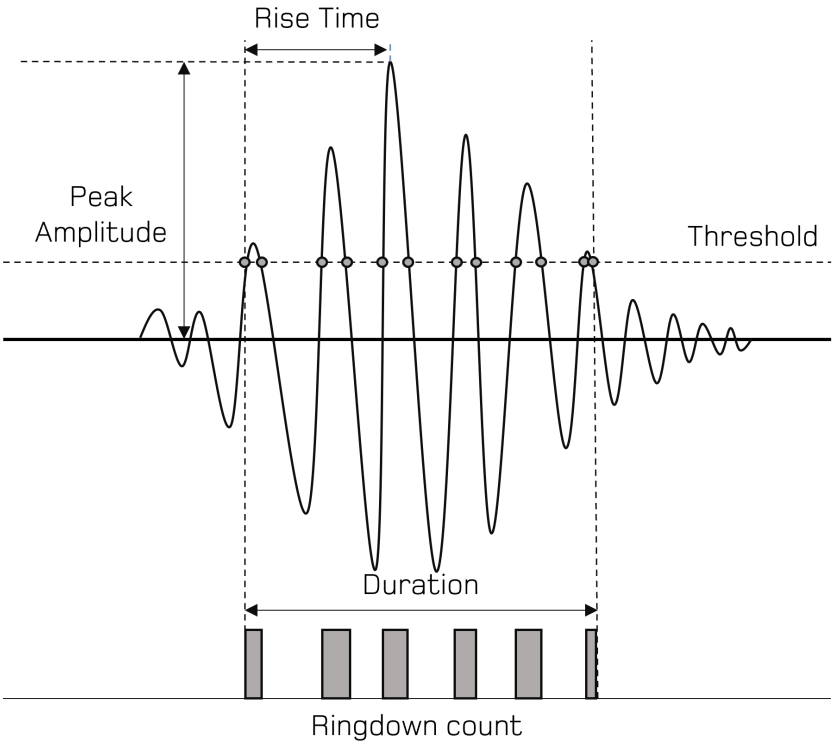


FIGURE 6.1: Features of a typical AE. Reproduced from [1].

Various means of feature extraction exist [172]; a common approach is to first identify and isolate the individual AE events/bursts that emerge throughout the recorded wave stream. The challenge of this approach, however, is to have an algorithm to reliably identify the relevant AE events for analysis, and distinguish them from background noise.

One of the first steps in an AE-based monitoring scheme is thus to define a threshold from which any voltage crossings are assumed to indicate the presence of some form of AE burst activity. This parameter is normally set via the data acquisition system. The issue is that deciding on the threshold is not as trivial as it may seem. In practice, the threshold may depend on the user experience. This approach is limited in that it is specific to the application at hand; that is, the operator must carefully select an optimal threshold to capture the AE events of interest. A suitable threshold value will depend on the properties and geometrical constraints of the propagating medium.

This crucial exercise is not only meaningful for reducing the overall size of the wave stream but also for retaining AE events carrying information relating to the health state of the structure. The underlying features that characterise the isolated AE events are somewhat influenced by their source, and some correlation will exist based on the type of damage or physical mechanism that generates them.

This complication is even more of an inconvenience in applications dominated by a rich collection of AE sources and/or excessive amounts of background noise. In such cases, statistical modelling and pattern recognition techniques can be used to facilitate the analysis of AE data. In [173], for example, a pattern-recognition approach was adopted to automatically classify AE signals obtained from a box girder of a bridge. The developed methods were able to identify three distinct classes that could distinguish crack-related signals from those generated by frictional processes. Another promising pre-processing technique that builds on this idea can be found in [174], where the proposed method outputs the best combination of AE-features that can provide the greatest separation of classes; therefore, yielding a dataset that is most sensitive to the various sources of AE.

While these studies address some of the challenges encountered in pre-processing AE data from a statistical point-of-view, the datasets used correspond to a collection of AE waves extracted from having defined user-specific settings in the data-acquisition system. The quality of the extracted waves thus relies on having these settings adjusted correctly, which is subject to the expertise of the operator. The issue addressed in this chapter thereby delves into making the extraction of AE waves less reliant on these variables, such as (manually) having to set a suitable threshold. Similarly, the aim here is to propose a robust method with the ability to autonomously extract AE events from streams that may ensure the effectiveness of subsequent analyses.

To this end, a probabilistic framework is developed by treating the generation of AE events as a random process. Adopting this perspective allows one to quantify the probability of finding an AE event in any arbitrary section of a raw wave stream. This implementation is fundamentally employed with a *Poisson* distribution, a *Probability Mass Function* (PMF) that will be used here as the building block for modelling counts extracted from AE recordings. Ultimately, a powerful nonparametric Bayesian approach is introduced and implemented, both to find the AE events in the raw signal and to group them based on their features in an online scheme. This approach is further demonstrated to be applicable for early damage identification in a structure using experimental AE data.

The proposed strategies are described in the following sections and then demonstrated with a rich AE dataset collected from a journal bearing in operation, and from an Airbus A320 main landing gear subjected to fatigue testing.

6.1 A Probabilistic Perspective for AE-Based Monitoring Techniques

As highlighted above, determining the threshold for AE events can be an assiduous task that may depend highly on the expertise of the operator. Alternatively, a potentially more robust approach could be employed to have the threshold defined statistically. One may assume the background noise in the signal to be the result of a generating process in which the observations are samples drawn from a Gaussian distribution, and the threshold could then be determined numerically using a Monte Carlo method [1]. Although this approach offers a threshold determined directly from the recorded signal, the resulting selection of AE events may still be heavily corrupted by background noise. For example, if a 99^{th}-percentile threshold is chosen on the recorded time-series, it means (at the risk of redundancy), having 1% of the data points estimated to lie above that threshold. This deceptively small percentage could be a problem in scenarios where the sampling frequency is high, such as an AE reading. In the span of a few seconds, millions of data points are recorded, and thus thousands of "events" would consequently be taken into consideration for further analysis. This issue means mistakenly identifying a vast number of events deriving from the background noise when it is likely that only a fraction of them will actually originate from damage.

An attempt to overcome these limitations is explored in this chapter by introducing a probabilistic framework; the idea is to quantify how certainly a threshold crossing is believed to derive from an AE event. Concretely, after establishing a threshold statistically, and then counting the number of times a signal exceeds the threshold in a given time frame, one can evaluate the probability of an AE event existing in that time frame. In the following sections, it will be demonstrated how the Poisson distribution can be conveniently used for this purpose.

6.1.1 The Poisson Distribution

The Poisson distribution is a discrete probability distribution used to model the number of events occurring in a given time or space interval. It can be used to model counts of rare occurrences, such as radioactive decays or traffic accidents [175]. Since the intention is to model the number of threshold crossings in a given time frame, and detect an unusually high number of threshold crossings attributed to an AE-burst, the Poisson distribution arises naturally for this type of problem. The Poisson PMF is defined as,

$$p(x|\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!} \tag{6.1}$$

where x is a positive integer that, in this application, corresponds to the counts of threshold crossings happening within a given time frame, and λ is the rate or expected number of these threshold crossings happening in the time interval.

The value of the parameter λ may be unknown, and a Bayesian framework is adopted here to infer the distribution over λ. The count-rate is therefore modelled according to some prior probability distribution, $p(\lambda)$, that quantifies a prior belief over all possible values of λ. Bayes' rule can then be applied for inference to recover the posterior $p(\lambda|x)$.

Fortunately, a conjugate prior to the Poisson distribution exists, allowing for a closed-form solution for $p(\lambda|x)$. In order to meet this condition, the chosen conjugate prior must be a Gamma distribution, resulting in another Gamma distribution when combined with the Poisson likelihood [175]. The Gamma distribution over λ can be defined as,

$$p(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda} \tag{6.2}$$

where $\Gamma(a)$ is the Gamma function that ensures Equation (6.2) is normalised, and the parameters $a > 0$ and $b > 0$ correspond to its shape and rate, respectively. These parameters define the functional form of the distribution. In this context, these parameters are intentionally chosen in a flexible manner to align the Gamma distribution with prior beliefs regarding the potential values of λ. Finally, the predictive distribution for a new observation x_{new}, given some available observations, can then be calculated by taking the expectation of the Poisson likelihood with respect to the posterior distribution,

$$p(x_{new}|\mathbf{x}) = \mathbb{E}_{p(\lambda|x)}[p(x_{new}|\lambda)] = \int p(x_{new}|\lambda)p(\lambda|\mathbf{x})\,\mathrm{d}\lambda \tag{6.3}$$

where a set of N observations $\mathbf{x} = \{x_1, \ldots, x_N\}$, is considered. Conveniently, equation (6.3) reduces to a negative-binomial distribution. In particular,

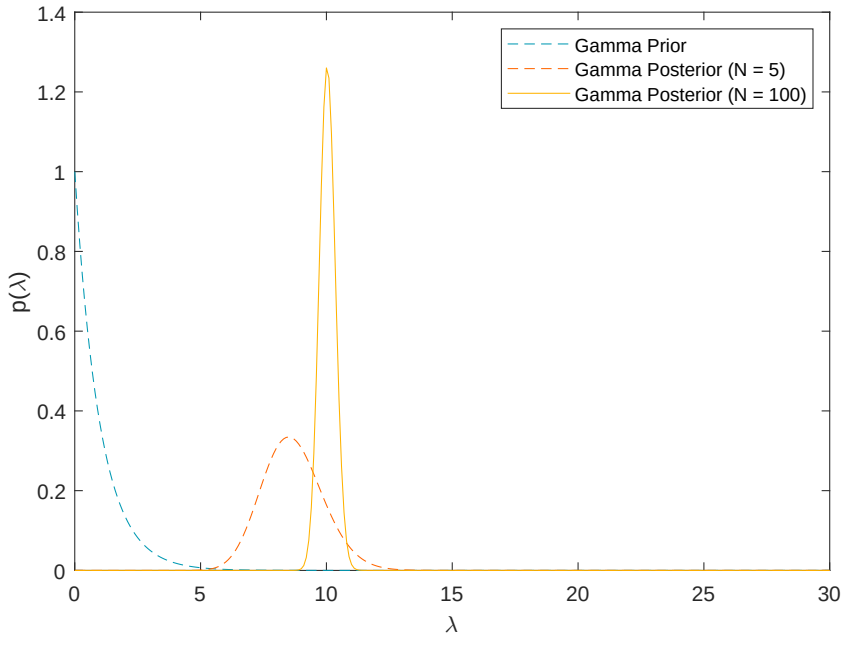
$$p(x_{new}|\mathbf{x}) = \frac{\Gamma(x_n + r)}{x_n!\Gamma(r)}p^r(1-p)^r \tag{6.4}$$

where $r = a^*$ and $p = b^*/(b^* + 1)$, with the updated parameters a^* and b^* given by,

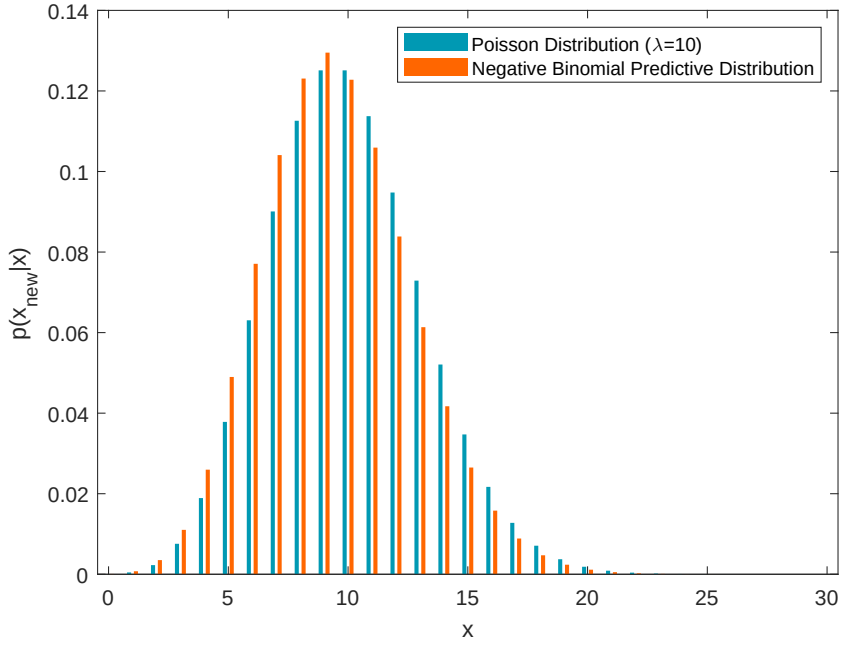
$$a^* = N\bar{\mathbf{x}} + a, \quad b^* = \frac{N+b}{N+b+1}$$

where $\bar{\mathbf{x}}$ is the sample mean of the number of available observations $\mathbf{x}$. The probability of observing a certain number of events in a given time interval can be easily calculated from (6.4), and a meaningful representation of the uncertainty over λ is given in the process. An illustration of the Bayesian updating process is shown in Figure 6.2. This process essentially returns some quantified uncertainty on the predictions of new observations,

which can be useful for the detection of anomalous data departing from the inferred distribution.



(a)



(b)

FIGURE 6.2: (a) Bayesian updating of the Gamma prior $p(\lambda)$, when presented with N samples drawn from a Poisson distribution parameterised by $\lambda = 10$. (b) Predictive posterior distribution $p(x_{new}|\mathbf{x})$, evaluated given $N = 100$ observations, and compared against the Poisson likelihood $p(\mathbf{x}|\lambda = 10)$.

6.1.2 Detection strategy and case study

The strategy that follows assumes that the extracted counts from the signal are distributed according to a Poisson distribution. In order to demonstrate how this reasoning can be implemented in practice, the considered AE time-series will be limited to unprocessed recordings in which distinct transient events can be observed occurring sporadically in a continuous stream of background noise. This type of waveform may be referred to as a *burst-type wave streams* [172]. Under this consideration, it is worth noting that a relatively-higher count-rate can be expected in sections encompassing an AE event.

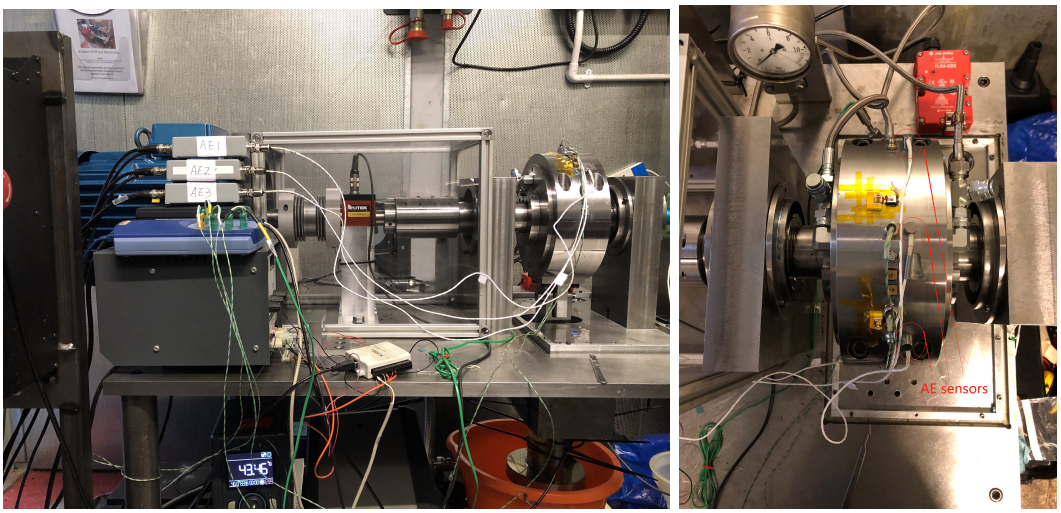
Now, given a high percentile threshold (in the range between 95% to 99.9%), one can let the Poisson distribution model the counts extracted from the background noise. The first step would be to construct a dataset $\mathbf{x}$, comprised of uni-dimensional features representing the counts observed in a windowed section that correspond uniquely to background noise, or sections where no obvious AE activity exists. It is then easy to imagine that the counts in each section will be low, and one would end up with an array comprised of a combination of zeros and/or random positive integers near zero. This approach may seem counter-intuitive at first since the background noise is being modelled rather than the actual AE events, but the results may be easier to interpret when evaluating the *Negative Log-Likelihood* (NLL), of new observations with respect to the already inferred distribution in (6.3),

$$\text{NLL}(x_{new}) = -\log(p(x_{new}|\mathbf{x})) \tag{6.5}$$

This strategy treats AE events as anomalies in the time-series, and evaluating (6.5) for new observations should therefore return higher-than-normal values whenever encountering sections in the time-series with AE events.

The case study considered for the following demonstration examines AE wavestreams recorded by Dr. Liqun Wu from the University of Sheffield. The AE equipment was installed on the same journal bearing test rig introduced in Subsection 5.1.2. Photographs of the experimental setup are shown in Figure 6.3. The sensors used were *WB Mistras AE sensors*, pre-amplified to 40dB. Three of these sensors were attached directly to the bearing housing using grease as the acoustic couplant. The data acquisition system used for the recording was the *Mistras Micro-II compact PCI AE Chassis*. Figure 6.4 shows a section of the AE signal recorded during the test. For this demonstration, a 99.5^{th}-percentile threshold was employed.

The problem, as highlighted earlier, is that counts extracted from the background noise would also be taken into consideration, when only the main bursts observed in the signal



(a) (b)

FIGURE 6.3: Photographs from experimental setup showing the (a) main test rig and (b) sensor arrangement on the bearing housing.

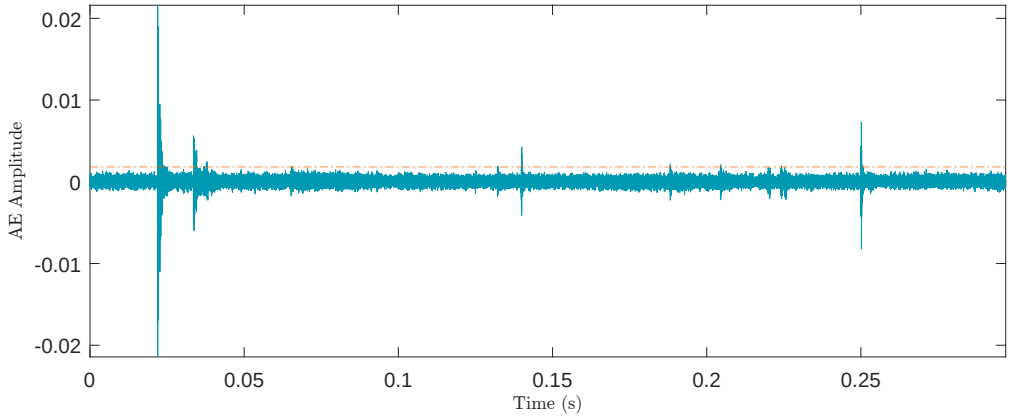
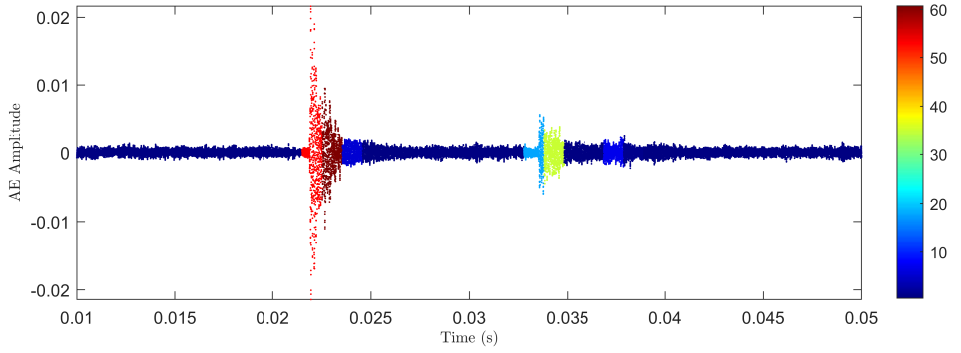


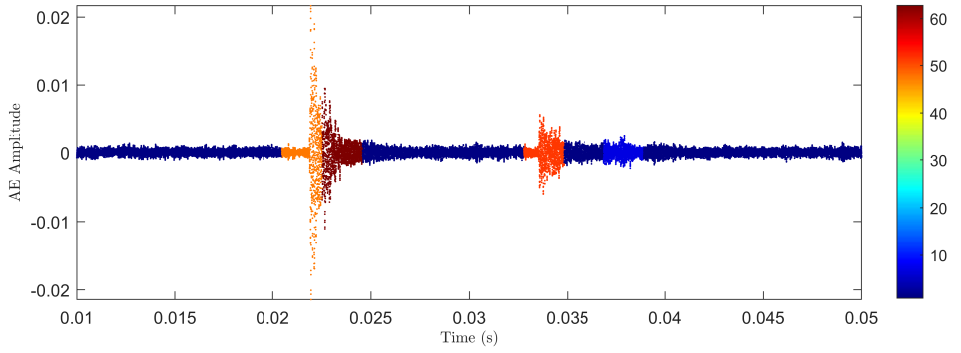
FIGURE 6.4: AE readings from a plain journal bearing operating under hydrodynamic lubrication at a rotational speed of 150rpm and applied static load of 5kN. The dashed line corresponds to a 99.5^{th} percentile threshold.

are of actual interest. Some progress can be achieved by adopting the Bayesian approach described above, where the count-rate λ, associated with the background noise, can be assumed to be distributed according to a Gamma distribution, with corresponding parameters set to units ($a = b = 1$). This assumption is based on the expectation that λ will be zero or near zero, given the relatively-high threshold established in this demonstration. The reasoning behind this assumption is again based on treating the actual AE bursts as "rare" events. However, it should be acknowledged that careful tuning of these parameters may be necessary to account for the expected counts given a specific threshold. While the consideration of priors in Bayesian modelling is crucial, exploring the choice of priors for this specific application would require additional extensive work, potentially diverting the reader's attention from the main purpose of this demonstration. Investigating the selection of priors in this context remains a subject for future research.

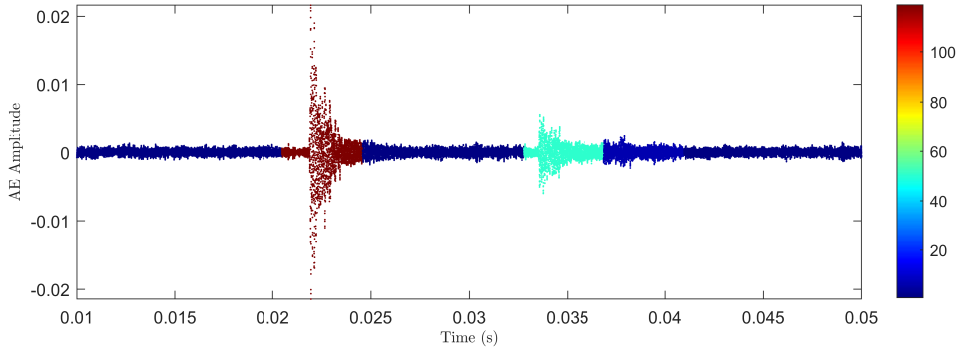
After windowing the entire time-series with a non-overlapping square function, 20 samples were extracted and used to infer the posterior distribution. It is reiterated here that these samples are counts from sections known to lack any form of meaningful AE activity. The probability of observing x number of threshold crossings in a given section was calculated using Equation (6.4). In order to quantify the confidence of an AE event existing within a section, Equation (6.5) was evaluated on each of the remaining sections. The results are shown in Figure 6.5, where colour-bars representing the NLL were included. The sudden increases in NLL can be interpreted as an anomaly being detected, which in turn happens to correspond to the AE events of interest.



(a)



(b)



(c)

FIGURE 6.5: Negative log-likelihood evaluation over section of wavestream in Figure 6.4. The colour bar represents the likelihood of an event existing in a given time interval. Window lengths covered (a) $n = 1024$, (b) $n = 2048$, (c) and $n = 4096$ data points.

An important point worth discussing is the immediate limitation that becomes evident when assessing the sensitivity of the model to the window-length. The results presented in Figure 6.5(c) seem more promising because a window-length comprised of 4096 sample-points somewhat matches the length of the main AE event. Unfortunately, knowing in advance the length of all potential AE events is unlikely to be the case, especially when dealing with more complex applications. The effectiveness of this method clearly depends on the length of the sliding window, as shown more clearly in Figures 6.5(a) and 6.5(b). By having a window-length shorter than the duration of the AE event, the model interprets several components in what should be a single event. The upside of establishing a shorter window-length is that the onsets of the events are captured more precisely. A validating step would become essential to determine the optimal window-length. In both cases, nonetheless, the background activity is mostly identified as noise.

Before concluding this section, it may be important to acknowledge that this exercise has been conducted under the assumption that the noise is purely acoustic. However, in practice, there may be other sources of background noise, such as AE activity induced by *Electronic Magnetic Interference* (EMI) [176], that could introduce complexities to the analysis and thereby necessitating more elaborate monitoring strategies. Exploring this issue further is beyond the scope of this study, but it certainly merits further investigation.

6.2 Towards a Nonparametric Clustering Approach for AE Event Detection

The case study above was presented merely as a demonstration to provide some insights into the implementation of a Poisson distribution for modelling AE data. Some involvement is still required in deciding whether the increase in NLL is, in fact, an indication of an AE wave; that is, one would need to determine how much the NLL must increase in some windowed sections of the wave stream to ascertain the presence of AE activity. Therefore, in the interest of enhancing the detection process and possibly identifying AE waves in the stream, it may be necessary to extend the model into a mixture of Poisson distributions. Such an extension would, indeed, introduce complexities to the model since the problem would then be that of density estimation. However, the mixture model would autonomously group the windowed sections based on their respective likelihood measures, thereby using ML to determine which sections correspond to AE waves and which to background noise, rather than having to do so manually. As stated

earlier, the aim here is to keep the number of variables to a minimum in order to reduce dependence on the expertise of the operator.

The premise of the mixture model in this case is to have the events of interest modelled by a set of independent distributions. Each of these distributions is assumed to represent a different mechanism responsible for the generation of the observable AE events. In doing so, the approach now aims to infer the parameters of the various components comprising the mixture model. While this consequence adds adjustable parameters that must be tuned (i.e. λ_i for each independent distribution $i = 1, 2, \dots$), these can be marginalised out, resulting in a nonparametric version of the model. This outcome is shown in the following section.

Unfortunately, it must be noted that this idea is still subject to the problem of finding a suitable window length, as demonstrated in Section 6.1.2. Before dwelling on ways in which the robustness of the model can be improved, some background on mixture models will be covered so that this chapter is as self-contained as possible. Interested readers can refer to [48] or [175] if they wish to learn more about the machine-learning methods covered in this chapter.

6.2.1 The parametric approach: Finite Poisson Mixture Model

Recalling the pre-processing steps followed so far, let a feature vector X be constructed by extracting the number of threshold crossings observed in subsections of a raw AE time-series, encapsulated by a sliding step-window of length n. As the window slides through the signal, the number of threshold crossings should vary with changes in the AE activity. The distribution of X will be such that a single Poisson distribution may poorly represent the observations. Therefore, a sensible choice may be to instead model the observations as draws from a combination of several independent Poisson distributions.

First, a multinomial distribution is proposed over a vector of mixing proportions $\boldsymbol{\pi}$, in which each element π_k, is a mixing coefficient defining the probability that an observation belongs to the k^{th} group. The total number of groups K, is predefined, and $\sum_K \pi_k = 1$. It should be noted that the collection of data-points assigned to a distribution in the mixture model will be referred to as a "group". In the literature, the term group is also used interchangeably with "cluster".

Part of the inference procedure is to derive a latent variable $\mathbf{z}$, representing the assignment of a data-point to one of the groups in the mixture model (i.e. its labels). If each group is defined by a unique Poisson distribution with count-rate λ_k, then each

observation x_n can be modelled in the following form,

$$\begin{aligned} z_n &\sim \text{Mult}(\boldsymbol{\pi}) \\ x_n|z_n &\sim \text{Poi}(\lambda_k) \end{aligned} \tag{6.6}$$

All the model parameters can be then determined efficiently via *Expectation-Maximisation* (EM) [48], resulting in a maximum-likelihood solution, given the data that have been observed so far. Although this model may be a better representation of the observations, it can still be limited in practice. If only a few data points were available, having the local parameters estimated empirically could result in the model not generalising well in the presence of new observations.

Another consideration is in deciding autonomously on the number of groups the mixture model should have. A possible solution is to assume that the observations can be represented by two unique groups; that is, one characterised by the relatively low count-rate corresponding to background noise, and another that accounts for all the possible higher count-rates that may correspond to AE events. The sequence of sections in the time-series should then be grouped into one in which some form of AE activity exists, and another in which minimal or no events are found. By discarding the latter, one is left with a series of AE events ready to be processed for feature extraction, as required by the application at hand.

However, it might be more convenient to take a step further and split the detected AE events into several groups, in order to get a better insight into their respective sources. For example, it might be better to know whether an AE event corresponds to a specific damaged component, rather than plainly categorising it together with all the rest. The problem, however, is in deciding *a priori* the number of groups, since one would need to consider every possible condition that a structure or machine might encounter during its lifespan. One way to circumvent this problem is by having the model autonomously create new groups whenever it may be deemed necessary. This approach will require yet another modification, to provide the mixture model the flexibility to account for a possibly *infinite* number of groups.

6.2.2 The nonparametric approach: Dirichlet Process Poisson Mixture Model (DP-PMM)

This subsection extends on the theory outlined in Chapter 4. In particular, the Dirichlet Process is used here for the derivation of an infinite mixture model composed of independent Poisson distributions. A graphical representation of the infinite Poisson mixture model is shown in Figure 6.6. By establishing a DP-prior on the mixture model, a

distribution over K can also be inferred from the observations directly as part of the learning procedure, making DPs an attractive solution in SHM/CM when faced with a dataset comprised of a collection of unknown conditions. A detailed introduction to the infinite mixture model can be found in [108], and examples of its use in SHM can be found in [109, 110].

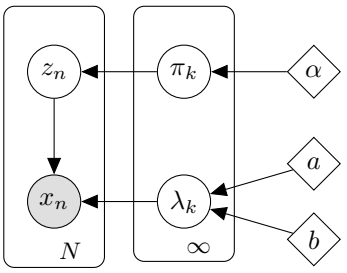


FIGURE 6.6: Graphical model of the infinite Poisson mixture model.

To begin, it will be necessary to make the inference over the parameters Bayesian. This step will lead nicely into the incorporation of a DP-prior over K. Firstly, the priors over λ_k are considered. As already reviewed in Section 6.1.1, the Gamma distribution is chosen once again to represent the prior belief on λ. Assuming that the rate values, λ_k, of each group are independent, allows the specification of the joint density,

$$p(\lambda_1, ..., \lambda_K | a, b) = \prod_{k=1}^{K} p(\lambda_k | a, b) \tag{6.7}$$

Now, the vector $\boldsymbol{z}_n$ is characterised according to a multinomial distribution parameterised by $\boldsymbol{\pi}$,

$$p(\boldsymbol{z}_n | \boldsymbol{\pi}) = \prod_k \pi_k^{z_{nk}} \tag{6.8}$$

where the normalising constant simplifies to unity, and the expression is reduced to the probability of assigning a group. A Dirichlet distribution is chosen as a suitable prior on π, and is also conjugate to the multinomial distribution. The probability density function of the Dirichlet distribution is given by,

$$p(\boldsymbol{\pi} | \boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)!}{\prod_k \Gamma(\alpha_k)} \prod_k \pi_k^{\alpha_k} \tag{6.9}$$

which is parameterised by $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_k\}$. Having defined the densities on the parameters, the joint posterior likelihood is given by,

$$p(\boldsymbol{Z}, \boldsymbol{\pi}, \lambda_1, \cdots, \lambda_k | X, \boldsymbol{\alpha}, a, b) \propto \left[\prod_n \prod_k (\pi_k p(x_n | \lambda_k))^{z_{nk}} \right] \left[\prod_k p(\lambda_k | a, b) \right] p(\boldsymbol{\pi} | \boldsymbol{\alpha}) \tag{6.10}$$

The conditional distributions for each parameter are proportional to the joint distribution and can be derived analytically to give a Gibbs sampler [48], to solve equation (6.10). However, it will be worth marginalising out $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ from the joint posterior to instead implement a collapsed Gibbs sampler [177]. This last step will not only improve the robustness of the sampler, but it will be a necessary one to have the model extend to an infinite number of components. The collapsed distribution is finally expressed as,

$$p(z_{nk} = 1|\boldsymbol{Z}^{-n}, X^{-n}, \boldsymbol{\alpha}, a, b) \propto \frac{c_k^{-n} + \alpha_k}{\sum_j^K c_j^{-n} + \alpha_j} p(x_n|\boldsymbol{Z}^{-n}, X^{-n}, a, b) \tag{6.11}$$

where,

$$p(x_n|\boldsymbol{Z}^{-n}, X^{-n}, a, b) = \mathrm{NB}(\boldsymbol{r}, \boldsymbol{p}) \tag{6.12}$$

and,

$$\boldsymbol{r} = a + \sum_{m \neq n} z_{mk} x_m, \quad \boldsymbol{p} = \frac{\sum_{m \neq n} z_{mk} + b}{\sum_{m \neq n} z_{mk} + b + 1}$$

The new notation c_k^{-n} corresponds to the count of all but the n^{th} element in the k^{th} group, $\boldsymbol{Z}^{-n}$ to the set of all assignments except for the n^{th} one, and X^{-n} to the set of all the other observations; the superscript $-n$ does not represent a power here. The expression in (6.11) is also simplified by considering the value where the k^{th} element of $\boldsymbol{z}_n$ is equal to unity, while all others are zero. A full derivation is provided in B.1. In this form, $\boldsymbol{z}_n$ can now be resampled directly from Equation (6.11). The two parts of this expression can be interpreted as comprising a new prior,

$$p(z_{nk} = 1|\boldsymbol{\alpha}, \boldsymbol{Z}^{-n}) = \frac{c_k^{-n} + \alpha_k}{\sum_{j=1}^K c_j^{-n} + \alpha_j} \tag{6.13}$$

and a likelihood defined by the negative binomial density with parameters $\boldsymbol{r}$ and $\boldsymbol{p}$. In this form, an infinite number of components can be managed, if the parameters α_k in (6.9) are all set to the same value. The Dirichlet prior can then be expressed as,

$$p(\boldsymbol{\pi}|\alpha) = \mathrm{Dir}(\alpha/K, ..., \alpha/K) \tag{6.14}$$

and the sampling prior for $z_{nk} = 1$ becomes,

$$p(z_{nk} = 1|\alpha, \boldsymbol{Z}^{-n}) = \frac{c_k^{-n} + \alpha/K}{\alpha + N - 1} \tag{6.15}$$

where the fact that $\sum_{j=1}^{K} \alpha/K = \alpha$ and $\sum_{j=1}^{K} c_j^{-n} = N - 1$ were used. Now, by letting $K \to \infty$, equation (6.15) reduces to,

$$p(z_{nk} = 1|\alpha, \boldsymbol{Z}^{-n}) = \frac{c_k^{-n}}{\alpha + N - 1} \tag{6.16}$$

This new expression defines the prior probability of the n^{th} data point going into the k^{th} group, and it is proportional to the number of members that currently exist in that group. So far, the new prior only accounts for groups in which members exist, but in this framework, there is always a non-zero probability for a data-point to be assigned to a new "empty" component that is yet to be created. The probability of a new member finding its place in one of these empty groups can be easily derived by computing 1 minus the total probability of it going into any of the non-empty components. Hence,

$$p(z_{nk_*} = 1|\alpha, \boldsymbol{Z}^{-n}) = 1 - \sum_{k=1}^{K} \frac{c_k^{-n}}{\alpha + N - 1} = \frac{\alpha}{\alpha + N - 1} \tag{6.17}$$

where the subscript $*$ in k_* denotes an empty component. One can notice that the prior defines a probability of a new member either going to a non-empty component proportional to the number of its members, or to one of an infinite number of empty components proportional to α. Implementing this reasoning may begin by assigning all observations to the same component, and then resampling the assignments of each observation with probability,

$$p(z_{nk} = 1|\alpha, \boldsymbol{Z}^{-n}) = \begin{cases} \frac{c_k^{-n}}{\alpha+N-1} & \text{for } c_k^{-n} > 0 \\ \frac{\alpha}{\alpha+N-1} & \text{for } c_k^{-n} = 0 \end{cases} \tag{6.18}$$

After incorporating the data, an observation is assigned by sampling from $p(z_{nk} = 1|...)$.

When accounting for an infinite number of components ($K \to \infty$), one can move from having a prior that determines how the observations group together among a fixed number of components, to one that can partition the data in any possible number of them. It should be noted at this point that the term "component" has been introduced to refer to the collection of all groups that are "empty" and "non-empty". The former corresponds to groups that are yet to have data-points assigned to them, and the latter to those that already exist in the model.

Finally, the result yields an infinite Gibbs sampling procedure that involves resampling each $\boldsymbol{z}_n$ according to the probabilities,

$$p(z_{nk} = 1|...) \propto \begin{cases} c_k^{-n} p(x_n|\boldsymbol{Z}^{-n}, X^{-n}, a, b) & \text{for non-empty components} \\ \alpha p(x_n|a, b) & \text{empty components} \end{cases} \tag{6.19}$$

where $p(x_n|\boldsymbol{Z}^{-n}, X^{-n}, a, b)$ is given by equation (6.12), and $p(x_n|a, b)$ by the expectation of the likelihood, $p(x_n|\boldsymbol{\lambda})$, with respect to the prior $p(\boldsymbol{\lambda}|a, b)$,

$$p(x_n|a, b) = \int p(x_n|\boldsymbol{\lambda})p(\boldsymbol{\lambda}|a, b)\,\mathrm{d}\boldsymbol{\lambda} \tag{6.20}$$

which, in this instance, is possible to evaluate because $p(x_n|\boldsymbol{\lambda})$ is a Poisson distribution parameterised by $\boldsymbol{\lambda}$, and $p(\boldsymbol{\lambda}|a, b)$ is its conjugate Gamma prior.

The resampling procedure can now be dictated as follows:

1. All elements are assigned to a non-empty component ($K = 1$).
2. One element is removed from its component and its likelihood is evaluated with equation (6.19). The component is eliminated if the element is unique in that component.
3. If the unassigned element is more likely to belong to an empty component, then a new component is created, and the data point is assigned to that component.
4. One sweep is completed when each data point has been evaluated. The whole process is then repeated from Step Two until the posterior converges to a solution.

It must be noted that inferring (6.19) is not limited to a collapsed Gibbs sampler and variational methods have been developed for inference of Dirichlet process mixture models [130]. The sampling method here was preferred for its simpler intuition. The DP-PMM will be now demonstrated to model the AE events in a time-series signal.

6.3 Automated AE Event Identification in Time-Series Signals

Recovering AE events from a raw time-series can be a challenging task, particularly as the behaviour of an event will differ from one to the next, and also because consecutive events may also overlap. As mentioned earlier, identifying and extracting these events may be an assiduous task, but a meaningful one, as it should reveal information on the health state of the machine being monitored.

By introducing a DP prior, the mixture model can now adapt to the complex nature of the AE analysis. Compared to the methods already described, an infinite mixture model may be better suited for the detection of individual AE events in the time-series. Even without knowing the exact nature of the generated AE signals, this approach should

assign the various events into groups based on their features. Given that the feature of interest here is the count of an event, these groups are therefore modelled by a set of independent Poisson distributions.

The AE wavestream shown in Figure 6.4 will be again considered for this demonstration. For feature extraction, a sliding window approach is again considered. As demonstrated in Section 6.1.2, the immediate challenge of this approach is finding a suitable window-length; if too long, it will be less likely to discern individual events and their counts. Conversely, if too short, a single event could mistakenly be interpreted as a collection of them. Under these considerations, the window-length should at least match the duration of the shortest event in the signal. However, it may be impossible to know this information beforehand, and some form of validation would be needed to find the best parameters. If the geometry and materials of the propagating medium are simple enough, some insight can be gained by simulating the potential events. Since this information was not readily available, a step function with a length of $n = 2048$ sample points ($\Delta t = 0.002048$s) was chosen for the windowing procedure. The choice of this specific length was based on [45], where the same window length was used to record individual AE waves.

Unfortunately, this step alone is limited by the fact that the sliding window is not guaranteed to align with the AE events in the signal. To address this issue, it is necessary to overlap the imposed windows. A window overlap of 87.5% was thus used for this demonstration. The following section dwells on the importance of overlapping and demonstrates how it improves the robustness of this approach. Having windowed the signal, the count was then extracted by taking the number of times the signal was found to go over a pre-defined threshold. In this case, a 99.5^{th} percentile threshold was defined over the entire waveform. It is worth reiterating that, although a hard threshold is established here, the probabilistic approach should alleviate the aforementioned shortcomings of having a statistically-determined threshold. A threshold crossing will no longer be taken to derive from an AE event; instead, a probability will quantify the (un)certainty of there being an AE event.

With the constructed feature set, the DP-PMM was implemented to infer a suitable partition of the extracted counts. The resulting clusters depend strongly on α, and even after having marginalised the underlying parameters in Equation (6.19), it is still necessary to find the optimal value for α. One way to achieve this is by running the Gibbs sampler multiple times, each time with different values of α, and evaluating the probability of data partitioning into K clusters. The probability mass function $p(K|X, \alpha)$, can be computed by keeping track of the number of inferred clusters at each iteration of the Gibbs sampler. The outcomes of this process, given different α values,

are presented in Figure (6.7), highlighting that the most likely partition was attained when $\alpha = 0.01$. This finding suggests that the data are optimally represented by four distinct clusters. Consequently, the following section presents the results obtained when $\alpha = 0.01$. In all cases, the Gibbs sampler was conducted over $10,000$ iterations.

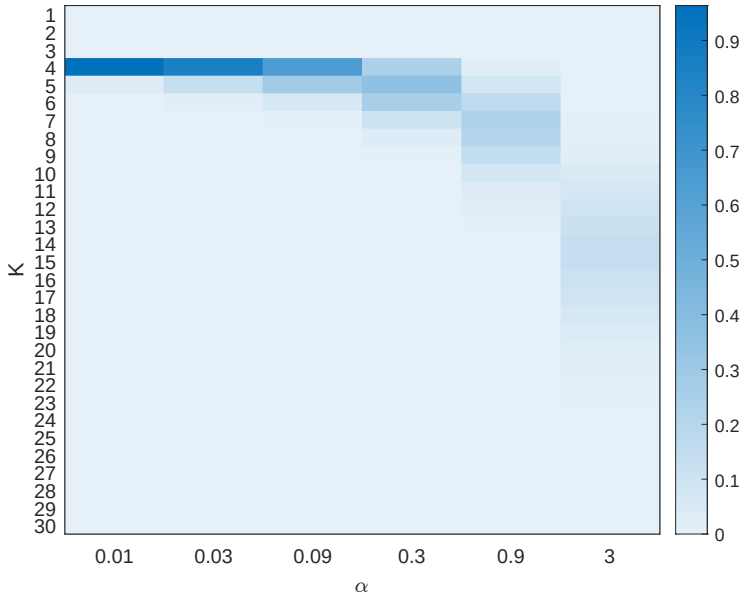
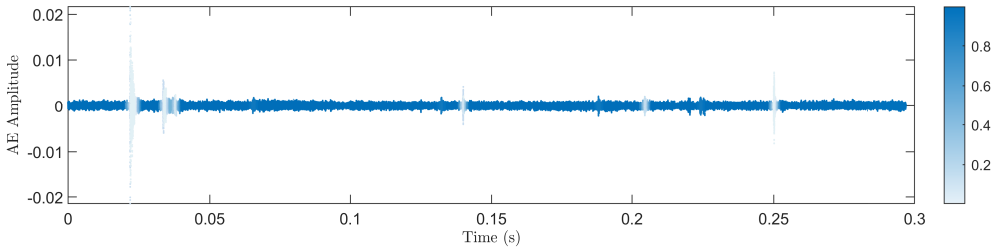


FIGURE 6.7: Predictive likelihood for the number of clusters K given α. The colour-bar indicates the outcome of $p(K|X, \alpha)$.

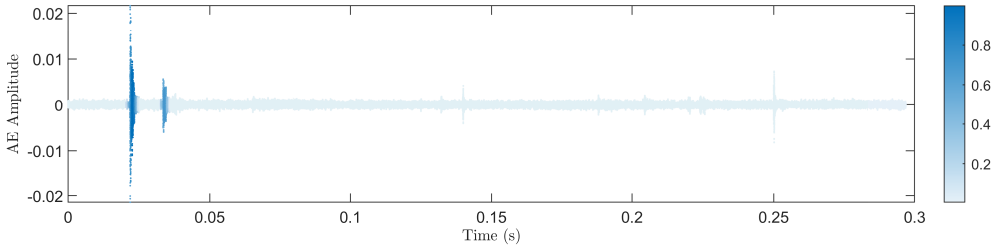
6.3.1 Results and Discussion

The probability of the assignments inferred by the proposed strategy is shown projected on the signal in Figure 6.8. Visually, it seems as if the main acoustic events were almost undoubtedly detected in all cases. A distinction among the clustered AE events can also be observed in the results. It appears as if the first group mostly captures the background noise, while the remaining groups capture the different AE events found in the signal. If it is wished to categorise the identified events, one can assign the regions in signal to the group that presents the highest probability. The result of this outcome is illustrated in Figure 6.9. Not only are these events identified with minimal intervention, but they are also clustered in a way that distinguishes the more "imposing" waves, such as the one found in Group 2, from the "smaller" ones, like those found in Groups 3 and 4.

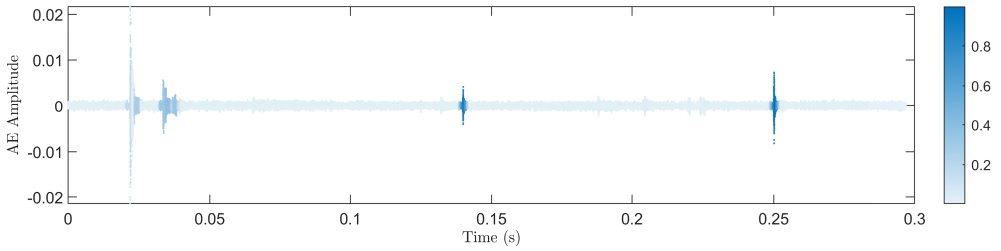
One of the advantages of having the AE events grouped in this way is that it provides an organised framework from which one can proceed with the analysis. For example, only events from Group 2 could be considered for extracting their remaining features, such as amplitude, rise-time and energy (among others). This approach significantly reduces the dimensionality of the original dataset and encodes most of the information



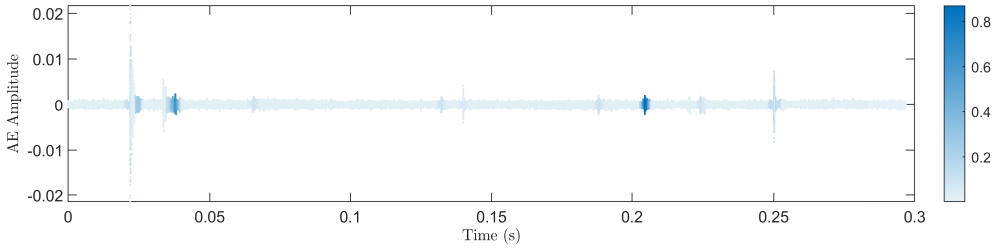
(a) $p(z = 1|\dots)$



(b) $p(z = 2|\dots)$



(c) $p(z = 3|\dots)$



(d) $p(z = 4|\dots)$

FIGURE 6.8: Visualising the assignment probabilities across the AE signal. The colour-bar corresponds to the predictive posterior probability inferred by the Gibbs sampler.

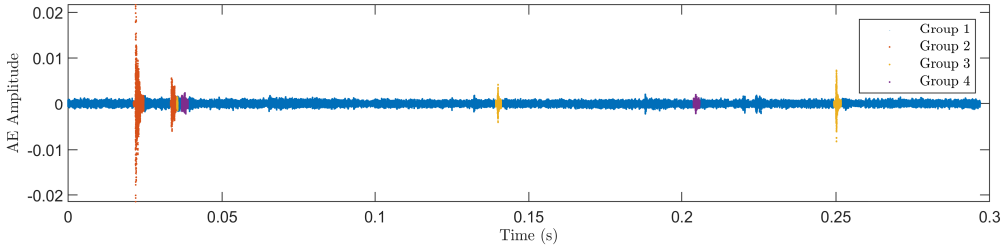


FIGURE 6.9: Assignment of sections in the signal to the groups that maximise the assignment probabilities. Four distinct groups of AE activity were inferred in this signal.

needed to diagnose the state of the structure or machine being monitored [1]. An emerging defect should affect these features to some extent, giving an anomaly detector an indication that the system has deviated from normal conditions. Alternatively, a new group could develop as damage becomes more prevalent, and one could then query this new component and decide whether it arises from further damage. The nature of the DP-PPM in this context could allow for online monitoring without having to initiate a new training period.

An additional advantage of this approach is that it provides a principled solution to the problem of deciding on the optimal parameters for feature extraction; most notably, the threshold, window length and amount of overlapping. It has been made evident that these parameters greatly influence the outcome of the model since these govern the outcome of the feature extraction. Given that the model is Bayesian, it is theoretically possible to also marginalise these parameters. Letting a particularly meaningful combination of these parameters define a model $\mathcal{M}_j$, and having a collection of J finite and discrete models $\mathbf{M} = \{\mathcal{M}_1, \dots, \mathcal{M}_J\}$, the predictive posterior can be evaluated by marginalising over each of the individual models. In particular,

$$p(\mathbf{z}_n|\dots) = \sum_{j=1}^{J} p(\mathbf{z}_n|\mathcal{M}_j, \dots)p(\mathcal{M}_j) \tag{6.21}$$

Assuming an equal contribution from each model, then $p(\mathcal{M}_j) = 1/J$, and the expression in (6.21) simplifies to the mean of the inferred assignment probabilities obtained from each $\mathcal{M}_j$. There are, indeed, some practical limitations to this proposal. The expense of having to evaluate an exhaustive combination of parameters would be computationally prohibitive. Nevertheless, having to overlap the windows inevitably necessitated the evaluation of Equation (6.21), and this step was in fact employed to attain the results presented in this section.

An 87.5% overlap (or 12.5% offset) led to the formation of seven distinct sets of feature vectors, with each set considered as a different model realisation $\mathcal{M}_j$. Therefore, this implication required inferring $p(\mathbf{z}_n|\mathcal{M}_j, \dots)$ for each model, before averaging the resulting probabilities. To illustrate this process, Figure 6.10 shows the effects of the overlapping offsets for the main AE event found in Figure 6.8(b), and the result of marginalising $\mathbf{M}$. The top two figures show the assignments inferred given two arbitrary models. It can be seen that somewhat of a good alignment is achieved in the first case, but when an offset is introduced, the partitioning "cuts" the event short. Figure 6.10(b) shows that the section preceding the actual event is assigned to Group 1. Most of this section is in fact attributed to noise, but an initial portion of the event is captured at the end-edge of the window, misleading the model into assigning this section incorrectly. Although

shown otherwise with this particular AE event, offsetting the windows is necessary for all existing events in the signal to be, at least once, somewhat aligned to the windows. Eventually, the evaluation of Equation (6.21) alleviates the potential discrepancies between models, yielding a balanced assignment and correct identification of the event, as shown in Figure 6.10(c). This reasoning can also be extended to incorporate different thresholds and/or window-lengths. However, as mentioned earlier, the inclusion of additional models requires inferring the DP-PMM over a larger dataset, which could be infeasible in practice. This issue will be the subject of further investigation in future work, but a brief example of marginalising the threshold can be found in Appendix (B.2).

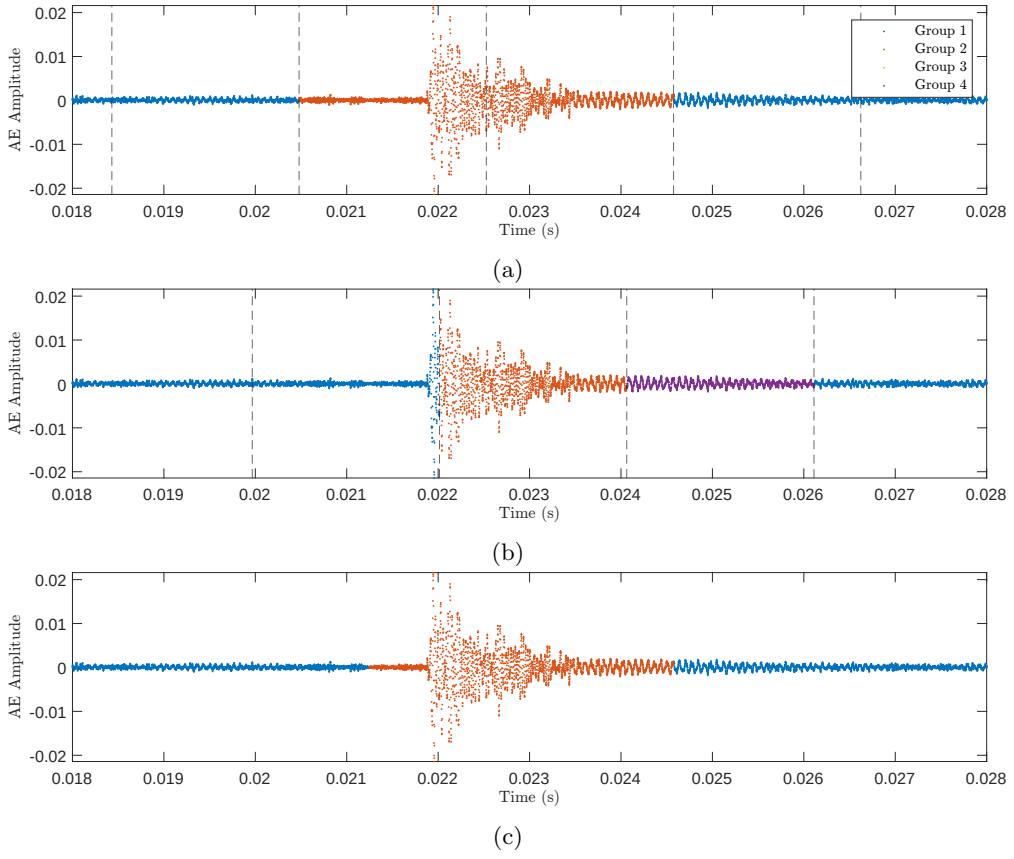


FIGURE 6.10: Close-up view of the main AE event found in Figure 6.8(b). The results demonstrate the effects of overlapping the windows for feature extraction, where the assignment of sections in the signal are determined with (a) no overlapping offset, and (b) a 62.5% overlapping offset - $(5/8)N_{\text{window}}$. (c) Resulting assignment of the AE event by averaging over each of the individual models.

Finally, an intriguing implication of implementing the probabilistic paradigm for AE identification is that it can naturally indicate the possible start and end points of the AE event by considering the region where the probabilities exceed some minimum value. As shown in Figure 6.11, the assignment-probability experiences gradual changes as the AE event unfolds and subsides. This consequence indirectly provides a window-length that

potentially adapts to the true duration of the AE event. The implication of this outcome offers an additional advantage in the AE wave extraction process. In the conventional approach, a window length must be carefully set in the data acquisition system. Given that the number of samples in the window is fixed, this adjustable variable may be too long for some AE events, or similarly, too short for others. The probabilistic approach seems to elegantly overcome this issue by providing a window length that autonomously adapts to each identified AE event.

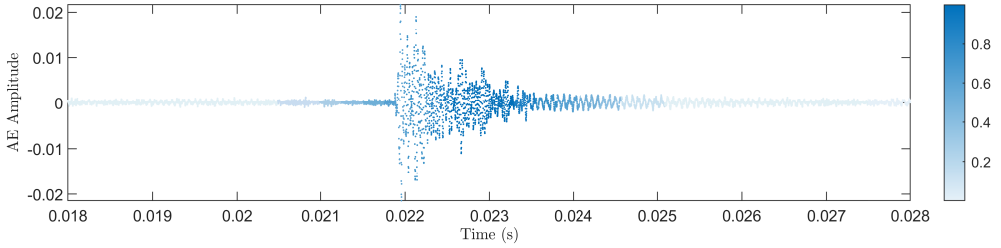


FIGURE 6.11: Close-up view of the main AE event found in Figure 6.8(b) displaying the probability of being assigned to Group 2.

6.4 Application to Landing Gear

Identifying AE events in a signal is an important pre-processing step when performing this type of AE-based monitoring. As demonstrated in the previous section, a probabilistic approach may offer several advantages over traditional methods in the identification of AE events, and with the added flexibility of the mixture model, the DP-PMM may be an effective alternative for managing what can be an overwhelming dataset.

An additional application of the DP-PMM in SHM is explored in this section where the modelling approach is employed directly for damage detection, rather than a practical pre-processing step. The experimental study under consideration is the fatigue testing of a 300M steel lug welded to an *Airbus A*320 main fitting. This study originally corresponds to one of a series of projects conducted by The University of Cardiff in collaboration with The University of Sheffield, and under the support of Messier-Dowty Ltd. The interested reader is encouraged to refer to the work of Holford et al. [178] for full details. The series of experiments carried out investigated the use of AE techniques for the detection and localisation of fractures in certified landing gears under fatigue testing.

This dataset has been analysed before in [45], where spatial-scanning statistics were used to localise the emerging source of damage in the landing gear. It is argued that the proposed localisation strategy works better because of the complications background activity introduces when monitoring sharp up-turns in the energy rate of AE events,

thereby suggesting that simply measuring the energy rate for the test is insufficient for detecting fractures. In the following, it is shown that the early identification of fracture is indeed possible by monitoring the energy rate, even when large amounts of background activity are present. The premise of this approach is to implement the nonparametric probabilistic techniques presented above to filter out background activity and "benign" AE events from the data and retain AE events that are relevant for the monitoring scheme.

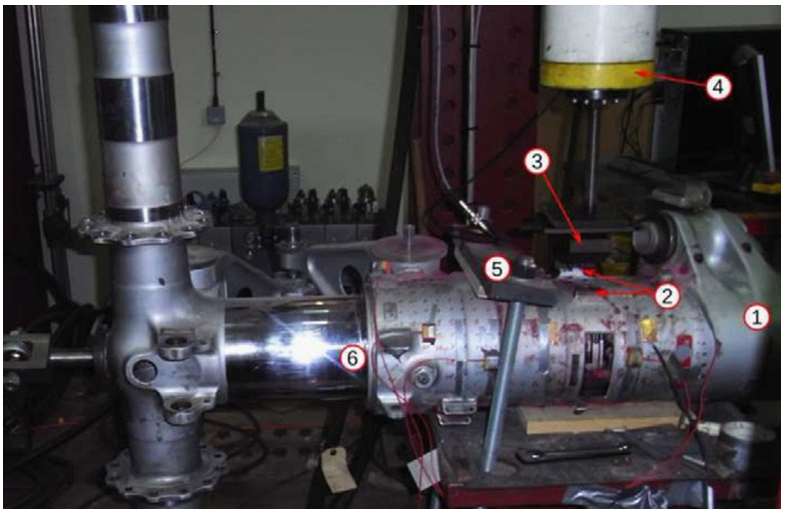
6.4.1 Experimental Method

The dataset explored here is an extensive collection of AE waveforms recorded throughout the fatigue test. A loading arm attached to the main-fitting lug was used to transmit a cyclic load at a rate of 1Hz, with an initial peak amplitude of 5.5kN, and R ratio 0.1. The load was later increased to 6kN, 6.5kN and 7kN after $90,000$ cycles, $110,000$ cycles and $138,500$ cycles, respectively. Gradually increasing the load was necessary to promote crack growth, eventually leading to rupture after $160,000$ cycles ($200,000$s). The test was made more "realistic" by also exciting the sliding tube of the main fitting, periodically at a rate of 0.4Hz and with a travel of 40mm. The added contribution of the sliding tube was included to promote the generation of benign AE events, as would be expected outside laboratory conditions.

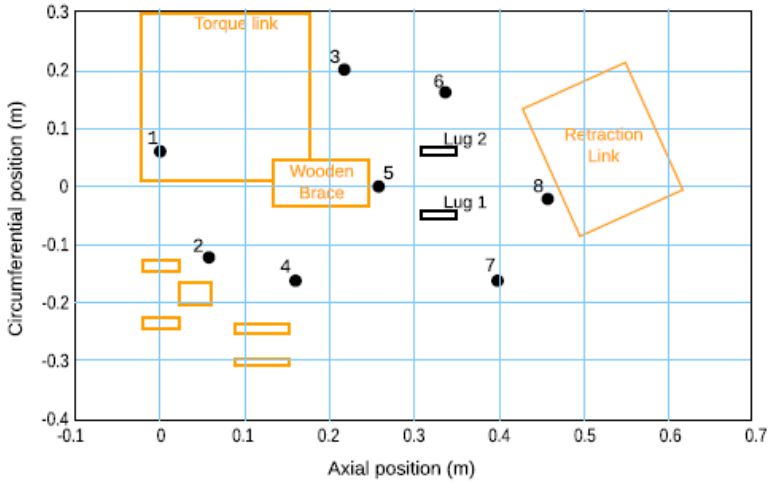
The main fitting was mounted with eight *Physical Acoustics Limited (PAL) Nano* 30 sensors around the cylindrical part of the landing gear (Figure 6.12(a)). These sensors have a frequency response in the range $125 - 750$kHz and with a resonance of 300kHz. All sensors were attached to the structure with magnetic clamps, and brown grease was used as an acoustic coupling. Using a $PALPCI - 2$ acquisition system, the measured data were pre-amplified and recorded at a sampling rate of 2MHz. Fixed-size windows of 2048 samples were recorded upon the signal crossing a pre-established threshold of 43dB. A memory buffer remained active to include 500 sample points before the trigger event of each waveform.

6.4.2 Feature Selection

From the sensor-arrangement presented in Figure 6.12(b), only the waveforms recorded from Channel Five were used in this analysis, given that the corresponding sensor was the one placed nearest to the lug subjected to the fatigue load. For the duration of the fatigue test, this sensor accounted for a total of $755,193$ recorded waveforms. Handling this amount of data quickly became a complication, and it was necessary to extract features from the recorded waves in order to proceed with the analysis. Following the



(a)



(b)

FIGURE 6.12: (a) Experimental set-up and (b) sensor arrangement (red dots) (photograph and schematic from Hensman et al. [45]). The main landing-gear fitting (1), can be seen to be comprised of the lugs (2), loading arm (3), load actuator (4), restraining wooden brace (5), and a sliding tube (6).

same reasoning as in the previous case study, the counts of each waveform were extracted and used as features for the analysis. In particular, a feature vector $X = \{x_1, ..., x_N\}$ representing the N extracted counts was constructed for the development of the damage-detection algorithm. Figure 6.13 shows the evolution of the counts as the waveforms were recorded during the fatigue test. The dataset is rich in information, and it is not all that clear whether an obvious pattern exists that could indicate a form of fracture emerging and evolving until failure. Nevertheless, some peaks are observed more frequently towards the end of the test, suggesting that a more elaborate monitoring mechanism could anticipate failure and warn an operator of the urgency for an inspection of the affected part.

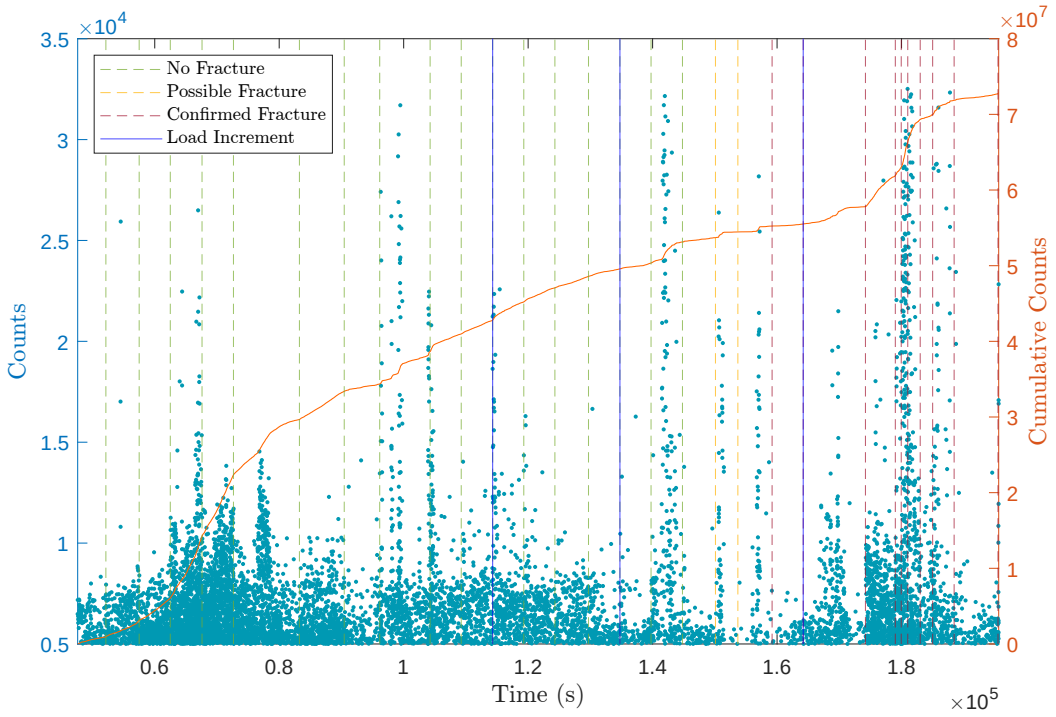


FIGURE 6.13: Experimental results: reduced dataset ($N = 10,000$). The retained data in the reduced set correspond to a subset starting at $\sim 47,000$s into the fatigue test.

Using a collapsed Gibbs sampler to infer the cluster assignment has the advantage of evaluating each sample-point marginally. The implication of this advantage is that the model can adapt in real-time as new sample-points are introduced, allowing for an online monitoring approach. However, as the dataset grows with the addition of new sample-points, inference becomes increasingly slower, to the point where inferring the posterior likelihood takes longer than the rate at which new observations are introduced. This drawback comes from having to reassess the assignment of every sample-point on each iteration of the Gibbs sampler. Because of this limitation, only a subset of the total number of recorded waveforms was considered. Indeed, an associated risk exists from potentially neglecting waveforms derived from structural damage, making this an important parameter to manage for this type of application. In this case, however, a reasonable balance between the time of computation and the quality of results was achieved by retaining $10,000$ samples.

An important step worth mentioning here is dye-penetrant visual inspections that were carried out periodically throughout the experiment. The outcome of each visual inspection is represented by the colour of the dashed vertical lines in Figure 6.13, where: (1) green is for no evident fracture, (2) yellow is for possible fracture, and (3) red is for a confirmed surface fracture. Carrying out this simple step adds special value to this dataset, since it gives one means to validate the performance of a novelty detector. That is, the monitoring system should be expected to flag an abnormal operation prior to the

first "possible fracture" given by the visual inspection. The blue lines here indicate the instances when the applied load was increased.

In an ideal scenario, where background AE activity is almost non-existent, a sudden upturn in energy should indicate anomalous activity, likely to have been caused by crack nucleation or growth. However, as demonstrated in this case, it is hard to discern any clear patterns in the dataset that can reliably indicate the presence of damage. The small load contribution administered by the sliding tube proves to have a substantial effect on the overall AE activity.

The idea here is to have the DP-PMM cluster the waveforms by their counts. It is assumed that count-rates corresponding to anomalous AE events will somewhat deviate from those corresponding to benign events. By clustering the AE events, closer attention can be paid to the inferred clusters that are characterised by the growth and development of fracture. Other features from these clusters, such as energy, can then be monitored to determine the health state of the lug.

6.4.3 Damage-Detection Strategy

There are different ways in which this method can be used for novelty detection. A simple mechanism could involve having the DP-based model trigger an alarm upon the formation of a new cluster. This approach would require having the model learn from a training set in which it is known that the structure is operating normally. During the training process, several clusters would form to account for the normal operation of the structure subjected to EoVs, and any new clusters forming after the training phase would suggest an anomalous behaviour that has not yet been observed.

One immediate complication of this approach is that the formation of new clusters may not necessarily be a direct consequence of some change occurring to the physical system, and it may instead be attributed to the stochastic nature of the sampler when inferring the number of non-empty components [112]. At each iteration, when the assignment of a sample-point is resampled, a finite probability of forming a new cluster exists, and the number of non-empty components may briefly grow as a result. These sporadic components may attract one or two members, but chances are that they will collapse again after a few iterations of the sampler. In practice, the new formation of clusters may only be an indicator of anomalous behaviour if the clusters "survive" to account for future observations. A possibly sensible alternative could be to monitor the rate at which the number of members from each component grows over time.

Cluster growth-rate can be a more robust indicator of damage, by working on the assumption that the generation of benign AE events will be steady. This assumption may hold since the cyclic load applied by the sliding tube remained unchanged for the duration of the experiment, and no additional inputs were introduced. Unless subjected to unforeseen variations, any clusters formed to characterise the benign AE events should therefore sustain a somewhat constant growth-rate. In contrast, a sudden release of energy caused by fracture will not only result in the additional formation of clusters, but also a more inconsistent growth-rate of these, since AE events of this type will likely occur sporadically with the progression of the original fracture, or from the generation of new fractures in the structure.

6.4.4 Results and Discussion

Following the strategy described above, the DP-based model clustered the recorded AE waveforms as shown in Figure 6.14(a). The dashed vertical lines indicate the moment when a new cluster was created. A different colour was used to represent each cluster and its corresponding members. The dataset was eventually split into $K = 9$ different non-empty components, by setting $\alpha = 0.01$.

One of the first observations that can be made from the results presented in Figure 6.14(a), is the rate at which new clusters appeared. At the beginning of the fatigue test, the model was exposed (for the first time), to a vast number of AE events defined by a variety of different features. It is therefore natural to expect most clusters to be created at the start when the model is learning to identify and group the AE events as they are being introduced. Most of the first-appearing clusters are likely to correspond to the benign generating mechanisms attributed to the friction introduced by the sliding tube. Some events exhibiting significantly higher counts can also be observed occurring somewhat early in the test, triggering the creation of new clusters.

These high-count events could potentially be attributed to sources of high energy, which may indicate the initial stages of crack initiation. Another plausible explanation is that these events might also be a result of internal fractures already existing prior to the start of the fatigue test. A point not yet mentioned is that the lug had previously been fractured from preceding fatigue tests, and then repaired by welding it back in its original place. The sudden release of the internal stresses induced by the weld could have justifiably been a source of AE events exhibiting high counts. It is, unfortunately, impossible to know for sure the sources of all the observed events, as they could have also been generated from a variety of other mechanisms, such as material plastification, crack closure and crack-face rubbing [179], among others.

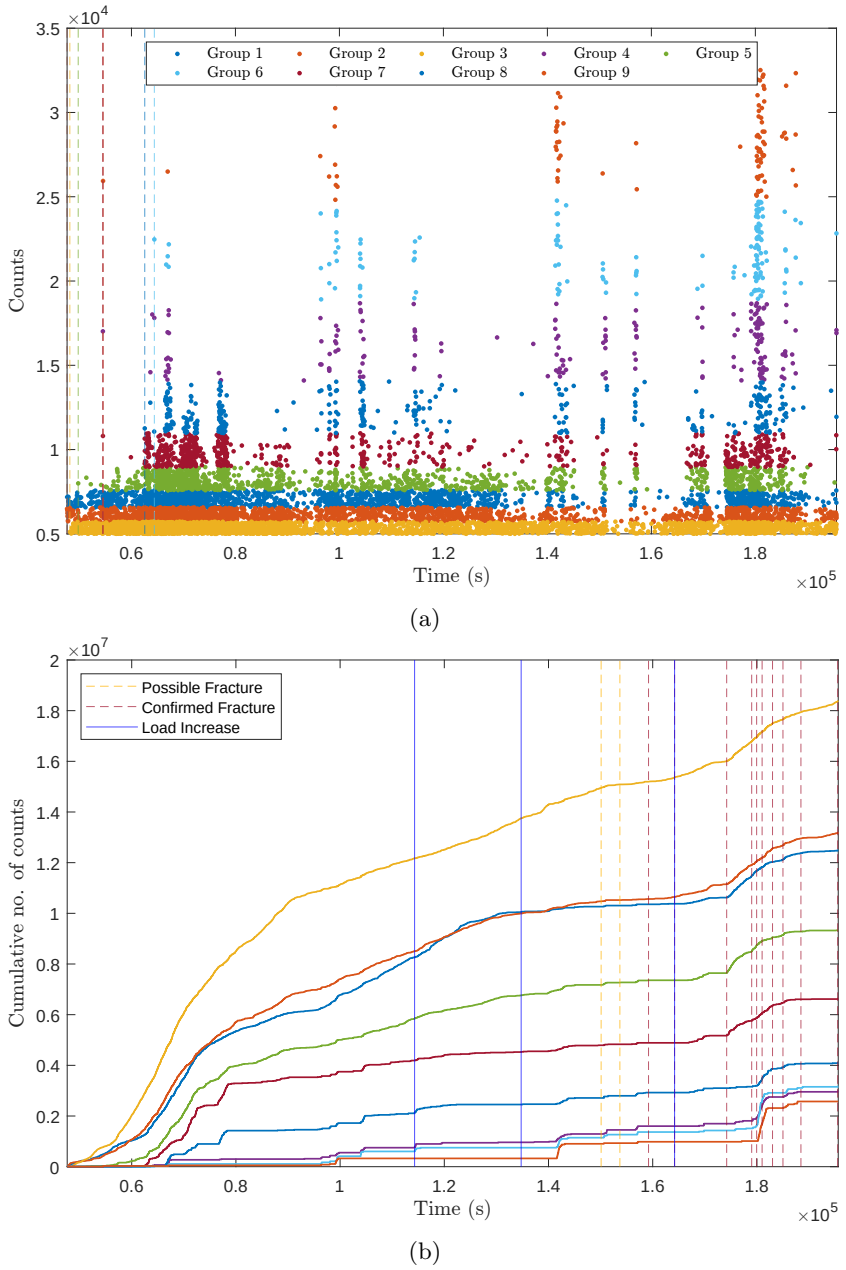


FIGURE 6.14: (a) AE counts clusters inferred by the DP-PMM and (b) Cumulative counts of all inferred groups. Colours assigned to the inferred groups are consistent between (a) and (b).

Nonetheless, some consolation can be found by recalling that AE events deriving from crack nucleation and propagation will tend to exhibit a higher energy content (and counts), than those from any other potential mechanism in this setting. As reviewed at the beginning of this chapter, the correlation of AE features in fatigue tests is a subject that has been thoroughly examined in the literature, and on the basis of this consideration, one can therefore simplify the monitoring scheme by only looking at clusters that are characterised by the highest count-rates.

Figure 6.14(b) shows the cumulative sum of events corresponding to each of the identified clusters. The counts of events were accumulated in this way to help visualise the growth-rate of the clusters throughout the test. Somewhat steady growth can be observed in almost all clusters, with a few exceptions exhibiting noticeable step increases. Since benign AE events are expected to emerge at a continuous rate, irrespective of the health-state of the lug, it may be reasonable to assume that clusters representing background activity are those that present a steady growth. Conversely, high-energy bursts manifesting from the growing crack are expected to be represented by the sudden step increases exhibited by some of these patterns.

The count rate corresponding to the 9^{th} and 3^{rd} clusters are shown in Figure 6.15. Group 9 is in fact characterised by the AE events with the highest energy content, and its selection was based on this parameter. Conversely, Group 3 corresponds to the AE events characterised by the lowest energy content and was included in Figure 6.15 for comparison purposes. The results now clearly show peaks emerging sporadically, with sharp up-turns manifesting throughout the experiment. By removing the background activity, mostly pertaining to Group 3, the data become clearer and easier to interpret. Additionally, the size of the dataset is drastically reduced, making it suitable for the development of a robust monitoring system. Groups exhibiting these sharp up-turns can therefore be extracted and inspected in isolation.

Visually, one can see peaks manifesting moments before the visual inspection indicated a possible presence of fracture, where a first substantial warning is provided in advance at $100,000$s, and the last by the prominent step increase occurring at approximately $140,000$s into the experiment. The observed peaks in this scenario may be interpreted as an indication of fracture extending to a detrimental length, and this form of warning is provided early enough for an operator to intervene and have the landing gear inspected for damage.

While the results are as good as those presented in [45], the advantage here is that the early presence of damage could be detected without needing to resort to source-localisation techniques, which can be cumbersome to implement and require at least three sensors to be placed within the propagating medium of the generated waves. Here, only a single sensor was needed, and although the data collected from this sensor alone were very rich, the processing strategies presented in this chapter appear to be enough to have the model successfully flag an abnormal event in real-time without performing an invasive procedure. If, however, the interest was not only to detect but also to localise the source of damage, then the strategy followed in [45] should be preferred. Unfortunately, there is no way to characterise the sources of AE in this configuration,

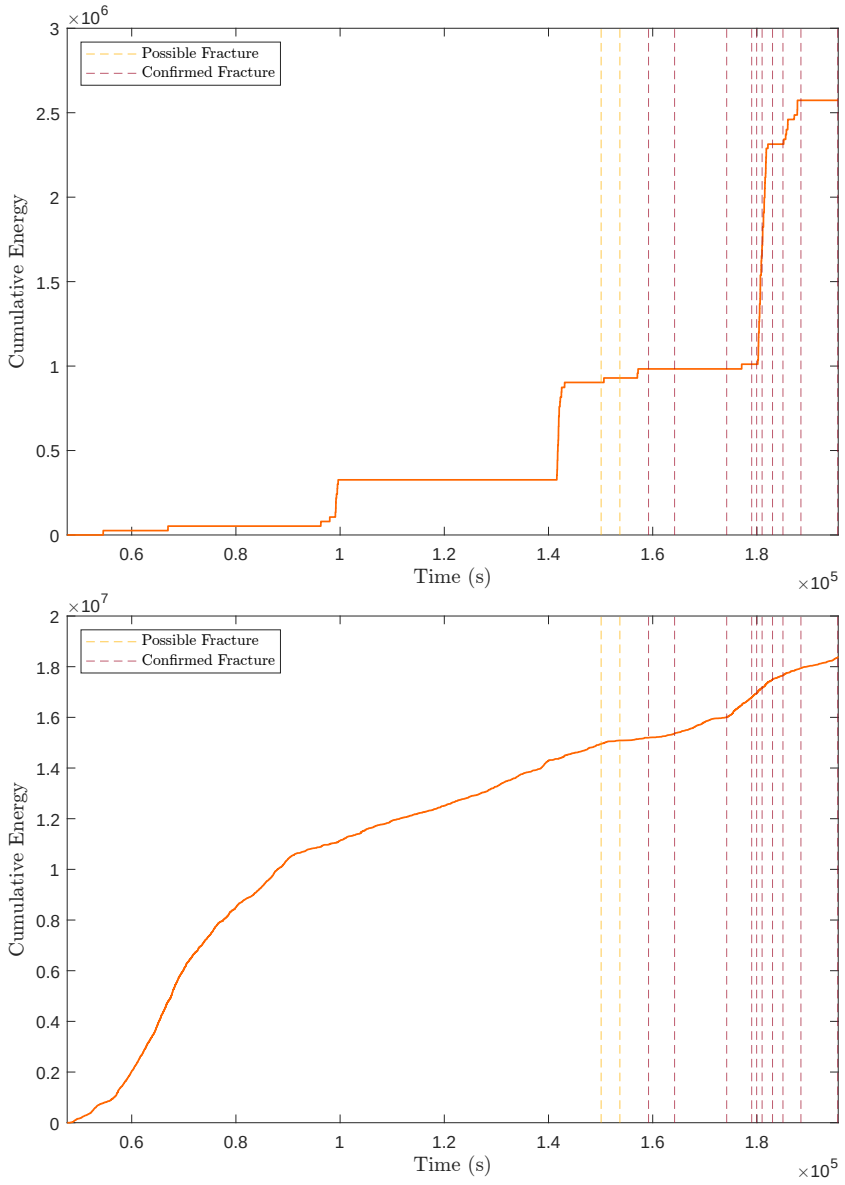


FIGURE 6.15: Rate of counts corresponding to AE events assigned to (a) Group 9 and (b) Group 3.

but the assumptions made here appear to satisfy the identification of AE waves that were more sensitive to damage.

6.5 Conclusions

The approach presented in this chapter demonstrated promising results in the probabilistic detection of AE events in time-series signals. Challenges associated with pre-defining an arbitrary hard threshold were addressed. In the presented method, threshold crossings in a given time window are no longer assumed to indicate, with absolute certainty, the existence of an AE event. Instead, the number of threshold crossings given in a time

window is treated as a random sample, which is then used to infer the probability of an event existing in that time window.

Given the probabilistic framework of the model, when an event is identified, a degree of (un)certainty on its assignment is provided. An operator could therefore decide to retain those AE events categorised with high certainty, after establishing a minimum criterion that is suitable for the application at hand.

Additionally, the proposed methodology gives some insight into scenarios where the nature of the AE activity is unknown, by not only detecting but also clustering all the observable events. Some form of validation would still be required to ensure that the threshold, window length, overlapped portion and hyperparameters are optimal. Nevertheless, it was demonstrated that these parameters can be marginalised during the inference of the infinite mixture model, thereby mitigating the dependency on the selection of these parameters.

Challenges related to this aspect of the model must be emphasised for future developments. One particular issue is regarding the computational resources required for inference of the DP-PMM posterior. To make the presented methods more practical and accessible, a more efficient approach to approximate the posterior would certainly be necessary. Addressing this issue is left as a subject for future research.

Another point worth noting is that the model relies solely on the AE count rates. This feature is assumed to represent the generating sources of AE, which might be an oversimplification of the involved physics. It may be necessary to adapt the model so that other relevant AE features are taken into consideration for clustering.

Finally, the promising capabilities of the DP-PMM in AE-based monitoring methods were explored further with experimental data gathered from a fatigue test of a landing gear. The implementation of the DP-PMM helped simplify the interpretation of an overwhelming dataset, allowing one to identify the groups of AE events that were more sensitive to damage, and therefore detect the early onset of fracture in the structure. Although this method yielded a successful outcome for this case study, exploring how well the DP-PMM generalises when faced with different applications is an exercise worth pursuing in the future.

Chapter 7

Insights Into Joint Input-State-Parameter Estimation for Journal Bearings

The statistical models presented thus far have been primarily driven by data. The premise of a data-driven approach to statistical modelling can be potentially enhanced with the incorporation of physics. This concept was introduced earlier in Chapter 1, and given the promising capabilities physics-informed machine learning has to offer for SHM, it seems fit to include an exploratory study of such an approach in this thesis. Therefore, this chapter is aimed at highlighting some implications one may encounter in pursuing this type of modelling for SHM. The journal bearing is once again considered for this purpose, building upon the case study outlined in Chapter 5.

Journal bearings are intriguing candidates for condition monitoring because of their complex dynamics. To grasp the extent of these complexities, one can envision the delicate balance that must prevail between externally-applied forces and the evolving fluid film in response to these forces. Disrupting this balance can lead to instabilities that pose a significant risk to the integrity of the entire rotor assembly.

This phenomenon gained importance in high-speed turbo-machinery systems, where unexpected vibrations raised concerns because of their potential to jeopardise the health of the turbomachinery. Eventually, it was discovered that this undesired phenomenon was in fact caused by the oil film-action deriving from the supporting journal bearings [157]. Addressing this issue and uncovering the underlying causes thus necessitated a profound comprehension of their dynamics.

This type of vibration in rotating shafts is now known as *self-excited* vibrations [180]. The amplitude and intensity of self-excited vibrations surpass those resulting from critical-speed excitation, rotor imbalances and cyclic stresses in the shaft [181]. This discovery prompted the need for explanations, leading to various theoretical solutions [182–184] and experimental investigations [181]. Discrepancies in their conclusions regarding the sources of self-excited vibrations revealed the complexity of the problem. Initially, the problem was thought to originate from resonances within the oil film [182]. However, the consistent increase in amplitude of self-excited vibrations, even after the shaft speed exceeded twice the critical speed, indicated that it should rather be attributed to the oil-film action [157].

Determining the onset of self-excited vibrations requires the computation of the internal forces exerted by the oil film, which is a challenging task given the strong nonlinearities these forces exhibit in relation to the response of the journal [159]. Nonetheless, exemplary models, such as that proposed by Lund in [185], bypassed this issue by calculating a *critical mass parameter* that determines stability based on the dynamic coefficients of the bearing. The premise of this idea is based on the concept of the fluid film dynamically acting like a spring-damper system [186, 187].

Nowadays, an extensive body of knowledge regarding the dynamics of rotors exists. Recent interests in this field have been channelled towards system identification problems involving rotating machinery, in which the response of journal bearings and their foundations are included in the analysis [188]. Part of the reason for the surge in this area of research is, perhaps, attributed to the prevalent role the dynamic coefficients have in calculating unbalanced responses, damped natural frequencies and stabilities [185].

While modern approaches to modelling rotor dynamics involve sophisticated *Finite Element* (FE) methods [133, 189], these models can overlook subtleties that often arise in practice. These discrepancies are quantified by differences in the estimates made by the model to data gathered from actual systems. A statistical model could, in principle, provide a more accurate representation of the rotor directly from the data, but at the risk of outputting poor predictions for newly unseen operational conditions.

In the light of these issues, a sensible solution may thus involve combining the extrapolating strengths of physics-based models with the learning flexibility of a data-based model. A variety of models capable of achieving this integration have been developed in the field of machine learning. An extensive survey on *physics-informed machine learning* (PIML) algorithms can be found in [190]. The outline of the current chapter, however, does not include an exhaustive comparison between these models for the problem at hand. Instead, the focus here is on exploring the use of the *Gaussian Process Latent Force Model* (GP-LFM) [191] for modelling journal-bearing dynamics.

The GP-LFM employs both the dynamics and observations to infer the states of a system (e.g. displacements and velocities) along with their driving forces. In this framework, forces are treated as latent states during inference, and to mitigate assumptions about the functional form of the force signals, these are modelled according to a set of independent GPs, each defined by their own covariance function. The fact that the response of the system and input force can be recovered simultaneously makes GP-LFMs an incredibly attractive method in various areas of engineering. Several recent studies have demonstrated the effectiveness of GP-LFMs for input-state estimation in wind turbines [192, 193], suspension bridges [194], and structural systems in general [195, 196].

Certainly, a notable advantage for applications in rotating machinery is that the GP-LFM framework enables the recovery of the dynamic coefficients of the bearing. However, achieving this involves dealing with additional complexities from the joint estimation of inputs, states, and parameters. This chapter aims to explore these implications via a simple numerical case study in which the bearing coefficients are unknown. However, before delving into this exercise, a background on GP-LFMs is covered, followed by the outline and analysis of a series of numerical case studies that consider various loading conditions that a journal bearing might experience in practice.

The background provided in the upcoming sections is largely based on the works of [197] and [198]. For a more comprehensive understanding of the subject, the interested reader is highly encouraged to refer to these references.

7.1 Continuous-Discrete State-Space Models

Fundamentally, the GP-LFM aims to solve *ordinary differential equations* (ODEs) driven by GPs. The stochastic nature of a GP input means that a given realisation yields a solution that is also random. Therefore, the interest is not on finding a particular solution to the ODE, but on finding the statistics of all possible realisations of the solution.

To demonstrate how one can achieve this while staying in the context of bearing dynamics, the corresponding equation of motion of an n_d *Degrees-Of-Freedom* (DOFs) system subjected to a forced excitation is first considered. This equation can be represented by the following second-order ODE,

$$\mathbf{M}\ddot{\mathbf{x}}(t) + \mathbf{C}\dot{\mathbf{x}}(t) + \mathbf{K}\mathbf{x}(t) = \mathbf{f}(t) \tag{7.1}$$

where, $\mathbf{x} \in \mathbb{R}^{n_d}$ is the vector of solutions of the ODE, $\mathbf{f} \in \mathbb{R}^{n_d}$ is the vector of external forces acting on the system, and the coefficients $\mathbf{M}$, $\mathbf{C}$ and $\mathbf{K}$ represent the mass, damping and stiffness matrices of the system, respectively. The overhead dots indicate

the order of differentiation with respect to time. For convenience, it is necessary to represent Equation (7.1) in its *state-space* form. In particular,

$$\dot{\mathbf{z}}(t) = \mathrm{F}\mathbf{z}(t) + \mathrm{L}\mathbf{f}(t) \tag{7.2}$$

where

$$\mathbf{z}(t) = \begin{pmatrix} \mathbf{x}(t) \\ \dot{\mathbf{x}}(t) \end{pmatrix}, \quad \mathrm{F} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{C} \end{pmatrix}, \quad \mathrm{L} = \begin{pmatrix} \mathbf{0} \\ \mathbf{M}^{-1} \end{pmatrix} \tag{7.3}$$

In this form, the second-order differential equations in (7.1) are now converted into a set of first-order differential equations, which are defined via the *state vector* $\mathbf{z} \in \mathbb{R}^{n_s}$, with $n_s = 2n_d$ elements.

In general, time-varying phenomena can be modelled by differential equations driven by white noise. In this example, $f(t)$ in Equation (7.2) can be replaced by a forcing function $w(t)$ modelled by white noise with spectral density Q. That is,

$$\dot{\mathbf{z}}(t) = \mathrm{F}\mathbf{z}(t) + \mathrm{L}w(t), \quad w(t) \sim \mathcal{N}(0, \delta(t-s)Q) \tag{7.4}$$

ODEs driven by white noise are referred to as *Stochastic Differential Equations* (SDEs), and if linear conditions apply, their solution can be interpreted as a linear transformation of a GP. It should be noted that the notation employed here is a heuristic representation of an SDE. In the traditional sense, a differential equation does not permit the discontinuities introduced by $w(t)$. A rigorous mathematical solution is possible by reducing the problem to the definitions of *Itô calculus* [199]. Given that the solution is now Gaussian,

$$\mathbf{z}(t) \sim \mathcal{GP}(\mathbf{z}(t)|\mathbf{m}(t), \mathbf{P}(t)) \tag{7.5}$$

where $\mathbf{m}(t)$ and $\mathbf{P}(t)$ denote the statistics of the solution; that is, the mean and covariance, respectively. The transition density can be recovered from this expression by conditioning the current state with itself at a preceding time instant $s < t$. By establishing the initial conditions $\mathbf{m}(s) = \mathbf{z}(s)$ and $\mathbf{P}(s) = 0$,

$$p(\mathbf{z}(t)|\mathbf{z}(s)) = \mathcal{N}(\mathbf{z}(t)|\mathbf{m}(t|s), \mathbf{P}(t|s)) \tag{7.6}$$

where,

$$\begin{aligned} \mathbf{m}(t|s) &= \psi(t,s)\,\mathbf{z}(s) \\ \mathbf{P}(t|s) &= \int_s^t \psi(t,\tau)\mathrm{L}(\tau)Q\mathrm{L}^\top(\tau)\psi^\top(t,\tau)\mathrm{d}\tau \end{aligned} \tag{7.7}$$

The definition in Equation 7.6 is only normally distributed if the set of observations is finite. While Equation (7.2) forms the continuous-time state-space model of the system

defined by Equation (7.1), in practice, outputs are obtained by sampling measurements from the response of the system at discrete time instants t_k. It is thus sensible to assume that a pair of processes exist such that one is observed while the other is hidden. An additional model can thus be constructed in which the measurements are expressed in terms of the states with some added noise. Concretely,

$$\mathrm{y}_k = \mathrm{H}_k \mathbf{z}(t_k) + \mathbf{r}_k, \quad \mathbf{r}_k \sim \mathcal{N}(0, \mathrm{R}_k) \tag{7.8}$$

where y_k corresponds to the k^{th} measurement gathered at the time instant t_k, H_k models the sensor, and the noise term $\mathbf{r}_k$ represents the uncertainties in the measurements. Overall, the *continuous-discrete state-space model* is defined as,

$$\begin{aligned} \dot{\mathbf{z}}(t) &= \mathrm{F}\mathbf{z}(t) + \mathrm{L}w(t) \\ \mathrm{y}_k &= \mathrm{H}_k \mathbf{z}(t_k) + \mathbf{r}_k \end{aligned} \tag{7.9}$$

The hidden process corresponds to the dynamics of the system, while the observed process corresponds to the measurements deriving from the system. This set of processes is connected in the sense that the distributions of y_k and $\mathbf{z}(t)$ coincide at the discrete-time samples t_k. It is thus possible to express Equation (7.4) in an equivalent discretisation form. The formalism of the transition densities (7.7) implies that the SDE is equivalent to the following discrete-time system,

$$\mathbf{z}(t_{k+1}) = \mathrm{A}_k \mathbf{z}(t_k) + q_k, \quad q_k \sim \mathcal{N}(0, \Sigma_k) \tag{7.10}$$

where,

$$\begin{aligned} \mathrm{A}_k &= \psi(t_{k+1}, t_k) \\ \Sigma_k &= \int_{t_k}^{t_{k+1}} \psi(t_{k+1}, t_k) \mathrm{L}(\tau) Q \mathrm{L}^\top(\tau) \psi^\top(t_{k+1}, \tau) \mathrm{d}\tau \end{aligned} \tag{7.11}$$

In the case of a *linear-time invariant* (LTI) SDE, the *transition matrix* A, is the matrix exponential function,

$$\mathrm{A}_k = \psi(t_{k+1}, t_k) = \exp(\mathrm{F}\Delta t) \tag{7.12}$$

where $\Delta t = t_{k+1} - t_k$. Similarly, the covariance of the discretised solution is given by [197],

$$\Sigma_k = \int_0^{\Delta t_k} \exp(\mathrm{F}(\Delta t - \tau)) \mathrm{L}(\tau) Q \mathrm{L}^\top(\tau) \exp(\mathrm{F}(\Delta t - \tau))^\top \mathrm{d}\tau \tag{7.13}$$

This expression for the covariance may be computed in closed form. When dealing with linear SDEs, however, this solution is conveniently simplified. As shown in [200, 201], Equation (7.13) can be reduced to the following expression,

$$\Sigma_k = \mathrm{P}_\infty - \mathrm{A}_k \mathrm{P}_\infty \mathrm{A}_k^\top \tag{7.14}$$

where P_∞ is the steady-state covariance that corresponds to the solution to the Lyapunov equation of the form,

$$\mathrm{F}\mathrm{P}_\infty + \mathrm{P}_\infty\mathrm{F}^\top + \mathrm{L}Q\mathrm{L}^\top = 0 \tag{7.15}$$

The steady-state covariance is a result of the assumption of a process that has been observed over an infinite duration of time. Therefore, provided that P_∞ and A_k can be computed, Σ_k is easily solved by equation (7.14).

By analysing the processes in (7.9) simultaneously, the states can be inferred from noisy measurements. The problem at hand is thus a statistical one, as it involves finding the conditional probability of $\mathbf{z}$ given the observed process $\mathbf{y}$. While $p(\mathbf{z}|\mathbf{y})$ is generally intractable, the reformulation of the state as a Gauss-Markov process means that filtering and smoothing techniques can be conveniently employed to find this solution. Before extending this idea, it is necessary to first look into the continuous-discrete state-space model when having the input forces modelled by a set of independent GPs other than white noises. More specifically, it is of interest in this context to assume some correlation in the structure of the driving noise.

7.2 Solution to LTI SDEs with GP Inputs

In the previous section, the equations of motion were assumed to be subjected to white-noise excitations. The objective of this section is to illustrate the generalisation of the SDE model to cases where the forces are modelled by GPs defined with stationary covariance functions. In particular, the state-space form of (7.1) can be expressed as,

$$\begin{aligned} \dot{\mathbf{z}}(t) &= \mathrm{F}\mathbf{z}(t) + \mathrm{L}\mathbf{g}(t) \\ g^{(j)}(t) &\sim \mathcal{GP}(0, \kappa_g^{(j)}(t,t')), \quad j = 1, \ldots, n_g \\ \mathrm{y}_k &= \mathrm{H}_k\mathbf{z}(t_k) + \mathbf{r}_k \end{aligned} \tag{7.16}$$

where the latent force vector $\mathbf{g}(t)$ has elements $g_j(t)$ modelled by GPs. The solution derived by Alvarez et al. [191] shows that $\mathbf{z}(\mathbf{t})$ is also a zero-mean GP with covariance matrix function,

$$\mathbf{K}_{zz}(t,t') = \int_0^t \int_0^{t'} \exp(\mathrm{F}(t-\tau))\mathrm{L}(\tau)\mathbf{K}_{gg}(\tau,\tau')\mathrm{L}^\top(\tau)\exp(\mathrm{F}(t'-\tau))^\top \mathrm{d}\tau\mathrm{d}\tau' \tag{7.17}$$

where zero initial conditions are assumed, and $\mathbf{K}_{gg}(\tau,\tau^\top)$ is the joint covariance matrix accounting for all latent forces,

$$\mathbf{K}_{gg}(t,t') = \mathbb{E}\left[\mathbf{g}(t)\mathbf{g}(t')^\top\right] = \mathrm{diag}\left[\kappa_1(t,t'), \ldots, \kappa_{n_g}(t,t')\right] \tag{7.18}$$

The states of $\mathbf{z}(t)$ thus form a multi-dimensional GP,

$$\mathbf{z}(t) \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_{zz}(t, t')) \tag{7.19}$$

While a closed-form solution may be obtained, the limitation of this approach is having to compute the covariance matrix function $\mathbf{K}_{zz}$, analytically, which may not often be possible. In such cases, it may be necessary to employ expensive numerical computations to solve the integration in equation (7.17).

A practical alternative to this approach is to instead embed a state-space representation of the GP into the ODE. This solution was proposed by Hartikainen and Sarkka [202], showing that the GP-driven ODE can be represented by an augmented form of the ODE driven by white noise; therefore, recovering a linear SDE framework such as that in Equation (7.4). The implication of this approach eventually leads to a representation in which the states, now including the latent forces, can be inferred with a *Kalman* filter [203] and *Rauch-Tung-Streibel* (RTS) smoother [204].

The state-space representation of a GP is summarised, according to [202], in the remainder of this section, and the interested reader is referred to their work for more details. To begin with the illustration of this process, the realisation of an arbitrary GP, $g(t)$, is allowed to be the solution to an m^{th}-order scalar LTI SDE driven by white noise. In particular,

$$\frac{d^m}{dt^m}g(t) + a_{m-1}\frac{d^{m-1}}{dt^{m-1}}g(t) + \cdots + a_1\frac{d}{dt}g(t) + a_o g(t) = w(t) \tag{7.20}$$

where $a_o, \ldots, a_{m-1}$ are known constants, and $w(t)$ is a white noise process with spectral density $S(w) = q$. In state-space form,

$$\frac{d}{dt}\mathbf{z}(t) = \mathrm{F}_g\mathbf{z}(t) + \mathrm{L}_g w(t) \tag{7.21}$$

where the state $\mathbf{z}(t)$ contains the derivatives of $g(t)$ up to order $m-1$; that is, $\mathbf{z}(t) = \left[g(t), dg(t)/dt, \ldots, d^{m-1}g(t)/dt^{m-1}\right]^\top$. The matrices $\mathrm{F}_g \in \mathbb{R}^{m\times m}$ and $\mathrm{L}_g \in \mathbb{R}^m$ are given as,

$$F_g = \begin{pmatrix} 0 & 1 & \ldots & & \\ \vdots & & \ddots & \ddots & \\ & & & 0 & 1 \\ -a_o & & \ldots & -a_{m-2} & -a_{m-1} \end{pmatrix}, \quad L_g = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \tag{7.22}$$

The solution $g(t)$ can be extracted from $z(t)$ by defining $\mathrm{H} = [1, 0, \ldots, 0]^\top$ and having $g(t) = \mathrm{H}\mathbf{z}(t)$. Now, the spectral density of $g(t)$ is calculated by replacing $z(t)$ with $\mathrm{H}\mathbf{z}(t)$ in (7.21), and taking the Fourier transform on both sides of the equation. This

calculation eventually leads to the following expression,

$$S_g(w) = \mathrm{H}(\mathrm{F} + i\omega\mathbb{I})^{-1}\mathrm{L}q\mathrm{L}^\top \left[(\mathrm{F} - i\omega\mathbb{I})^{-1}\right]^\top \mathrm{H}^\top \tag{7.23}$$

The covariance function is finally recovered by taking the inverse Fourier transform of the spectral density,

$$\kappa_g(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} S_g(\omega)e^{i\omega\tau}\mathrm{d}\omega \tag{7.24}$$

In practice, however, it is generally the case that one assumes the functional form of the input, and hence, a covariance function that can accommodate such assumptions is chosen. Since the desire is to represent $g(t)$ in a state-space form, the problem is on finding F_g, L_g and q such that the first component in $\mathbf{z}(t)$ is defined by the covariance function $\kappa_g(\tau)$. In other words, the problem is now inverted, whereby the derivation of (7.21) is approached from knowing $\kappa_g(\tau)$ *a-priori*.

The inverted approach to this problem involves determining the coefficients $a_o, \ldots, a_{m-1}$, in F_g, from the spectral density of the covariance function. In the special case where the covariance function is stationary, and has a spectral density represented by a rational function of the form,

$$S(\omega) = \frac{q}{\text{polynomial in } \omega^2} \tag{7.25}$$

the coefficients of the polynomial in the denominator happen to correspond to the coefficients in (7.20), and the numerator corresponds to the spectral density of $w(t)$. Although this definition may be somewhat abstract, a more concrete illustration can be envisioned with the Matérn family of covariance functions. It should be noted that the following demonstration applies to a number of different covariance functions, but the Matérn family is used here because (1) they can yield a closed-form solution and (2) are used for the case studies outlined in this chapter.

Recall the definition of the Matérn covariance function from Chapter 4,

$$k(r) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu r}}{l}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}r}{l}\right), \qquad r = ||t - t'|| \tag{7.26}$$

where ν and l are positive parameters, and K_ν is a modified Bessel function. For a one-dimensional process, the spectral density of (7.26) is,

$$S(\omega) = \frac{2\sigma^2\pi^{1/2}\Gamma(\nu + 1/2)}{\Gamma(\nu)}\lambda^{2\nu}(\lambda^2 + \omega^2)^{-(\nu+1/2)} \tag{7.27}$$

where σ^2 is a scaling parameter and $\lambda = \sqrt{2\nu}/l$. In the special case where $\nu = p + 0.5$ and p is a non-negative integer,

$$\begin{aligned} S(\omega) &\propto (\lambda^2 + \omega^2)^{-(p+1)} \\ &\propto (\lambda + i\omega)^{-(p+1)}(\lambda - i\omega)^{-(p+1)} \end{aligned} \tag{7.28}$$

which has the desired rational functional form in (7.25), with corresponding spectral density of the white noise process given by,

$$q_c = \frac{2\sigma^2 \pi^{1/2} \lambda^{2p+1} \Gamma(p+1)}{\Gamma(p+1/2)} \tag{7.29}$$

Finally, the state-space representation of (7.26) can be recovered by choosing non-negative integers for p. For example, having $p = 1$ reduces to the state-space representation of the GP defined by a Matérn 3/2 covariance function. Concretely,

$$\frac{d}{dt}\mathbf{z}(t) = \begin{pmatrix} 0 & 1 \\ -\lambda^2 & -2\lambda \end{pmatrix} \mathbf{z}(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} w(t), \quad w(t) \sim \mathcal{N}(0, q_c) \tag{7.30}$$

where $\lambda = \sqrt{3}/l$ and $q_c = 12\sqrt{3}\sigma^2/l^3$.

Having the inputs modelled as GPs, and in state-space form, means that the force components can now be seamlessly incorporated in the hidden process for inference. The trick is to augment the transient model to include the force components as additional latent states. This can be achieved by extending $\mathbf{z}(t)$ to account for the latent forces and their corresponding derivatives. The augmented state-space form of Equation (7.1) can thus be expressed as,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{z}_{ag}(t) &= \mathrm{F}_{ag}\mathbf{z}_{ag}(t) + \mathrm{L}w(t) \\ \mathrm{y}_k &= \mathrm{H}\mathbf{z}_{ag}(t_k) + r_k \end{aligned} \tag{7.31}$$

where $w(t) \sim \mathcal{N}(0, Q_c)$ and $r_k \sim \mathcal{N}(0, R_k)$. The diagonal elements of the matrices Q_c and R_k, correspond to the spectral densities evaluated for each force component, and the noise of the measured states, respectively. The augmented vector of states now accounts for the latent forces, i.e. $\mathbf{z}_{ag}(t) = \left[\mathbf{x}, \dot{\mathbf{x}}, \mathbf{f}, \dot{\mathbf{f}}\right]^\top$, and the corresponding augmented drift matrix F_{ag} is now expressed as,

$$\mathrm{F}_{ag} = \begin{pmatrix} \mathrm{F} & \mathrm{G} \\ \mathbf{0} & \mathrm{F}_g \end{pmatrix}, \quad \mathrm{G} = \begin{pmatrix} \begin{bmatrix} 0 & 0 \\ m_1^{-1} & 0 \end{bmatrix} & \cdots & 0 \\ \vdots & \ddots & \\ 0 & \cdots & \begin{bmatrix} 0 & 0 \\ m_{n_f}^{-1} & 0 \end{bmatrix} \end{pmatrix} \tag{7.32}$$

The solution to Equation (7.31) can now be obtained by forming an equivalent discrete system, followed by the implementation of a Kalman filter to infer the states in $\mathbf{z}_{ag}(t)$. The inclusion of GPs as inputs of the SDE implies finding the hyperparameters of the Matérn function such that the functional form of the inferred forces is accurate and meaningful. This implication requires the addition of a method that operates in parallel with the Kalman filter so that the set of optimal parameters can be estimated. In the setting considered here, the parametric form of the SDE is governed by the set of parameters $\Theta = \left[\sigma^2, l, r\right]$.

7.3 Parameter Estimation

The dependencies of the model on σ^2, l, and r, unfortunately, complicate the analysis a little, as these parameters are typically unknown. Nevertheless, their values can be somewhat determined in the light of measurements. In the scenario where a finite number of observations are available, parameter estimation for SDEs can be achieved with *Maximum Likelihood* (ML) methods.

As discussed earlier, the solutions to LTI SDEs are Markov Processes since they are characterised by Gaussian transition densities $p(x(t)|x(s))$. Given the Markov properties of SDEs, the likelihood of the observations, given the parameters, can be defined as,

$$p(x(t_1), \ldots, x(t_T)|\Theta) = \prod_{k=0}^{T-1} p(x(t_{k+1})|x(t_k), \Theta) \tag{7.33}$$

and finding an optimal set of parameters $\hat{\Theta}$ can be achieved by,

$$\hat{\Theta} = \arg\min_{\Theta} - \log p(x(t_1), \ldots, x(t_T)|\Theta) \tag{7.34}$$

or, if the model is Bayesian, inference can be made directly on the posterior distribution. That is,

$$p(\Theta|y(t_1), \ldots, y(t_T)) \propto p(\Theta) \prod_{k=0}^{T-1} p(y(t_{k+1})|y(t_k), \Theta) \tag{7.35}$$

where $p(\Theta)$ is the prior distribution over Θ. The likelihood in either case is determined recursively as the transition densities are made available during the filtering and smoothing process. The recursion for the marginal posterior of a linear Gaussian state space is given as,

$$\varphi_k(\Theta) = \varphi_{k-1}(\Theta) + \frac{1}{2}\log|2\pi S_k(\Theta)| + \frac{1}{2}\mathrm{v}_k^\top(\Theta) S_k^{-1}(\Theta)\mathrm{v}_k(\Theta) \tag{7.36}$$

where the terms $\mathrm{v}_k = \mathbf{y}_k - \mathrm{H}_k\left[A_k m_{k-1}\right]$ and $S_k = \mathrm{H}\left[A_k \mathrm{P}_{k-1} A_k^\top + \Sigma_k\right]\mathrm{H}^\top + R_k$, are calculated during the updating step of the Kalman filter with the parameters fixed to

Θ. Equation (7.36) is also referred to as the *energy function* [198], and by inspection, one may notice that its functional form corresponds to the negative log-likelihood of a Gaussian evaluated up to the kth observation. The recursion is initialised with $\varphi_o(\Theta) = -\log p(\Theta)$. The evaluation of $\varphi_T(\Theta)$ at time step $k = T$, returns the (unnormalised) likelihood posterior density at point Θ,

$$p(\Theta|y_{1:T}) \propto \exp(-\varphi_T(\Theta)) \tag{7.37}$$

The recursion can then be repeated for a new selection of proposals until equation (7.33) is minimised, or $p(\Theta|y_{1:T})$ is approximated. A selection of different approximation strategies is apt for this problem, such as *Maximum-A-Posteriori* (MAP) estimation or *Laplace approximation*. While these techniques provide an optimal combination of these parameters, in many cases, it may be necessary to know the PDF of the posterior, and off-the-shelf MCMC methods may be preferred to approximate $p(\Theta|y(t_1), \ldots, y(t_T))$.

7.4 Simulating Journal-Bearing Dynamics

Let the response of a journal bearing, operating under stable conditions, be defined by the coordinate system shown in Figure 7.1. The coordinates of the shaft centre, O_j, are given by the *eccentricity* e and *attitude angle* ϕ, relative to the bore centre O_b. The reaction of the journal to some external loading is provided by the sum of the load components F_e and F_ϕ. One may note that F_e acts along the line of centres, while F_ϕ acts perpendicular to the line of centres. The static load of the journal is denoted by W, and the direction and magnitude of the reaction load vector $Fr = \left(F_\phi^2 + F_e^2\right)^{1/2}$, adjust to keep the journal balanced upon changes imparted by external loads. The eccentricity and attitude angle are enough to completely characterise the motion of the journal, and from Figure 7.1, the equations of journal motion are defined as,

$$\begin{aligned} m\ddot{\mathrm{x}}_1 &= W - F_e\cos\phi - F_\phi\sin\phi - W_r\cos\phi_r \\ m\ddot{\mathrm{x}}_2 &= F_\phi\cos\phi - F_e\sin\phi + W_r\sin\phi_r \end{aligned} \tag{7.38}$$

where m denotes the mass contribution supported by the bearing, $\mathrm{x}_1 = e\cos\phi$ and $\mathrm{x}_2 = e\sin\phi$.

However, even when W_r and ϕ_r are known, some difficulties arise in attempting to solve equation (7.38) because of the strong nonlinear dependencies in the response of the journal. This issue means that the current form of the equations of motion is incompatible with the linear framework established for the GP-LFM. It is thus necessary to linearise the reaction forces about a quasi-steady-state journal location so that the

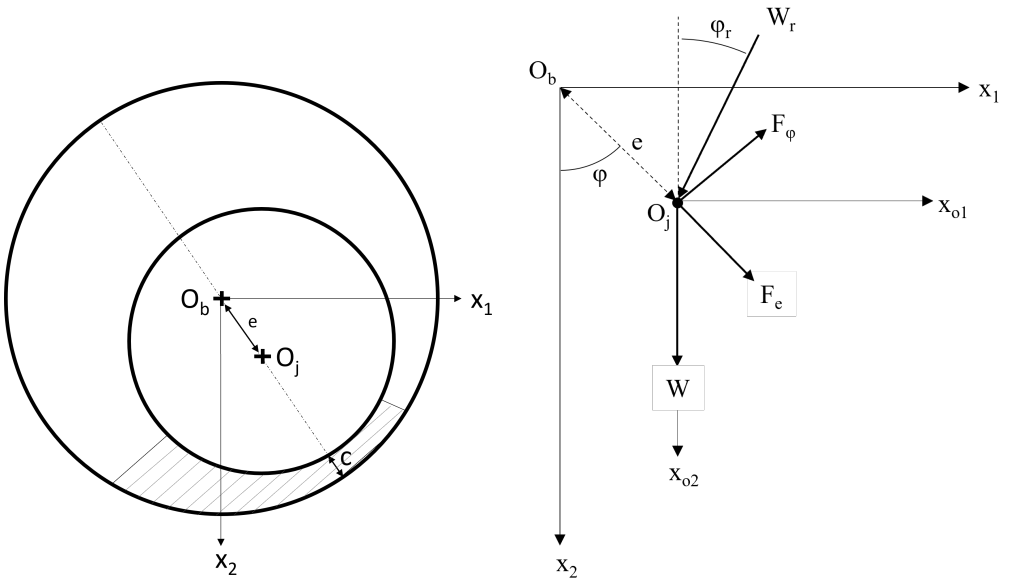


FIGURE 7.1: Diagram showing forces components acting on the journal bearing. Details of the notation are provided in the text.

system becomes linear under small displacements. The linearised equations of motion reduce to [159],

$$\begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix} \begin{bmatrix} \Delta\ddot{\mathrm{x}}_{o1} \\ \Delta\ddot{\mathrm{x}}_{o2} \end{bmatrix} + \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix} \begin{bmatrix} \Delta\dot{\mathrm{x}}_{o1} \\ \Delta\dot{\mathrm{x}}_{o2} \end{bmatrix} + \begin{pmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{pmatrix} \begin{bmatrix} \Delta\mathrm{x}_{o1} \\ \Delta\mathrm{x}_{o2} \end{bmatrix} = \begin{bmatrix} \Delta f_{o1} \\ \Delta f_{o2} \end{bmatrix} \tag{7.39}$$

where $c_{11}, c_{12}, c_{21}, c_{22}$ denote the linearised damping coefficients and $k_{11}, k_{12}, k_{21}, k_{22}$ denote the linearised stiffness coefficients of the bearing. All eight dynamic coefficients are characterised by the static-equilibrium eccentricity and attitude angle. The equilibrium position of the journal depends on a series of operational parameters, and under the steady-state assumption, these parameters remain constant. Equation (7.39) is generally a good approximation for many practical purposes [159, 185]. Additionally, linearisation of the bearing reaction has the added advantage of decoupling the rotor dynamics from the bearing, which can impose some difficulties in the overall analysis.

In the scenario where an exact analysis is required, the effects of the rotor reaction must be taken into consideration, whereby the rotor equations are integrated with the equations of journal motion. Making the reaction forces linear, however, allows the response to be solved without needing to account for the rotor. While the importance of a nonlinear analysis is worth investigating, it is a task left for future research.

For small displacements, the response of the journal is thus recovered by solving equation (7.39). A fifth-order Runge-Kutta time-step integration method is employed to

calculate the states of the journal subjected to dynamic loading. The bearing parameters are based on the test-rig designed in [131] and are thus the same key values used in Chapter 5 (Table 5.2). The rotational speed, however, in all of the following simulations is fixed at 115rpm.

As mentioned above, measurements are modelled as the states corrupted by noise at discrete time instances. Therefore, measurements were simulated as a subset of evenly-spaced discrete samples of the states with the addition of artificial noise modelled by independent and identically distributed random samples from a Gaussian distribution with zero mean and standard deviation equivalent to 10% of the standard deviation of the noise-free signals.

7.5 Input-State Estimation with Known Parameters

A series of three simulated case studies are considered in this section, and in all cases, it is assumed that the dynamic coefficients are known in advance. The values of the dynamic coefficients can be calculated analytically from the linearisation of the journal response under a *short-width* bearing assumption [159]. Derivations of these equations are quite involved and are not covered in the present work. However, these can be found in [157, 159], or alternatively, one may refer to their solutions from the extensive collection of tables produced in the *journal-bearing databook* [205].

Table 7.1 includes the values of the computed dynamic coefficients in relation to the operating parameters outlined in Table 5.2. These coefficients are adopted in all numerical simulations.

TABLE 7.1: Journal bearing dynamic coefficients used for the simulation.

Coefficient	Value	Units
m	200	kg
c_{11}	$10,253$	kNsm^{-1}
c_{12}	7468.4	kNsm^{-1}
c_{21}	7468.4	kNsm^{-1}
c_{22}	$21,990$	kNsm^{-1}
k_{11}	$88,555$	kNm^{-3}
k_{12}	$-35,155$	kNm^{-3}
k_{21}	$159,000$	kNm^{-3}
k_{22}	$115,810$	kNm^{-3}

Since modelling is employed in a Bayesian framework, the prior distributions on the hyperparameters are defined as shown in Table 7.2. These priors are employed in all three

case studies, and the MAP estimates of the posterior are found with the *Quantum-Behaved Particle-Swarm Optimiser* (QPSO) [121], such that equation (7.36) is minimised. Because the hyperparameters are required to be positive, the search space of the energy function is transformed by taking the logarithm of the hyperparameters, i.e. $\hat{\Theta} = \log(\Theta)$.

TABLE 7.2: Priors for GP-LFM hyperparameters.

Hyperparameter	Prior
$\hat{l}$	$\hat{l} \sim \mathcal{N}(-2, 1)$
$\hat{\sigma}^2$	$\hat{\sigma}^2 \sim \mathcal{N}(8, 1)$
$\hat{r}$	$\hat{r} \sim \mathcal{N}(-30, 1)$

7.5.1 Numerical Case Study 1: Harmonic Excitation

The predictive capabilities of the GP-LFM are evaluated in this section by first considering the scenario in which the journal is subjected to harmonic excitation. In real applications, a harmonic response may be driven by out-of-balance forces deriving from the rotor. That is, assuming a perfectly rigid rotor, the components of the external forcing can be defined by,

$$\begin{aligned} \Delta f_{o1} &= m e_\mu \omega^2 \sin(\omega t - \phi) \\ \Delta f_{o2} &= m e_\mu \omega^2 \cos(\omega t - \phi) \end{aligned} \tag{7.40}$$

where $me_\mu\omega^2$ denotes the magnitude of the out-of-balance forces, with e_μ being the unbalance eccentricity and $\omega = 2\pi N/60$. Solving Equation (7.39) with the external forces defined by (7.40), results in the response shown in Figure 7.2.

Detailed views of displacement responses are included in the figures, showing the simulated measurements used for inferring the states. Figure 7.2(c) shows a synchronous whirl forced upon the journal as it reacts to out-of-balance forces equal to one-quarter of the static load (i.e. $me_\mu\omega^2 = 250N$). Although this simulation is somewhat of an idealised representation, this orbiting effect is characteristic of journal bearings when their operation is stable.

Because of the smooth aspect of the forcing signal employed in this exercise, the covariance function chosen for the input-state estimations is the Matérn 5/2 ($p = 2$). Alternatively, for this particular case, the periodic covariance function may be arguably a choice better suited to this problem, as a sinusoidal behaviour can be (potentially) expected from a rotor that is out of balance. However, the periodic function would be limited to purely periodic force histories, which might not necessarily be true in real applications. For example, additional external forces may come into play, contributing

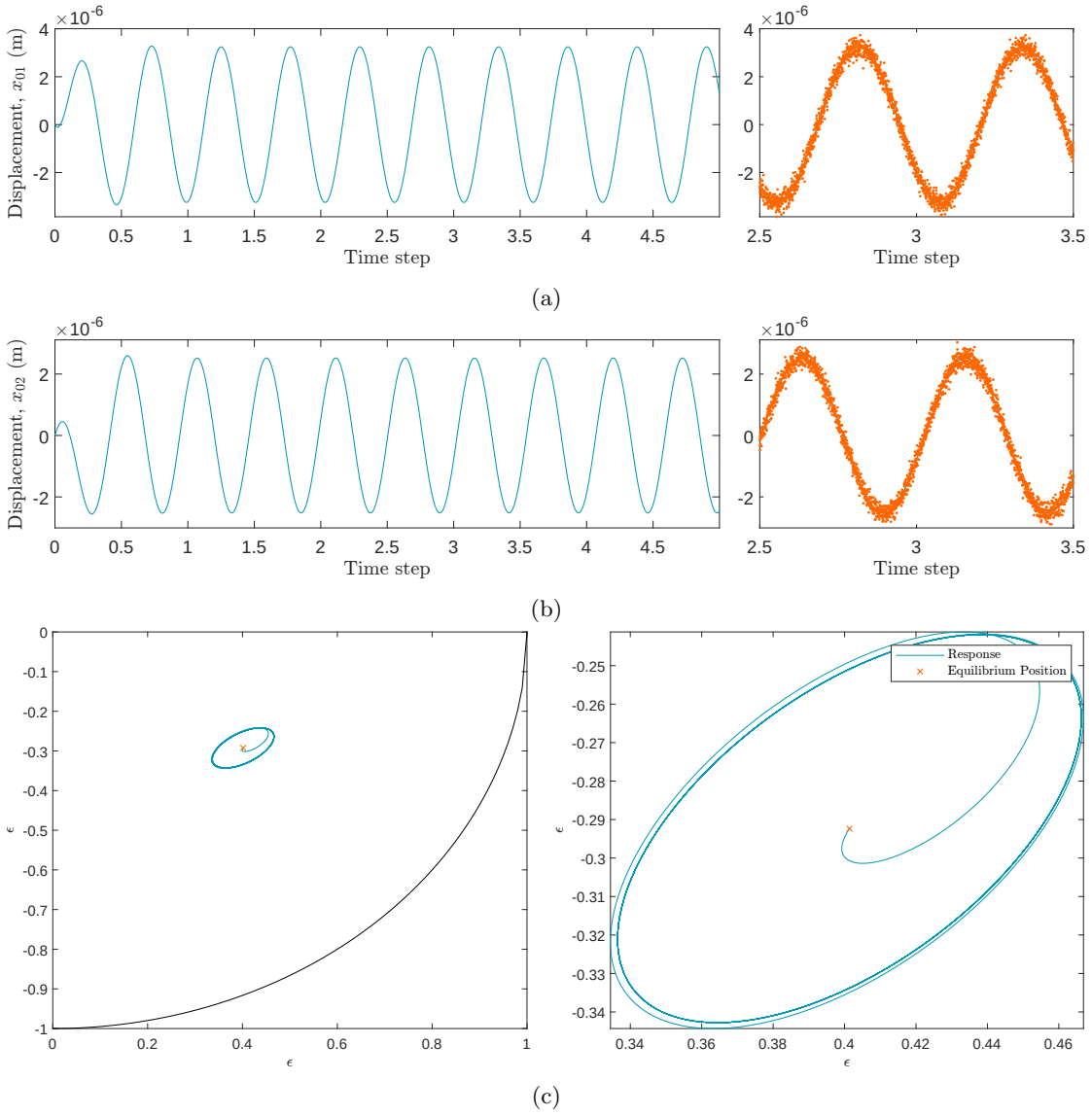


FIGURE 7.2: Journal response to out-of-balance forces: illustrating the histories of (a) x_{o1}, (b) x_{o2}, and (c) orbiting within the bearing. Detailed views are provided to show the noisy measurements.

to the overall signal to the extent that strictly periodic function may no longer accurately capture the underlying driving forces. The Matérn 5/2 may, at least, adapt to smooth transitions that depart from a strictly sinusoidal behaviour, while still being able to capture the periodic nature of the out-of-balance forces simulated in this numerical study.

Using the displacement measurements, the states and latent force histories are estimated as shown in Figure 7.3. It is found that the predictions of the states, in both the vertical and horizontal directions, are very accurate. The expected values of the estimates closely match the ground truth, encompassed within a 3σ confidence interval. Visually, the

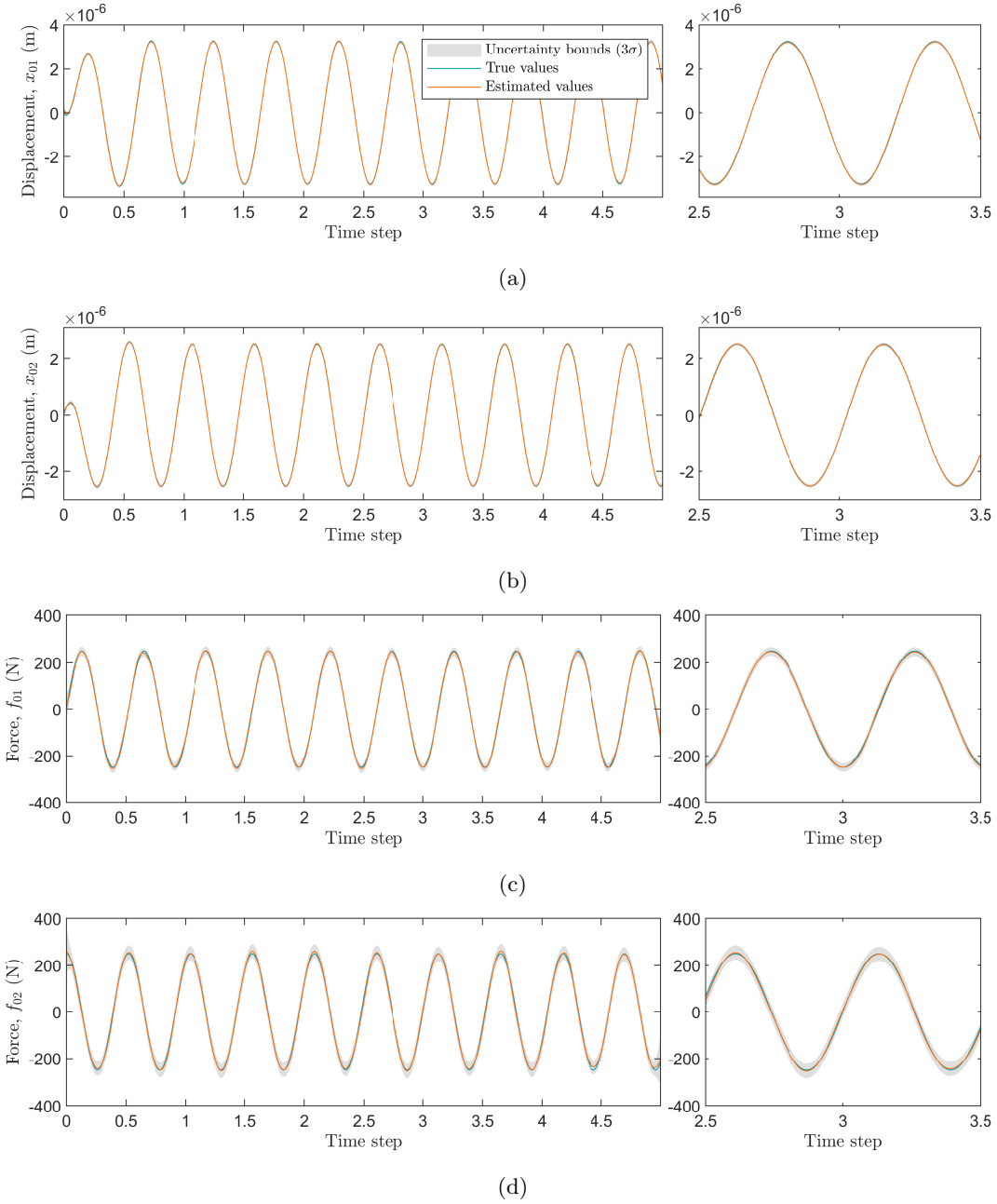


FIGURE 7.3: GP-LFM estimates of (a) x_{o1}, (b) x_{o2}, (c) f_{o1} and (d) f_{o1} for simulated harmonic excitation. The MAP estimates of hyperparameters are $\hat{l} = -1.3941$, $\hat{\sigma}^2 = 11.4498$, and $\hat{r} = -30.8297$.

estimates of the latent forces seem to also be in good agreement with the ground truth, and the periodicity forced by the out-of-balance forces is indeed captured by the GPs.

7.5.2 Numerical Case Study 2: Multi-Sine Force Excitation

A somewhat more challenging force estimation exercise is considered in this second case study, where the inputs are simulated by generating a set of narrowband multi-sine

signals. These signals are generated with a sampling rate of $f_s = 2048$Hz over the span of five seconds and are characterised by a sum of 10 different sinusoidal waves within a frequency range from 20.48Hz to 40.96Hz. The purpose of this particular case study is to evaluate the performance of the GP-LFM at estimating the loading that may derive from external factors.

The estimation of this type of loading may be of interest in determining the remaining useful life of the structure housing the journal bearing. For example, in marine applications, the complex system of stresses that ships experience from different sea states can be incredibly difficult to determine accurately. Measuring these stresses can greatly help prevent catastrophic failures caused by prolonged cyclic loading, to which large container ships are particularly susceptible [206].

Given that fluid-film bearings are commonly preferred as the main supporting components of the propeller shafts, wave-forcing histories could, in theory, be indirectly determined via measurements of the journal response. It would, indeed, be unreasonable to expect this approach to provide a complete representation of the stresses experienced by the ship structure as a whole, but some meaningful information about the loading experienced towards the stern may be recovered.

Following the same procedure as in the previous case study, the response of the journal to a set of independent multi-sine wave functions, applied in the vertical and horizontal direction, results in the journal displacements shown in Figure 7.4.

Much like the response caused by harmonic excitation, the orbiting effect illustrated in Figure 7.4(c) can also be seen to become apparent from this type of loading, but with a traced path that is more irregular. The extent of the displacements may still be considered to be within the linear range; since, according to Lund [185], the journal-response estimation is fairly accurate as long as the amplitudes remain within 40% of the bearing radial clearance.

This condition is enforced by scaling the generated forces to small enough amplitudes. In this case, the force signals were generated with a maximum amplitude of 1kN, which resulted in displacements contained within 3.3% of the clearance. In the previous case study, while the displacements were slightly higher in amplitude, these were still within 6.6% of the clearance.

For the latent force estimation, a Matérn 3/2 ($p = 1$) covariance function is employed rather than the Matérn 5/2 used in the previous case study. This choice aims to reduce the smoothing effect that the GP may have over high-frequency components in the forcing signal. The estimations of the states and force histories are shown in Figure 7.5. The states and force signals in this exercise are, once again, very accurately predicted by

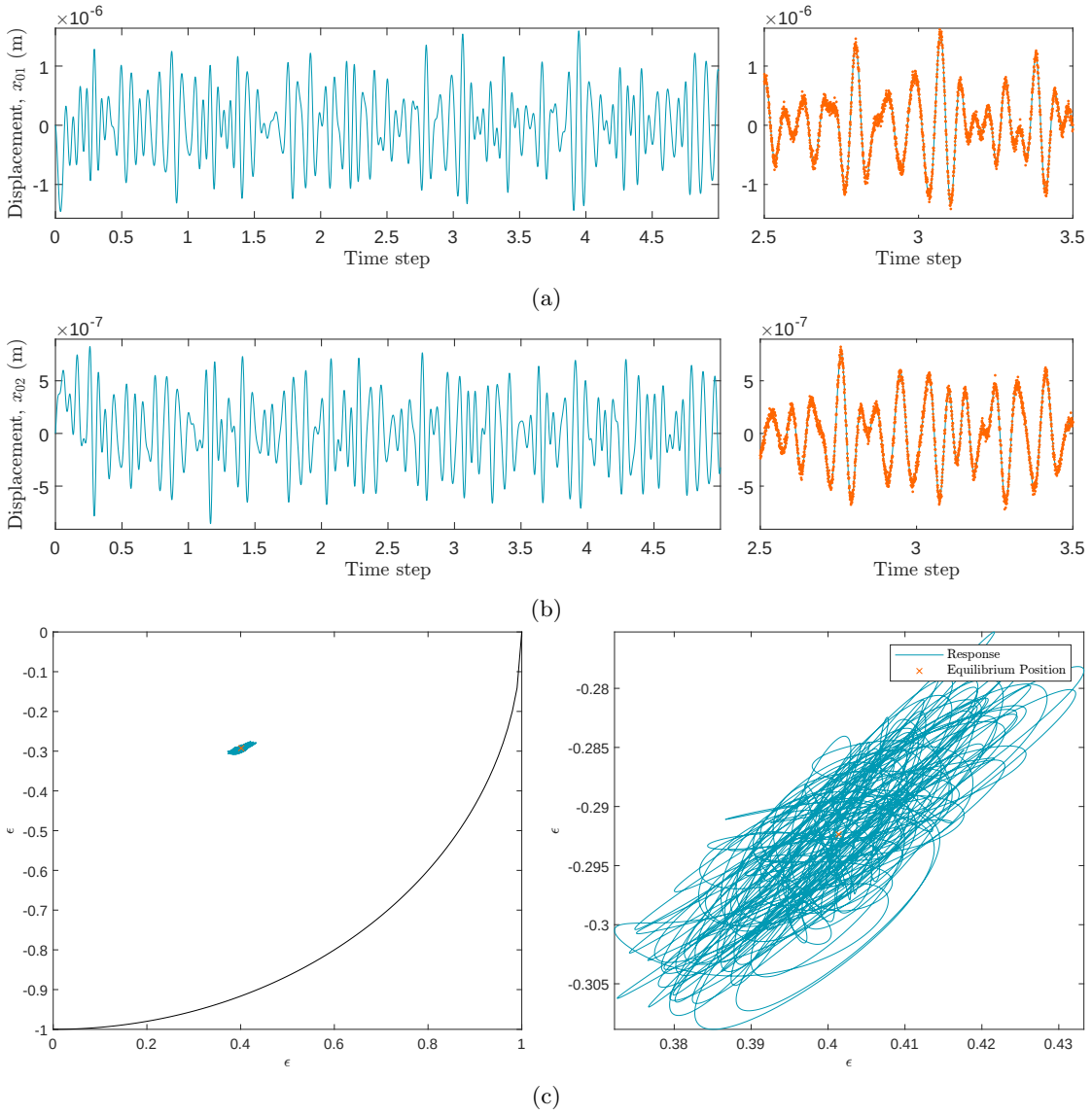


FIGURE 7.4: Journal response to multi-sine forces: illustrating the histories of (a) x_{o1}, (b) x_{o2}, and (c) orbiting within the bearing. Detailed views are provided to show the noisy measurements.

the Kalman filter and RTS smoother. All values fall within 3σ away from the expected displacement estimates, and the predictions are almost indistinguishable from the ground truth.

7.5.3 Numerical Case Study 3: Impulse Excitation

To finalise this series of numerical studies, the journal response to an impulse excitation is now considered. This numerical case study is conducted to simulate the scenario in which the rotor experiences an unprecedented impact. The aim here is to assess how well the GP-LFM can predict the instant and magnitude of such impact. The force

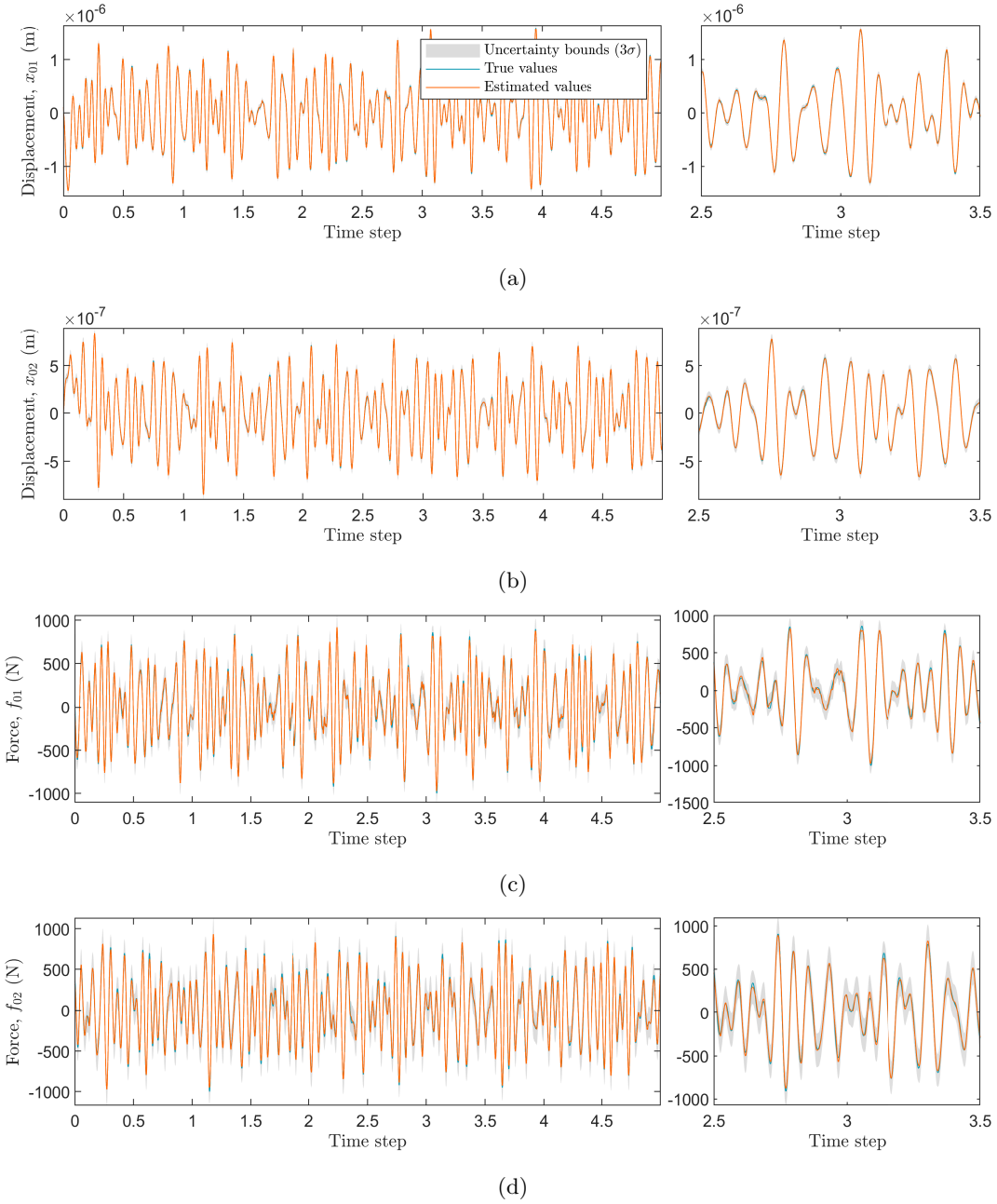


FIGURE 7.5: GP-LFM estimates of (a) x_{o1}, (b) x_{o2}, (c) f_{o1} and (d) f_{o1} for simulated multi-sine excitation. The MAP estimates of hyperparameters are $\hat{l} = -3.6811$, $\hat{\sigma}^2 = 12.4130$, and $\hat{r} = -33.7743$.

signals are simulated with a 100kN impulse excitation applied at $\phi = 45°$. In addition to the impact excitation, the force signal is overlaid with random white noise to simulate possible random forces the bearing might also experience during operation. The random excitation is generated with i.i.d. samples from a zero-mean Gaussian with a standard deviation of 316.2N, and the impulse is triggered at 1.1s into the recording. The solution of the journal response to an impulse excitation is shown in Figure 7.6.

The effect caused by the impact excitation is depicted more evidently in Figure 7.6(c). A

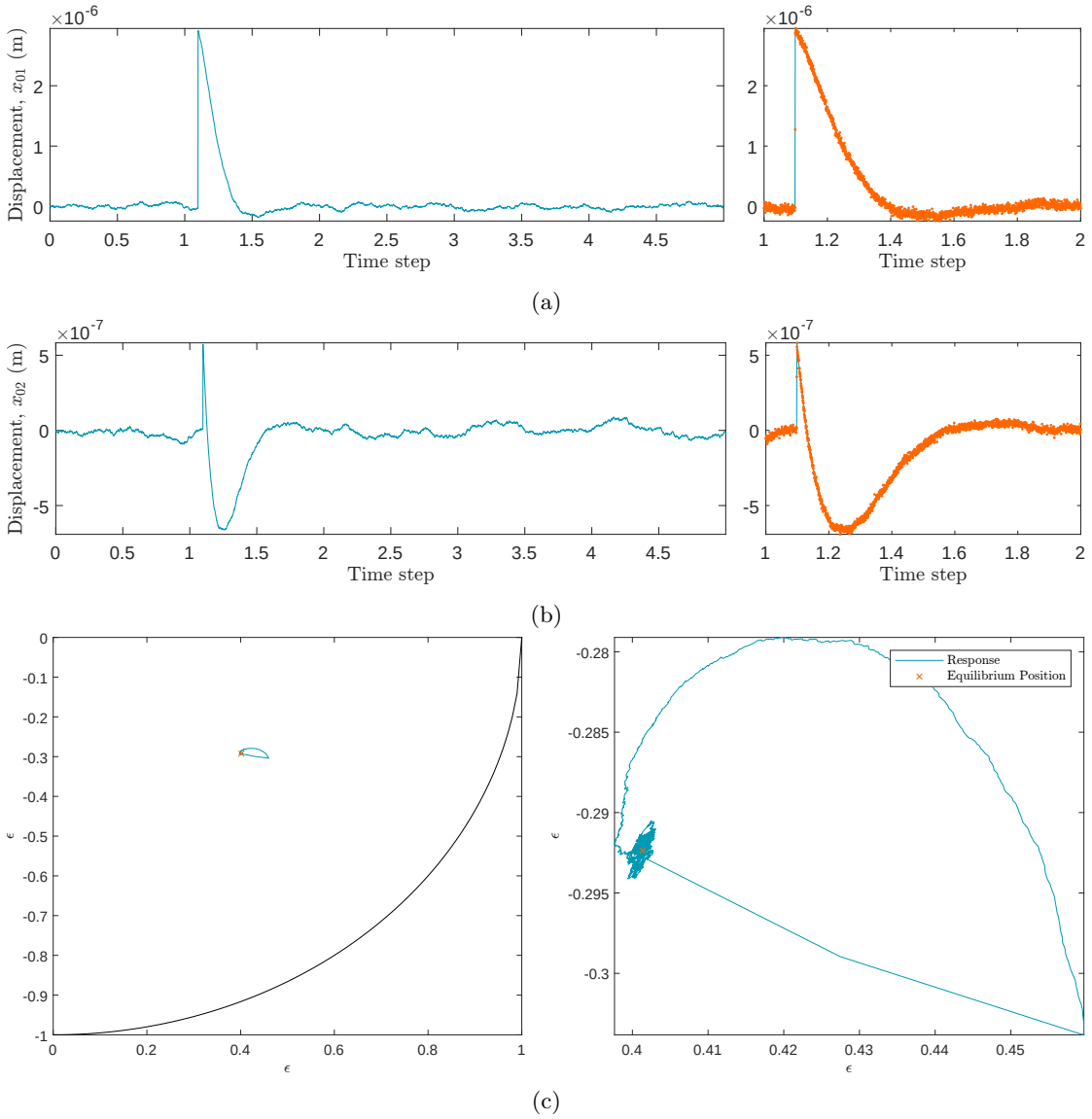


FIGURE 7.6: Journal response to an impulse: illustrating the histories of (a) x_{o1}, (b) x_{o2}, and (c) orbiting within the bearing. Detailed views are provided to show the noisy measurements.

sudden change in the trajectory of the journal is shown with the journal getting "knocked out" off its equilibrium position before gradually finding its way back and adopting an orbiting motion similar to those exhibited in the preceding case studies but in an even more irregular manner. Intuitively, the orbiting about the equilibrium position is caused by the internal forces reacting to the random excitations.

The response to an impulse excitation is of particular importance because it can lead to metal-to-metal contact between the journal and outer race, thereby causing potential damages and/or instabilities [181]. As shown in this case study, the magnitude of the impulse is not high enough to cause the journal to come into contact with the bore; that is, the journal motion remains within the geometrical bounds of the bearing. A

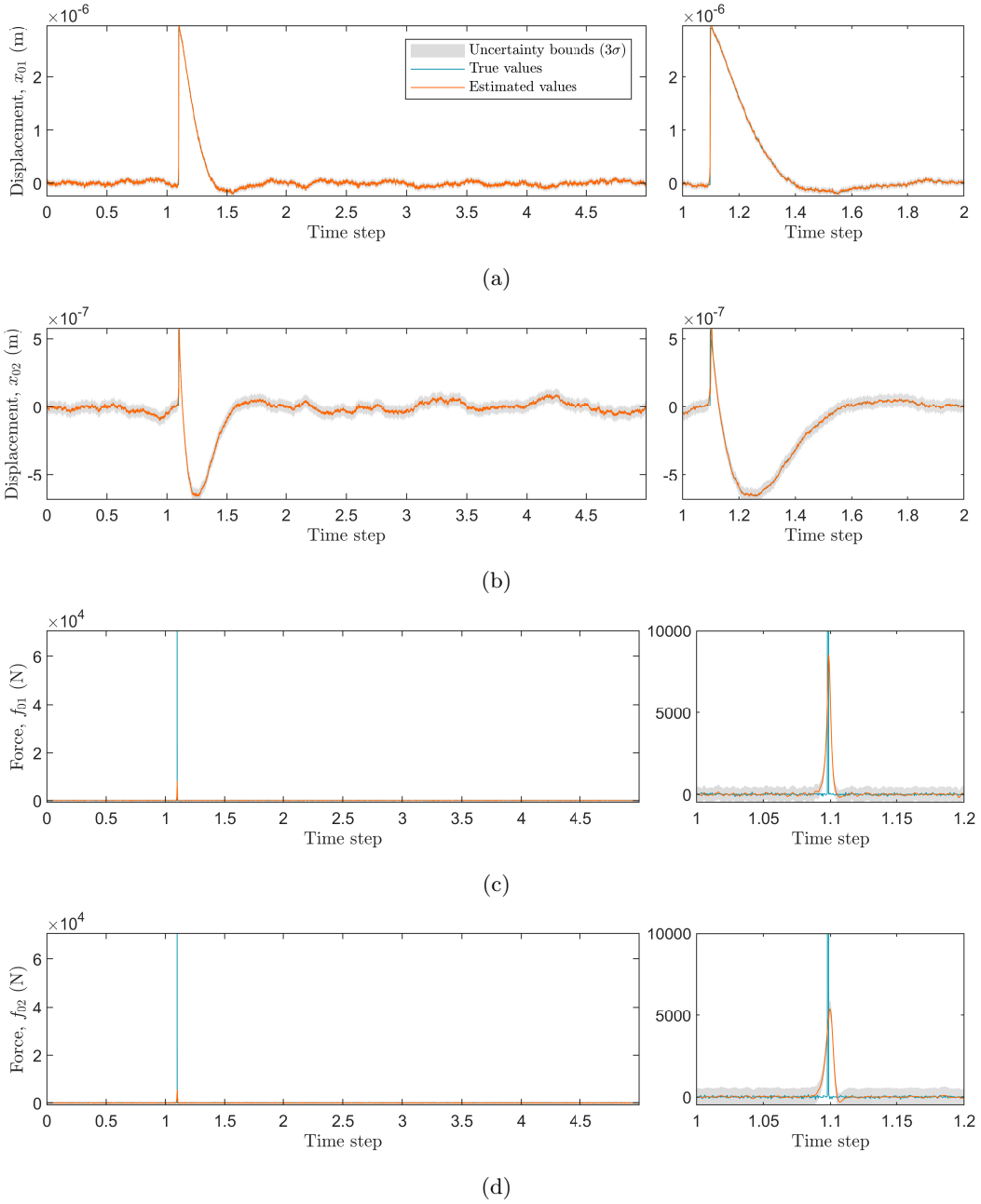


FIGURE 7.7: GP-LFM estimates of (a) x_{o1}, (b) x_{o2}, (c) f_{o1} and (d) f_{o1} for simulated impulse excitation. The MAP estimates of hyperparameters are $\hat{l} = 0.8745$, $\hat{\sigma}^2 = 17.3064$, and $\hat{r} = -34.6335$.

potential concern, nonetheless, is that abrupt changes in the response can lead to instabilities, which are characterised by strong nonlinearities that the present linearised model disregards. For now, the response is assumed correct, given that the intention here is to determine the onset of the impact in the force history.

Unlike the previous case studies, the expected latent forces present discontinuities that might never be captured with a smooth GP. It is thus reasonable to set $p = 0$ in the generalised definition of the Matérn function, in order to recover the exponential

covariance function, which renders a rougher process. The results of implementing the GP-LFM with a Matérn 1/2 are shown in Figure 7.7. Similar to the previous case studies, the latent states are predicted with a high degree of accuracy.

On the other hand, while the predictions of the latent forces seem to capture the instant at which the impulse occurs, the amplitude is quite significantly underestimated. A detailed view of the force histories illustrates the shortcomings of the model in capturing the discontinuity imposed by the impulse. Despite using a covariance function that should, in theory, capture sharp up-turns like the present one, there are a few considerations that prevent a more accurate representation of such. The first is the fact that the GP attempts to model an inherently non-stationary signal. In other words, the covariance function is simply not expressive enough to account for the sudden and abrupt change in amplitude. Additionally, the majority of the signal is dominated by random excitations, thereby promoting the GP-LFM to learn over the entire signal at the expense of predictive accuracy within the relatively short section where the impulse manifests. In the interest of detecting the impact, however, the estimated forcing signals do present a peak that is prevalent at the instant in which the impulse occurs, providing information that is still meaningful for analysis.

7.5.4 Overall Remarks and Discussion

The results obtained from the numerical case studies raise some compelling points that are worthy of discussion. A prevalent consideration that is shared among all cases is the choice of covariance function; indeed, this is a key aspect that requires careful attention and is of particular importance in the present analysis. Generally, the choice of a suitable covariance function may be assessed by comparing the fit of the rendered models to a set of training data. The problem in this case, however, is clear, since the "training data" would correspond to the measurements of the true forces driving the response, which are unknown. It is thus unfortunate that the means to validate the reliability of the latent force predictions warrant their corresponding measurements, a requirement which one may argue would make this modelling approach redundant once force measurements are available. Nevertheless, whenever possible, the value of measuring force signals for validating GP-LFMs is in the insights they may provide about the system, which can help ensure accurate predictions in other situations where force measurements are simply prohibited.

Since the only source of data in these case studies was derived from the measured journal displacements, the choice of a suitable covariance function was based merely on the nature of the expected excitations. The justification for employing the different

variations of the Matérn function is briefly outlined in each simulated scenario, where the Matérn 5/2, 3/2 and 1/2 were selected for modelling out-of-balance, multi-sine and impulse excitations, respectively. Given the absence of input measurements, the smoothing parameter ν was determined in anticipation of the potential excitations the journal bearing could experience.

By this premise, the correct selection of a covariance function is, certainly, application-dependent, and relies heavily on the expertise one has about the system. However, on a more pragmatic note, it is worth highlighting that some challenges may limit the selection process to a handful of options, as the state-space representation of a covariance function is not readily available in most cases. The Matérn family is a special case in which a closed-form state-space representation exists for non-negative integer values of p, but it is often the case that the state-space representation of a covariance function requires an approximation of its spectral density that can be somewhat cumbersome to evaluate. This consideration is the subject of ongoing research in the field of machine learning. Although reviewed briefly in the above studies, a comprehensive analysis of the choice of covariance function for predicting a richer selection of loading conditions is an exercise that requires further investigation, and it is left as research that will be pursued in the future.

For now, the outcome in all three case studies can be considered to be satisfactory in terms of the GP-LFM correctly identifying the unique signatures of the simulated forces. Unlike the first two case studies, the impulse excitation was the least accurate estimation, as shown by the root-mean-squared error of the expected estimates with the true values in Table 7.3. This outcome, however, comes as no surprise, since the high error is likely attributed to the underestimation of the impulse amplitude. Nonetheless, if the overall aim is to recover accurate load estimates, then a few measures can be enforced to address potential shortcomings. For example, within the established Bayesian framework, priors can be defined to favour hyperparameters that are expected to provide a better representation of the states and latent forces. This consideration was, in fact, indispensable in attaining the results presented above. Changing the mean and covariance of the priors had a substantial effect on the predictions, and the best results were obtained when imposing the priors in Table 7.2.

TABLE 7.3: Root-mean-squared error of input-state predictions from numerical case studies with known parameters.

Case Study	f_{o1}(N)	f_{o2}(N)	x_{o1}(m)	x_{o2}(m)
1	4.5792	5.7515	6.3535×10^{-8}	1.3497×10^{-8}
2	15.5027	15.1927	1.1608×10^{-8}	5.4718×10^{-9}
3	963.9761	981.6669	6.0019×10^{-8}	2.6457×10^{-8}

An element of interpretability comes into play when determining the values of the hyperparameters, which are not readily interpretable in a physical sense. Nevertheless, by acknowledging the role these parameters have in defining the Matérn family of functions, a sensible initial guess might involve setting the prior over σ^2, and the prior over l, around the expected variance and frequency content of the force signal, respectively. If such information is inaccessible, inference may become more challenging, potentially requiring the definition of improper priors. In such cases, finding optimal parameters may only be possible with sufficient representative data to mitigate the biases introduced via the prior.

While the exact dynamic coefficients were provided, it is important to note that a unique solution is not necessarily guaranteed. This argument is primarily based on the complexities that the posterior distribution might present over the space of hyperparameters. It is unlikely for the posterior to be characterised by an unimodal functional form, implying that the optimiser is susceptible to finding a solution in one of potentially several local minima. Therefore, achieving a precise estimation of the ground truth may be possible by finding the correct combination of hyperparameters for a particular problem; assuming, of course, that a suitable covariance function has been chosen beforehand.

The caveats discussed thus far are, unfortunately, exacerbated when the parameters of the dynamic model are also unknown. However, pursuing this exercise with the GP-LFM has the potential to correctly identify the dynamic coefficients, in addition to the driving forces, *in-situ*. As highlighted in the introduction, this outcome is clearly desirable in the analysis of rotor-machinery systems. An exploratory study of this approach is thus conducted in the following section.

7.6 Input-State Estimation with Unknown Parameters

Determining all eight dynamic coefficients can be a challenging task, both analytically and experimentally. The reasons that amount to this challenge are vast, including assumptions made about the solution to the Navier-Stokes equations [207], which are further simplified in defining the Reynolds' equation [159] governing fluid-film bearings. The accuracy of the solutions based on these equations depends on operational conditions and can be affected by multiple factors, including changes in oil temperature, oil-flow supply, and turbulence. The introduction of rotor dynamics further complicates the analysis, and finite-element modelling often becomes necessary for accurate results. Numerical modelling ranges in complexity, with the *Elasto-Hydrodynamic* (EHD) and *Thermo-Elasto-Hydrodyanmic* (TEHD) models [133] being prevalent examples used to overcome the limitations of a purely analytical solution.

Despite the advances in modelling journal-bearing dynamics, the agreement between theoretically-evaluated and experimentally-measured dynamic coefficients are generally limited, typically within $10 - 20\%$ [208].

The GP-LFM offers the advantage of leveraging data to mitigate modelling errors. By incorporating observations gathered from the bearing (e.g. displacement, velocity, acceleration), the model seeks to compensate for the assumptions made about the system. The previous section introduced the use of the GP-LFM to recover the latent forcing from different simulations, under the assumption that dynamic coefficients were known. However, these coefficients might not be readily available or could be subject to errors in practical applications.

The GP-LFM can be easily extended to incorporate the dynamic coefficients as unknown parameters during inference, making the problem at hand one of finding an optimal set of parameters $\Theta = \left[\sigma^2, l, c_{11}, c_{12}, c_{21}, c_{22}, k_{11}, k_{12}, k_{21}, k_{22}, r\right]$, such that Equation (7.36) is minimised during the implementation of the Kalman filter for input-state estimation. However, this introduces an ill-posed problem since any two of the estimates - inputs, state, or parameters - are necessary to determine the remaining third.

In the implementation above, for example, the parameters were known, and information about the states was provided by the observations, facilitating the estimation of the latent-force signals. Indeed, the GP hyperparameters had to be inferred in the process, but finding their true values is often of little interest as long as the predictions made by the GPs are reasonable. Therefore, adding the dynamic coefficients as an unknown appears to present a problem that is seemingly impossible to solve.

Some progress towards solving this ill-posed problem involves the use of well-informed priors on the parameters. These priors are necessary to constrain the range of possible solutions to those that are physically meaningful. In practice, finite-element model solutions or experimental measurements can inform these priors, allowing for a more realistic estimation of the dynamic coefficients.

To illustrate the intricacy of the problem, the conditions outlined in Case Study 2 are replicated in the following demonstration. The main difference, however, is in the fact that the dynamic coefficients are no longer assumed to be known in advance. To constrain the problem, the mass of the journal is assumed to be entirely supported by the bearing, and it is thus known in advance.

Table 7.4 provides insight into the parameters, including their original values and their corresponding prior distributions. The value $\hat{c}_{22}$ is disregarded since $\hat{c}_{12} = \hat{c}_{22}$. As all the parameters are required to be positive, a logarithmic transformation is employed to map

the search space, denoted by $(\hat{\cdot})$, following the consistent notation from the preceding section.

TABLE 7.4: Priors for GP-LFM hyperparameters and dynamic coefficients.

Parameter	Prior	Transformed original value	MAP estimate
$\hat{l}$	$\hat{l} \sim \mathcal{N}(-2, 1)$	–	-3.9295
$\hat{\sigma}^2$	$\hat{\sigma}^2 \sim \mathcal{N}(8, 1)$	–	11.7961
$\hat{c}_{11}$	$\hat{c}_{11} \sim \mathcal{N}(16.1, 0.1)$	16.1431	15.8411 (-1.87%)
$\hat{c}_{12}$	$\hat{c}_{12} \sim \mathcal{N}(15.4, 0.1)$	15.8262	15.2789 (-3.46%)
$\hat{c}_{22}$	$\hat{c}_{22} \sim \mathcal{N}(15.9, 0.1)$	16.9061	16.4598 (-4.00%)
$\hat{k}_{11}$	$\hat{k}_{11} \sim \mathcal{N}(18.3, 0.1)$	18.2991	18.4147 $(+0.63\%)$
$\hat{k}_{12}$	$\hat{k}_{12} \sim \mathcal{N}(17.7, 0.1)$	17.3753	19.0135 $(+9.43\%)$
$\hat{k}_{21}$	$\hat{k}_{21} \sim \mathcal{N}(19.1, 0.1)$	18.8844	19.2562 $(+1.97\%)$
$\hat{k}_{22}$	$\hat{k}_{22} \sim \mathcal{N}(18.4, 0.1)$	18.5675	18.4947 (-0.39%)
$\hat{\sigma}_n^2$	$\hat{\sigma}_n^2 \sim \mathcal{N}(-30, 1)$	-30.7310	-31.9998 (-0.04%)

The prior distributions established for the dynamic coefficients are designed to roughly encompass the feasible range of physical values that these coefficients might exhibit for the given bearing configuration. For example, Figure 7.8 illustrates the prior distribution on $\hat{c}_{11}$, which prioritises values that are more likely to explain the journal response when its equilibrium position is defined by values near $\epsilon = 0.5$.

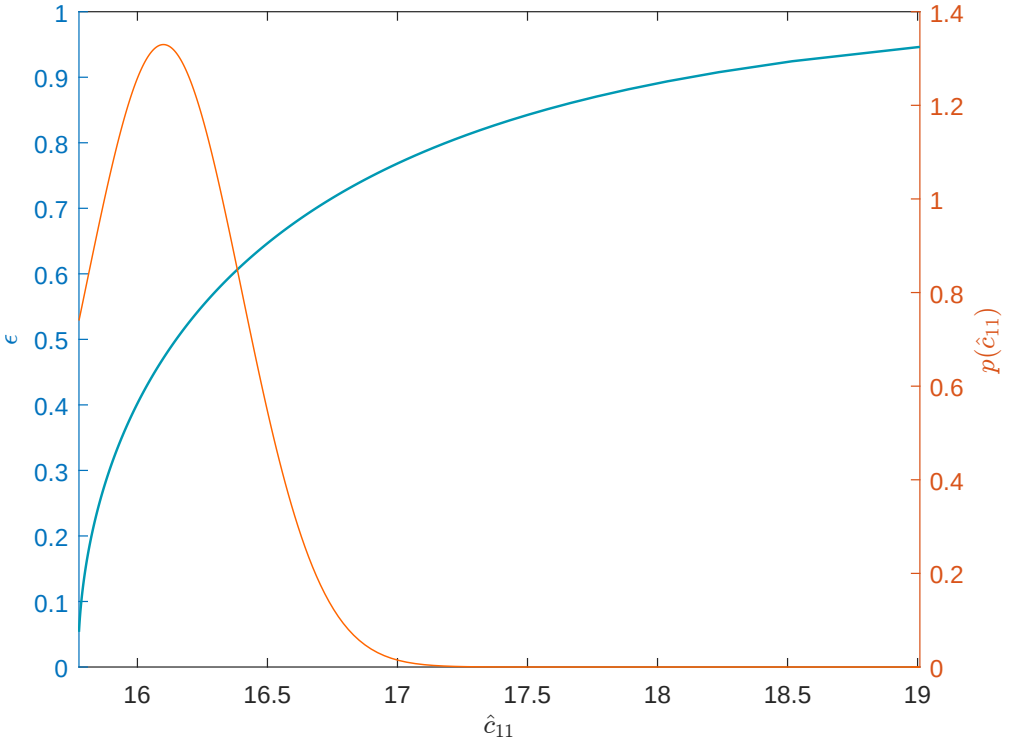


FIGURE 7.8: Illustration of prior distribution, $p(\hat{c}_{11}) = \mathcal{N}(16.1, 0.1)$, defined for the dynamic coefficient $\hat{c}_{11}$.

The MAP estimates of the parameters are again identified using the QSPO, and the results have been included in Table 7.4 alongside their original values. While the MAP

estimates seem to closely approximate the target values, it should be noted that small alterations in the log space amount to substantial deviations in the actual coefficients. To put this notion into perspective, the 4.00% difference of the estimated $\hat{c}_{22}$, equates to a 36% difference upon transformation to the original space.

The extent of these differences is evident in the quality of the latent-force predictions, illustrated in Figure 7.9. What is interesting about these results, however, is the fact the states are accurately estimated regardless of the elicited parameters. Quantitatively, the root-mean-squared error of the predictions is provided in Table 7.5. The performance of the GP-LFM in estimating the states remains comparable to that when the dynamic coefficients are known, even when the error of the force estimates increases by almost an order of magnitude.

TABLE 7.5: Root mean-squared error of state and input predictions.

f_{o1}(N)	f_{o2}(N)	x_{o1}(m)	x_{o2}(m)
110	149	1.6630×10^{-8}	9.1712×10^{-9}

This outcome can be attributed to the flexibility of the GPs, which enables them to adapt to the biased parameters employed in the Kalman filter. Consequently, even if the proposed parameters diverge from their true values, the GPs can adjust to produce state estimates that align well with measurements. This phenomenon is rooted in *non-identifiability*, whereby sets of distinct optimal parameter, e.g. $\hat{\Theta}_1$ and $\hat{\Theta}_2$, yield equivalent likelihoods for the observations ($p(y_{1:T}|\hat{\Theta}_1) = p(y_{1:T}|\hat{\Theta}_2)$). These shortcomings are, in fact, alleviated with priors that promote the search for optimal minima that approximate closely to the true values of the parameters.

In terms of inferring the accurate dynamic coefficients using GP-LFM, it is important to acknowledge that without precise prior knowledge, this task proves remarkably difficult. The high dimensionality of the posterior distribution renders the problem susceptible to the curse of dimensionality, thereby necessitating substantial amounts of data to isolate the true parameters.

It is worth highlighting that, although the true dynamic coefficients might not be recovered precisely from the MAP estimates, the true force signals can be found within a 3σ confidence bounds from the expected values. Furthermore, the probabilistic framework implies that the true parameters are encompassed within the confidence bounds of the posterior distribution. These bounds can be inferred by means of MCMC algorithms, but with the caveat that for a ten-dimensional space, the complexity of the posterior distribution might hinder the sampler from converging to a meaningful approximation. This subject requires further investigation, and it will be pursued as an advance on the work presented in this chapter.

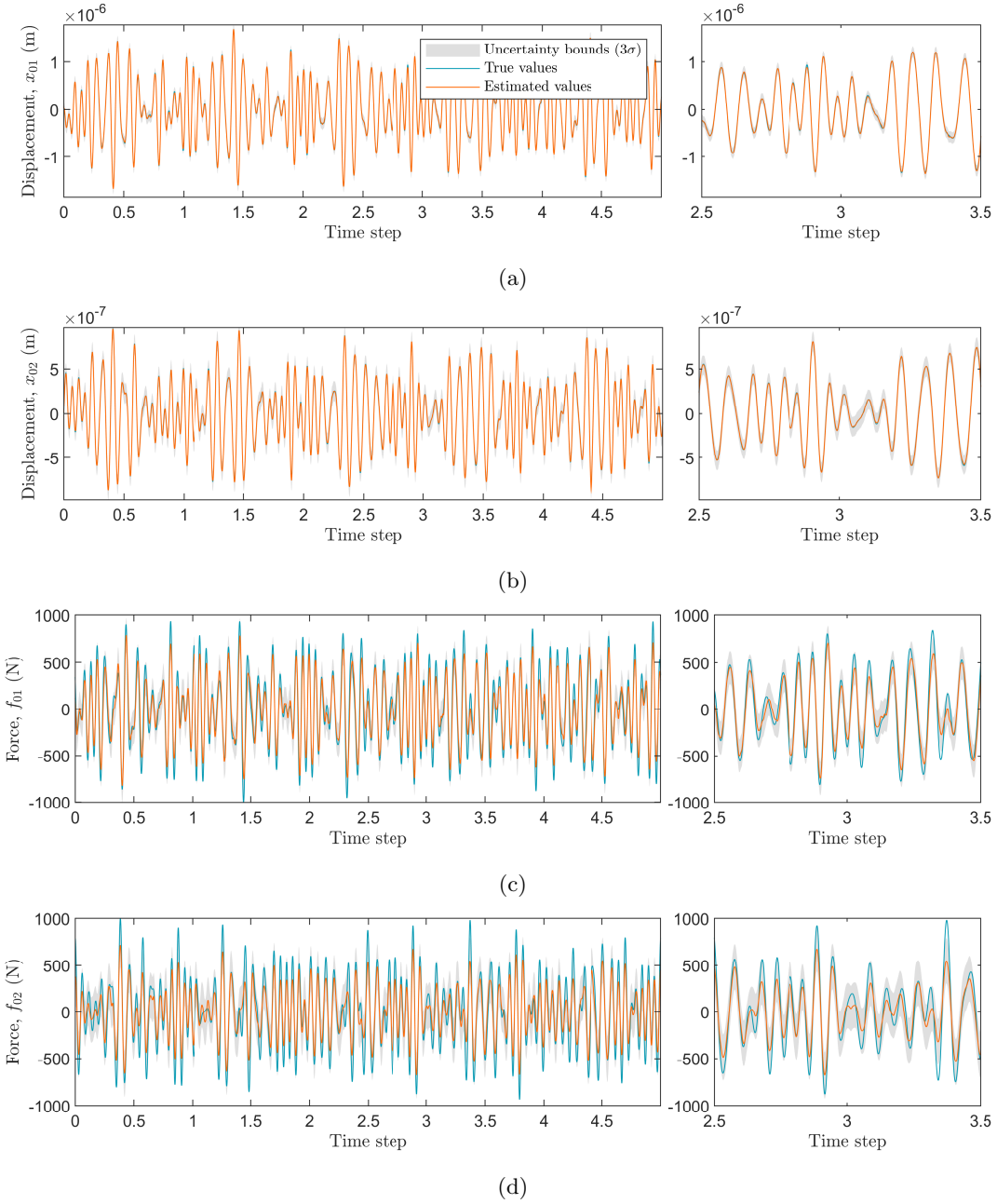


FIGURE 7.9: GP-LFM estimates of (a) x_{o1}, (b) x_{o2}, (c) f_{o1} and (d) f_{o1} for simulated multi-sine excitation with unknown parameters.

7.7 Conclusions

The objective of this chapter has been to explore whether the GP-LFM can offer a promising approach to address common challenges encountered in monitoring rotating machinery; namely, in estimating driving forces and dynamic coefficients with measurements obtained from within a journal bearing. Instead of providing a comprehensive solution, an exploratory approach was taken, presenting results from simulated case studies.

The results highlighted that the success of the GP-LFM in making predictions is heavily reliant on selecting appropriate covariance functions. Equally important is the definition of well-informed prior distributions, which are vital for attaining meaningful estimates of the inputs, states and parameters. Although the outcomes were analysed with respect to simple numerical case studies, the insights elicited from them are also likely to be applicable to more complex scenarios.

Arguably of more significance, however, is to conclude this chapter by discussing potential directions for further exploration. Four main concepts remain open-ended and are certainly worthy of further investigation, suggesting possible directions for advancing these ideas:

1. Validating the GP-LFM using measurements collected from actual systems. This task might require more sophisticated models than the simplified linearised short-width bearing assumption to ensure accurate predictions on real-world systems.

2. Analysing responses in which nonlinear effects become prevalent. While the intricacies of acknowledging this consideration may involve relying on models with added complexities, a more realistic representation of the journal responses would further enhance the applicability of the model in practical scenarios.

3. Consideration of including confounding factors in the analysis, such as the impact of the rotor on the response and operational variations that were omitted in the presented case studies. Incorporating these factors would lead to a more comprehensive understanding of the bearing's behaviour.

4. Developing a systematic methodology to address the ill-posed problem of jointly estimating inputs, states, and parameters for journal bearings. This direction might involve addressing the complexities of parameter estimation in high-dimensional spaces and exploring prior distributions that might be better suited for this problem.

Chapter 8

On the Use of Variational Auto-Encoders for Preprocessing Data in SHM Applications

In Chapter 2, the importance of feature selection was highlighted in the development of statistical models. This chapter delves deeper into the concept of feature selection by examining data from two distinct experimental case studies.

Given the importance of data pre-processing in SHM, this chapter serves as a brief independent study of how some discussed concepts in Chapter 2 manifest in practical applications. In particular, it highlights the challenges that may arise during damage identification, such as dealing with high-dimensional features, or benign variations that can potentially mislead novelty detectors.

The first case study investigates rolling-element element bearing operation for damage detection. Although this is a CM-oriented problem, the concepts presented are equally applicable to SHM applications. The second case study explores the effects of environmental variations on damage-sensitive features, introducing complexities that must be addressed to ensure the effectiveness of a novelty detector.

An exploratory study is pursued offering a novel view of the implementation of autoencoders for data preprocessing. Specifically, the use of VAEs is investigated as a means of dimensionality reduction before conducting damage detection analysis. As a type of latent space detection-based model [209], VAEs have recently shown promise in generating low-dimensional data representations [210–213]. Their potential in the context of SHM is yet to be fully explored.

A brief theoretical background related to VAEs is covered in the following section. The first of the case studies is then outlined and examined before advancing to the second. The structure of this chapter is intended to demonstrate systematically the advancements required to address the added complexities introduced by environmental variations, and thus illustrate how VAEs can be adapted to effectively address them accordingly.

8.1 A Brief Background on VAEs

Introduced by Kingma and Welling [72], VAEs make use of neural networks to efficiently derive approximate posteriors over continuous latent variables. The premise of VAEs is based on a set of N observable variables $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \in \mathbb{R}^{D \times N}$, assumed to have been generated from a distribution conditioned on an underlying set of latent variables $\mathbf{Z} = \{\mathbf{z}_1, ..., \mathbf{z}_N\} \in \mathbb{R}^{J \times N}$. That is, for every input vector $\mathbf{x}_n \in \mathbb{R}^D$, a latent vector $\mathbf{z}_n \in \mathbb{R}^J$ exists. The inputs are projected into a reduced latent space with a nonlinear transformation $e : \mathbb{R}^D \to \mathbb{R}^J$, where $J < D$.

A VAE infers an approximate posterior distribution $q(\mathbf{Z}|\mathbf{X})$, from which the latent variables are generated. The need to approximate the posterior is motivated by the potential intractability involved in solving for the true posterior $p(\mathbf{Z}|\mathbf{X})$. The approximate posterior derives when attempting to maximise the log marginal likelihood $p(\mathbf{X})$, leading to the following decomposition [48, 175],

$$\log p(\mathbf{X}) = \mathcal{L}(q(\mathbf{Z}|\mathbf{X})) + D_{KL}(q(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z}|\mathbf{X})) \tag{8.1}$$

where the terms in the expression are defined by,

$$\mathcal{L}(q(\mathbf{Z}|\mathbf{X})) = \int q(\mathbf{X}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z}|\mathbf{X})} \right) \mathrm{d}\mathbf{Z} \tag{8.2}$$

$$D_{KL}(q(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z}|\mathbf{X})) = -\int q(\mathbf{X}) \log \left(\frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z}|\mathbf{X})} \right) \mathrm{d}\mathbf{Z} \tag{8.3}$$

The first term in Equation (8.1) corresponds to the *Evidence Lower-Bound* (ELBO), while the second term is the *Kullback-Leibler* (KL) divergence between the approximate and true distributions $q(\mathbf{Z}|\mathbf{X})$ and $p(\mathbf{Z}|\mathbf{X})$, respectively. The KL-divergence term is always positive and reduces to zero if, and only if, $q(\mathbf{Z}|\mathbf{X}) = p(\mathbf{Z}|\mathbf{X})$. Therefore, an estimate of the approximate posterior $q(\mathbf{Z}|\mathbf{X})$ can be found by minimising $D_{KL}(q(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z}|\mathbf{X}))$. Given that $\log p(\mathbf{X})$ is constant in (8.1), this process is equivalent to maximising $\mathcal{L}(q(\mathbf{Z}|\mathbf{X}))$, which is easier to evaluate since one now deals with the joint probability $p(\mathbf{X}, \mathbf{Z})$, rather than the intractable posterior $p(\mathbf{Z}|\mathbf{X})$.

The optimisation objective of the VAE is thus defined by the ELBO, which can be conveniently expressed in the following way,

$$\mathcal{L}(q(\mathbf{Z}|\mathbf{X})) = \mathbb{E}_{q(z|x)}\left[\log p(\mathbf{X}|\mathbf{Z})\right] - D_{KL}(q(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z})) \tag{8.4}$$

where the prior distribution over the latent variables $p(\mathbf{Z})$, now appears in the KL divergence term. Details on the derivations of equations (8.1) and (8.4) are provided in Appendix C.

The VAE computes the parameters that define $q(\mathbf{Z}|\mathbf{X})$ with a neural network encoder $e_\phi(\mathbf{X})$, which is parameterised by the set of weights ϕ. If the posterior is assumed Gaussian with diagonal covariance, the encoder outputs the mean and variance corresponding to each independent distribution. Concretely, for a single datapoint $\mathbf{x}_n$, the approximate posterior is defined as,

$$q_\phi(\mathbf{z}_n|\mathbf{x}_n) = \mathcal{N}(\mathbf{z}_n|\boldsymbol{\mu}_\phi(\mathbf{x}_n), \boldsymbol{\sigma}^2_\phi(\mathbf{x}_n)\mathbb{I}) \tag{8.5}$$

where $\boldsymbol{\mu}_\phi(\mathbf{x}_n) = e_{\phi,\mu}(\mathbf{x}_n;\phi)$ and $\boldsymbol{\sigma}^2_\phi(\mathbf{x}_n) = e_{\phi,\sigma}(\mathbf{x}_n;\phi)$. As illustrated by Figure 8.1, the probabilistic encoder – or *recognition network* –, encodes the inputs into a stochastic bottleneck before passing samples from $q_\phi(\mathbf{z}_n|\mathbf{x}_n)$ to the probabilistic *decoder* – or generative network –, which attempts to approximately reconstruct the original inputs.

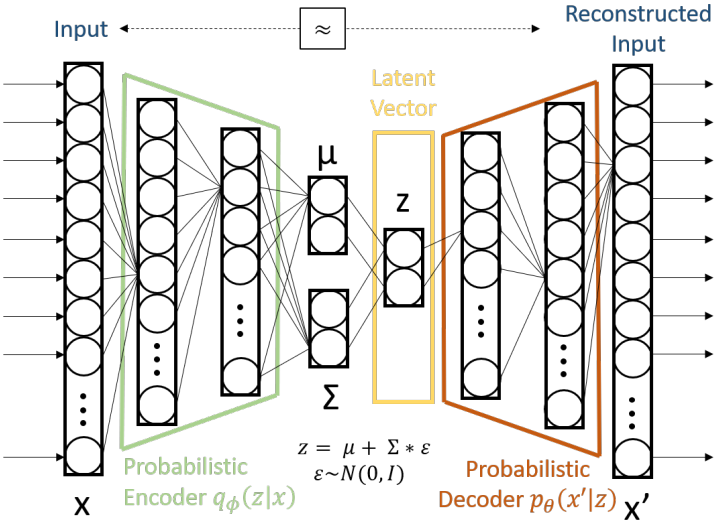


FIGURE 8.1: Diagram illustrating the structure of a VAE.

Therefore, optimising the ELBO over $q(\mathbf{z}_n|\mathbf{x}_n)$ is achieved by optimising the ELBO over the neural-network parameters (ϕ,θ). The optimal values of these sets of parameters can be found jointly using *Stochastic Gradient Descent* (SGD). The problem is that the gradients of the ELBO with respect to ϕ are not readily solvable because the expectation with respect to $q_\phi(\mathbf{z}_n|\mathbf{x}_n)$ is also a function of ϕ.

Fortunately, this issue can be addressed with the *reparameterisation trick* [214], whereby the expectation with respect to $q_\phi(\mathbf{z}_n|\mathbf{x}_n)$ is replaced with one with respect to a distribution of an auxiliary random variable ϵ, that is independent of $\mathbf{x}_n$ and ϕ. Rewriting the ELBO expectation in terms of ϵ thus yields,

$$\mathbb{E}_{q_\phi(\mathbf{z}_n|\mathbf{x}_n)}\left[\log p(\mathbf{x}_n|\mathbf{z})\right] = \mathbb{E}_{p(\epsilon)}\left[\log p(\mathbf{x}_n|\mathbf{z}_n)\right] \tag{8.6}$$

where $\mathbf{z}_n$ is defined by some vector-valued function $g(\epsilon, \phi, \mathbf{x}_n)$, parameterised by the corresponding $\mathbf{x}_n$, the set of weights ϕ, and the auxiliary variable now defined as a random noise sample $\epsilon \sim p(\epsilon)$. A Monte Carlo estimator $\tilde{\mathcal{L}}(\phi, \theta; \mathbf{x}_n)$ of the ELBO can then be defined as follows,

$$\begin{aligned} \epsilon^{(l)} &\sim p(\epsilon) \\ \mathbf{z}_n^{(l)} &= g(\epsilon^{(l)}, \phi, \mathbf{x}_n) \\ \tilde{\mathcal{L}}(\phi, \theta; \mathbf{x}_n) &= \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}_n|\mathbf{z}_n^{(l)}) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}_n)||p_\theta(\mathbf{z})) \end{aligned} \tag{8.7}$$

where L denotes the number of samples per datapoint.

Given that the posterior is approximated with a diagonal Gaussian, the sampling process defined as $\mathbf{z}_n^{(l)} \sim \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}_n), \boldsymbol{\sigma}_\phi^2(\mathbf{x}_n)\mathbb{I})$ can be reparameterised as $\mathbf{z}_n^{(l)} = \boldsymbol{\mu}_\phi(\mathbf{x}_n) + \boldsymbol{\sigma}_\phi(\mathbf{x}_n) \cdot \boldsymbol{\epsilon}^{(l)}$, where $\boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, \mathbb{I})$. This derivation is replicated from [72], where it is also demonstrated that the gradient $\nabla_{\phi,\theta}\tilde{\mathcal{L}}(\phi, \theta; \mathbf{x}_n)$ can be used to optimise the ELBO with minibatch SGD. Further details of this computation can be found in [215].

Solving for the KL-divergence term can be achieved without the estimation required for the log-likelihood in (8.7). An analytical closed-form solution is possible when having the prior distribution defined by a centred isotropic multivariate Gaussian; i.e. $p_\theta(\mathbf{z}_n) = \mathcal{N}(0, \mathbb{I})$. By using this prior, the ELBO estimator now yields,

$$\tilde{\mathcal{L}}(\phi, \theta; \mathbf{x}_n) = \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(\mathbf{x}_n|\mathbf{z}_n^{(l)}) + \frac{1}{2}\sum_{j=1}^{J}(1 + \log(\sigma_{\phi,j}^2(\mathbf{x}_n)) - \mu_{n,j}^2(\mathbf{x}_n) - \sigma_{n,j}^2(\mathbf{x}_n)) \tag{8.8}$$

where the subscript j denotes the j^{th} element in vectors $\boldsymbol{\mu}_\phi(\mathbf{x}_n)$ and $\boldsymbol{\sigma}_\phi^2(\mathbf{x}_n)$. Having the objective function estimated as in (8.8) allows it to be differentiable with respect to all parameters in the network.

Finally, the *reconstruction* part of the objective function can be defined by allowing the inputs to be i.i.d. samples of a Gaussian distribution. Much like in the encoding process, the likelihood $p_\theta(\mathbf{x}_n|\mathbf{z}_n)$ can be defined by the neural network decoder $d_\theta(\mathbf{z}_n)$, which is parameterised by the set of weights θ. In particular, having the decoder output the

mean and variance of the Gaussian likelihood results in the following,

$$p_\theta(\mathbf{x}_n|\mathbf{z}_n) = \mathcal{N}(\boldsymbol{\mu}_\theta(\mathbf{z}_n), \boldsymbol{\sigma}^2_\theta(\mathbf{z}_n)\mathbb{I}) \tag{8.9}$$

where $\boldsymbol{\mu}_\theta(\mathbf{z}_n) = d_\theta(\mathbf{z}_n;\theta)$ and $\boldsymbol{\sigma}^2_\theta(\mathbf{z}_n) = d_\theta(\mathbf{z}_n;\theta)$. By taking the log of equation (8.9) and omitting all additive constants, the following expression is obtained,

$$\log p_\theta(\mathbf{x}_n|\mathbf{z}_n) = -\sum_{d=1}^{D}\left[\log \sigma_{\theta,d}(\mathbf{z}_n) + \frac{(\mu_{\theta,d}(\mathbf{z}_n) - x_{n,d})^2}{2\sigma^2_{\theta,d}(\mathbf{z}_n)}\right] \tag{8.10}$$

where the subscript d denotes the d^{th} element in vectors $\boldsymbol{\mu}_\theta(\mathbf{z}_n)$, $\boldsymbol{\sigma}^2_\theta(\mathbf{z}_n)$, and $\mathbf{x}_n$. While not explicitly depicted in Figure 8.1, the variance of the data can also be learnt by extending the output of the decoder accordingly.

The definitions covered in this section pertain specifically to the special case involving Gaussian distributions only. This VAE configuration is the one adopted for the two case studies that follow. The exploration of more extensive applications of the VAE in this context is a task reserved for future research.

8.2 Experimental Case Study 1: Rolling-Element Bearing Subjected to Damage

The first set of experimental data examined here corresponds to a case study from Worden et al. [216], using data provided by Mr. M. Tabaszewski and Prof. C. Cempel, from Poznan Technical University. The original experiment involved monitoring a ball bearing (type 6024 with a steel cage), operating under various health states. The raw data were recorded from accelerometers placed on the outer casing of the bearing. Five different health states were monitored:

1. New ball bearing – Normal operation
2. Broken outer race – Damage 1.
3. Broken cage with one loose element – Damage 2.
4. Damaged case with four loose elements – Damage 3.
5. Badly worn ball bearing with no evident damage – Damage 4.

Each time signal was divided into 64-point intervals and then Fourier transformed to record the magnitude of each spectral line. The resulting dataset thus yielded a series

(1797 samples) of 32-point vectors. A training set was constructed by concatenating 200 random objects from each state, leaving 150 samples for a validation set and 647 for a test set. The architecture and hyperparameters of the networks were also adjusted by implementing a *random-grid search cross-validation* over a range of possible values. At the risk of worsening performance, however, the latent space was restricted to two dimensions in all methods. Compressing the data to only two dimensions makes the results easier to visualise and also to identify clusters of their underlying features.

8.2.1 Dimensionality Reduction Strategy on Experimental Data

Arguably, the most commonly-employed technique for dimensionality reduction is PCA [70]. This choice is, perhaps, mostly attributed to its simplicity. While PCA remains a practical technique for managing high-dimensional data, it is limited to linear transformations, thereby disregarding any potential nonlinear dependencies in the principal component projections. Being able to recover nonlinear features is in many cases necessary, warranting more elaborate dimensionality reduction techniques [217]. As highlighted in Chapter 2, one alternative is to have a multi-layer perceptron output the desired features by implementing some form of nonlinear activation in its layers.

The compressed embeddings of datapoints correspond to reduced-size arrays, where most of the information carried by the data is encoded. The collection of embeddings is projected on a latent space from which the desired analysis can be performed. In the case of AANN, the latent space is deterministic, and the embeddings are implied to be unique to their original representations. Conversely, the VAE embeds data into latent distributions, encouraging local smoothness and thereby the ability to generate sensible outputs from random noise. Additionally, VAEs treat the dimensionality-reduction problem from a Bayesian perspective by defining a prior distribution over the latent variables. This aspect of VAEs can be advantageous since the embeddings are somewhat enforced to adopt interpretable distributions, and thus facilitate subsequent analysis. Additionally, uncertainties are quantified for each latent variable, which can be of use when dealing with "borderline" datapoints.

To assess their performance in this context, the AANN and VAE are employed to transform the processed data from the experimental case study into a reduced dimensional representation. Armed with the transformed datasets, a simple probabilistic classification exercise is followed to determine how likely it was for a new data object to derive from a healthy operating condition. This procedure involves taking the transformed training sets and adopting empirical Gaussian distributions on the latent variables, such

that,

$$p(\mathbf{z}_n|\mathbf{x}_n) = \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \tag{8.11}$$

where $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ are the Gaussian mean and covariance based on the training points associated with a new ball bearing. In this case, these parameters are found such that the likelihood of the observations was maximised. Omitting the details of their derivation, the maximum likelihood estimates of $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ are given by the following equations [48],

$$\boldsymbol{\mu}_h = \frac{1}{N_h}\sum_{n=1}^{N_h}\mathbf{x}_n \tag{8.12}$$

$$\boldsymbol{\Sigma}_h = \frac{1}{(N_h-1)}\sum_{n=1}^{N_h}(\mathbf{x}_n - \boldsymbol{\mu}_h)(\mathbf{x}_n - \boldsymbol{\mu}_h)^\top \tag{8.13}$$

where the summations are only over the data instances from the normal condition. As an alternative, one could use a Bayesian approach by defining a prior density on these parameters, $p(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$, and then compute the posterior using Bayes' equation. This approach, however, is not adopted here as it was assumed that enough data objects were present in order to obtain sensible estimates from Equations (8.12) and (8.13).

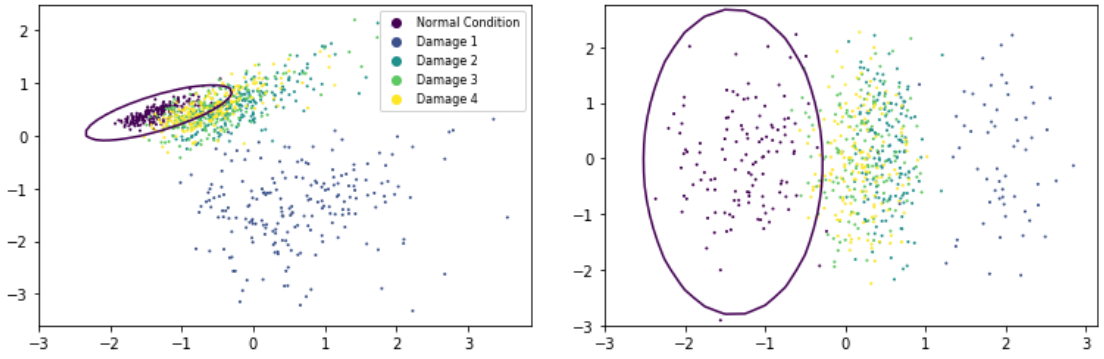
While multiple classes exist in the training set, the aim here is not to classify the type of damage. Instead, the problem is narrowed to the development of a simple novelty detector capable of distinguishing healthy condition features from any of the recorded damaged ones. With the estimated parameters of the Gaussian density, the negative log-likelihood (NLL) can then be calculated for every sample in the set.

Finally, it is necessary to establish a threshold from which an observation would be classed as abnormal. A 99^{th}-percentile threshold is thus chosen based on the NLL calculated from the normal condition objects associated with the training set. For a given new input x_n, indices captured above this threshold are flagged as novel, indicating that these may be treated as abnormalities deriving from a damaged state.

8.2.2 Results and Discussion

The results obtained given each respective transformed test set are illustrated by Figure 8.2. The 99^{th}-percentile thresholds corresponding to each model were also included and are depicted by the contour lines surrounding the normal condition data points. In all cases, most of the confusion can be noted to have occurred when presented with data objects corresponding to the bearing operating with either four loose elements (Damage Three) or wear (Damage Four). It could be that these types of damage introduce vibrations that are not as easily detectable as when the bearing exhibits a broken outer race

or a broken cage with loose elements. Consequently, the variability of the spectral lines may have not been as pronounced and thereby only slightly deviating from the normal condition spectrum. A more elaborate pre-processing stage could be required to further improve the current performance.



(a) (b)

FIGURE 8.2: Data objects projected onto: (a) AANN latent space and (b) VAE latent space.

Nevertheless, the area under the *Receiver Operating Characteristic* (ROC-AUC) curve [218] evaluated on each test set (Table 8.1) suggests that, in spite of the overlaps, good classification performances were achieved. A few interesting remarks can be made from these results. Indeed, the VAE was generally better at correctly distinguishing a normal operation from each of the damaged states. This outcome can be envisioned in Figure 8.2, where the transformation conducted by the VAE appears to yield greater separability between the formed clusters.

TABLE 8.1: Performance on test set - ROC-AUC results.

Anomaly	**AANN**	**VAE**
Damage 1	0.976	**0.992**
Damage 2	0.972	**0.992**
Damage 3	0.895	**0.921**
Damage 4	0.856	**0.921**

Perhaps of more importance, however, is to highlight the distinction observed in the shape of the clusters generated by the AANN and VAE. Compared to the former, the VAE enforces embeddings into sets of elliptical Gaussian distributions that seem to exhibit very little correlation between the latent features. This outcome may be explained by the inductive biases introduced from having a prior defined with a diagonal covariance matrix. While this assumption simplifies the derivation of the objective function (8.8), it should be noted that there is no guarantee for true generative distributions to align with it. In fact, this potential discrepancy is an important caveat that must be considered when implementing VAEs in the manner followed here. Despite this particular

shortcoming, however, the consequences of it seem to benefit the intended purpose of extracting damage-sensitive features in a space of reduced dimensionality. The transformed representation of the VAE may thus be better suited for the implementation of elaborate mixture models since the shape of distributions would be readily provided by this approach.

An important consideration to make in concluding this section is that the presented results may have been partly attributed to the nature of the data used for this demonstration. To further validate the insights made from this study, data from different case studies should be considered, in order to have their results compared for the assessment of a more stringent methodology. This direction is touched upon in the following section, where the data under examination are driven by underlying nonstationarities not present in this case study.

8.3 Experimental Case Study 2: Composite Plate Subjected to Damage Under Changing Environmental Conditions

The following case study refers to a comprehensive experimental procedure conducted during the Brite-Euram project DAMASCOS (BE97 4213). Specifically, the case study examined here corresponds to a Lamb-wave inspection of a 300mm $\times$ 300mm composite plate subjected to cyclic temperature variations in a controlled environmental chamber. The setup of this experiment consisted of having two identical piezoceramic disks bonded to the centre of opposite edges of the plate (i.e. 300mm apart). The disks were used to transmit and receive fundamental symmetric and antisymmetric Lamb waves. These waves were launched with a 5 cycle toneburst from a signal generator at 300kHz and 80kHz, respectively.

Overall, the experimental procedure comprised three main parts. In **Part One**, the ambient temperature was held constant at 25°C. In **Part Two**, the temperature was varied cyclically between 10°C to 30°C every 9h. Finally, in **Part Three**, the temperature kept varying with the addition of damage introduced to the plate, which was simulated by drilling a 10mm hole between the sensors. The conditions the plate was subjected to in the final part of the experiment aimed at simulating scenarios in which both benign environmental variations and damage manifest simultaneously in the features of the recorded Lamb waves.

In this case, the extracted features for analysis correspond to the first 50 spectral lines from the frequency spectrum of each time-domain recording. This series of spectral lines

can be observed in Figure 8.3, with the initiation of each experimental part denoted by the vertical lines at sample-point numbers 1355 and 2482, respectively.

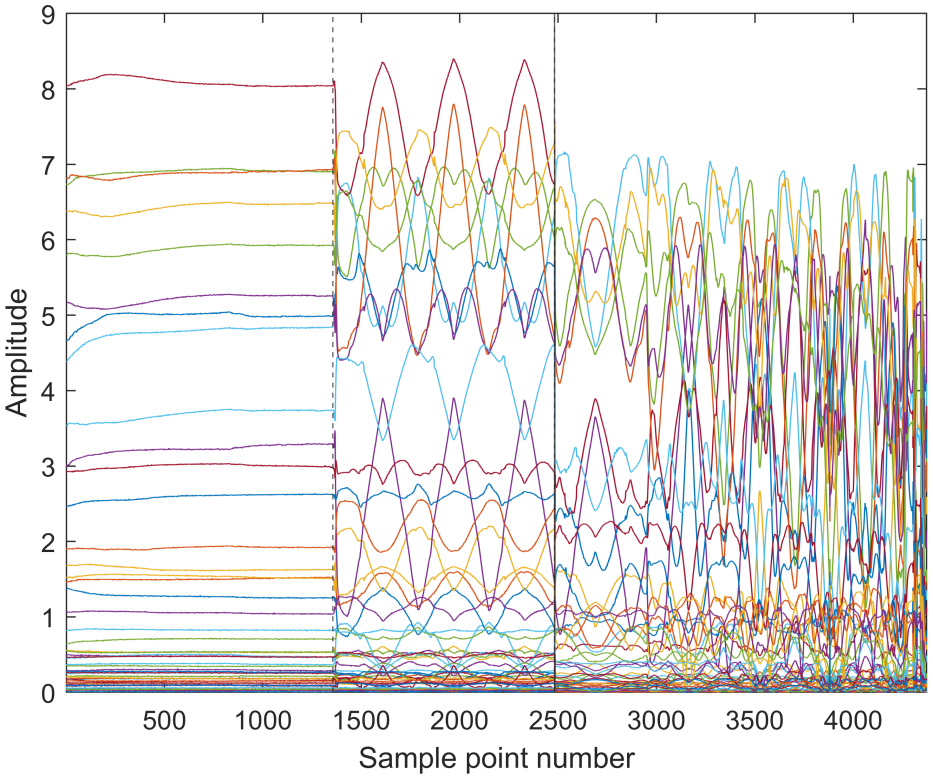


FIGURE 8.3: Time history of 50-dimensional feature over the whole experimental procedure. The dashed line indicates the introduction of the temperature variation phase and the bold line the introduction of damage.

A training set is defined by including every second datapoint from Parts One and Two of the experiment (i.e. taking half of the constant temperature and varying temperature datapoints as the normal condition), and the testing set to include the remaining datapoints accompanied by those from Part Three (i.e. the damaged condition with varying temperature).

8.3.1 Influence of Confounding Factors in SHM

While damage detection is a popular research topic, many proposed techniques in the literature are often tested in controlled environments that may not accurately replicate real-world operations. Part of the reason for this issue is, perhaps, because of the challenges in anticipating all potential environmental conditions a structure may encounter during its lifespan. Furthermore, replicating actual environmental conditions in a laboratory setting is typically expensive and difficult to achieve. Indeed, dealing with confounding effects presents a significant challenge in SHM. The goal is thus to design

monitoring systems capable of reliably distinguishing changes caused by damage from those from benign EoVs.

In relation to the current case study, a number of attempts have been made to address this issue. Manson et al. [219], for example, proposed focusing on the minor components of a PCA projection, as these were expected to be sensitive to damage while remaining unaffected by temperature variations. Although this approach showed some success, it was concluded that there was no compelling reason to believe the selected minor components would uniquely detect damage. An advancement in this study was introduced in [220], where *cointegration* was employed as a powerful technique to remove long-term trends from time-series data. This concept was further explored in [221], where a *discrete wavelet multiresolution decomposition* approach to cointegration was applied. An important conclusion made in this study is that the multiresolution decomposition alone may not effectively remove long-term trends, but could enhance sensitivity for damage detection when combined with cointegration. More intricate studies involving EoVs in SHM have also been explored in [222], where robust regression analysis was used to improve novelty detection.

One limitation shared across these methods is that they rely on linear relationships among the extracted features. The activation function employed in the VAE hidden layers may overcome this issue. Therefore, this section aims to investigate the capabilities of VAEs in identifying the underlying mechanisms driving the observed changes in the data, which, in this case, correspond to the temperature variations in the environmental chamber. This section, however, simply aims to demonstrate the effectiveness of the VAE in removing confounding effects from the presented data, in hopes that it may do the same for features with strong nonlinear relationships. This notion will be further pursued in the future.

A *data-normalisation* strategy [51] is developed with the VAE to project temperature variations out of the damage-sensitive features. While it is worth noting that data normalisation is a concept not tied to any particular technique, it should be acknowledged that, for the remainder of this chapter, the term "normalisation" will be informally employed to refer to the proposed approach using the VAE.

8.3.2 VAE for Data Normalisation and Damage Detection

The training set comprises dynamic features subjected to both constant and varying environmental conditions. It is worth noting that the data used to train the VAE exclusively correspond to the undamaged plate.

The hyperparameters of the VAE in this case study were optimised using the randomised *grid-search cross-validation* approach, resulting in a model with a nine-dimensional latent space, and *Scaled Exponential Linear Unit* (SELU) functions [223] employed as the nonlinear activations for the hidden layer.

Hence, the probabilistic encoder produces nine sets of parameters that define the approximate posterior $q_\phi(\mathbf{Z}|\mathbf{X})$. Similarly, $q_\phi(\mathbf{Z}|\mathbf{X})$ can be thought of as a decomposed collection of nine independent Gaussian distributions, each characterised by a mean and variance determined by the encoder when presented with data. The bottle-neck effect enforced by the VAE leads to a decomposed representation of the underlying mechanisms, where these individual Gaussian distributions capture different aspects of the data.

The key idea here is that some of these Gaussian distributions may exhibit varying statistics in response to temperature changes, while others remain somewhat consistent. Those that consistently maintain the same statistics across all data in the training set can be considered as representing effects uncorrelated to temperature variations. Any deviation observed in the statistics of these stationary distributions in the presence of damage can, therefore, be attributed to damage-sensitive features.

To put this notion into practice, the trained probabilistic encoder is used to identify non-stationary trends in the test dataset. These trends are then removed from the original data by setting their corresponding Gaussian distributions to zero-mean delta distributions. Subsequently, modified latent vectors are generated and passed through the probabilistic decoder to reconstruct a version of the original data devoid of temperature variations.

With a normalised dataset, a straightforward novelty detector can be used to identify damage. The *Mahalanobis Squared-Distance* (MSD) [224] is employed for this purpose, quantifying the discordancy of new observations while accounting for both environmental variations and potential damage. The MSD, D_ζ^2, is given by,

$$D_\zeta^2 = (x_\zeta - \boldsymbol{\mu}_h)^T \boldsymbol{\Sigma}_h^{-1} (x_\zeta - \boldsymbol{\mu}_h) \tag{8.14}$$

where x_ζ is the potential outlier, and $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ are now the sample mean and sample covariance of the current training set, respectively. The MSD is a metric that has been used recursively in SHM for damage detection; several examples of its implementation can be found in [1].

Having determined the estimates of the statistics using Equations (8.12) and (8.13), the MSD values of potential outliers can be computed and compared against a threshold

to determine their status. In this exercise, a threshold corresponding to the 97^{th} percentile is chosen, based on the MSD calculated from normal condition datapoints in the training set. For a given new input x_ζ, indices exceeding this threshold are flagged as abnormalities that could be indicative of a damaged state.

The importance of normalising data during the pre-processing stage is illustrated in the following subsection by comparing the outcome of computing the MSD with the original data against data that are transformed with the VAE.

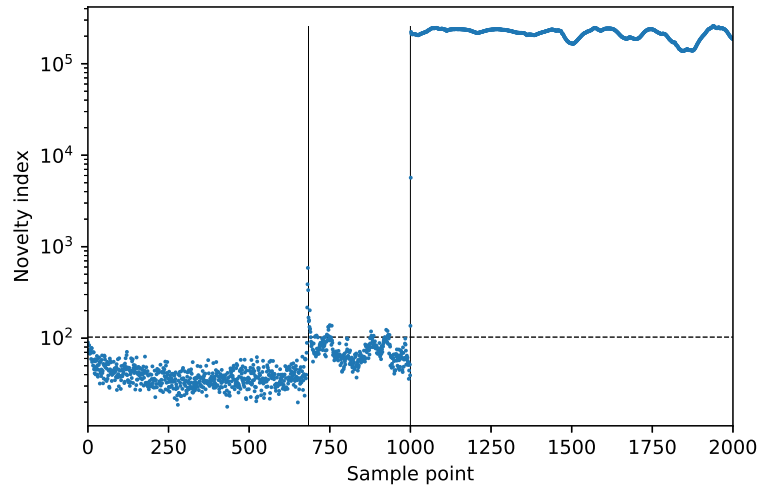
8.3.3 Results and Discussion

To begin, the MSD was first computed with the original data (unnormalised), using the training set to determine the corresponding empirical mean, $\boldsymbol{\mu}_h$, and covariance, $\boldsymbol{\Sigma}_h$. The result of this process is shown in Figure 8.4(a). The vertical lines in this figure, and subsequent ones, separate the three parts of the experimental procedure, while the horizontal dashed line indicates the threshold.

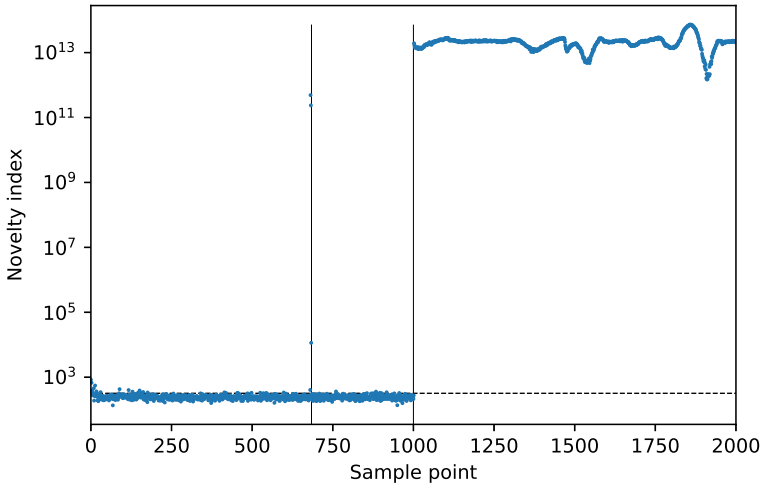
After drilling the hole (at sample-point 1000), the novelty detector can effectively identify damage by producing high novelty indices with respect to the threshold. However, many other indices in the second stage of the experiment are also flagged as anomalies, even when they correspond to an undamaged condition. Ideally, the trend that manifests in the second stage should be removed to mitigate false-positive results.

The results shown in Figure 8.4(a) demonstrate the risks associated with overlooking the influence of benign variations, such as temperature changes. In practical applications, using this outlier analysis for damage detection would be inappropriate, as it might lead to unnecessary maintenance, and shutting down the system when it may not be strictly required.

Some improvement is achieved by first normalising the data using the strategy described above. The results of the outlier analysis for the normalised data are shown in Figure 8.4(b). To illustrate the effects of data normalisation more clearly, detailed views of the results corresponding to the first two parts of the experiment are shown in Figures 8.5(a) and 8.5(b). The cyclic behaviour observed in the previous scenario appears to have been removed, resulting in novelty index measures that closely resemble those corresponding to the undamaged plate. A few exceptions can be observed at the time the plate began to experience temperature variations. These anomalous points may have been the result of some complex relationships existing between temperature variations and recorded signals at the time of switching between Parts 1 and 2 of the experimental test. Nonetheless,



(a)



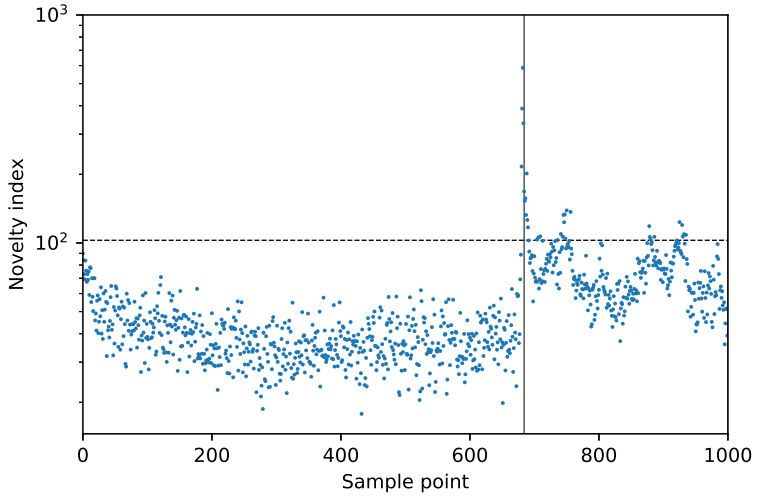
(b)

FIGURE 8.4: Outlier analysis results using (a) original features and (b) normalised features.

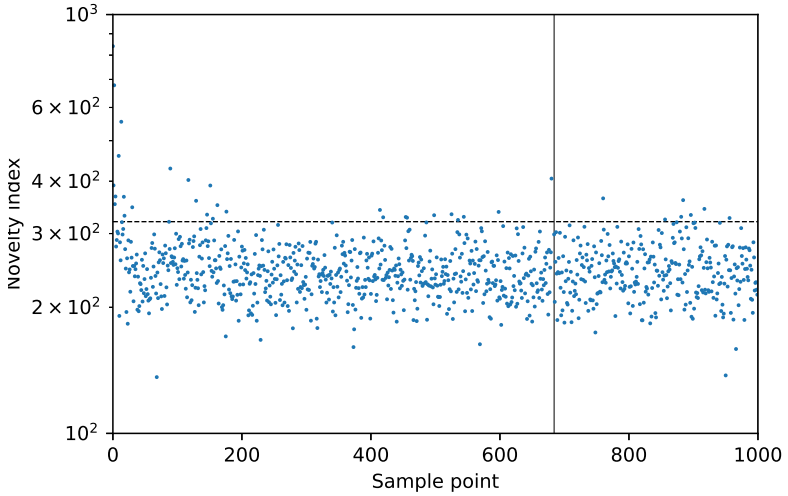
these points can be ignored since the features returned to equilibrium after the transition period.

A noteworthy observation arises when assessing the classification metrics after normalising the data. This observation is quantified by the confusion matrices presented in Figure 8.6, which show a slight increase in the false-positive rate following data normalisation. Based solely on these results, it may seem superfluous to conduct this type of normalisation in the first place. This argument is further supported when comparing the respective metrics provided in Table 8.2.

However, the value of this transformation becomes evident when noting that 100% of



(a)



(b)

FIGURE 8.5: Outlier analysis results of test set using: (a) original features and (b) normalised features.

TABLE 8.2: Classification metrics on test set.

Test Set	**F1 Score**	**ROC-AUC**
Unnormalised	0.985	0.984
Normalised	0.983	0.984

the false positives in the unnormalised case occur within Part Two of the experiment. In contrast, when normalising the data, the expected proportion of false positives – approximately 3% based on the established threshold –, is distributed across both Parts One and Two. This outcome demonstrates that the novelty detector is somewhat more robust to the effects of temperature variations when data is normalised with the VAE.

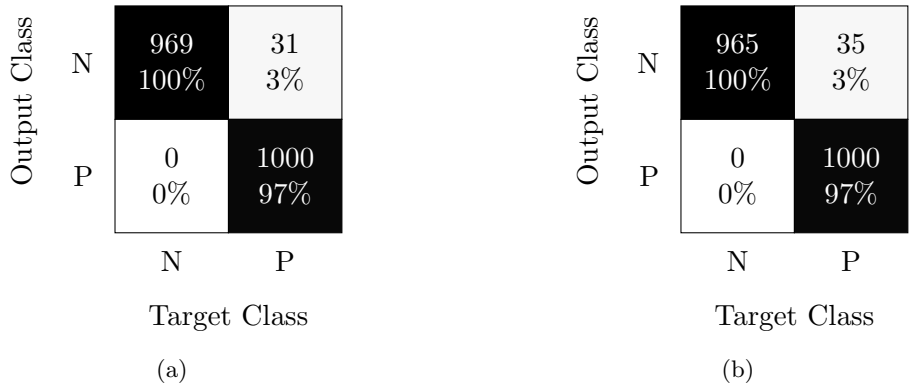


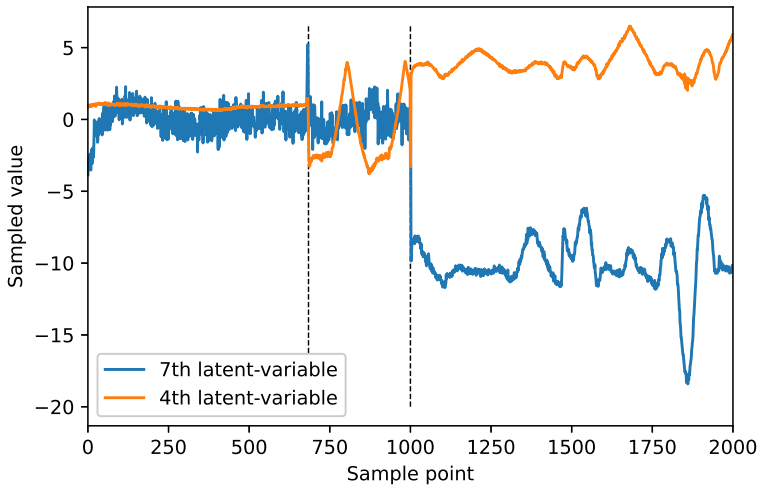
FIGURE 8.6: Confusion matrices for outlier detection using (a) unnormalised data and (b) normalised data.

Additionally, it may be important to highlight that the normalisation process was applied to all sample points in the test set, including those corresponding to the damaged plate. This observation is significant as it demonstrates that the health-state features in Part Three remained sufficiently distinctive for the novelty detector to effectively distinguish damage and normal conditions.

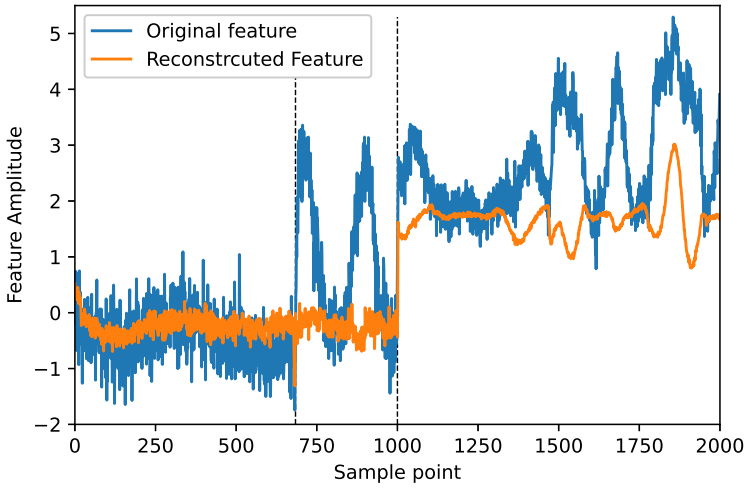
The identification of damage-sensitive features in the latent space can be better envisioned in Figure 8.7(a), where samples drawn from the inferred parameters of the Gaussian distributions, represented by the seventh and fourth latent variables, are shown. The fourth latent variable seems to have captured the unmeasured temperature embedded in the data, as the cyclic trend manifests in the results. Conversely, the seventh latent variable appears somewhat impervious to temperature changes, but sensitive to damage. By retaining the samples drawn from the more stationary Gaussian, and replacing the remaining ones with zero-mean delta distributions, the inputs are reconstructed from information exclusively regarding the health state of the plate.

The aftermath of this reconstruction is shown in Figure 8.7(b). Specifically, an arbitrary element in the reconstructions $\hat{\mathbf{x}}$ is shown alongside their original counterparts $\mathbf{x}$, at all instances in chronological order. Visually, the trend induced by the temperature variation appears to have been almost entirely removed from $\mathbf{x}$ in the second stage of the experiment.

A final remark worth making is that it may be unreasonable to expect the VAE to accurately reconstruct the original inputs in this framework. However, this consequence is the desired intent of the presented exercise. Instead of having features defined by temperatures, the VAE now reconstructs vectors without the influence of this benign effect. This notion is driven by the generative capabilities of VAEs, allowing sensible reconstructions even when manipulating the latent space in this manner. Nonetheless,



(a)



(b)

FIGURE 8.7: (a) Samples drawn from each realisation of the Gaussian parameters corresponding to the seventh and fourth latent variables. (b) The novelty index corresponds to an arbitrary feature before (blue) and after (orange) the normalisation procedure.

this strategy is certainly in need of a more rigorous and systematic approach. Potential directions to this end are detailed in the following section.

8.4 Conclusions

This chapter aimed to provide some insights into some of the challenges encountered in the implementation of novelty detectors for damage identification in structures. The

implementation of the VAE to address these issues was proposed in light of data derived from two comprehensive experimental case studies.

Indeed, the concepts covered in this chapter are not exhaustive, and further investigation is required to determine more assertive conclusions. Therefore, the following directions may be explored to ensure the reliability of VAE for preprocessing data in SHM:

1. Validating the latent representation of data with additional case studies could help determine whether well-defined Gaussian distributions do manifest as a result of establishing an isotropic Gaussian prior. This approach could be extended by implementing advanced clustering algorithms, and assessing whether VAE transformations do in fact enhance statistical models generated from the processed data.

2. Identifying the relevant distributions in the latent space was conducted visually, and there was no hard proof to support their elimination. A systematic and quantifiable way of identifying these distributions is thus required to avoid a subjective input in the procedure. A possible approach could be to employ a normality test, such as the Shapiro-Wilk test [225], to identify stationary Gaussian distributions in the latent space.

3. Another important shortcoming that might be worth pursuing is addressing the somewhat arbitrary step of replacing the time-varying embeddings with zero-mean delta distributions. A more reasonable manipulation of the latent space could, perhaps, exist for this type of application.

4. Finally, VAEs may not completely decorrelate time-varying parameters. Therefore, exploring other data-based models capable of disentangling correlations in a reduced latent space, such as β-VAEs [226] or info-GANs [227], could enhance the presented approach for addressing EoVs.

Chapter 9

Conclusions and Future Work

The surge of research involving the use of nonparametric Bayesian models for SHM reflects the growing interest in this area. By considering several challenging scenarios, this thesis involved exploring aspects of this modelling framework to address key issues encountered in SHM. In particular, special emphasis was given towards the selection of priors, and how the efficacy of the statistical model depends substantially on this important consideration. The advantages of employing nonparametric models in a Bayesian setting is another matter highlighted throughout this thesis. The outlined concepts were framed in the context of SHM, in an attempt to demonstrate how abstract concepts in machine learning can be naturally adapted to real-life problems.

In short, the primary highlights of this thesis are as follows. Firstly, the notion of "large" models was presented in relation to the Bayesian paradigm, elucidating what is meant by *complexity* in terms of modelling. The importance of this concept was argued to be fundamental in justifying the existence of nonparametric models, which were discussed to arise naturally in a Bayesian framework, regardless of the size and inherent complexity of the data. With the foundations set to support the use of nonparametric Bayesian models, the GP and DP were introduced as powerful models of this sort, and implemented to aid with the mapping of intricate relationships that may have been too cumbersome to determine otherwise. Furthermore, the added advantage of these models in providing quantified uncertainty alongside their predictions was exploited for a comprehensive analysis of the available datasets.

Following the conceptually-oriented chapters, the main contributions of this thesis were presented by translating the discussed methodologies into practice via a series of different case studies. The initial application involved using GPs to enhance localisation systems in both rotating machinery and composite structures. The insights gained from these case studies were then pursued further in subsequent chapters.

Notably, a significant limitation associated with AE-based techniques for damage identification was addressed with the proposal of an infinite mixture of Poisson distributions, which was only possible by incorporating a DP prior in the mixture. Additionally, this thesis addresses the design of a model for estimating driving forces applied to a journal bearing, leveraging GPs to predict these forces from journal displacement measurements.

In addition to these contributions, a small independent study was conducted to explore the use of VAEs for preprocessing data in SHM applications. This study was presented at the latest stage of the thesis, where the importance of the feature selection step in the organising principle for SHM was highlighted. Despite not being strictly defined as a nonparametric model, the VAE was argued to effectively provide transformed representations of data that could enhance the subsequent development of a statistical model.

A summary of the work presented in this thesis is provided in the following section, where the main conclusions drawn from each chapter are outlined in more detail.

9.1 Thesis Summary

The fundamental concepts of SHM were introduced first in Chapter 1, highlighting the necessity and challenges involved in its practical implementation. Building upon these challenges, the motivation to explore intelligent data-driven algorithms was conducted from a Bayesian perspective. Specifically, the intention of using nonparametric models within this framework was established, indicating the benefits they provide in addressing complex problems encountered in SHM.

Details of foundational concepts outlined in the introduction were elaborated in Chapter 2. A brief background on SHM was covered, emphasising advancements in sensing technologies for data acquisition, and outlined the systematic methodology necessary for the success of SHM. Within this framework, special attention was given to the development of statistical models, illustrating common practices in machine learning. Having set the stage, the role corresponding to nonparametric models could be defined, where salient aspects were provided with intuitive demonstrations. Furthermore, Bayesian reasoning was defined mathematically via Bayes' Theorem.

In Chapter 3 a comprehensive review of related work in the field was presented, with a particular focus on GPs in SHM. While the referred literature cannot be said to be exhaustive, it was extensive enough to demonstrate the ingenuity and extent to which GPs have been employed in health-monitoring systems. In contrast, the use of DPs in SHM was less prevalent, indicating a lack of rich literature in this area.

In Chapter 4, the GP and DP were scrutinised, offering an engineering perspective on these nonparametric Bayesian models. Considerations for practical implementation were highlighted, especially for DPs, where relevant metaphors were described in line with mathematical definitions to provide insight into their operation.

The contents in Chapter 5 focused on the use of GPs for developing localisation-based models. Two experimental case studies were presented, each with a unique approach to implementing GPs for localisation. This distinction was dictated in terms of the sensing technologies employed and the mechanisms involved. During the analysis of the first case study, it was argued that a journal bearing poses challenges that cannot be easily addressed with vibration-based techniques. Therefore, the proposed methodology for an effective monitoring system involved employing a GP to learn the mapping between operational parameters and shaft-centre location. Anomalous behaviour could then be identified by monitoring departures of the journal from the predicted equilibrium position. Conversely, in the second study, a more direct approach for localising damage was pursued; in particular, using AE waves to pinpoint the source of damage in composite blades. The GP was used, in this case, to learn the complex relationship between the source location of an AE wave and the resulting differences in time of arrival. An advancement of this method was then proposed in an attempt to minimise the required data for training. By simulating a scenario limited to 100 measurements, the proposed method yielded more accurate estimations when compared to relying on measurements made at random locations. It was concluded that this method may facilitate the implementation of localisation technologies for SHM.

The uses of AE-based techniques for damage detection were then investigated in Chapter 6. Here, the shortcomings of gathering AE data were discussed, highlighting the complications involved in their storing and manipulation. Common practices in the field were outlined to motivate the problem at hand; that is, to find an effective way to identify and isolate AE events in real-time recordings. The key to addressing this problem translated into answering the following question: *What is the probability of an AE wave existing in any given section of the signal?* Given the nature of the features being analysed, the Poisson distribution was chosen as the building block of the statistical model, designed precisely to provide an explicit answer to this question. In order to account for the variety of AE waves that could emerge in a time-series signal, it was assumed that an inherent correlation exists between the generating mechanisms and corresponding features of the AE waves. By having a unique Poisson distribution representing each group of generated AE waves – or underlying mechanisms –, a mixture model could be constructed to identify, in a probabilistic sense, distinct "levels" of acoustic activity in the signal. The implications of this method were illustrated in the results, where sources of background noise were autonomously distinguished from

actual AE waves. Additionally, the identified waves were further separated concerning their features, distinguishing the most imposing waves from the rest. Deciding on the number of Poisson distributions to employ was eliminated by having a DP prior in the mixture model, granting enough flexibility for the model to infer this parameter directly from the data. This chapter concluded with the implementation of the DP-PMM for the early identification of damage in an Airbus A320 landing gear. The results revealed that the DP-PMM was successful at identifying the onset of fracture during the fatigue test in advance of the scheduled visual inspections.

In Chapter 7, vibration problems involving rotating machinery were explored. The journal bearing was once again recalled for the sake of the established challenge. Here, the compelling aspects of journal-bearing dynamics were discussed, and how this fundamental component in turbo-machinery systems has motivated several studies on the subject. The exercises conducted involved the examination of a series of numerical case studies; each designed to simulate different conditions a journal bearing might experience in practical applications. Given the growing interest in system identification for rotors, the use of the *Gaussian Process Latent Force Model*(GP-LFM) was explored for this purpose. The advantages supporting this approach were provided, with special emphasis given to the fact that physical knowledge can be incorporated into this framework to improve the extrapolating estimates of an algorithm that is purely reliant on data. Having defined the set of linearised equations governing the response of a journal bearing, the GP-LFM was employed with simulated measurements to estimate (latent) driving forces. Furthermore, an exercise was conducted involving the joint estimation of the inputs, states and dynamic coefficients of the bearing. While the outcome in all cases proved satisfactory under the right conditions, the conclusions drawn from these exploratory studies remained open-ended. However, potential courses of action were outlined in light of the obtained results. These future directions, alongside those corresponding to the previous chapters, are discussed thoroughly in the following section.

Finally, in Chapter 8, an independent investigation into advanced data-driven techniques for SHM was explored. A VAE was used to extract damage-sensitive features from data in two experimental case studies. The first case study involved analysing vibration signals measured from a rolling-element bearing in operation. By using the VAE to project the data into a lower-dimensional space, it was shown that the resulting latent features clustered into well-defined Gaussian distributions, an outcome that could be promising in enhancing the performance of an anomaly detector. Thereafter, the second case study was outlined, referring to a Lamb-wave inspection of a composite plate subjected to cyclic environmental variations. The VAE was used to remove the (benign) effects of temperature on damage-sensitive features, which, when disregarded, mislead an anomaly detector despite no damage being present.

9.2 Limitations and Future Work

Placing complete reliance on intelligent systems to assess the health of a structure is, understandably, met with reluctance. Erroneous predictions are simply prohibited in this context; primarily, because of the risk this dependency poses to human safety. The current state of SHM still carries the risk of overlooking the susceptibility of structures to experience unprecedented events, remaining an issue that is yet to be addressed extensively.

While the methods presented in this thesis were intended as small contributions towards the overarching goal of breaching this gap, there are, of course, several associated limitations that must be acknowledged. Arguably, the most prominent limitation shared across all methods proposed in this work is that of establishing an informed prior distribution. Indeed, selecting an appropriate prior took precedence over other aspects regarding nonparametric Bayesian models. Despite these efforts, however, it should be noted that there is no standardised method for defining informative priors in engineering applications. This unfortunate reality cannot be disregarded in Bayesian modelling. Ultimately, the elicitation of a suitable prior is application-dependent, requiring the recognition of the involved physics, which is often an assiduous task.

Another issue that may have not been explicitly addressed here is attributed to the high computational demands of the data-driven approaches used in this thesis. While this is often one of the main limitations of nonparametric Bayesian models, it may be safe to state that this particular problem is not exclusive to the work presented in this thesis. Nevertheless, substantial research in the fields of machine learning and computer science strives to provide elegant solutions to algorithms for high-order computations. Exploring some of these solutions may be prudent to ensure the success of the presented methods in real-life scenarios.

A significant portion of this thesis focused on AE-based techniques. Chapters 5 and 6 delve into the limitations of adopting this approach for SHM. While the proposed methods aimed to facilitate the practical implementation of AE-based monitoring, demonstrating promising capabilities, it is evident that these methodologies still require further refinement.

For example, as discussed in Chapter 5, it was concluded that pencil-lead breaks may not be the most effective method to construct ΔT maps. This recurring issue in AE-based methods is arguably one of the hardest challenges currently hindering the use of statistical models for damage localisation. The exploration of novel techniques for simulating convincing AE waves remains an ongoing challenge, whereby the creation of a pragmatic, cost-effective, and readily available technique for generating artificial

AE waves could significantly bolster the effectiveness of data-driven approaches in this context.

Even when simplifying the problem to focus solely on damage detection, the challenges associated with AE-based monitoring, unfortunately, remain prevalent enough to dissuade one from choosing this type of monitoring approach. A noteworthy issue, as discussed in Chapter 6, was in relation to the high sampling frequencies required when gathering AE data. In an effort to facilitate the analysis of very rich data, a powerful nonparametric Bayesian model was employed. However, attempting to overcome this issue introduced several others, including the high computational costs associated with inferring a model driven by a DP.

Since the premise of this approach is based on learning in an online setting, fast computations may be essential to detect the early onset of fracture promptly. Therefore, addressing this computational efficiency concern may arguably be the most crucial aspect to address in future research efforts.

Considering the practical nature of these methods, assessing their robustness and effectiveness may best be achieved via their implementation in numerous real-world case studies. Ideally, these case studies ought to involve structures (or machines) operating outside controlled experimental setups; that is, while experiencing naturally occurring events *in situ*, which may be nearly impossible to reproduce artificially. Certainly, this form of validation is required for all methods outlined in this thesis, but there were two particular occasions where comprehensive validation was particularly required.

The first instance relates to the detection of oil starvation in journal bearings, as presented in Chapter 5. Here, the notion of using the GP likelihood maps for detecting a simulated anomalous operation might be harder to demonstrate in practice. Numerous unaccounted factors may impose additional complexities, warranting further improvements to the proposed method, which may have been overlooked in the concluding section of this chapter.

Similarly, the second instance pertains to the methods examined in Chapter 7, whereby the applicability of the GP-LFM was assessed merely with numerically-simulated case studies. The complex nature of external loads exerted on a journal bearing may have been oversimplified in the design of these case studies. Therefore, a more stringent validation is essential to truly determine the effectiveness of the GP-LFM in estimating relevant force histories.

On both occasions, the design, construction and execution of comprehensive experimental procedures were beyond the available resources during the development of the

corresponding statistical models. Overcoming this limitation is thus a direction that should be pursued in future research efforts.

Finally, it is worth noting that there is an opportunity to explore other families of nonparametric Bayesian models beyond GPs and DPs for the development of enhanced health-monitoring systems. In Chapter 3, it was discussed that GPs currently dominate as the preferred nonparametric Bayesian model for regression problems in SHM. While some research has explored salient aspects of DPs for damage detection, the literature on this subject remains limited compared to those related to GPs.

The theoretical foundation of DPs outlined in Chapters 4, and their application demonstrated in Chapter 6, aimed to promote the use of DPs in SHM applications and demonstrate their adaptability in density-estimation problems. Both the GP and DP are, indeed, characterised by their flexibility, but they operate in significantly different ways, offering advantages in various areas of SHM. The same principle may hold for other nonparametric Bayesian models, which could potentially offer solutions to some of the most challenging problems in SHM.

For example, the *Beta Process* [228] could be employed as a prior in clustering problems, particularly when dependencies between mixing coefficients are a limiting constraint. Another direction worth exploring is the *Hierarchical Dirichlet Process* (HDP) [229], which may be desired in cases where groups consist of clusters while allowing for the sharing of clusters between groups.

It should be noted, however, that the choice of a nonparametric Bayesian model should be based on its suitability for the specific problem at hand. Identifying the most appropriate model for a given problem is the first step in deciding which family of models to employ.

9.3 Concluding Remarks

The limitations discussed throughout this thesis seem to indicate that the current state of statistical modelling for SHM has a way to go before becoming standard practice in real-life applications. However, optimism is rather found in taking these limitations as opportunities to learn. Recent advances in the literature already offer promising solutions to address many of the most challenging issues encountered in practical SHM.

It can be argued that the ultimate goal here is to make SHM a standard practice in the near future. With the growing availability of data from comprehensive monitoring campaigns, the development of powerful statistical models may no longer be confined to

experimental case studies. This progress in statistical modelling holds the potential to greatly enhance the safety and efficiency of various structures and systems.

The work presented in this thesis represents a contribution towards the broader aim of advancing SHM. It is hoped that future developments will further enhance the methodologies outlined in the preceding chapters and that their practical utility becomes increasingly evident in real-world applications.

Appendix A

Supplementary Background for GPs and DPs

A.1 Bayesian Linear Regression

The standard linear regression model can be defined as follows,

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} \tag{A.1a}$$

$$y = f(\mathbf{x}) + \epsilon \tag{A.1b}$$

where $\mathbf{x} \in \mathbb{R}^D$ is an input vector, $\mathbf{w} \in \mathbb{R}^D$ is a vector of weights, $f(\cdot) \in \mathbb{R}$ is the function value, and $y \in \mathbb{R}$ denotes the noisy function value (target). The additive Gaussian noise term ϵ, follows a normal distribution $\mathcal{N}(0, \sigma^2)$.

With the inclusion of $\epsilon \sim \mathcal{N}(0, \sigma^2)$ in the predictions, a probabilistic framework is established, and the likelihood can be expressed as the probability density of the observations conditioned on the parameters. Concretely,

$$p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2) \tag{A.2}$$

If N number of independent observations are considered, then this expression can be rewritten as,

$$p(y|\mathbf{x}, \mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n^\top \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{X}^\top \mathbf{w}, \sigma^2 \mathbb{I}) \tag{A.3}$$

where $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \in \mathbb{R}^{D \times N}$ is a matrix in which the column inputs are aggregated, and $\mathbf{y} = \{y_1, \ldots, y_N\}^\top \in \mathbb{R}^N$ is a vector with its entries corresponding to the respective output targets. The probability of all observations can be factorised in this way because

these are assumed to be conditionally independent with respect to $\mathbf{w}$. In other words, while conditioned on some underlying correlation, the outcome of an observation will not have an effect on those remaining.

The model is made Bayesian by introducing a prior distribution over the parameters $\mathbf{w}$. For simplicity, the parameters are assumed to be Gaussian distributed with zero mean and covariance Σ_p. That is,

$$p(\mathbf{w}) = \mathcal{N}(0, \Sigma_p) \tag{A.4}$$

Although this assumption serves as more of a mathematical convenience, as it allows for a closed-form solution, it is still a sensible choice for the plausible range of values that $\mathbf{w}$ can take. By applying Bayes' Theorem (2.1), the posterior distribution can be calculated as follows,

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y})} \propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}^\top \mathbf{w})^\top (\mathbf{y} - \mathbf{X}^\top \mathbf{w})\right) \exp\left(-\frac{1}{2}\mathbf{w}\Sigma_p^{-1}\mathbf{w}\right) \tag{A.5}$$

and by *completing the square* of this expression, the posterior is found to be a Gaussian distribution,

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \mathcal{N}\left(\frac{1}{\sigma^2}A^{-1}\mathbf{X}\mathbf{y}, A^{-1}\right) \tag{A.6}$$

where $A = \sigma^{-2}\mathbf{X}\mathbf{X}^\top + \Sigma_p^{-1}$. The posterior distribution represents an updated belief that quantifies how much the prior information is supported by the observations. Unlike deterministic methods, the solution to $\mathbf{w}$ is no longer unique. In other words, the posterior distribution indicates that a unique value for $\mathbf{w}$ cannot be determined with absolute certainty. This concept is fundamental when aiming to assess the confidence of a model in making predictions.

The primary goal of the model is to predict the outcome of new, unseen observations, and thus the concern is not on finding the specific values of the parameters $\mathbf{w}$. Instead, the posterior distribution is used to evaluate the *predictive distribution*, which involves marginalising the parameters out of the likelihood. The predictive distribution represents the uncertainty in making predictions for new observations. By integrating over all possible values of $\mathbf{w}$, weighted by their corresponding posterior probabilities, the predictive distribution is returned. This process allows one to account for the uncertainty associated with the model parameters and obtain a more comprehensive representation of the predictions.

By defining $\mathbf{X}$ as the collection of training inputs, and $\mathbf{y}$ as the vector that collects all training targets, the predictive distribution for an arbitrary test input $\mathbf{x}_* \in \mathbb{R}^D$ – given

the training data – is computed as follows,

$$p(f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \int p(f_*|\mathbf{x}_*, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{y}) \mathrm{d}\mathbf{w} = \mathcal{N}\left(\frac{1}{\sigma^2}\mathbf{x}_*^\top A^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}_*^\top A^{-1}\mathbf{x}_*\right) \tag{A.7}$$

where the subscript $*$ denotes a new observation from a test set. Since the likelihood and posterior are both Gaussian, the predictive distribution is also a Gaussian with mean $\sigma^{-2}\mathbf{x}_*^\top A^{-1}\mathbf{X}\mathbf{y}$ and covariance $\mathbf{x}_*^\top A^{-1}\mathbf{x}_*$.

In the generalised case defined by (4.3), the predictive distribution becomes,

$$f_*|\mathbf{x}_*, \mathbf{X}, \mathbf{f} \sim \mathcal{N}\left(\frac{1}{\sigma^2}\phi(\mathbf{x}_*)^\top A^{-1}\Phi(\mathbf{X})\mathbf{y}, \phi(\mathbf{x}_*)^\top A^{-1}\phi(\mathbf{x}_*)\right) \tag{A.8}$$

where $\Phi(\mathbf{X}) \in \mathbb{R}^{M\times N}$ is the *design matrix* comprised of the aggregation of columns $\phi(\mathbf{x}_n)$, and $A = \sigma^{-2}\Phi(\mathbf{X})\Phi(\mathbf{X})^\top + \Sigma_p^{-1}$.

A.2 Beta Distribution

The assignment of a given observation can be represented by the following distribution,

$$p(\mathbf{y}|\boldsymbol{\rho}) = \prod_{k=1}^{K} \rho_k^{y_k} \tag{A.9}$$

where $\mathbf{y} \in \mathbb{R}^K$ is a K-dimensional vector in which one of its elements y_k equals one, while the remaining elements are equal to zero, with $\rho \in \mathbb{R}^K$ also being a K-dimensional vector with elements ρ_k constrained to satisfy $\rho_k \geq 0$ and $\sum \rho_k = 1$. The expression takes the following form when accounting for a dataset of N independent observations,

$$p(\mathbf{y}|\boldsymbol{\rho}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \rho_k^{y_{nk}} = \prod_{k=1}^{K} \rho_k^{\sum y_{nk}} \tag{A.10}$$

In the special case of a bivariate random variable $\mathbf{y} = \{0, 1\}$. The Bayesian treatment to this model is implemented by assuming ρ to be distributed according to a *Beta* distribution, (hyper)parameterised by the values α and β,

$$\rho \sim \mathrm{Beta}(\alpha, \beta) \tag{A.11}$$

$$p(\rho|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\rho^{\alpha-1}(1-\rho)^{\beta-1} \tag{A.12}$$

where $\Gamma(.)$ is the *Gamma function*. A single element ρ is only needed as the probability of the remaining category can be easily computed as $1 - \rho$, given that $0 \leq \rho \leq 1$.

The Beta distribution has useful analytical properties, and it is easy to interpret for this type of modelling. Some examples of the Beta distribution, evaluated with different valued combinations of α and β, are shown in Figure A.1. The functional form of the PDFs is controlled by these parameters, and ρ is evaluated over the continuous space of real values between zero and one. Because of this bound, the random variable ρ can be interpreted as the probability of a binary random variable taking one of the two possible outcomes.

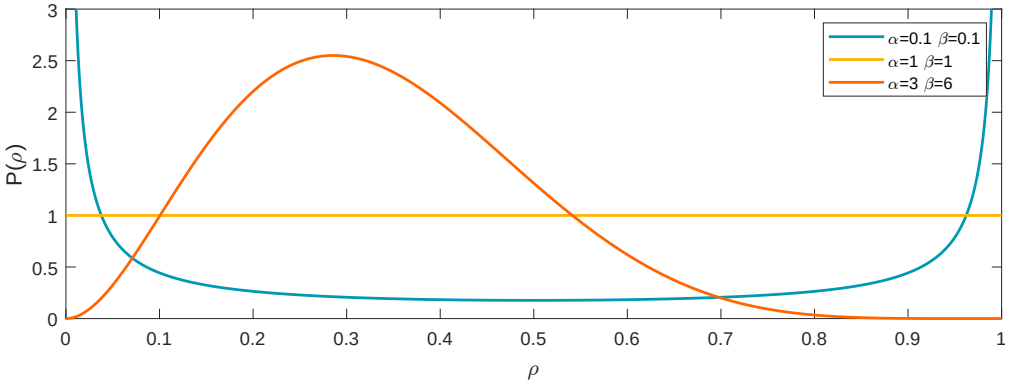


FIGURE A.1: Plots of the Beta distribution given by (A.12), as a function of ρ for different values of the hyperparameters α and β.

Now, if chosen as a prior, the Beta distribution is conjugate to the categorical distribution (A.10), yielding a closed-form solution for the posterior. When taking $\mathbf{y}$ to be a bivariate random variable, the categorical distribution returns a Bernoulli-likelihood of the form,

$$p(\mathbf{y}|\rho) = \prod_{n=1}^{N} \rho^{y_n}(1-\rho)^{1-y_n} = \rho^{\sum y_n}(1-\rho)^{N-\sum y_n} \tag{A.13}$$

The posterior can be easily determined by inspection of the product of the likelihood and prior,

$$\begin{aligned} p(\rho|\mathbf{y}) &\propto \rho^{\sum y_n}(1-\rho)^{N-\sum y_n} \times \rho^{\alpha-1}(1-\rho)^{\beta-1} \\ &\propto \rho^{(\sum y_n+\alpha)-1}(1-\rho)^{(N-\sum y_n+\beta)-1} \end{aligned}$$

Hence,

$$p(\rho|\mathbf{y}) = \text{Beta}(\gamma, \delta) \tag{A.14}$$

where

$$\gamma = \sum y_n + \alpha, \quad \delta = \left(N - \sum y_n\right) + \beta$$

Predicting the assignment of new observations is possible by evaluating the posterior predictive distribution, which in this case, can be determined by marginalising ρ. That

is,

$$\begin{aligned} p(y_* = 1|\mathbf{y}) &= \int p(y_*|\rho)p(\rho|\mathbf{y})\mathrm{d}\rho \\ &= \int \rho^{y_*}(1-\rho)^{1-y_*} \times \frac{\Gamma(\gamma+\delta)}{\Gamma(\gamma)\Gamma(\delta)}\rho^{\gamma-1}(1-\rho)^{\delta-1}\mathrm{d}\rho \\ &= \frac{\gamma}{\gamma+\delta} = \frac{\sum y_n + \alpha}{N + (\alpha+\beta)} \end{aligned} \tag{A.15}$$

which is simply the fraction of the number of times $y_n = 1$ over the total number of trials, modulated a little by the prior parameters α and β.

A.3 Dirichlet Distribution

The multivariate generalisation of the Beta distribution is the *Dirichlet* distribution. A draw from a Dirichlet distribution is thus a vector of K-random variables that add up to one. That is, $0 \leq \rho_k \leq 1$ and $\sum \rho_k = 1$. The Dirichlet distribution is given by,

$$\mathrm{Dir}(\boldsymbol{\rho}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_o)}{\Gamma(\alpha_1),\ldots,\Gamma(\alpha_k)} \prod_{k=1}^{K} \rho_k^{\alpha_k - 1} \tag{A.16}$$

where the vector $\boldsymbol{\alpha} \in \mathbb{R}^K$ controls the shape of the PDF over a confined simplex, and $\alpha_o = \sum \alpha_k$.

Some examples of Dirichlet distributions with $K = 3$ and different values of α are shown in Figure A.2. Although a three-dimensional case is shown here, the Dirichlet distribution can account for any finite number of random variables.

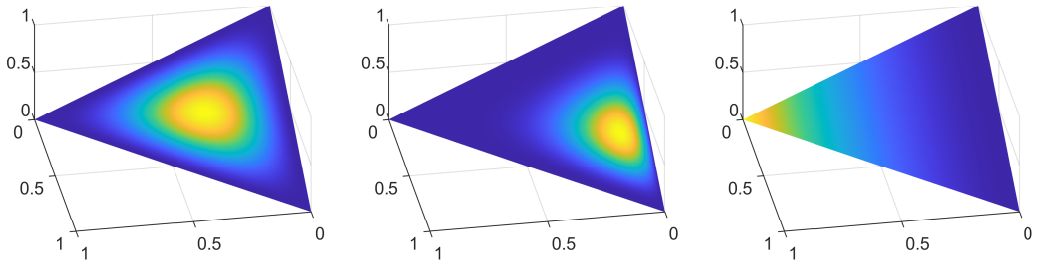


FIGURE A.2: Plots of Dirichlet distributions over a confined simplex, as given by (A.16) as a function of $\boldsymbol{\rho}$ for various values of $\boldsymbol{\alpha}$: (a) $\boldsymbol{\alpha} = [3, 3, 3]$, (b) $\boldsymbol{\alpha} = [2, 6, 4]$, and (c) $\boldsymbol{\alpha} = [3, 1, 1]$,

Having the categorical distribution represent a K-dimensional random variable, the predictive distribution in this case can be calculated in the same way as for the bivariate

case. The posterior can be first defined as follows,

$$\begin{aligned} p(\boldsymbol{\rho}|\mathbf{y}) &\propto \prod_{n=1}^{N}\prod_{k=1}^{K}\rho_k^{y_{nk}} \times \prod_{k=1}^{K}\rho_k^{\alpha_k-1} \\ &\propto \prod_{k=1}^{K}\rho_k^{(\sum y_{nk}+\alpha_k)-1} \end{aligned}$$

By inspection, the posterior is found to be a Dirichlet distribution parameterised by $\alpha_k^* = \sum y_{nk} + \alpha_k$, where $\sum y_{nk}$ denotes the number of elements assigned to the k^{th} category. Therefore,

$$p(\boldsymbol{\rho}|\mathbf{y}) = \text{Dir}\left(\alpha_1 + \sum y_{n1}, \ldots, \alpha_k + \sum y_{nK}\right) \tag{A.17}$$

The predictive distribution is then given by,

$$\begin{aligned} p(y_* = k|\mathbf{y}) &= \int p(y_*|\boldsymbol{\rho})p(\boldsymbol{\rho}|\mathbf{y})\mathrm{d}\boldsymbol{\rho} \\ &= \int \prod_{k=1}^{K}\rho_k^{y_{*k}} \times \frac{\Gamma(\alpha_o^*)}{\Gamma(\alpha_1^*),\ldots,\Gamma(\alpha_K^*)}\prod_{k=1}^{K}\rho_k^{\sum(y_{nk}+\alpha_k)-1}\mathrm{d}\boldsymbol{\rho} \\ &= \frac{\alpha_k^*}{\alpha_o^*} = \frac{\sum y_{nk}+\alpha_k}{\sum \alpha_k^*} \end{aligned} \tag{A.18}$$

The probability of assigning a new observation to the k^{th} group is the proportion of previous observations already assigned to the k^{th} group over the total number of observations (excluding y_*), modulated by the prior parameters.

Appendix B

Infinite Mixture of Poisson Distributions

B.1 Derivation of $p(z_{nk} = 1|\boldsymbol{Z}^{-n}, X^{-n}, \boldsymbol{\alpha}, a, b)$

Given the joint distribution defined by equation (6.10), it is possible to derive the conditional distributions over $\boldsymbol{z}_n$, $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$. A conventional Gibbs sampler could then be employed to infer the joint distribution over these parameters. However, in this case, the probability distribution of interest is that defined over $\boldsymbol{z}_n$, conditioned on all other parameters. The motivation to derive an expression for $p(\boldsymbol{z}_n|...)$ begins by marginalising $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$ out of equation (6.10). If $\boldsymbol{\pi}$ is first considered, then one can approach the marginalisation process in a two-step decomposition, where sampling from $p(\boldsymbol{z}_n, \boldsymbol{\pi})$ is said to be equivalent to drawing a sample of $\boldsymbol{\pi}$ first, and then using this sample to draw $\boldsymbol{z}_n$. That is,

$$\begin{aligned} p(\boldsymbol{z}_n, \boldsymbol{\pi}|\boldsymbol{Z}^{-n}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, x_n) &\propto p(\boldsymbol{z}_n|\boldsymbol{\pi}, x_n, \boldsymbol{\lambda}) p(\boldsymbol{\pi}|\boldsymbol{Z}^{-n}, \boldsymbol{\alpha}) \\ &\propto \prod_{k=1}^{K} \left[\pi_k p(x_n|\lambda_k)\right]^{z_{nk}} p(\boldsymbol{\pi}|\boldsymbol{Z}^{-n}, \boldsymbol{\alpha}) \end{aligned}$$

Here, $\boldsymbol{Z}^{-n}$ is the set of all assignments except the n^{th} one. The notation of this expression can then be simplified by instead considering the value where the k^{th} element of z_n is equal to unity,

$$\begin{aligned} p(z_{nk} = 1, \boldsymbol{\pi}|\boldsymbol{Z}^{-n}, \boldsymbol{\lambda}, \boldsymbol{\alpha}, x_n) &\propto \pi_k p(x_n|\lambda_k) p(\boldsymbol{\pi}|\boldsymbol{Z}^{-n}, \boldsymbol{\alpha}) \\ &\propto p(z_{nk} = 1|\boldsymbol{\pi}) p(x_n|\lambda_k) p(\boldsymbol{\pi}|\boldsymbol{Z}^{-n}, \boldsymbol{\alpha}) \end{aligned}$$

In this form, it becomes possible to marginalise $\boldsymbol{\pi}$ out,

$$\begin{aligned} p(z_{nk}=1|\boldsymbol{Z}^{-n},\boldsymbol{\lambda},\boldsymbol{\alpha},x_n) &\propto \int p(z_{nk}=1|\boldsymbol{\pi})p(x_n|\lambda_k)p(\boldsymbol{\pi}|\boldsymbol{Z}^{-n},\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\pi} \\ &\propto p(x_n|\lambda_k)\int p(z_{nk}=1|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{Z}^{-n},\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\pi} \\ &\propto p(x_n|\lambda_k)p(z_{nk}=1|\boldsymbol{Z}^{-n},\boldsymbol{\alpha}) \end{aligned}$$

where,

$$p(z_{nk}=1|\boldsymbol{Z}^{-n},\boldsymbol{\alpha}) = \int p(z_{nk}=1|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{Z}^{-n},\boldsymbol{\alpha})\,\mathrm{d}\boldsymbol{\pi}$$

and the posterior, $p(\pi|\boldsymbol{Z}^{-n},\boldsymbol{\alpha})$, is the result of having a Dirichlet prior, $p(\boldsymbol{\pi}|\boldsymbol{\alpha})$, as the conjugate of the multinomial likelihood, $p(\boldsymbol{Z}^{-n}|\boldsymbol{\pi})$. Concretely,

$$p(\boldsymbol{\pi}|\boldsymbol{Z}^{-n},\boldsymbol{\alpha}) \propto \prod_{k=1}^{K}\pi_k^{\beta_k-1}, \quad \beta_k = \alpha_k + c_k^{-n}$$

where c_k^{-n} is the number of objects assigned to the k^{th} group, excluding the assignment on x_n. Hence,

$$\begin{aligned} p(z_{nk}=1|\boldsymbol{Z}^{-n},\boldsymbol{\alpha}) &= \frac{\Gamma(\sum_{j=1}^{K}\beta_j)}{\prod_{j=1}^{K}\Gamma(\beta_j)}\int \pi_k \prod_{j=1}^{K}\pi_j^{\beta_j-1}\,\mathrm{d}\boldsymbol{\pi} \\ &= \frac{\Gamma(\sum_{j=1}^{K}\beta_j)}{\prod_{j=1}^{K}\Gamma(\beta_j)}\int \prod_{j=1}^{K}\pi_j^{\beta_j+\delta_{jk}-1}\,\mathrm{d}\boldsymbol{\pi} \\ &= \frac{\Gamma(\sum_{j=1}^{K}\beta_j)}{\prod_{j=1}^{K}\Gamma(\beta_j)}\frac{\prod_{j=1}^{K}\Gamma(\beta_j+\delta_{jk})}{\Gamma(\sum_{j=1}^{K}(\beta_j+\delta_{jk}))} \end{aligned}$$

where $\delta_{jk}=1$ when $j=k$, and zero otherwise. Since $\sum_{j=1}^{K}\delta_{jk}=1$,

$$\begin{aligned} p(z_{nk}=1|\boldsymbol{Z}^{-n},\boldsymbol{\alpha}) &= \frac{\Gamma(\sum_{j=1}^{K}\beta_j)}{\prod_{j=1}^{K}\Gamma(\beta_j)}\frac{\prod_{j=1}^{K}\Gamma(\beta_j+\delta_{jk})}{\Gamma(\sum_{j=1}^{K}\beta_j+1)} \\ &= \frac{\beta_k}{\sum_{j=1}^{K}\beta_j} \end{aligned}$$

Now, because $\beta_k = c_k^{-n}+\alpha_k$,

$$p(z_{nk}=1|\boldsymbol{Z}^{-n},\boldsymbol{\alpha}) = \frac{c_k^{-n}+\alpha_k}{\sum_{j=1}^{K}c_j^{-n}+\alpha_j}$$

Following the same reasoning, it is also possible to collapse λ_k from the sampler. The joint density over $z_{nk}=1$ and λ_k is derived by carrying out an equivalent two-step

decomposition,

$$p(z_{nk}=1, \lambda_k|\boldsymbol{Z}^{-n}, X^{-n}, \boldsymbol{\alpha}, a, b) \propto p(z_{nk}=1|\boldsymbol{Z}^{-n}, \boldsymbol{\alpha})p(x_n|\lambda_k)p(\lambda_k|\boldsymbol{Z}^{-n}, X^{-n}, a, b)$$

where X^{-n} corresponds to all objects except for x_n. Marginalising λ_k out from this expression gives,

$$\begin{aligned} p(z_{nk}=1|\boldsymbol{Z}^{-n}, X^{-n}, \boldsymbol{\alpha}, a, b) &\propto \int p(z_{nk}=1|\boldsymbol{Z}^{-n}, \boldsymbol{\alpha})p(x_n|\lambda_k)p(\lambda_k|\boldsymbol{Z}^{-n}, X^{-n}, a, b)\,\mathrm{d}\lambda_k \\ &\propto p(z_{nk}=1|\boldsymbol{Z}^{-n}, \boldsymbol{\alpha}) \int p(x_n|\lambda_k)p(\lambda_k|\boldsymbol{Z}^{-n}, X^{-n}, a, b)\,\mathrm{d}\lambda_k \\ &\propto p(z_{nk}=1|\boldsymbol{Z}^{-n}, \boldsymbol{\alpha})p(x_n|\boldsymbol{Z}^{-n}, X^{-n}, a, b) \end{aligned}$$

where,

$$p(x_n|\boldsymbol{Z}^{-n}, X^{-n}, a, b) = \int p(x_n|\lambda_k)p(\lambda_k|\boldsymbol{Z}^{-n}, X^{-n}, a, b)\,\mathrm{d}\lambda_k$$

and, like in the previous case, the posterior, $p(\lambda|\boldsymbol{Z}^{-n}, X^{-n}, a, b)$, is the result of having a Gamma prior $p(\lambda_k|a, b)$, as conjugate of the Poisson likelihood, $p(X^{-n}|\lambda_k, \boldsymbol{Z}^{-n})$. Concretely,

$$p(\lambda_k|\boldsymbol{Z}^{-n}, X^{-n}, a, b) \propto \lambda_k^{\delta-1}e^{-\gamma\lambda_k},$$

$$\delta = a + \sum_{m\neq n} z_{mk}x_m, \quad \gamma = \sum_{m\neq n} z_{mk} + b$$

Hence,

$$\begin{aligned} p(x_n|\boldsymbol{Z}^{-n}, X^{-n}, a, b) &= \frac{\gamma^\delta}{x_n!\Gamma(\delta)} \int \lambda_k^{(x_n+\delta)-1}e^{-(\gamma+1)\lambda_k}\,\mathrm{d}\lambda_k \\ &= \frac{\gamma^\delta}{x_n!\Gamma(\delta)} \frac{\Gamma(x_n+\delta)}{(1+\gamma)^{(x_n+\delta)}} \end{aligned}$$

Re-arranging and having $r = \delta$ and $p = \gamma/(\gamma+1)$,

$$\begin{aligned} p(x_n|\boldsymbol{Z}^{-n}, X^{-n}, a, b) &= \frac{\Gamma(x_n+r)}{x_n!\Gamma(r)} \frac{\left(\frac{p}{1-p}\right)^r}{\left(1+\frac{p}{1-p}\right)^{(x_n+r)}} \\ &= \frac{\Gamma(x_n+r)}{x_n!\Gamma(r)} p^r(1-p)^r \\ &= NB(r, p) \end{aligned}$$

Finally,

$$p(z_{nk} = 1|\boldsymbol{Z}^{-n}, X^{-n}, \boldsymbol{\alpha}, a, b) \propto \frac{c_k^{-n} + \alpha_k}{\sum_j^K c_j^{-n} + \alpha_j} p(x_n|\boldsymbol{Z}^{-n}, X^{-n}, a, b)$$

B.2 Marginalisation of the Threshold

A salient aspect of Bayesian modelling is the ability to compare models in a principled manner. In this case, the threshold can be interpreted as the realisation of a model $\mathcal{M}_j$. By following the same approach as in Section 8.3.3, the threshold can be marginalised out of the DP-PMM, thereby improving its overall robustness. In this demonstration, three sets of features are considered, and extracted using three different thresholds. The thresholds used correspond to the 95^{th}, 99.5^{th}, and 99.9^{th} percentiles of the signal, respectively. For each model, the overlapping scheme from Section 8.3.3 is applied, resulting in a total of 21 sets of features. The assigned sections of the signal, for each threshold, are shown in Figure B.1. The results clearly demonstrate that a lower threshold leads the DP-PMM to cluster sections in the background noise, while a higher threshold disregards potential events in the signal. By averaging the probabilities obtained from each model, as described in Equation (6.21), the discrepancies across models are mitigated, and return a more balanced outcome that is less sensitive to the chosen threshold. Figure B.2 illustrates this outcome, where two distinct groups appear to dominate across all the clusters inferred from each model.

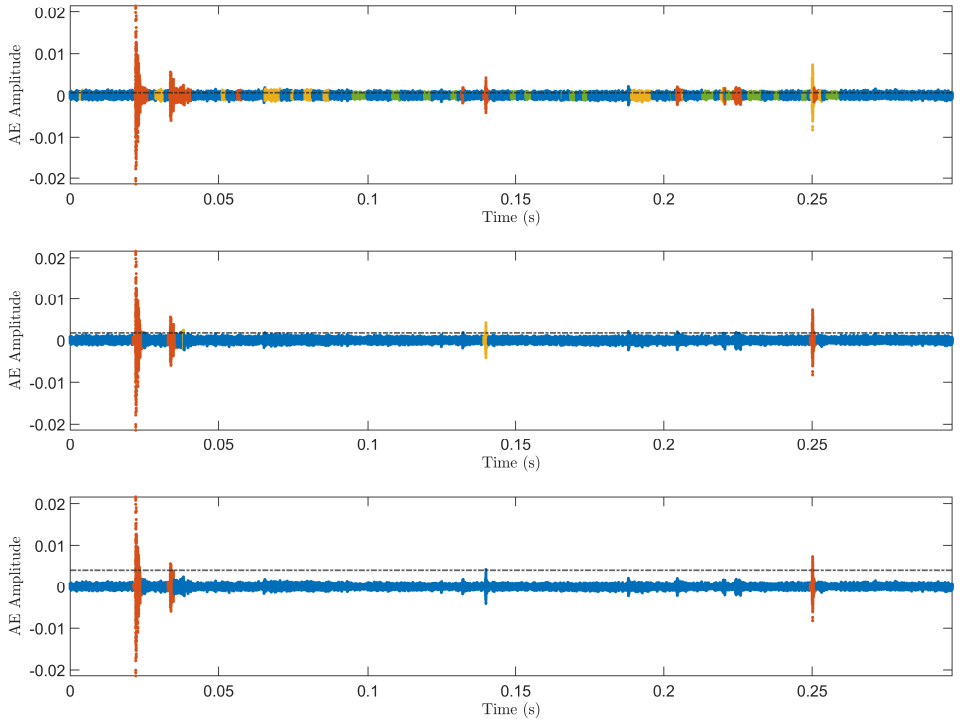


FIGURE B.1: Assignment of sections in the signal to the groups that maximise the assignment probabilities, given thresholds corresponding the (Top) 95^{th}, (Middle) 99.5^{th}, and (Bottom) 99.9^{th} percentiles, respectively.

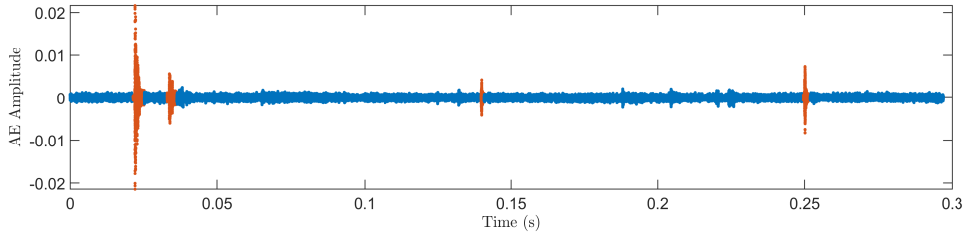


FIGURE B.2: Assignment of sections in the signal to the groups that maximise the assignment probabilities after marginalising the threshold. From all the clusters inferred over each model, only two distinct groups of AE activity are retained from the marginalisation.

Appendix C

Derivation of ELBO for VAEs

To derive the lower bound of the marginal likelihood, a general case is considered where some observations $\mathbf{X}$ and a model that implies some latent variables $\mathbf{Z}$. The marginal likelihood $p(\mathbf{X})$ is defined as,

$$p(\mathbf{X}) = \int p(\mathbf{X}, \mathbf{Z}) \mathrm{d}\mathbf{Z} \tag{C.1}$$

Maximising this expression with respect to all the parameters it is conditioned on is often difficult because of the potentially high-dimensional parameter space involved in evaluating the integral. The lower bound on the marginal likelihood can facilitate solving for $p(\mathbf{X})$.

To begin, it is necessary to define the log marginal likelihood,

$$\log p(\mathbf{X}) = \log \int p(\mathbf{X}, \mathbf{Z}) \mathrm{d}\mathbf{Z} \tag{C.2}$$

and introduce an auxiliary distribution $q(\mathbf{Z})$ into the RHS,

$$\log p(\mathbf{X}) = \log \int q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \tag{C.3}$$

The lower bound appears when using *Jensen's inequality* in terms of the current form of the log marginal likelihood. Jensen's inequality is defined by the following relationship,

$$\log \mathbb{E}\left[f(x)\right] \geq \mathbb{E}\left[\log f(x)\right] \tag{C.4}$$

stating that the log of the expected value of $f(x)$ is always greater than or equal to the expected value of the log of $f(x)$. Since the expression in (C.3) defines the log marginal likelihood as the expected value of $\frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})}$, with respect to $q(\mathbf{Z})$, the following relationship

applies according to Jensen's inequality,

$$\begin{aligned}\log p(\mathbf{X}) &= \log \int q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \\ &\geq \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} = \mathcal{L}(q(\mathbf{Z}))\end{aligned} \tag{C.5}$$

where $\mathcal{L}(q(\mathbf{Z}))$ denotes the lower bound on $\log p(\mathbf{X})$.

In the interest of computing the approximate posterior, subtracting the true log marginal likelihood with the lower bound reveals how it can be obtained. In particular,

$$\begin{aligned}\log p(\mathbf{X}) - \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} &\geq 0 \\ \log p(\mathbf{X}) - \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}) p(\mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} &\geq 0 \\ \log p(\mathbf{X}) - \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} - \int q(\mathbf{Z}) \log p(\mathbf{X}) \mathrm{d}\mathbf{Z} &\geq 0 \\ \log p(\mathbf{X}) - \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} - \log p(\mathbf{X}) \int q(\mathbf{Z}) \mathrm{d}\mathbf{Z} &\geq 0 \\ - \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} &\geq 0\end{aligned}$$

Therefore,

$$\log p(\mathbf{X}) = \mathcal{L}(q(\mathbf{Z})) + D_{\mathrm{KL}} \left[q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{X}) \right] \tag{C.6}$$

where the *Evidence Lower Bound* (ELBO) and *Kullback-Leibler* (KL) divergence terms are defined, respectively, as,

$$\mathcal{L}(q(\mathbf{Z})) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \tag{C.7a}$$

$$D_{\mathrm{KL}} \left[q(\mathbf{Z}) || p(\mathbf{Z}|\mathbf{X}) \right] = - \int q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \tag{C.7b}$$

Finally, for VAEs, it is convenient to rewrite the ELBO as follows,

$$\begin{aligned}\mathcal{L}(q(\mathbf{Z})) &= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \\ &= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X}|\mathbf{Z}) p(\mathbf{Z})}{q(\mathbf{Z})} \mathrm{d}\mathbf{Z} \\ &= \int q(\mathbf{Z}) \left[\log p(\mathbf{X}|\mathbf{Z}) + \log p(\mathbf{Z}) - \log q(\mathbf{Z}) \right] \mathrm{d}\mathbf{Z} \\ &= \int q(\mathbf{Z}) \log p(\mathbf{X}|\mathbf{Z}) \mathrm{d}\mathbf{Z} - \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z})} \mathrm{d}\mathbf{Z} \\ &= \mathbb{E}_{q(\mathbf{Z})} \left[\log p(\mathbf{X}|\mathbf{Z}) \right] - D_{\mathrm{KL}} \left[q(\mathbf{Z}) || p(\mathbf{Z}) \right]\end{aligned}$$

Appendix D

Publications

D.1 Journal Papers

- C.A. Lindley, M.R. Jones, T.J. Rogers, E.J. Cross, R.S. Dwyer-Joyce, N. Dervilis, K. Worden (2024). A probabilistic approach for acoustic emission based monitoring techniques: with application to structural health monitoring. *Mechanical Systems and Signal Processing*, *208*, 110958.
- C.A. Lindley, S. Beamish, R.S. Dwyer-Joyce, N. Dervilis, K. Worden (2022). A Bayesian approach for shaft centre localisation in journal bearings. *Mechanical Systems and Signal Processing*, *174*, 109021.

D.2 Conference Papers

- C.A. Lindley, M.R. Jones, T.A. Dardeno, R.S. Mills, N. Dervilis, K. Worden (2023). Acoustic emission source location using Bayesian optimisation for a composite helicopter blade. *Proceedings of the Fourteenth International Workshop on Structural Health Monitoring, IWSHM 2023* (pp. 1181-1191).
- C.A. Lindley, R.S. Dwyer-Joyce, N. Dervilis, K. Worden (2023). On the application of variational auto-encoders (VAE) for damage detection under changing environmental conditions. *IMAC-XLI 2023*.
- C.A. Lindley, M.R. Jones, T.J. Rogers, R.S. Dwyer-Joyce, N. Dervilis, K. Worden (2022). A nonparametric Bayesian approach for the detection of acoustic emission events in time series signals. *Proceedings of the International Conference on Noise and Vibration Engineering, ISMA 2022* (pp. 4777-4787).

- C.A. Lindley, T.J. Rogers, R.S. Dwyer-Joyce, N. Dervilis, K. Worden (2021). On the application of variational autoencoders (VAE) for damage detection in rolling element bearings. *Proceedings of the Thirteenth International Workshop on Structural Health Monitoring, IWSHM 2021* (pp. 388-397).

Bibliography

[1] C.R. Farrar and K. Worden. *Structural Health Monitoring : a Machine Learning Perspective*. Wiley, 2013.

[2] A. Rytter. *Vibrational Based Inspection of Civil Engineering Structures*. PhD thesis, Aalborg University, 1993.

[3] C.R. Farrar, T.A. Duffey, S.W. Doebling, and D.A. Nix. A statistical pattern recognition paradigm for vibration-based structural health monitoring. *Structural Health Monitoring*, 2000:764–773, 1999.

[4] K. Worden and J.M Dulieu-Barton. An overview of intelligent fault detection in systems and structures. *Structural Health Monitoring*, 3(1):85–98, 2004.

[5] R.B. Randall. *Vibration-Based Condition Monitoring: Industrial, Automotive and Aerospace Applications*. John Wiley & Sons, 2021.

[6] K. Worden and G. Manson. The application of machine learning to structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1851):515–537, 2006.

[7] G. Toh and J. Park. Review of vibration-based structural health monitoring using deep learning. *Applied Sciences*, 10(5):1680, 2020.

[8] E.J. Cross, S.J. Gibson, M.R. Jones, D.J. Pitchforth, S. Zhang, and T.J. Rogers. Physics informed machine learning for structural health monitoring. *Structural Health Monitoring Based on Data Science Techniques*, pages 347–367, 2022.

[9] T.J. Rogers, P. Gardner, N. Dervilis, K. Worden, A.E. Maguire, E. Papatheou, and E.J. Cross. Probabilistic modelling of wind turbine power curves with application of heteroscedastic Gaussian process regression. *Renewable Energy*, 148:1124–1136, 2020.

[10] L. Bornn, C.R. Farrar, G. Park, and K. Farinholt. Structural health monitoring with autoregressive support vector machines. *Journal of Vibration and Acoustics*, 131(2):021004, 02 2009.

[11] W. Kurt. *Bayesian Statistics The Fun Way: Understanding Statistics and Probability with Star Wars, Lego, and Rubber Ducks*. No Starch Press, 2019.

[12] S.M. Stigler. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press: Cambridge, 1986.

[13] H. Jeffreys. *Theory of Probability*. (Third Edition 1998) Clarendon Press, 1939.

[14] W.H. Jefferys and J.O. Berger. Ockham's razor and Bayesian analysis. *American Scientist*, 80(1):64–72, 1992.

[15] B. de Finetti. Foresight: Its logical laws, its subjective sources. In *Breakthroughs in Statistics: Foundations and Basic Theory*, pages 134–174. Springer, 1937.

[16] C.K.I Williams and C.E. Rasmussen. *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA, 2006.

[17] L.P. Swiler, M. Gulian, A.L. Frankel, C. Safta, and J.D. Jakeman. A survey of constrained Gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2), 2020.

[18] A. Pensoneault, X. Yang, and X. Zhu. Nonnegativity-enforced Gaussian process regression. *Theoretical and Applied Mechanics Letters*, 10(3):182–187, 2020.

[19] E. Papatheou, N. Dervilis, A.E. Maguire, I. Antoniadou, and K. Worden. A performance monitoring approach for the novel Lillgrund offshore wind farm. *IEEE Transactions on Industrial Electronics*, 62(10):6636–6644, 2015.

[20] J.H. Mclean, M.R. Jones, B.J. O'Connell, A.E. Maguire, and T.J. Rogers. Physically meaningful uncertainty quantification in probabilistic wind turbine power curve models as a damage sensitive feature, 2023.

[21] M.R. Jones, T.J. Rogers, and E.J. Cross. Constraining Gaussian processes for physics-informed acoustic emission mapping. *Mechanical Systems and Signal Processing*, 188:109984, 2023.

[22] W. Qiao and D. Lu. A survey on wind turbine condition monitoring and fault diagnosis-part II: Signals and signal processing methods. *IEEE Transactions on Industrial Electronics*, 62(10):6546–6557, 2015.

[23] S.W. Doebling, C.R. Farrar, M.B. Prime, and D.W. Shevitz. Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review. *Los Alamos National Laboratory*, (LA-13070-MS), 1996.

[24] H. Sohn, C.R. Farrar, F.M. Hemez, D.D. Shunk, D.W. Stinemates, B.R. Nadler, and J.J. Czarnecki. A review of structural health monitoring literature: 1996–2001. *Los Alamos National Laboratory*, 1(LA-UR-02-2095), 2003.

[25] K. Nienhaus, F.D. Boos, K. Garate, and R. Baltes. Development of acoustic emission (AE) based defect parameters for slow rotating roller bearings. *Journal of Physics: Conference Series*, 364(1):012034, 2012.

[26] J. Kaiser. *An investigation on the occurrence of noises in tensile tests or a study of acoustic phenomena in tensile tests.* PhD thesis, Tech Hochsch, 1950.

[27] J. Hanchi and B.E. Klamecki. Acoustic emission monitoring of the wear process. *Wear*, 145(1):1–27, 1991.

[28] Y. Fan, F. Gu, and A. Ball. Modelling acoustic emissions generated by sliding friction. *Wear*, 268:811–815, 02 2010.

[29] A. Hase, H. Mishina, and M. Wada. Correlation between features of acoustic emission signals and mechanical wear mechanisms. *Wear*, 292-293:144–150, 2012.

[30] C.J. Hellier. *Handbook of Nondestructive Evaluation.* McGraw-Hill Education, 2013.

[31] P. Feng, P. Borghesani, W.A Smith, R.B. Randall, and Z. Peng. A review on the relationships between acoustic emission, friction and wear in mechanical systems. *Applied Mechanics Reviews*, 72(2), 10 2019.

[32] J. Ma, H. Zhang, Z. Shi, F. Chu, F. Gu, and A.D. Ball. Modelling acoustic emissions induced by dynamic fluid-asperity shearing in hydrodynamic lubrication regime. *Tribology International*, 153:106590, 2021.

[33] T.M. Morton, R.M. Harrington, and J.G. Bjeletich. Acoustic emissions of fatigue crack growth. *Engineering Fracture Mechanics*, 5(3):691–697, 1973.

[34] M.N. Bassim, S.St. Lawrence, and C.D. Liu. Detection of the onset of fatigue crack growth in rail steels using acoustic emission. *Engineering Fracture Mechanics*, 47(2):207–214, 1994.

[35] A. Berkovits and D. Fang. Study of fatigue crack characteristics by acoustic emission. *Engineering Fracture Mechanics*, 51(3):401–416, 1995.

[36] M. Rabiei, M. Modarres, and P. Hoffman. Structural integrity assessment using in-situ acoustic emission monitoring. *Annual Conference of the PHM Society*, 3(1), 2011.

[37] P. Paris and F. Erdogan. A critical analysis of crack propagation laws. *Journal of Basic Engineering*, 85(4):528–533, 12 1963.

[38] A. Choudhury and N. Tandon. Application of acoustic emission technique for the detection of defects in rolling element bearings. *Tribology International*, 33(1):39–45, 2000.

[39] Y. He, X. Zhang, and M. Friswell. Defect diagnosis for rolling element bearings using acoustic emission. *Journal of Vibration and Acoustics*, 131, 12 2009.

[40] P. Feng, P. Borghesani, H. Chang, W.A. Smith, R.B. Randall, and Z. Peng. Monitoring gear surface degradation using cyclostationarity of acoustic emission. *Mechanical Systems and Signal Processing*, 131:199–221, 2019.

[41] S. Poddar and N. Tandon. Detection of journal bearing vapour cavitation using vibration and acoustic emission techniques with the aid of oil film photography. *Tribology International*, 103:95–101, 2016.

[42] S. Poddar and N. Tandon. Detection of particle contamination in journal bearing using acoustic emission and vibration monitoring techniques. *Tribology International*, 134:154–164, 2019.

[43] C. Chen, X. Chen, and S. Guo. Experimental study on acoustic emission characteristic of fatigue crack growth of self-compacting concrete. *Structural Control and Health Monitoring*, 26(4):e2332, 2019.

[44] K.M Holford, R. Pullin, S. Evans, M. 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*, 223:525–532, 2009.

[45] 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*, 25(7):2462–2474, 2011.

[46] 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*, 373(2035):20140075, 2015.

[47] H. Sohn, C.R. Farrar, N.F. Hunter, and K. Worden. Structural health monitoring using statistical pattern recognition techniques. *Journal of Dynamic Systems, Measurement, and Control*, 123(4):706–711, 2001.

[48] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[49] S.K. Mitra. *Digital Signal Processing: A Computer-Based Approach*. McGraw-Hill Higher Education, 2001.

[50] K. Worden, C.R. Farrar, J. Haywood, and M. Todd. A review of nonlinear dynamics applications to structural health monitoring. *Structural Control and Health Monitoring*, 15(4):540–567, 2008.

[51] H. Sohn. Effects of environmental and operational variability on structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1851):539–560, 2007.

[52] V. Cherkassky and F.M Mulier. *Learning From Data: Concepts, Theory, and Methods*. John Wiley & Sons, 2007.

[53] L.A. Bull, P.A. Gardner, J. Gosliga, T.J. Rogers, N. Dervilis, E.J. Cross, E. Papatheou, A.E. Maguire, C. Campos, and K. Worden. Foundations of population-based shm, part I: Homogeneous populations and forms. *Mechanical Systems and Signal Processing*, 148:107141, 2021.

[54] J. Gosliga, P.A. Gardner, L.A. Bull, N. Dervilis, and K. Worden. Foundations of population-based shm, part II: Heterogeneous populations–graphs, networks, and communities. *Mechanical Systems and Signal Processing*, 148:107144, 2021.

[55] P. Gardner, L.A. Bull, J. Gosliga, N. Dervilis, and K. Worden. Foundations of population-based shm, part III: Heterogeneous populations–mapping and transfer. *Mechanical Systems and Signal Processing*, 149:107142, 2021.

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

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

[58] L.A. Bull, P. Gardner, T.J. Rogers, E.J. Cross, N. Dervilis, and K. Worden. Probabilistic inference for structural health monitoring: New modes of learning from data. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 7(1):03120003, 2021.

[59] M.P Deisenroth, A.A Faisal, and C.S. Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020.

[60] V.N. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[61] G.E. Hinton and T.J. Sejnowski. Learning and relearning in Boltzmann machines. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1(282-317):2, 1986.

[62] D.J.C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, 1999.

[63] D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[64] C. Rasmussen and Z. Ghahramani. Occam's razor. *Advances in Neural Information Processing Systems*, 13, 2000.

[65] R.M. Neal. *Bayesian Learning for Neural Networks*, volume 118. Springer Science & Business Media, 2012.

[66] H. Sohn, K. Worden, and C.R. Farrar. Novelty detection under changing environmental conditions. In *Smart Structures and Materials 2001: Smart Systems for Bridges, Structures, and Highways*, volume 4330, pages 108 – 118. International Society for Optics and Photonics, SPIE, 2001.

[67] H. Sohn, K. Worden, and C.R. Farrar. Statistical damage classification under changing environmental and operational conditions. *Journal of Intelligent Material Systems and Structures*, 2002.

[68] C.R. Farrar, W.E. Baker, T.M Bell, K.M. Cone, T.W. Darling, T.A. Duffey, A. Eklund, and A. Migliori. Dynamic characterization and damage detection in the I-40 bridge over the Rio Grande. Technical report, Los Alamos National Laboratory, 1994.

[69] D.A. Pomerleau. Input reconstruction reliability estimation. *Advances in Neural Information Processing Systems*, 5, 1992.

[70] R. Bro and A.K. Smilde. Principal component analysis. *Analytical Methods*, 6(9):2812–2831, 2014.

[71] 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*, 333(6):1833–1850, 2014.

[72] D.P. Kingma and M. Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[73] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational Bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, pages 2391–2400. PMLR, 2017.

[74] A. Farhidzadeh, S. Salamone, and P. Singla. A probabilistic approach for damage identification and crack mode classification in reinforced concrete structures. *Journal of Intelligent Material Systems and Structures*, 24(14):1722–1735, 2013.

[75] P.R. Prem and A.R. Murthy. Acoustic emission monitoring of reinforced concrete beams subjected to four-point-bending. *Applied Acoustics*, 117:28–38, 2017.

[76] S.G. Walker, P. Damien, P.W. Laud, and A.F.M. Smith. Bayesian nonparametric inference for random distributions and related functions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 61(3):485–527, 1999.

[77] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Routledge, 2018.

[78] K. Worden, S.G. Pierce, G. Manson, W.R. Philp, W.J. Staszewski, and B. Culshaw. Detection of defects in composite plates using Lamb waves and novelty detection. *International Journal of Systems Science*, 31(11):1397–1409, 2000.

[79] M.L. Fugate, H. Sohn, and C.R. Farrar. Unsupervised learning methods for vibration-based damage detection. In *Proceedings of 18th International Modal Analysis Conference–IMAC*, volume 18, 2000.

[80] L. Yu and Z. Su. Application of kernel density estimation in Lamb wave-based damage detection. *Mathematical Problems in Engineering*, 2012(406521), 2012.

[81] H.P Wan and Y.Q.˙Bayesian modeling approach for forecast of structural stress response using structural health monitoring data. *Journal of Structural Engineering*, 144(9):04018130, 2018.

[82] J. Herp, M.H. Ramezani, M. Bach-Andersen, N.L. Pedersen, and E.S. Nadimi. Bayesian state prediction of wind turbine bearing failure. *Renewable Energy*, 116:164–172, 2018.

[83] Y. Pang, X. Zhou, W. He, J. Zhong, and Q. Hui. Uniform design-based Gaussian process regression for data-driven rapid fragility assessment of bridges. *Journal of Structural Engineering*, 147(4):04021008, 2021.

[84] J. Paixao, S. da Silva, E. Figueiredo, L. Radu, and G. Park. Delamination area quantification in composite structures using Gaussian process regression and autoregressive models. *Journal of Vibration and Control*, 27(23-24):2778–2792, 2021.

[85] R.K. Pandit and D. Infield. Comparative analysis of Gaussian process power curve models based on different stationary covariance functions for the purpose of improving model accuracy. *Renewable Energy*, 140:190–202, 2019.

[86] R. Pandit and D. Infield. SCADA - based wind turbine anomaly detection using Gaussian process models for wind turbine condition monitoring purposes. *IET Renewable Power Generation*, 12, 05 2018.

[87] Y. Li, S. Liu, and L. Shu. Wind turbine fault diagnosis based on Gaussian process classifiers applied to operational data. *Renewable Energy*, 134, 11 2018.

[88] K. Chandrasekhar, N. Stevanovic, E.J. Cross, N. Dervilis, and K. Worden. Damage detection in operational wind turbine blades using a new approach based on machine learning. *Renewable Energy*, 168:1249–1264, 2021.

[89] K. Worden and E.J. Cross. On switching response surface models, with applications to the structural health monitoring of bridges. *Mechanical Systems and Signal Processing*, 98:139–156, 2018.

[90] E.J. Cross, K. Worden, K.Y. Koo, and J.M.W. Brownjohn. Filtering environmental load effects to enhance novelty detection on cable-supported bridge performance. In *Bridge Maintenance, Safety, Management, Resilience and Sustainability—Proceedings of the Sixth International Conference on Bridge Maintenance, Safety and Management*, pages 745–752, 2012.

[91] L.D. Avendano-Valencia, E. Chatzi, and D. Tcherniak. Gaussian process models for mitigation of operational variability in the structural health monitoring of wind turbines. *Mechanical Systems and Signal Processing*, 142:106686, 2020.

[92] J. Kullaa. Distinguishing between sensor fault, structural damage, and environmental or operational effects in structural health monitoring. *Mechanical Systems and Signal Processing*, 25(8):2976–2989, 2011.

[93] 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*, 24(1):211–223, 2010.

[94] M.R. Jones, T.J. Rogers, K. Worden, and E.J. Cross. A Bayesian methodology for localising acoustic emission sources in complex structures. *Mechanical Systems and Signal Processing*, 163:108143, 2022.

[95] J. Hensman, R. Pullin, M. Eaton, K. Worden, K.M. Holford, and S.L. Evans. Detecting and identifying artificial acoustic emission signals in an industrial fatigue environment. *Measurement Science and Technology*, 20(4):045101, 2009.

[96] S.J. Gibson, T.J. Rogers, and E.J. Cross. Distributions of fatigue damage from data-driven strain prediction using Gaussian process regression. *Structural Health Monitoring*, 2023.

[97] G. Holmes, P. Sartor, S. Reed, P. Southern, K. Worden, and E.J. Cross. Prediction of landing gear loads using machine learning techniques. *Structural Health Monitoring*, 15(5):568–582, 2016.

[98] S. Mohanty, A. Chattopadhyay, P. Peralta, and S. Das. Bayesian statistic based multivariate Gaussian process approach for offline/online fatigue crack growth prediction. *Experimental Mechanics*, 51:833–843, 2011.

[99] Y. Ling and S. Mahadevan. Integration of structural health monitoring and fatigue damage prognosis. *Mechanical Systems and Signal Processing*, 28:89–104, 2012.

[100] H. Teimouri, A.S. Milani, J. Loeppky, and R. Seethaler. A Gaussian process-based approach to cope with uncertainty in structural health monitoring. *Structural Health Monitoring*, 16(2):174–184, 2017.

[101] J. Chen, S. Yuan, and H. Wang. On-line updating Gaussian process measurement model for crack prognosis using the particle filter. *Mechanical Systems and Signal Processing*, 140:106646, 2020.

[102] J. Kocijan. Dynamic GP models: an overview and recent developments. In *Proceedings of 6th International Conference on Applied Mathematics, Simulation and Modelling*, pages 38–43, 2012.

[103] K. Worden, W. Becker, T.J. Rogers, and E.J. Cross. On the confidence bounds of Gaussian process NARX models and their higher-order frequency response functions. *Mechanical Systems and Signal Processing*, 104:188–223, 2018.

[104] D.J. Pitchforth, T.J. Rogers, U.T. Tygesen, and E.J. Cross. Grey-box models for wave loading prediction. *Mechanical Systems and Signal Processing*, 159:107741, 2021.

[105] J.R. Morison, J.W. Johnson, and S.A. Schaaf. The force exerted by surface waves on piles. *Journal of Petroleum Technology*, 2(05):149–154, 1950.

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

[107] D.A. Reynolds. Gaussian mixture models. *Encyclopedia of Biometrics*, 741(659-663), 2009.

[108] C. Rasmussen. The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems*, volume 12, pages 554–560. MIT Press, 1999.

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

[110] C.T. Wickramarachchi, T.J. Rogers, T.E. McLeay, W. Leahy, and E.J. Cross. Online damage detection of cutting tools using Dirichlet process mixture models. *Mechanical Systems and Signal Processing*, 180:109434, 2022.

[111] P. Cheema, M.M. Alamdari, G.A. Vio, F.L. Zhang, and C.W. Kim. Infinite mixture models for operational modal analysis: An automated and principled approach. *Journal of Sound and Vibration*, 491:115757, 2021.

[112] S. Rogers and M. Girolami. *A First Course in Machine Learning.* Chapman and Hall/CRC, 2nd edition, 2016.

[113] M.L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer Science & Business Media, 1999.

[114] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, volume 55. US Government printing office, 1948.

[115] G.E. Uhlenbeck and L.S. Ornstein. On the theory of the Brownian motion. *Physical review*, 36(5):823, 1930.

[116] D.J.C. MacKay. Introduction to Gaussian processes. *NATO ASI series F Computer and Systems Sciences*, 168:133–166, 1998.

[117] M. Haywood-Alexander, N. Dervilis, K. Worden, E.J. Cross, R.S. Mills, and T.J. Rogers. Structured machine learning tools for modelling characteristics of guided waves. *Mechanical Systems and Signal Processing*, 156:107628, 2021.

[118] R. Fletcher. *Practical Methods of Optimization.* John Wiley & Sons, 2000.

[119] L.M. Rios and N.V. Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56:1247–1293, 2013.

[120] D. Wolpert and W.G Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.

[121] J. Sun, W. Xu, and B. Feng. A global search strategy of quantum-behaved particle swarm optimization. In *IEEE Conference on Cybernetics and Intelligent Systems, 2004.*, volume 1, pages 111–116. IEEE, 2004.

[122] W.J. Ewens. Population genetics theory-the past and the future. In *Mathematical and Statistical Developments of Evolutionary Theory*, pages 177–227. Springer, 1990.

[123] F.M. Hoppe. Pólya-like urns and the Ewens' sampling formula. *Journal of Mathematical Biology*, 20(1):91–94, 1984.

[124] P. Diaconis. Finite forms of de Finetti's theorem on exchangeability. *Synthese*, 36:271–281, 1977.

[125] D.J. Aldous, I.A. Ibragimov, and J. Jacod. *Exchangeability and Related Topics*. Springer, 1985.

[126] David Blackwell and James B MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[127] J. Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4(2):639–650, 1994.

[128] H. Ishwaran and L.F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[129] R.M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

[130] D.M. Blei and M.I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121 – 143, 2006.

[131] S. Beamish, X. Li, H. Brunskill, A. Hunter, and R.S. Dwyer-Joyce. Circumferential film thickness measurement in journal bearings via the ultrasonic technique. *Tribology International*, 148:106295, 2020.

[132] M.R. Jones. *On Novel Machine Learning Approaches for Acoustic Emission Source Localisation: A Probabilistic Perspective*. PhD thesis, University of Sheffield, March 2023.

[133] G.W. Stachowiak and A.W. Batchelor. *Engineering Tribology*. Elsevier Science, 2013.

[134] M.M. Muhammad, S.D. Yati, N.A. Yahya, and N.A. Sa'at. Fault diagnosis of rolling element bearings in a generator using envelope analysis. *Defence S & T Technical Bulletin*, 4(2):131–143, 2011.

[135] A. Roque, T. Silva, J. Calado, and J. Dias. An approach to fault diagnosis of rolling bearings. *WSEAS Transactions on Systems and Control*, 4, 04 2009.

[136] T. Hossein, R. Parno, G. Fengshou, and A. Ball. Characterization of acoustic emissions from journal bearings for fault detection. pages 92–103. British Institute of Non-Destructive Testing, 2013.

[137] H. Sadegh, A.N. Mehdi, and A. Mehdi. Classification of acoustic emission signals generated from journal bearing at different lubrication conditions based on wavelet analysis in combination with artificial neural network and genetic algorithm. *Tribology International*, 95:426–434, 2016.

[138] M.M. Khonsari and E.R. Booser. Effect of contamination on the performance of hydrodynamic bearings. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, 220(5):419–428, 2006.

[139] A. Dadouche and M.J. Conlon. Operational performance of textured journal bearings lubricated with a contaminated fluid. *Tribology International*, 93:377–389, 2016.

[140] Q. Wang. Seizure failure of journal-bearing conformal contacts. *Wear*, 210(1-2):8–16, 1997.

[141] A. Hase, H. Mishina, and M. Wada. Fundamental study on early detection of seizure in journal bearing by using acoustic emission technique. *Wear*, 346-347:132–139, 2016.

[142] R.W. Bruce. *Handbook of Lubrication and Tribology. Volume II, Theory and Design*. CRC Press, 2^{nd} edition, 2012.

[143] M.H. He, J.M. Byrne, C.H. Cloud, and J.A. Vazquez. Fundamentals of fluid film journal bearing operation and modeling. *Texas A&M University - Turbomachinery Laboratories*, 2016.

[144] Z. Cui, C. Yang, B. Sun, and H. Wang. Liquid film thickness estimation using electrical capacitance tomography. *Measurement Science Review*, 14(1):8–15, 2014.

[145] S.B. Glavatskih, Ö. Uusitalo, and D.J. Spohn. Simultaneous monitoring of oil film thickness and temperature in fluid film bearings. *Tribology International*, 34(12):853–857, 2001.

[146] Y. Fang, J. He, and P. Huang. Experimental and numerical analysis of soft elastohydrodynamic lubrication in line contact. *Tribology Letters*, 65(42), 2017.

[147] N. Marx, J. Guegan, and H.A. Spikes. Elastohydrodynamic film thickness of soft EHL contacts using optical interferometry. *Tribology International*, 99:267–277, 2016.

[148] K. Zhang, P. Dou, T. Wu, K. Feng, and Y. Zhu. An ultrasonic measurement method for full range of oil film thickness. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, 233(3):481–489, 2019.

[149] R.S. Dwyer-Joyce, B.W. Drinkwater, and C.J. Donohoe. The measurement of lubricant film thickness using ultrasound. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 459(2032):957–976, 2003.

[150] A. Hunter, R.S. Dwyer-Joyce, and P. Harper. Calibration and validation of ultrasonic reflection methods for thin-film measurement in tribology. *Measurement Science and Technology*, 23(10):105605, 2012.

[151] S. Kasolang, D.I. Ahmed, R.S. Dwyer-Joyce, and B.F. Yousif. Performance analysis of journal bearings using ultrasonic reflection. *Tribology International*, 64:78–84, 2013.

[152] T. Reddyhoff, S. Kasolang, R.S. Dwyer-Joyce, and B.W. Drinkwater. The phase shift of an ultrasonic pulse at an oil layer and determination of film thickness. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, 219:387–400, 2005.

[153] T. Pialucha and P. Cawley. The detection of thin embedded layers using normal incidence ultrasound. *Ultrasonics*, 32(6):431–440, 1994.

[154] A.A. Raimondi and J. Boyd. A solution for the finite journal bearing and its application to analysis and design: I. *ASLE Transactions*, 1(1):159–174, 1958.

[155] E. Padonou and O. Roustant. Polar Gaussian processes for predicting on circular domains. *Society for Industrial and Applied Mathematics and American Statistical Association*, 4:1014–1033, 2015.

[156] S. Kasolang and R.S. Dwyer-Joyce. Observations of film thickness profile and cavitation around a journal bearing circumference. *Tribology Transactions*, 51(2):231–245, 2008.

[157] Y. Hori. *Hydrodynamic Lubrication*. Springer Tokyo, 2006.

[158] A.M. Molinaro, R. Simon, and R.M. Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 21(15):3301–3307, 05 2005.

[159] B.J. Hamrock. *Fundamentals of Fluid Film Lubrication*. McGraw-Hill, 1994.

[160] T. Kundu, H. Nakatani, and N. Takeda. Acoustic source localization in anisotropic plates. *Ultrasonics*, 52(6):740–746, 2012.

[161] A. Tobias. Acoustic-emission source location in two dimensions by an array of three sensors. *Non-Destructive Testing*, 9(1):9–12, 1976.

[162] H. Yamada, Y. Mizutani, H. Nishino, M. Takemoto, and K. Ono. Lamb wave source location of impact on anisotropic plates. *Journal of Acoustic Emission*, 18:51–60, 2000.

[163] T. Kundu, S. Das, S.A. Martin, and K.V. Jata. Locating point of impact in anisotropic fiber reinforced composite plates. *Ultrasonics*, 48(3):193–201, 2008.

[164] M.G. Baxter, R. Pullin, K.M. Holford, and S.L. Evans. Delta T source location for acoustic emission. *Mechanical Systems and Signal Processing*, 21(3):1512–1520, 2007.

[165] M.R. Jones, T.J. Rogers, K. Worden, and E.J. Cross. A Bayesian methodology for localising acoustic emission sources in complex structures. *Mechanical Systems and Signal Processing*, 163:108143, 2022.

[166] A. Carpinteri, J. Xu, G. Lacidogna, and A. Manuello. Reliable onset time determination and source location of acoustic emissions in concrete structures. *Cement and Concrete Composites*, 34(4):529–537, 2012.

[167] N.N. Hsu. Acoustic emission simulator, 1977. US Patent 4,018,084.

[168] J.H. Kurz, C.U. Grosse, and H. Reinhardt. Strategies for reliable automatic onset time picking of acoustic emissions and of ultrasound signals in concrete. *Ultrasonics*, 43(7):538–546, 2005.

[169] J.R. Donald, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[170] R. Fuentes, P. Gardner, C. Mineo, T.J. Rogers, S.G. Pierce, K. Worden, N. Dervilis, and E.J. Cross. Autonomous ultrasonic inspection using Bayesian optimisation and robust outlier analysis. *Mechanical Systems and Signal Processing*, 145:106897, 2020.

[171] J.Z. Sikorska and D. Mba. Challenges and obstacles in the application of acoustic emission to process machinery. *Journal of Process Mechanical Engineering*, 222(1):1–19, 2008.

[172] A. Terchi and Y.H.J. Au. Acoustic emission signal processing. *Measurement and Control*, 34(8):240–244, 2001.

[173] S. Rippengill, K. Worden, K.M Holford, and R. Pullin. Automatic classification of acoustic emission patterns. *Strain*, 39(1):31–41, 2003.

[174] M.G.R. Sause, A. Gribov, A.R. Unwin, and S. Horn. Pattern recognition approach to identify natural clusters of acoustic emission signals. *Pattern Recognition Letters*, 33(1):17–23, 2012.

[175] K.P. Murphy. *Machine Learning: a Probabilistic Perspective*. MIT Press, 2013.

[176] C.B. Scruby. An introduction to acoustic emission. *Journal of Physics E: Scientific Instruments*, 20(8):946, 1987.

[177] S. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics - Simulation and Computation*, 23(3):727–741, 1994.

[178] K.M. Holford, M.J. Eaton, J.J. Hensman, R. Pullin, S.L. Evans, N. Dervilis, and K. Worden. A new methodology for automating acoustic emission detection of metallic fatigue fractures in highly demanding aerospace environments: An overview. *Progress in Aerospace Sciences*, 90:1–11, 2017.

[179] T.M. Roberts and M. Talebzadeh. Acoustic emission monitoring of fatigue crack propagation. *Journal of Constructional Steel Research*, 59(6):695–712, 2003.

[180] D.M. Smith. The dynamics of synchronous whirl in turbine rotors. *IUTAM SYMPOSIUM Dynamics of Rotors*, pages 524–545, 1975.

[181] A. Tondl. *Some Problems of Rotor Dynamics*. Publishing House of the Czechoslovak Academy of Sciences, 1965.

[182] B.L. Newkirk and H.D. Taylor. Shaft whipping due to oil action in journal bearings. *General Electric Review*, 28(8):559–568, 1925.

[183] O. Pinkus, B. Sternlicht, and E. Saibel. Theory of hydrodynamic lubrication. *Journal of Applied Mechanics*, 29(1):221–222, 03 1962.

[184] J.S Rao, R.J. Raju, and K.V.B. Reddy. Experimental investigation on oil whip of flexible rotors. *Tribology*, 3(2):100–103, 1970.

[185] J.W. Lund. Review of the concept of dynamic coefficients for fluid film journal bearings. *Journal of Tribology*, 109(1):37–41, 01 1987.

[186] A. Stodola. Kritische wellenstörung infolge der nachgiebigkeit des oelpolsters im lager. *Schweizerische Bauzeitung*, 85(21):265–266, 1925.

[187] C. Hummel. *Kritische Drehzahlen als folge der Nachgiebigkeit des Schmiermittels im Lager*. Number 287. VDI-Verlag, 1926.

[188] A.W. Lees. *Vibration Problems in Machines: Diagnosis and Resolution*. CRC Press, 2020.

[189] H.D. Nelson. A finite rotating shaft element using Timoshenko beam theory. *Journal of Mechanical Design*, 102(4):793–803, 1980.

[190] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[191] M. Alvarez, D. Luengo, and N.D. Lawrence. Latent force models. In *Artificial Intelligence and Statistics*, pages 9–16. PMLR, 2009.

[192] J. Bilbao, E. Lourens, S. Schulze, and L. Ziegler. Virtual sensing in an onshore wind turbine tower using a Gaussian process latent force model. *Data-Centric Engineering*, 3:e35, 2022.

[193] J. Zou, E. Lourens, and A. Cicirello. Virtual sensing of subsoil strain response in monopile-based offshore wind turbines via Gaussian process latent force models. *Mechanical Systems and Signal Processing*, 200:110488, 2023.

[194] Ø.W. Petersen, O. Øiseth, and E. Lourens. Wind load estimation and virtual sensing in long-span suspension bridges using physics-informed Gaussian process latent force models. *Mechanical Systems and Signal Processing*, 170:108742, 2022.

[195] T.J. Rogers, K. Worden, and E.J. Cross. On the application of Gaussian process latent force models for joint input-state-parameter estimation: With a view to Bayesian operational identification. *Mechanical Systems and Signal Processing*, 140:106580, 2020.

[196] R. Nayek, S. Chakraborty, and S. Narasimhan. A Gaussian process latent force model for joint input-state estimation in linear structural systems. *Mechanical Systems and Signal Processing*, 128:497–530, 2019.

[197] S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*, volume 10. Cambridge University Press, 2019.

[198] S. Särkkä and L. Svensson. *Bayesian Filtering and Smoothing*, volume 17. Cambridge University Press, 2023.

[199] B. Dupire. Functional Itô calculus. *Quantitative Finance*, 19(5):721–729, 2019.

[200] P. Axelsson and F. Gustafsson. Discrete-time solutions to the continuous-time differential Lyapunov equation with applications to Kalman filtering. *IEEE Transactions on Automatic Control*, 60(3):632–643, 2014.

[201] N. Wahlström, P. Axelsson, and F. Gustafsson. Discretizing stochastic dynamical systems using Lyapunov equations. *IFAC Proceedings Volumes*, 47(3):3726–3731, 2014.

[202] J. Hartikainen and S. Särkkä. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 379–384. IEEE, 2010.

[203] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

[204] H.E. Rauch, F. Tung, and C.T. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, 1965.

[205] T. Someya, J. Mitsui, J. Esaki, S. Saito, Y. Kanemitsu, T. Iwatsubo, M. Tanaka, S. Hisa, T. Fujikawa, and H. Kanki. *Journal-Bearing Databook*. Springer Science & Business Media, 2013.

[206] X. Chen, T. Okada, Y. Kawamura, and T. Mitsuyuki. Estimation of on-site directional wave spectra using measured hull stresses on 14,000 TEU large container ships. *Journal of Marine Science and Technology*, 25:690–706, 2020.

[207] R. Temam. *Navier-Stokes equations: Theory and Numerical Analysis*, volume 343. American Mathematical Society, 2001.

[208] T.N. Babu, T.M. Raj, and T. Lakshmanan. A review on application of dynamic parameters of journal bearing for vibration and condition monitoring. *Journal of Mechanics*, 31(4):391–416, 2015.

[209] K. Otomo, S. Kobayashi, K. Fukuda, and H. Esaki. Latent variable based anomaly detection in network system logs. *IEICE Transactions on Information and Systems*, E102.D(9):1644–1652, 2019.

[210] J. Sun, Z. Wang, N. Xiong, and J. Shao. Learning sparse representation with variational autoencoder for anomaly detection. *IEEE Access*, 6:33353–33361, 2018.

[211] T. Song, J. Sun, B. Chen, W. Peng, and J. Songa. Latent space expanded variational autoencoder for sentence generation. *IEEE Access*, 7:1–1, 2019.

[212] K.L. Lim, X. Jiang, and C. Yi. Deep clustering with variational autoencoder. *IEEE Signal Processing Letters*, 27:231–235, 2020.

[213] Q. Yu, M.S. Kavitha, and T. Kurita. Mixture of experts with convolutional and variational autoencoders for anomaly detection. *Applied intelligence*, 51(6):3241–3254, 2021.

[214] D.J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286. PMLR, 2014.

[215] D.P. Kingma and M. Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

[216] K. Worden, S. Patsias, and W.J. Staszewski. Damage detection in machines using neural networks. pages 207–214, 1 1998.

[217] H. Hoffmann. Kernel PCA for novelty detection. *Pattern Recognition*, 40(3):863–874, 2007.

[218] J.N. Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.

[219] G. Manson, B.C. Lee, and W.J. Staszewski. Eliminating environmental effects from Lamb wave-based structural health monitoring. pages 471–484, 1 2005.

[220] E.J. Cross, G. Manson, K. Worden, and S.G. Pierce. Features for damage detection with insensitivity to environmental and operational variations. *Proceedings of the Royal Society. A, Mathematical, physical, and engineering sciences*, 468(2148):4098–4122, 2012.

[221] K. Worden, E.J. Cross, I. Antoniadou, and A. Kyprianou. A multiresolution approach to cointegration for enhanced SHM of structures under varying conditions -– an exploratory study. *Mechanical Systems and Signal Processing*, 47(1-2):243–262, 2014.

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

[223] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. *Advances in Neural Information Processing Systems*, 30, 2017.

[224] R. Maesschalck, D. Jouan-Rimbaud, and D.L. Massart. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000.

[225] S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[226] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

[227] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, 29, 2016.

[228] E. Cinlar. *Probability and Stochastics*, volume 261. Springer, 2011.

[229] Y. Teh, M. Jordan, M. Beal, and D. Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. *Advances in Neural Information Processing Systems*, 17, 2004.