Autonomous Ultrasound-based Inspection using Bayesian Optimisation and Robust Outlier Analysis

Ramon Fuentes[1], Paul Gardner[1], Carmelo Mineo[2],
Tim Rogers, Gareth Pierce[2], Keith Worden[1], Nikolaos Dervilis[1], Elizabeth J. Cross[1]

***Abstract*— The use of robotics is begining to play a key role in automating the data collection process in Non Destructive Testing (NDT). Increasing the use of automation quickly leads to the gathering of large quantities of data, which makes it inefficient, perhaps even unfeasable for a human to parse the information contained in it. This paper presents a solution to this problem by making the process of NDT data collection an autonomous one. In order to achieve this, the robotic data acquisition task is treated as an optimisation problem, where one seeks to find the location within a component being inspected, with the greatest damage index. The resulting algorithm is a combination of ideas borrowed from uncertainty quantification which enable the optimisation routine to be probabilistic and damage detection from the field of data-driven Structural Health Monitoring (SHM). The algorithm operates sequentially, where a decision is made in every iteration regarding the next most optimal physical location for placing an observation. This is achieved by modelling the two-dimensional field of novelty indexes, which are derived from a robust outlier analysis procedure. More specifically, the model of the novelty index involves the widely-used probabilistic and nonlinear regression scheme of Gaussian Processes (GPs). The value that performing inspection in this autonomous fashion adds is that its output is not just data, it is the required information from an NDT inspection: the probability that a component contains damage. Most importantly, it does so while also minimising the number of observations that it takes to give such estimate, thus minimising the time and cost of data gathering.**

I. INTRODUCTION

Non Destructive Testing (NDT) forms an integral part in the assessment of quality and structural integrity in engineering components in-service and post-manufacture. Whilst in the present day, the majority of NDT is carried out manually, recent increased attention to robotic-based inspection [1] promises to change matters. The use of robotics in the context of NDT is key when inspecting high value and/or safety critical assets; it also serves the purpose of removing humans from harm, when an inspection needs to be carried out in harsh and dangerous environments.

Whilst the motivations for robotics-based NDT are clear, and the relevant research is under-way [1], [2], little or nothing has been done in the way of performing *autonomous* as opposed to *automated* inspections. In an automated inspection scheme, the robot path is programmed prior to the inspection, and the robot makes no decisions regarding what areas of a component it should prioritise. The data collected has to be reported back to a human, or to further post-processing algorithms in order to assess the state of the component. In an autonomous scheme on the other hand, a robot would make its own decisions regarding the path it should take, and should also continuously perform its own critical assessment of the component: does it contain a defect or not, and with what probability? Has enough area been inspected, and are there critical areas that need more attention? The purpose of this paper is to formulate an algorithmic framework for such autonomous inspection.

This paper will focus on the particular use of ultrasound-based NDT, due to its current widespread use within engineering industry. However, the reader will recognise that the ideas presented are applicable to other types of testing. The proposed algorithm is demonstrated here using an example dataset, originating from a carbon fibre wing panel specimen with two known areas of de-lamination, acquired using a robotic system, which is described in detail in [1].

The rest of this introduction will provide a general introductory overview of the algorithm, with the following subsections providing a general background to ultrasound-based NDT, the challenges of autonomous inspection, as well as spatial sampling, which is closely related to the algorithm presented here.

A. Algorithm Overview

The first question to be asked is, what is the *objective* of the robot? Here, this objective shall be clearly defined as: determining, with high confidence, whether the component being inspected contains a defect, flaw or damage of a given minimum target size, ℓ. Given such a clear objective, it is then possible to cast the autonomous inspection problem, as an optimisation problem. After an observation is taken, the optimiser should decide what position (in two or three dimensional space), should an observation be taken next, so as to maximise the given objective. These decisions are thus carried out sequentially as information is being gathered. One method that deals particularly well with this sequential design problem is *Bayesian optimisation*, which will be covered in detail in Section III. The general idea behind it is to fit a Bayesian nonlinear regression model, such as a Gaussian Process (GP), to the objective function. This allows one to infer the objective at unobserved locations. The useful aspect of GP regression is that being a Bayesian method, it can not only predict the objective at unobserved

*This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC)

[1]University of Sheffield
[2]University of Strathclyde

locations, but also give an estimate of the uncertainty over this prediction. Armed with this, it is then possible to place observations in areas of both high uncertainty and with high values of the objective function. This strategy maximises the information gain of every single observation that is taken, it can be carried out sequentially and in an autonomous fashion.

For this type of scheme to work, a good objective function is required. As discussed above, the objective of the problem is to detect defects, so the objective should capture information regarding whether a particular location contains a defect or not. This task can be cast as a problem of *novelty detection*, which is commonly used as a method of damage detection in the field of data-driven Structural Health Monitoring (SHM) [3]. The basic idea behind performing novelty detection is to model the probability density of data that is known to belong to an un-damaged class, and to then compute a distance between new (unknown) observations and the centre of that probability density. When data is considered to be Gaussian, the relevant distance is the Mahalanobis squared-distance, and it provides one with a *novelty index*. One particularly useful aspect of the Mahalnobis squared-distance is that it can deal well with multivariate data, and this is useful in the general context of analysing NDT data.

In the problem of autonomous inspection, a novelty index provides an excellent objective for the optimisation scheme. It is desirable to maximise it: if the highest values of a novelty index are found, these correspond to the most abnormal observations, with respect to the rest.

So, combining Bayesian optimisation, with a novelty index such as a Mahalanobis squared-distance seems like a reasonable idea so far. However, a major caveat appears when one considers that the novelty detection must be computed both on-line (since it must guide a robot sequentially) and must also, for the purposes of capturing a reference un-damaged set, ignore abnormal observations. If outlying observations are taken into account for training of the reference set, the novelty detector will become biased towards them. This would in turn, result in potential damage sites being classed as benign.

To remedy this, the authors have turned to robust statistics, and in particular, the Minimum Covariance Determinant (MCD): a robust method for estimating the parameters of a Gaussian distribution. The MCD is discussed in detail in Section II. In this introduction, it is important to simply highlight that using a method that ignores outliers present in the reference set is paramount to the success of the autonomous inspection algorithm presented here, as it automatically discriminates between outlying and inlying observations when outliers are present in the training set, and thus not biasing the reference set toward the outliers. This enables an autonomous inspection algorithm that is entirely data-driven - it does not use external information to form a model of a healthy component.

An overview of the autonomous inspection scheme is illustrated in Figure 1. Step one involves data collection and feature extraction from raw data. Step two involves modelling these features using robust outlier analysis. Steps three and four form the Bayesian optimisation step, first modelling the objective function with a GP, and then choosing a location that maximises information gain.

A major assumption being made by this autonomous inspection strategy is that the physical area being scanned is nominally similar, in terms of material and geometrical characteristics. It is a reasonable assumption that normal variability in these characteristics plus any noise arising from their physical observation and computation of features, could be modelled as a (multivariate) Gaussian distribution. This implies modelling the data as uni-modal. If a component had two sections of different characteristics, this would add multiple modes to the underlying probability distribution of undamaged class features. In terms of pure novelty detection, this could still be dealt with via a clustering analysis, and appropriate novelty indices exist that capture this situation well [4]. However, in this scenario of actively looking for data that is dissimilar, how could one distinguish between a new data cluster from an undamaged component section, and a new data cluster arising from damage? In the absence of any prior knowledge about the location or characteristics of either the known undamaged sections or the expected damage, it is effectively impossible to tell these two apart. The key here is prior knowledge. It would be undesirable to impose prior knowledge over the damaged state, as the algorithm may simply fail to find any damage that does not "look like" the expected prior assumption. A better use of prior knowledge is to assume the robot knows which sections of a component are expected to be nominally the same. This can be achieved by performing independent inspections over areas that are known to be homogeneous in their material and geometrical properties. If a component happens to be homogeneous, then this presents no issue. If not, this can simply be achieved through elicitation of Computer Aided Design (CAD) models, which has already been applied to robotic path planning in NDT inspections [2]. In this paper, the emphasis is placed on the algorithmic processes required to achieve autonomous inspection, through the use of Bayesian optimisation and robust outlier analysis.

B. Ultrasound signal features

This paper will use example data from ultrasound-based NDT, due to its current widespread use in practice. For this reason, this section provides a brief background to this type of testing. The objective is to provide the reader with a clear understanding of the type of signals, and the features extracted from such signals, as well as the type of physical testing required to collect ultrasound data. Readers familiar with ultrasound-based NDT can safely skip this sub-section.

A typical ultrasound pulse is shown in Figure 2a, with two time indices marked as t_a and t_b. These times correspond to reflections from the front and back wall of the inspected specimen respectively. A pulse of this kind effectively constitutes an A-scan. The information extracted from this is the time difference $t_f = t_b - t_a$, and this is often referred to as the ultrasonic Time of Flight (ToF). This can be related to the thickness of the specimen, if the propagation

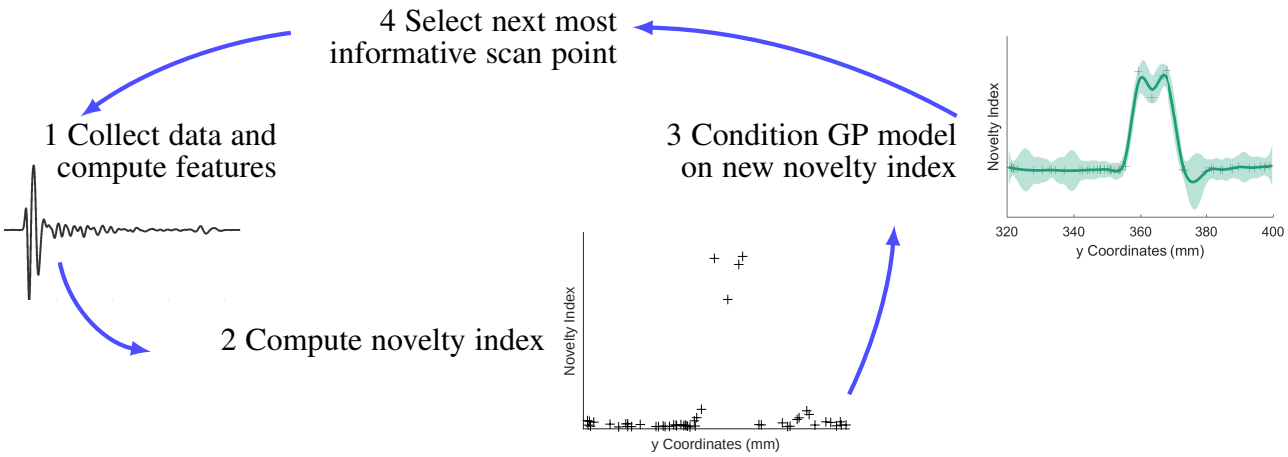

Fig. 1: Overview of the steps involved in the autonomous inspection algorithm.

speed of bulk waves for the material is known. Another feature of interest is the ratio $x(t_a)/x(t_b)$ (where $x(t)$ is the measured amplitude of the ultrasound pulse), as this contains information about the attenuation of the wave as it travelled through the thickness of the component. An A-scan thus gives information about a single physical coordinate on a surface.

A B-scan can be formed by collecting a series of A-scans along a line (illustrated in Figure 2b), while a C-scan is formed by collecting a series of B-scans, to give a two-dimensional grid of ultrasound pulse information (illustrated in Figure 2c). Note that higher times of flight in Figure 2c imply wider plate thickness. The salient features in Figure 2c are the stringers and the variable thickness of the wing panel. The stringers are evident by a low ToF index, due to their thickness being higher than the maximum value that can be captured within the collected data. There are two regions of delamination evident in Figure 2c.

C. Towards practical autonomous inspection

In practice, these A, B and C-scans are often performed manually, with the aid of engineering expertise, in order to carry out three key tasks:

1) Physically scan the structure with an ultrasound probe.
2) Interpret the results of the collected dataset and form a decision over the quality of the data, and the state of the component.
3) Decide if any which areas of the component need re-scanning.
4) Provide an assessment of the best course of action, and recommend an appropriate repair strategy.

Recently, there has been a move towards an increased use of robotics for automating the process of acquiring the necesary data, where ultrasound data acquisition hardware is combined with robotic interfaces in order to automate the physical scanning of the structure; task one, in the hierarchy above. Examples of this can be found in [5]–[7] and more recently in [1], [2], where problems such as path planning and integrating Computer Aided Design (CAD) geometries into the scanning process have been addressed. Other important issues, such as dealing with uncertainty in the positioning of autonomous robots, have been addressed in [8] using Bayesian filtering methods.

The interpretation of results, the second task in the hierarchy, has received a noteworthy amount of attention in the NDT research community. Research has evolved from investigation into the required signal processing methods to extract useful features from the raw data [9]–[11], and into the use of machine learning and novel statistical inference methods applied to those features in order to detect, cluster and classify defects [12]–[16].

The third task in the hierarchy involves deciding appropriate places to scan, based on information collected so far about the component. The continuous path planning problem using optimisation has been investigated before in a general context [17], but significantly less research has been carried out towards this end in NDT. The state of the art in path planning for robotic ultrasound inspection, at the moment, is to use CAD models to plan the scanning trajectory, where research has focused around issues with the feedback control of the ultrasonic probe head [2].

The objective of performing a scan is to find defects that are larger than a minimum given size. The current solution in pulse-echo ultrasound measurements is to define a spatial grid with a high enough resolution to capture this minimum expected flaw size. Whilst this approach has worked so far, there are two major issues with it,

1) The scan can potentially take a long time, and there are also cost implications associated with this.
2) A large quantity of data needs to be stored and post-processed later.

These effects are exacerbated when large scanning areas are involved, as is typical in aerospace industries. Depending on the motivation for carrying out the inspection, the scan time may not present an issue; if inspecting a high-value, safety critical engineering component, large scanning times are acceptable. However, the data storage and processing requirements of a dense grid over a large area can make

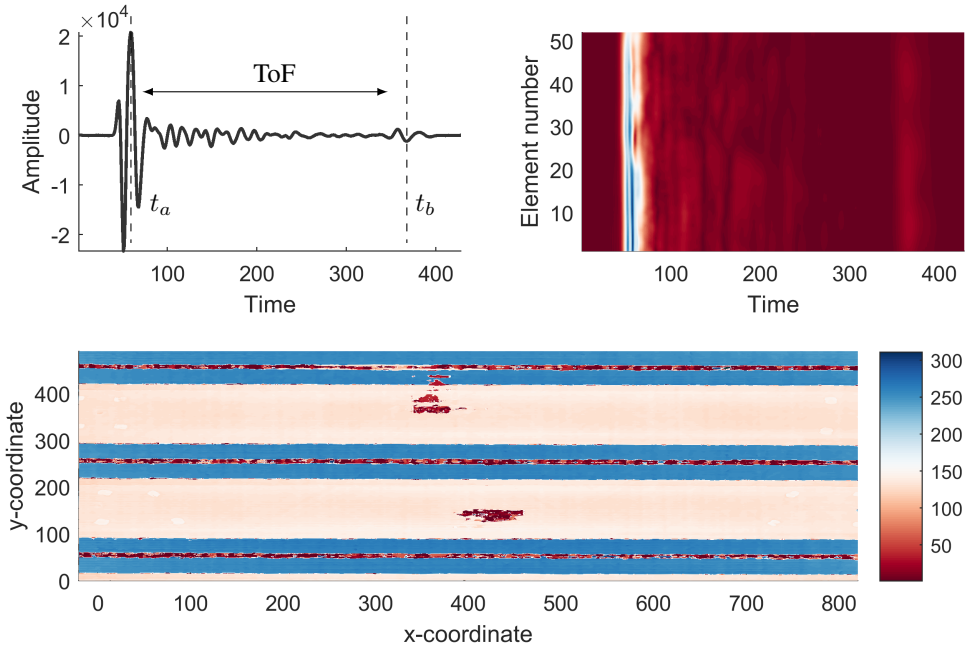

Fig. 2: Illustration of a) single ultrasound reflection (A-scan), b) reflections along a phased array probe (B-scan) and c) time of flight map along the two dimensions of a composite plate (C-scan).

the problem intractable.

The data storage requirement can be alleviated by compressing the raw data for later post-processing, leveraging advanced signal processing techniques such as wavelet transforms [18], sparse signal modelling [19], and more recently compressive sensing [20].

D. Spatial sampling

Evidently, efforts in developing compression algorithms for time domain data will do little to address the problem of dealing with large quantities of data in the spatial domain. The question being asked here is: what is the minimum number of observations required to capture the statistics of the data collected from the component, in the spatial domain? This spatial sampling problem has been studied for centuries, though a recent twenty-first century review can be found in [21]. The most active research communities in this field are those of geo-statistical modelling, environmental modelling [22], and disease mapping [23], where the motivations for studying spatial sampling schemes arise from the fact that these measurements can be sparse, which raises the need to interpolate data in regions with low observation density, while quantifying the uncertainty around those predictions.

In spatial statistics, the features of interest are treated as random variables across space; they can be separated into two categories, those which can be assumed to be identically and independently distributed (iid), and those that are autocorrelated across space. Intuitively, NDT data falls within the second category, where values of NDT data features will tend to be similar to those features observed a small distance away (there is some experimental backing to this notion [24]).

Modelling the spatial correlation of stochastic processes has been central to the task of interpolation in spatial analysis [25]. Spatial correlation quantifies the influence that a variable has on its own value at a distance r away. In spatial statistical analysis, this influence is modelled through the use of a covariance, or correlation function. The quality of spatial interpolation, based on correlations or covariates, therefore relies heavily on the properties of the covariance function used. The covariance function should encode the overall smoothness of the random field being modelled, as well as the overall length-scale over which this random field influences itself. Analysis of spatial variation has been of interest in NDT, and has been used as a tool to assess optimal resolution and scanning strategies [24], [26].

Under the general context of spatial data interpolation, when the application is risk-sensitive which is the case in NDT data analysis, it is critical to use methods that are probabilistic. This idea has led to the use of Gaussian Process (GP) regression, a flexible tool for performing probabilistic, nonlinear and nonparametric regression. The use of the GP is central to the modelling of spatial data in this paper, and it it will be discussed in more detail in Section III-C. Here, it is worth noting that this method relies on the use of covariance functions to model variation in space (or any other coordinate in fact). In [26], for example, GPs are used to replace the more tradition empirical variograms, when smoothing NDT data features across space. One interesting application of spatial modelling to outlier analysis is presented in [27] where transformations to the standard formulation of the GP were used to derive "extreme spatial processes", based on standard extreme value distributions. In the present paper, the flexibility of the GP is used to fit a surrogate model to

novelty scores from a robust outlier analysis procedure. This approach is novel compared to other work in this area, where GPs are often used to fit data features directly.

One important aspect of spatial variation and NDT data analysis is that of *Probability of Detection* (PoD). This describes the likelihood that an object will be detected in a particular measurement scenario. Its origin can be traced back to the development of radar technology, but in the present day it is widely used to quantify the reliability of NDT processes [28]. PoD curves describe the probability that a particular NDT system will detect a flaw as a function of the flaw size. Traditionally, these curves are built experimentally, by performing blind scans of calibration pieces with flaws of various sizes, and quantifying the false negative rates [29]. PoD is affected by a number of factors: measurement noise, spatial scanning resolution and the scattering characteristics of the expected flaw with respect to the ultrasound wave. Evidently, the spatial resolution of a scan will affect the overall probability of detection of the procedure. The problem of evaluating PoD curves from spatial random fields that exhibit autocorrelation has been investigated in [30]. This is interesting from the point of view of the work presented in the present paper, as it links the idea of the surrogate model to the computation of the more classical assessment of reliability of the NDT process based on PoD. While in [30], the surrogate being used is Polynomial Chaos model, this paper models spatial variation using a GP; the analysis of PoD, however, would follow similar lines to those of [30].

E. Outline of the paper

The focus of this paper is given to the development of an autonomous scheme to perform a search for outliers over a spatial field. As explained in Section I-A, and as illustrated in Figure 1, the autonomous inspection algorithm consists of two main ingredients: 1) robust outlier analysis, and 2) Bayesian optimisation. The details of each individual method matter in terms of understanding the full algorithm, so for the benefit of the reader, they are both given thorough discussions individually. Section II covers the details of robust outlier analysis, while III will discuss Bayesian optimisation. Their use in the context of the autonomous inspection algorithm is discussed in Section IV where issues such as initialisation, convergence and adjustment of risk according to expected minimum flaw sizes are discussed. Throughout the paper, illustrations of the various aspects of the procedure are drawn from ultrasound data collected from an aerospace panel. These will be subsets of the data illustrated in Figure 2. Section V provides a more detailed experimental investigation of the autonomous inspection algorithm, using this data set, while Section VI will present some and conclusions directions of future work.

II. ROBUST OUTLIER ANALYSIS

The process of outlier analysis often involves the evaluation of a distance metric, that measures how far away any given observation lies from the centre of the data mass, relative to its spread. In the case of the Gaussian-distributed data assumption, the relavant metric is the Mahalanobis Squared-Distance (MSD),

$$d_i = (\mathbf{y}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \tag{1}$$

where the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ are the mean and variance, respectively; the two parameters that characterise a Gaussian density. The multivariate observation vector $\mathbf{y}$ represents the data features. Novelty detection with an MSD involves two steps. The first is training, and involves estimating the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ from a training data set derived from a normal (un-damaged) condition set. Part of the first step also involves the estimation of a novelty threshold for the novelty indexes, T, above which observations are to be classed as novel. The second step is prediction, and involves evaluating a novelty index at new observations, and determining whether they fall under the novelty threshold. This outlier analysis procedure has been extensively used for the purposes of data-driven Structural Health Monitoring (SHM) [3].

The top-level requirement in the data-driven autonomous inspection scheme is for the system to make a decision about whether the component contains outliers or not, based only on the data collected on that component alone. The implication of this is that outlying observations can potentially appear in the training set used for estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. This problem is well-suited for the application of *robust outlier analysis*, where robust measures of the mean and covariance of the data are estimated, which ignore outlying observations when computing these Gaussian distribution parameters. Outlier analysis based on robust estimates of location and scatter in multivariate data has recently seen applications in the field of SHM [31], which, in terms of statistical inference, is a closely related problem of NDE inspection. This paper thus closely follows the ideas presented in [31].

A. Minimum Covariance Determinant Estimators

Two key ideas exist in the field of robust estimation of Gaussian distribution parameters: the Minimum Volume Ellipse (MVE) [32], [33] and the Minimum Covariance Determinant (MCD) [34] (see [35] for a recent review). MVE schemes seek to find an ellipsoid of minimum volume, which encloses at least half of the total number of data points. On the other hand, MCD schemes seek to find a covariance matrix $\boldsymbol{\Sigma}$ with the smallest possible determinant that leaves out h observations from the training set, where h is the number of outliers. One of the advantages of the MCD over the MVE is that the MCD is affine equivariant; location and scale parameters will remain constant under rescaling and affine transformations of the data. Furthermore, the estimates of $\boldsymbol{\mu}$ under MCD schemes have been shown to be asymptotically normal [36]. Beyond these theoretical guarantees, one useful practical advantage comes from the existence of an efficient way to compute MCD parameter estimates: the *Fast-MCD* algorithm [37]. Exact MCD estimates are numerically expensive to compute, requiring an evaluation of $n \times h$ subsets of size h. Fast-MCD uses an

iterative scheme in order to achieve an MCD estimate with efficient computation. The Fast-MCD algorithm is relatively simple, and is described in Algorithm 1. The key step is to iteratively form h-subsets of the data matrix $\mathbf{Y}$, which contain values of low Mahalanobis distance (computed using equation (1)) against the current estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$. The full algorithm is relatively simple, and it is described in Algorithm 1.

Algorithm 1 Fast-MCD Algorithm

procedure FAST MCD($\mathbf{Y}, h$)
  set $\boldsymbol{\pi}$ to be a random sample of h indices of $\mathbf{Y}$
  $\mathbf{H}_1 \leftarrow \mathbf{Y}(\boldsymbol{\pi})$ ▷ form an h-subset of the data
  compute $\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1$ from mean and covariance of $\mathbf{H}_1$
  while $\det(\boldsymbol{\Sigma}_k) \neq \det(\boldsymbol{\Sigma}_{k-1})$ **do**
    $d \leftarrow mahal(\mathbf{Y}, \boldsymbol{\mu}_{k-1}, \boldsymbol{\Sigma}_{k-1}))$
    sort d in ascending order, store indices in $\boldsymbol{\pi}$
    $\boldsymbol{\pi} \leftarrow \boldsymbol{\pi}(1:h)$ ▷ Discard $n-h$ largest distances
    $\mathbf{H}_k \leftarrow \mathbf{Y}(\boldsymbol{\pi})$
    compute $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ from mean and covariance of $\mathbf{H}_k$
    $k \leftarrow k+1$
  end while
  return $\boldsymbol{\Sigma}_k, \boldsymbol{\mu}_k$
end procedure

While the Fast-MCD algorithm is naturally greedy, it does enjoy the theoretical guarantee of either diminishing the determinant of the covariance, $\det(\boldsymbol{\Sigma})$, or leaving it the same, at every iteration [37]. This is why the stopping criteria for the algorithm is such that it continues to run as long as $\det(\boldsymbol{\Sigma}_k) \neq \det(\boldsymbol{\Sigma}_{k-1})$.

An application of the resulting covariance estimation of the Fast-MCD algorithm is given in Figure 3 for an illustrative synthetic data set. The data being fitted are samples from a two-dimensional Gaussian random variable, contaminated with outliers (on the top right). The original distribution of the un-contaminated data is shown in light-grey, the maximum-likelihood covariance is plotted with a dashed line, while the Fast-MCD fit is shown with a solid line. The bias that the outliers introduce to the covariance is clear: they increase the scatter. The robust covariance on the other hand gives a better approximation to the original density.

B. Application to NDT data

The computation of Mahalanobis distances, using robust estimates of the data density, is central to the scanning procedure presented in this paper, which combines Bayesian optimisation with outlier analysis. The application of robust outlier analysis to NDT data will be briefly discussed here, in order to make clearer the application of Bayesian optimisation, presented in Section III, to the problem of outlier analysis.

The goal is to assemble a mean and covariance of a reference, undamaged condition, even when there are outliers present on the data, which as discussed above, is often the case.

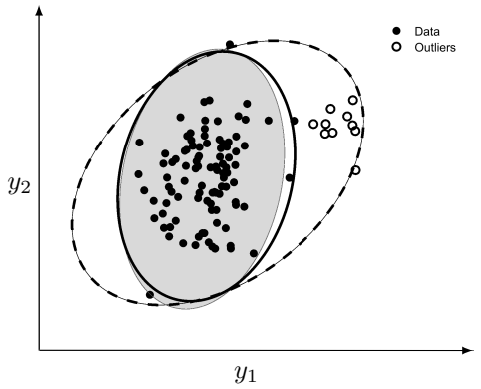

Fig. 3: Illustration of robust estimation of Gaussian distribution parameters, in the presence of inclusive outliers.

Two common features used in NDT data analysis are the Time of Flight (ToF) and attenuation, as discussed in Section I-B. Here, these will be used to demonstrate the idea of novelty detection, using robust measures. In order to simplify the problem, the process is demonstrated as a one-dimensional problem. For this, data from a cross section along the y-axis of the aerospace panel shown in Figure 2 has been used [1]. The ToF, together with the robust estimates (using Fast-MCD) for its mean and variance are shown in Figure 4a. This cross-section was selected as it contains two small regions where delamination of the carbon fibre composite has occurred; these are evident as areas of much lower ToF compared to the rest of the specimen, due to the reflection of the wave from the delaminated section. On the bottom, Figure 4b shows the resulting robust Mahalanobis distance for all of the observations. The red horizontal line shows the threshold, which has been defined as the Mahalanobis distance of the 3σ values of $\mathbf{Y}$ (just the ToF in this particualr case). Any observations with a Mahalanobis distance above this threshold are defined as abnormal, compared to the rest of the data. The two regions where there is delamination have clear excursions above the Mahalanobis distance threshold. There are other outliers visible on the left side of the plot; these are not known damaged locations, so it is likely that they have arisen from electrical noise or bad contact between the ultrasound probe and the specimen.

In any case, one would like to know about any observations that lie far from the bulk of the data, in order to later be able to assert whether those observations are a genuine flaw or defect, or simple noise. The next section describes the strategy for spatially scanning a component based on this robust Mahalanobis distance.

III. BAYESIAN OPTIMISATION

This section will describe the Bayesian optimisation scheme used to select appropriate scanning points, based on the robust outlier analysis results. Bayesian optimisation is a relatively advanced technique, used primarily in

[1]The actual coordinates used in this example (with respect to Figure 2), were constrained to $401 > y > 400$ and $400 > x > 320$

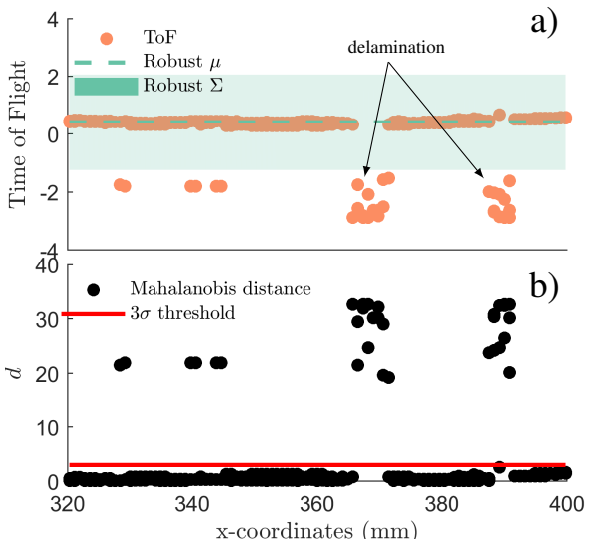

Fig. 4: Illustration of outlier analysis on ultrasound-NDT data, using time-of-flight as a feature, and a robust mean and variance to compute a Mahalanobis distance.

situations where it is expensive to evaluate the function being optimised, so a small number of function evaluations are required. This is the case in the field of Design of Experiments (DoE). This optimisation technique, the authors find, tackles the problem of searching for spatial outliers rather well. This section outlines the necessary background theory for Bayesian optimisation, and this includes a brief introduction to the underlying nonlinear regression scheme used within it: the Gaussian Process. However, before delving into mathematical details, the motivation and reasoning for selecting this type of optimisation algorithm over the multitude of others will be described first, in Section III-A. The following sub-sections will then provide a background on general Bayesian optimisation, and on Gaussian Process regression.

A. *Reasoning and motivation*

So far, the discussion has centred mostly around performing outlier analysis on NDT data, and specifically on using robust estimates of the location and scatter of key features extracted from the raw data. These enable the computation of robust novelty indices, which separate the outliers from the bulk of the observations. The objective is to generate an estimate of whether the component being scanned contains anomalies, using the most efficient strategy in terms of both time spent scanning and data storage. These two requirements both translate to scanning the minimum number of points in order to maximise the confidence in the estimate of whether the component contains anomalies. Armed with a method for separating outliers from the bulk of the observed data, the question is now: what are the optimal points to scan, or "query", in order to achieve an estimate of the condition of the component, to within a specified confidence level? In this paper, this question is addressed as a problem of optimisation.

In general, the optimisation task seeks to find the arguments of a function that maximise (or minimise) an objective, or cost, function. If at the end of a scan, one analyses the data and finds that the greatest novelty score falls within a novelty threshold, then confidence can be established over the absence of anomalies. If, on the other hand, one, or a significant number of observations lie above the novelty threshold, this indicates a problem. The task of inspection is thus a task of searching for the regions that are most different from the rest; that contain the highest novelty score. Therefore, the problem of searching for outliers in a spatial domain can be cast as an optimisation problem, where the objective, or cost function being maximised is the novelty score.

Optimisation is used to solve problems across the realm of disciplines. In science and engineering, one of its most popular uses is arguably that of estimating the parameters of a model, be it a statistical or a physical one. Gradient descent algorithms have been a popular solution to the parameter learning problem since the day models started to be fit to data. The general idea is that if a search for a better parameter is carried out in the direction of greatest change of the objective function, it will lead to the argument that maximises it. The problem is analogous to finding the peak of a mountain by walking in the uphill direction. The underlying assumption in gradient-based methods is that the parameters will cause smooth changes to the objective function, with a positive gradient towards the optimum. This may be true in some cases, but certainly not all. Gradient-based algorithms are not suitable for the problem at hand; there is no reason to believe that selecting the direction of greatest change in novelty score will lead to finding the location of greatest damage. The spatial location of damage can be discontinuous in nature, and so may not not necessarily cause a smooth increase in novelty score as the optimiser approaches it. Therefore, this problem requires some form of gradient-free global optimisation approach.

Because of the nature of the problem of inspection, it is desireable for the optimisation method to be able to quantify uncertainty on its predictions. This would allow the user to tune the search according to the risk he or she is willing to accept in a given inspection problem. Bayesian techniques are well suited to the task of global optimisation where a quantification of uncertainty is required. The underlying principle of Bayesian inference is to solve for the probability of the quantity of interest, $\boldsymbol{\theta}$, given some data, $\boldsymbol{\mathcal{D}}$. The term $\boldsymbol{\theta}$ is used here as a general parameter vector, which could describe, for example, a defect's position, size, shape, orientation, and other parameters. Using a combination of the sum and product rules of probability leads to Bayes' theorem, which provides a formulation for the probability distribution of $\boldsymbol{\theta}$ given $\boldsymbol{\mathcal{D}}$, or $p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$,

$$p(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}) = \frac{p(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{\mathcal{D}})}. \tag{2}$$

There are three probabilities on the right-hand side of equation (2): the prior, the likelihood and the marginal. The prior, $p(\boldsymbol{\theta})$, should represent a prior belief about the process

before it is observed. The likelihood, $p(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$ represents the distribution of the model error, with respect to the parameters. Finally, the marginal, $p(\boldsymbol{\mathcal{D}})$, can be expanded using the sum rule of probability to yield the following integral,

$$P(\mathbf{Y}) = \int_{-\infty}^{\infty} p(\mathbf{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{3}$$

which sums the product of the prior and likelihood (often called the marginal), over all possible parameter values $\boldsymbol{\theta}$. This is often an intractable integral, with no closed form solution available. The solution of the marginal integral often leads to one of two paths: approximations, or sampling schemes. The Laplace approximation, and variational inference lie in the approximation paths, while Markov Chain Monte Carlo (MCMC) methods lie in the sampling path. Sampling techniques present one of the most robust ways of performing Bayesian inference, but also one of the most expensive. In the context of optimising efficient search paths, given some underlying physical problem (ultrasound waves in this case), MCMC methods are well suited to the case when a suitable forward-model of the problem exists, that takes $\boldsymbol{\theta}$ as a input, and simulates the observations. In that setting, MCMC is particularly well suited for the problem of deriving a posterior over $\boldsymbol{\theta}$. In fact, this approach has previously been investigated for ultrasound test parameter optimisation [38]. This is a related problem to that of an efficient search path, with the key difference that the aim here is to achieve this both sequentially, and without the aid of a forward-model. MCMC generally does not satisfy these two requirements.

Additionally, it is necessary for the search algorithm to yield an answer as to the state of the component with a low number of observations. This requirement has led to the development of Bayesian approaches to global optimisation, that use the idea of fitting a surrogate statistical model to the function being maximised [39], [40]. The motivation has arisen both from the side of conducting efficient optimisations when using expensive computer simulation models, and when collecting experimental data, as well as the calibration between these two. The surrogate model approach is particularly well suited for optimising over models for which no analytic expression can be written down (black-box models), and where its derivatives are not available. This is true for the novelty scores across the surface of a component: there is no closed-form expression for them, as they are completely dependent on the data, and as discussed above, computation of their derivatives would help little towards finding the point of highest novelty score.

The surrogate models used to fit the function being maximised are typically nonlinear and *GPs probabilistic regression models*. Allowing the model to be GPs gives flexibility as to the shape of functions that can be modelled, while a probabilistic model implies that it makes predictions over the whole probability distribution of the model outputs, so it effectively quantifies the uncertainty in the model predictions. The particular flavour of probabilistic regression used within Bayesian optimisation is the Gaussian Process (GP) model.

The use of probabilistic regression models, and the GP in particular, is particularly suited to the problem at hand. It can be used to leverage some of the features that physical damage can exhibit, encoded into the model through the Bayesian formalism of the prior. One of the key aspects of physical damage is that it is usually localised in space, and this, in turn, results in a spatial correlation of NDT data features. A single outlying data point surrounded by data with low novelty scores should be discarded as a noisy observation. On the other hand, a cluster of outliers is indicative of a problem area. The prior belief that a surrogate model of the novelty scores can encode is: how much is the novelty score expected to change, within a given distance, x? The GP model can encode this information in a fairly straight-forward manner. The necessary background for GP regression will be discussed in Section III-C.

B. Bayesian Optimisation Background

Bayesian optimisation solves the sequential design problem: given all the observed data so far, $\mathcal{D}_{1:t}$ [2], what is the optimal point to observe next such that maximum information gain is achieved? The information gain is estimated directly from the surrogate model, by combining the prior and the likelihood of equation (2) to form a posterior distribution over the objective function, denoted here as f, given the data observed so far. However, this does not involve a parametric model any more; the parameter vector, $\boldsymbol{\theta}$, is thus replaced with all the observed data so far, $\mathcal{D}_{1:t}$. This includes both the arguments $\mathbf{x}$ and observations $\mathbf{y}$, so $\mathcal{D}_{1:t} = \{\mathbf{y}_t, \mathbf{x}_t\}$. This posterior can be written down as,

$$p(f|\mathcal{D}_{1:t}) \propto p(f)p(\mathcal{D}_{1:t}|f) \tag{4}$$

This posterior distribution contains useful information about where it is best to place the next observation, in terms of both areas of high uncertainty as well as areas of high values of f. The sequential search needs to balance between the tasks of exploration and exploitation. Exploration involves searching in areas of high uncertainty, whilst exploitation involves placing observations in areas of high function values, in order to narrow down to the values of $\mathbf{x}$ that lead to globally maximising f. In Bayesian optimisation, the balance between exploration and exploitation is achieved through the use of an acquisition function, that evaluates an expected utility $u(\mathbf{x})$.

The general algorithm for performing Bayesian optimisation is relatively straight-forward, and is described in Algorithm 2. It involves iterations over evaluations of the posterior over the objective, $p(f|\mathcal{D}_t)$ in order to find the points that maximise a utility function $u(\mathbf{x})$, and acquisitions of new data at those points.

Mockus established that a Bayesian optimisation technique will converge to the optimum if [41], [42]:

1) The acquisition function is continuous and approximately minimises the risk.

[2]Note that the subscript t in $\mathcal{D}_t$, and other relevant variables is used here to denoted observations *up to* t

Algorithm 2 Bayesian Optimisation Algorithm

for $t = 1, 2, ...$ **do**
  $\mathbf{x}_t \leftarrow \arg\max_{\mathbf{x}} u(\mathbf{x}|\mathcal{D}_{1:t-1})$ ▷ Maximise utility
  $y_t \leftarrow f(\mathbf{x}_t) + \varepsilon$ ▷ Sample the objective
  $\mathcal{D}_{1:t} \leftarrow \{\mathcal{D}, (\mathbf{x}_t, y_t)\}$ ▷ Include new observations
end for

2) The conditional variance converges to zero, or to an appropriate measurement noise floor.

where the risk is defined as the discrepancy between the inferred maximum and the true maximum. The continuity condition is clearly satisfied in the outlier search problem, given that the spatial field over which novelty scores are sampled is continuous. The second condition relates to the variance of the posterior distribution, $p(f|\mathcal{D})$, and it states the intuitive notion that once the variance of $p(f|\mathcal{D})$ reaches the variance of the true objective, no further improvement can be achieved in the optimisation. This criteria is particularly important in this sequential design context; it answers the question: when should one stop searching? In practice, one may actually want to stop the search once the variance reaches a user-set confidence level, but this upper bound on possible improvement is clearly necessary to perform efficient, yet still thorough searches.

Having now discussed the general framework of Bayesian optimisation from a general point of view, two details need discussion: the computation of both $p(f|\mathcal{D})$ and acquisition functions. These are discussed in Sections III-C and III-E respectively.

C. Gaussian Process regression

This section discusses the estimation of the posterior density over the objective function given past observations, through the use of Gaussian Process (GP) regression. The GP model is one of the most widespread forms of nonparametric and probabilistic regression models, both in machine learning and statistical inference as well as in the specific field of Bayesian optimisation. The use of GPs in the context of Bayesian optimisation dates back to O'Hagan [43] and later Zilinskas [44], as a better alternative to quadratic estimation of cost functions in the case where it may be multi-modal.

At this point, it is now necessary to formally define the GP, and to discuss the computation of the posterior $p(f|\mathcal{D})$. In the same way that a Gaussian distribution fully describes the distribution over a random variable $\mathbf{y}$, using a mean and a variance, a Gaussian process defines a probability distribution over a function $f(\mathbf{x})$, using a mean $m(\mathbf{x})$, and a covariance function $k(\mathbf{x}, \mathbf{x}')$,

$$f(\mathbf{x}) = \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{5}$$

The GP could be viewed as an extension of a Gaussian distribution to the infinite-dimensional space of functions. This infinite-dimensional space represents a stochastic process, with a covariance represented by $k(\mathbf{x}, \mathbf{x}')$. In this space, any finite combination of dimensions can be represented by a Gaussian distribution, which will be defined by a covariance matrix $\mathbf{K}$, whose entries are defined by an evaluation of the covariance function on a finite set of points on $\mathbf{x}$, $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. This link between an intangible infinite-dimensional space described by the covariance function, and the practical and tangible space of finite dimensions described by the covariance matrix, is a fairly powerful idea that allows for the practical computation of probabilities over functions. A GP is fully defined by both its mean $m(\mathbf{x})$, and covariance function $k(\mathbf{x}, \mathbf{x}')$, although in practice the mean function is often set to zero. This works well provided one removes the mean of the data being modelled, and this shall be assumed in the discussions that follow.

The covariance function plays a central role in the predictive process of the GP, as it captures the influence that any given point in the space of $\mathbf{x}$ has on all the other points. In other words, the covariance function models the fact that given a value pair for (x_1, y_1), other values of y at nearby values of x are likely to be similar to y_1. Exactly how similar, depends on the specific form of the covariance function. Two of these will be discussed later in Section III-D.

Now that the GP has been defined, the next step is to show how one could use it to make predictions of the posterior probability over the function given all of the observed data so far. Observed data so far are represented by $\mathcal{D}_t$[3], while the candidate test locations, for which the posterior probability will be computed, are denoted as $\mathbf{x}^\star$. Likewise, the function values at those candidate locations are denoted as $f^\star$. Note that all observations are assumed to be contaminated by Gaussian noise, so that one never directly observes the actual function values, only the contaminated version:

$$\mathbf{y} = f(\mathbf{x}) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma_n) \tag{6}$$

where $\mathcal{N}(0, \sigma_n^2)$ denotes a Gaussian distribution with zero mean and variance σ_n^2.

The predictive distribution of the GP conditioned on observations up to t is defined as a Gaussian distribution, with mean $\mathbf{m}_t$ and covariance $\mathbf{v}_t$ given by [45],

$$\mathbf{m}_t = K(\mathbf{x}^\star, \mathbf{x}_t)\mathbf{P}^{-1}\mathbf{y} \tag{7}$$

$$\mathbf{v}_t = K(\mathbf{x}^\star, \mathbf{x}^\star) - K(\mathbf{x}^\star, \mathbf{x}_t)\mathbf{P}^{-1}K(\mathbf{x}_t, \mathbf{x}^\star) \tag{8}$$

where the matrix $\mathbf{P}$ is defined as:

$$\mathbf{P} = K(\mathbf{x}_t, \mathbf{x}_t) + \sigma_n^2\mathbf{I} \tag{9}$$

This predictive mean and covariance fully define the posterior predictive distribution of the GP, and thus have great practical value when performing Bayesian optimisation, and subsequently autonomous inspection.

D. Covariance functions

As alluded to before, the covariance function plays an important role in defining the region of influence between points in the input space, $\mathbf{x}$. Covariance between pairs of observations can be represented by the pairwise distance between observations:

$$\mathbf{r} = |\mathbf{x} - \mathbf{x}'| \tag{10}$$

[3]Note that this shorter notation will be used from here on, for clarity of presentation

In his seminal work investigating the issue of spatial variation, applied to the problem of forest surveying, Matérn developed a family of functions that are suitable for modelling spatial random fields. The family of correlation functions he developed encode the ability to control both the smoothness and the lengthscale of the random field through turning parameters v and k. The functional form of this family is given by

$$k_{\text{Mat}}(\mathbf{r}) = \sigma_f^2 \frac{2^{1-v}}{\Gamma(r)} \left(\frac{\sqrt{2v}\mathbf{r}}{\ell} \right)^v \mathcal{K}_v \left(\frac{\sqrt{2v}\mathbf{r}}{\ell} \right) \tag{11}$$

where $\mathcal{K}_v$ is a modified Bessel function of order v, and ℓ and v are scale and smoothness parameters respectively. The process variance is defined as the scaling parameter σ_f^2. There are certain properties of the Matérn family that make it very useful. One of these is that it is $(v-1)$-times mean-square differentiable, which relates to its ability to model both smooth and non-smooth processes. This is in comparison to the often used squared-exponential covariance function in spatial interpolation applications, which is *infinitely* mean-square differentiable and thus imposes too much smoothness to characterise real-world processes [25]. One of the interesting points of the Matérn class is the recurring theme of its ability to describe physical processes that can be described by differential equations.

Whilst there is no appropriate differential equation to describe the spatial correlation of novelty scores across the spatial field of NDT data, this could be appropriately modelled as a random field, with the characteristic that it will generate extreme, non-smooth changes in the boundary between the normal condition and damage. The Matérn famility is thus well suited to model this problem. The particular flavour that will be used in this paper is the case where $v = \frac{3}{2}$, which is given by,

$$k_{\text{Mat3/2}}(\mathbf{r}) = \sigma_f^2 \left(1 + \frac{\sqrt{3}\mathbf{r}}{\ell} \right) \exp\left(-\frac{\sqrt{3}\mathbf{r}}{\ell} \right) \tag{12}$$

Figure 5 illustrates GP predictions with the same data set used in the robust outlier illustration of Figure 3. For comparison purposes, the GP predictions are shown using two different covariance functions; the Matérn $3/2$ of equation (12) and the squared-exponential kernel which is popular and often the first choice of kernel when modelling smooth processes [45].

Figure 5 zooms into a region around the boundary between an undamaged and damaged condition. Whilst outside this region, the data may be represented well as a smooth process, the boundary introduces a discontinuity. The purpose is to illustrate how the Matérn kernels deal well with this step change, as well as with the smooth process in non-damaged regions. The length-scales in this case were fixed to $\ell = 10$ and the noise variance was set $\sigma_n^2 = 10$, for this particular example. It is clear that the squared-exponential kernel imposes too much smoothness to the data, whereas the Matérn $3/2$ kernel is far superior at modelling the step between undamaged and damaged conditions.

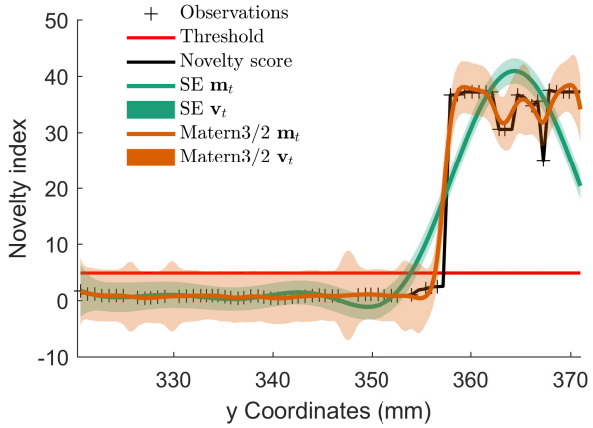

Fig. 5: Illustration of GP posterior fit to novelty scores across a one-dimensional slice of ultrasound NDT data, comparing the use of squared-exponential and Matérn $3/2$ covariance functions.

E. Acquisition functions

Having discussed how to compute the posterior probability over the objective, represented by $\mathbf{m}_t$ and $\mathbf{v}_t$, the last step is to use these to guide the search in an intelligent manner, that balances between exploring areas of high uncertainty and exploiting regions of high function values, by assembling an *acquisition function*. Maximising the acquisition function should lead to the next best point in the space of $\mathbf{x}$ to place an observation at. Thus, an acquisition function is designed to be high in areas of high uncertainty, $\mathbf{v}_t$, high estimates $\mathbf{m}_t$, or both. In the discussion that follows, the highest observed objective will be denoted as $\mathbf{y}_t^+$, while the highest point of the GP mean (with observations up to t) will be represented as $\mathbf{m}_t^+$.

Early work in Bayesian optimisation focused around the use of probability of improvement (PI) [46], over the current best point, denoted as $\mathbf{x}^+$, that previously maximised the acquisition function. PI is defined as the probability that $f(\mathbf{x})$ is greater than $f(\mathbf{x}^+)$,

$$\begin{aligned} PI(\mathbf{x}) &= p(f(\mathbf{x}) > f(\mathbf{x}^+)) \\ &= \Phi\left(\frac{\mathbf{m}_t - f(\mathbf{x}^+)}{\mathbf{v}_t} \right) \end{aligned} \tag{13}$$

where Φ represents a normal cumulative distribution function, and as before, $\mathbf{m}_t$ and $\mathbf{v}_t$ represent the mean and variance predictions of the GP. The drawback of the PI is that it performs pure exploitation and no exploration. Once the search finds areas of $\mathbf{x}$ with high values, PI will choose points around that region. A tuning parameter can be introduced in order to force PI to explore the space more, but this can be considered to be more of a hack that yields optimisation algorithms that are "extremely sensitive" to the choice of tuning parameter [47].

Expected Improvement (EI) [48] provides a better alternative to PI; it takes into account the probability of improvement, but also accounts for the expected magnitude of improvement that candidate points would give. PI is defined as the quantity,

$$I(\mathbf{x}) = \max(f(\mathbf{x}^+) - f(\mathbf{x}^\star), 0) \tag{14}$$

The expected improvement is simply the expectation over the improvement function given above, $E[I(\mathbf{x})]$, which, under the case of a Gaussian density assumption, and using integration by parts yields [48],

$$E[I(\mathbf{x})] = (f^+ - \mathbf{m}_t)\Phi(\mathbf{z}) + \mathbf{v}_t\phi(\mathbf{z}) \qquad \mathbf{z} = \frac{(f^+ - \mathbf{m}_t)}{\mathbf{v}_t} \tag{15}$$

where ϕ and Φ denote the probability density and cumulative probability density functions of a normal Gaussian distribution, respectively.

Other acquistion functions exist for the Bayesian optimisation problem, ranging from the simple, such as the upper/lower confidence bound method [49], and more recently developed entropy search methods [50]. In this paper, EI will be used as an acquisition function within the Bayesian optimisation framework. While other methods exist, the justification for this choice in this case is given by the balance between exploration and exploitation that EI provides, which is sufficient for the problem at hand. There is no reason other acquisition functions could not be used within the framework being presented. However, a full comparison of different functions is outside the scope of this paper.

IV. AUTONOMOUS INSPECTION

The two main components of the autonomous inspection framework have been discussed so far: robust outlier analysis and Bayesian optimisation. In this section, the strategy for combining these two elements in order to create an efficient outlier search algorithm is laid out.

A. Initialisation

Both in the context of robust outliers as well as Bayesian optimisation, it would not be a good idea to initialise the algorithm with zero observations.

Considering first the robust estimates of the mean and covariance, one of the issues arises from the dimensionality of the features extracted from the data. In general, one would require at least twice as many observations as there are dimensions in the feature vector for the covariance to be well defined. From this point of view, at least $2d$ observations are required prior to commencing a scan. The issue is exacerbated by the fact that some of the observations may be potentially discarded from the set used to compute the covariance, during the FastMCD algorithm. This may bring the number of observations available to compute the feature covariance matrix down to too low a number.

On the other hand, one has the problem of the number of initial observations required to start a Bayesian optimisation scheme. In general, it is not a wise idea to initialise this type of sequential optimisation scheme with zero observations, as this would lead to undefined covairnace matrices, and the GP would be completely unable to make any predictions of the objective functions. For the GP to be rendered usable, at least one observation is required. Event in the limit of such low observations, this *one shot* sequential design results in an initial GP model that is performing pure extrapolation at the initial stage of the algorithm. Its predictions of where high function values and uncertain regions lie will tend to be poor in this extrapolation setting. The result would be an optimiser that takes much longer to converge to an optimum. Here, this means requiring more observations to reach a conclusion over the state of the component.

It is beneficial to initialise the algorithm with a small number of points, by performing a low resolution scan with a space-filling design. In the field of Optimal Experimental Design (OED) a popular choice for this is Latin Hypercube Sampling (LHS) [51]. Its main feature is that it ensures that samples are well distributed over the space of each individual dimension. The result is a GP that performs interpolation, as much as possible, right from the start of the sequential steps. One of the drawbacks of LHS is that it tends to underestimate the variance of the objective function [52], and this can have a serious impact on the performance of objective function as a novelty detector. For this reason, random sampling from a uniform distribution, in the interval between the minimum and maximum x, y coordinates is used. This is sub-optimal in terms space-filling, but is known to provide an un-biased estimate of objective function variance [52], which is ultimately more useful here.

The last, and perhaps most important reason to use a space-filling design such as LHS in the autonomous inspection scheme is so that the objective function being optimised is clearly defined. In this case, the objective function being maximised is the MSD given by equation (1). The MSD is a function of the data features and the Guassian distribution parameters, $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. These parameters are themselves a function of the data, estimated through FastMCD. The problem is that the optimiser decides where to place observtions based on the objective function, but these observations eventually change the definition of the objective function. This is an undesireable situation, but one that can be easily remedied by initialising the inspection with a space fillling design, of an appropriate sample size that ensures. Initialisation provides a stable objective function for which to optimise over. After initialisation, further updates to the mean and covariance of the data features are possible as new data arrives, as these parameters will converge to a stable solution as more observations are gathered.

Given the discussion above, the problem of exactly how many samples are required for initialisation translates to a problem finding the number of samples n required to correctly capture the parameters of a Gaussian distribution of dimension p. This problem has been investigated since the early days of application of pattern recognition algorithms to one-class classification problems [53] (to which novelty detection belongs), and continues to be a problem of interest in the present day [54]. The recommendation in [54] is to use a sample size of at least $14p$ in order to avoid non-convexities in the log-likelihood function used for estimation of the data feature covariance matrix. This specifies the minimum necesary to have a well-determined problem. The minimum number of samples required to correctly capture the variance of the novelty scores is a different one, and will in practice tend be greater than $14p$. The important aspect is

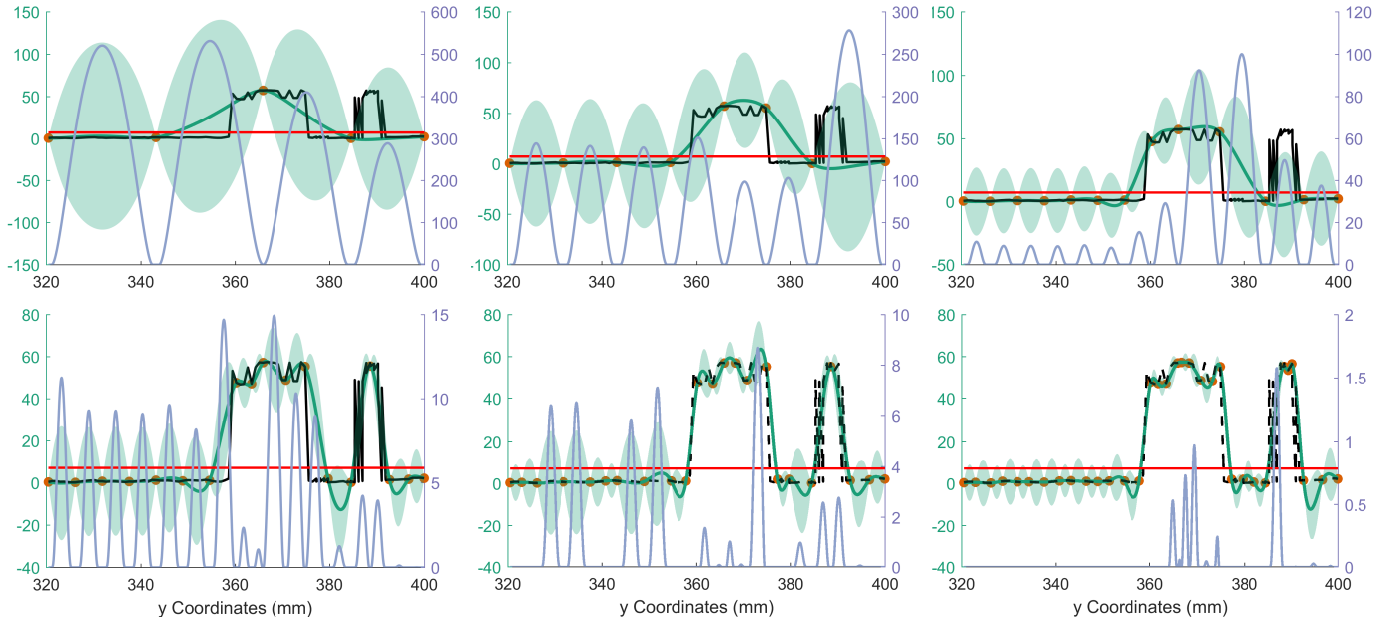

Fig. 6: Illustration of outlier search using Bayesian optimisation, on a one-dimensional slice of NDT data. The novelty scores (black) and its GP mean and 3σ-bounds (green), and acquired observations (orange circles) are shown on the left axis. The red horizontal line represents a 3σ threshold on the novelty scores. The expected improvement (blue) is overlaid and shown on the right axis.

gather enough samples to correctly capture the variance of the novelty indexes, so in this paper Monte Carlo samples are gathered until the novelty index variance converges to a stable value. In the wing panel ultrasound data used in this paper, in practice, 100 samples tend to yield a stable variance. However, this is domain-dependent.

B. Algorithm Iterations

After the algorithm is initialised, the main iterations follow Figure 1, where the four main elements of the iteration are:

1) Collect data and evaluate features.
2) Update robust mean and covariance including new observations and evaluate novelty indices for the entire set.
3) Condition the GP model on the new novelty indices.
4) Evaluate the Expected Improvement to find next best suitable location to gather data.

The overall search algorithm, incorporating the robust estimates of mean and covariance of NDT data features into the Bayesian optimisation framework is shown in Algorithm 3. Note that in this case, the multi-dimensional features of raw data are denoted as the vector $\mathbf{z}$, whilst the novelty scores are denoted as y (and $\mathbf{y}$ for arrays), in order to keep consistency with the input-output notation usually used in GP literature. The feature vector $\mathbf{z}$ could consist of features that are sensitive to flaws, such as ToF or attenuation, as discussed in the introduction of this paper. It is worth reiterating that the Mahalanobis distance computes novelty scores for multi-dimensional data, so it is possible to use high dimensional, informative features such as an autocorrelation function or an entire ultrasound pulse sequence. The use of such high-dimensional features in the context of ultrasound NDT and SHM has been investigated in the past [55].

An illustration of the application of Algorithm 3 is presented in Figure 6. The autonomous search was carried out using the same ultrasound data as the one utilised in Section II to demonstrate the application of robust outliers to NDT data. The data was taken along a one-dimensional cross-section of an aerospace carbon fibre plate specimen, that includes an area of delamination. The cross-section selected for the example shown in Figure 6 was selected as it cuts across two regions that contain damage. The example shows the output of the iterations for increasing number of observations, and initialised with five random samples. At the point of initialisation, the uncertainty close to observed values is low, and grows at points far away between observations. Overlaid on the plots, in light blue is the EI. Its maximum indicates the location where the next observation is to be taken. It can be observed from Figure 6 that as expected, the algorithm carries out a balancing act between exploration and exploitation. In the initial stage, it places slightly more observations at high novelty index values, but then moves away to explore areas of low novelty index values but high uncertainty. As more observations are gathered and areas of low novelty are explored (and thus its variance reduced), EI starts to weight more heavily those locations with high function values. On the last instance shown in Figure 6, observations have been placed evenly in areas of low novelty indexes, and there is a higher density of observations placed

around the areas of delamination. Furthermore, on this last instance the EI indicates a clear priority for high novelty indices, meaning that if the algorithm were to run more iterations it would concentrate around areas of damage.

The mean and covariance of the data are re-estimated at every iteration of the algorithm in a robust manner using FastMCD. It has been discussed above how the autonomous inspection algorithm presented here tends to place more observations in regions of damage. On the other hand there is the fact that the MCD method defines inliers as those representing at least 50% of the data. From these two points, it is easy to see how placing more than 50% of the observations around damaged regions will result in a shift where the damage area defines the data inliers, and the data for rest of the undamaged component are defined as outliers. This is of course undesired. In order to remedy this, any observations flagged as outliers after the initialisation step, have to be excluded from the set used to re-estimate the mean and covariance of the data. For this reason, the initialisation step is also very important. The LHS sampling scheme used for initialisation is agnostic to whether a region is damaged or not, so the locations where observations are placed are not biased in this sense. The use of the MCD is most critical at this stage. After initialisation, outliers can be removed through simple use of a threshold. However, MCD is an excellent way to to separate outliers in the type of batch data generated by the initial scan.

Algorithm 3 Autonomous Inspection

Initialise $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and $\mathbf{z}_{in}$ with low resolution C-scan
while $\mathbf{v}_{t-1} \geq \alpha\sigma_n^2 \vee t \geq budget$ **do**
$\mathbf{x}_t \leftarrow \arg\max_{\mathbf{x}} u(\mathbf{x}|\mathcal{D}_{t-1})$ ▷ Maximise utility
$\mathbf{z}_t \leftarrow$ Acquire data at $\mathbf{x}_t$
$\mathbf{y} \leftarrow mahal(\mathbf{z}_{in}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ ▷ Compute novelty scores
$\mathbf{z}_{in} \leftarrow \mathbf{z}[\mathbf{y}_t < T]$ ▷ Update outlier-excluding set
$\boldsymbol{\mu}, \boldsymbol{\Sigma} \leftarrow FastMCD(\mathbf{z}_{in})$ ▷ Update robust statistics
$\mathcal{D}_{1:t} \leftarrow \{\mathcal{D}, (\mathbf{x}_t, y_t)\}$ ▷ Include new observations
end while

C. Threshold estimation

One important aspect of the algorithm, in particular the novelty detection part, is the definition of a threshold over the MSD above which observations are defined as outliers. There is more than one way of tackling the threshold selection problem. Methods commonly used in SHM for this include the use of percentiles, Monte Carlo sampling [3], and extreme value statistics [**?**], [**?**]. In this paper, a simple threshold-selection method was used, for the sake of simplicity. The threshold was determined by computing the MSD of points six robust standard deviations away from the robust mean, along each dimension. This thresholding methodology makes use of the robust covariance computed by FastMCD, which has already been tuned for the separation of inclusive outliers.

D. Estimating probability of damage

The desired output from the search algorithm is a summary of the probability that the component being inspected contains flaws. An observation is flagged as abnormal if its novelty score is higher than the damage threshold. The GP provides a probabilistic estimate of the novelty scores over a two-dimensional spatial field. The probability that a given novelty score at spatial coordinate $\mathbf{x}$ corresponds to damage, is given by the probability that its novelty score is under the threshold T (computed as discussed in Section IV-C). This can be obtained using the mean and variance estimates from the GP, which define a local Gaussian distribution for the novelty scores through $\mathbf{m}_t$ and $\mathbf{v}_t$.

The probability that this uncertain measurement lies above the threshold, is thus given by the integral of the Gaussian probability density function between the threshold and infinity:

$$p(y > T) = \int_T^{\infty} \mathcal{N}(\mathbf{m}_t, \mathbf{v}_t)dy \tag{16}$$

which can be evaluated in a straightforward manner from the cumulative density function of the Gaussian:

$$p(y > T) = \Phi\left(\frac{\mathbf{m}_t - T)}{\mathbf{v}_t}\right) \tag{17}$$

where, as before, Φ represents a Gaussian normal cumulative density function. Note that this is effectively the same quantity as the PI acquisition function, used in Bayesian optimisation, but where the maxima is given by the novelty threshold. This quantity is then assessed at every spatial coordinate for which a GP prediction is available. When the variance of the GP is high, implying it is predicting at points far away from any observations, the natural tendency will be for the probability of damage to be high, as there will be a high probability mass between the T and ∞. Note that this measure is sensitive to the definition of T, so an estimate of the threshold should be used that truly captures the variability of the maxima of the novelty indexes. The probability of damage assumes the meaning of the probability that the GP prediction exceeds the highest novelty index that would be attributed to a normal-condition process.

Figure 13 provides an illustration of the problem using a small subset of the ultrasound NDT data from the wing panel. On the left Figure 13a shows a set of MSDs corresponding to the training data. Note that no values above the threshold are shown because, as per Algorithm 3 and the above discussion, these observations are discarded for the purpose of building a reference covariance model. On the right, Figure 13b shows four different predictions of a GP, corresponding to 1) a low mean with low variance, 2) a low mean with high variance, 3) a high mean with high variance and 4) a high mean with low variance. Overlaid on the probability densities is the probability of damage, computed using equation (17). This provides an intuitive view of this quantity; it is simply the amount of total probability mass above the threshold. In the case of an inlier GP prediction with low variance, this quantity is negligible, and thus close to zero; one can be confident that the observation is not an outlier. If variance

increases, a more significant amount of probability mass lies above the threshold, and hence the probability of damage rises, even though the mean may still be within the threshold. In the case of outlying means, a high variance will tend to lower the probability of damage, while a low variance will have the effect of getting it closer to one.

Note that the GP prediction is characterised by a Gaussian distribution. Its tails decay to infinity, so there will always be some small probability mass either under or over the threshold, therefore never allowing a probability of damage of zero or one. In practice, if either a very low, or very high variance exist, finite numerical machine precision means it is possible that a probability of one or zero to results from equation (17).

One important aspect to highlight is that there is a lower-bound on the probability of damage that can be predicted by the GP, and it is given by σ_n^2 (see equation (9)). This noise variance sets the minimum variance to be predicted by the GP at any given input location. In principle the lowest achievable MSD of any given observation is zero, as that means one has an observations that exactly matches the mean of the data. By construction, the lower-bound on damage probability can be obtained by setting $\mathbf{m}_t = 0$ and $\mathbf{v}_t = \sigma_n$ in equation (17). This lower-bound is of particular practical value when considering the termination criteria for the algorithm. This is discussed further in Section IV-G.

E. Two-dimensional example

An illustration of the autonomous inspection procedure in a two-dimensional subset is shown in Figure 8. Note that this is a subset of the data shown in Figure 2. As in the previous illustrations, ToF and ultrasound signal attenuation were used as features. The section shown in Figure 8 is the region around the delaminated area of the composite, on the top section. Algorithm 3 was applied to this data set.

The figure shows the evolution of the location of the observations as data is gathered, illustrating the mean and variance of the GP, and the probability of damage derived from them. Note that as soon as a region of high novelty values is discovered, the search prioritises exploitation around this region, until the uncertainty around it is significantly reduced, so it becomes more worth-while to explore somewhere else. Note that the result of the search is close to a grid search in regions of low novelty, with a resolution given by ℓ, while a higher resolution of observations is placed around potentially damaged regions.

F. Adjusting the level of risk via the GP prior

Even though the GP is often branded as a "nonparametric" regression technique [45], it does depend on a small number of hyperparameters, and their values will greatly influence the result of the optimisation. The interesting point is that their values will effectively dictate the level of uncertainty one is willing to accept about finding a flaw of a specific size. It has already been discussed in Section III-D how a covariance function can be interpreted as modelling the expected correlation between two pairs of observations, separated by a Euclidean distance r apart. By specifying the length-scale parameter of the covariance function, one effectively specifies the expected distance from observed points, where observations will be correlated.

As previously discussed in III-D, the Matérn kernel will be used in the experimental investigations of this paper. Referring back to (12), the relevant hyper-parameters in this case are the characteristic length-scale, ℓ^2, and the scaling parameter, σ_f^2, of the covariance function. Since observations are assumed to be corrupted with noise, a variance term, σ_n^2, is also required to model the additive noise on top of the functional input-output relationship.

The length-scale, ℓ controls the rate, in the input space, at which the variance grows as one moves away from observed values. In this problem of spatial sampling, it means that small length-scales will result in variance growing at high rates between observations far apart, while long length-scales will result in the GP predicting a lower variance between those same observations. This point is illustrated in Figure 9 where the variance is shown for a GP conditioned on two observations for a range of growing length-scales.

This point of the length-scale controlling the variance rate of growth is emphasised here because *it effectively controls the risk one is willing to accept regarding detecting a specific flaw size*. This is the tuning parameter of this algorithm that will have the most impact on the resulting data acquisition, and overall probability of damage estimates. It will influence the acquisition of data, as this is decided based on the acquisition function (EI in this case), and this will be guided partly to place observations in areas of high uncertainty. Short length-scales will mean observations will be placed closer together, and vice-versa.

The recommendation here is that the length-scale used should match the minimum expected flaw size. The form of the covariance function used can shed some light into what this implies. For instance, the (unscaled) Matérn $3/2$ function used in this paper in particular, yields a correlation of $1/2$ at a distance ℓ apart, and a sharp decay in correlation beyond that. This means that good interpolation is expected from the GP at a distance of at least ℓ and placing observations closer to this is inefficient, whilst placing observations further away increases risk.

If a component were to be scanned that contained no anomalies whatsoever, this autonomous inspection scheme based on Bayesian optimisation should result in a scan with an average spatial resolution of ℓ (the acquisition of data would change when high novelty scores are found).

The other two hyperparameters also play an important role in the uncertainty quantification given by the GP; the process variance σ_f^2 and the noise variance σ_n^2. The process variance will place an upper bound on the total possible uncertainty. It effectively specifies prior information on the uncertainty and defines the variance in the absence of any observations. The noise variance does exactly the opposite and places a lower-bound on the uncertainty by specifying the variance around regions where plenty of observations exist. In other words, the noise variance provides an indication of the inherent variability in novelty indexes that is to be expected by taking

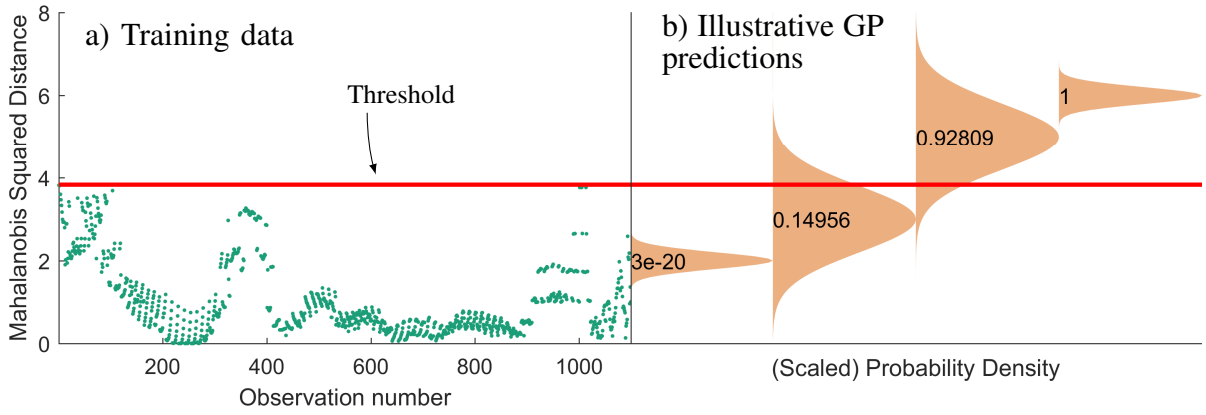

Fig. 7: Illustration of probability of damage, illustrating a) threshold placed over MSDs from a training data set and b) four illustrative GP predictions, two inliers and two outliers, each with low and high variance.

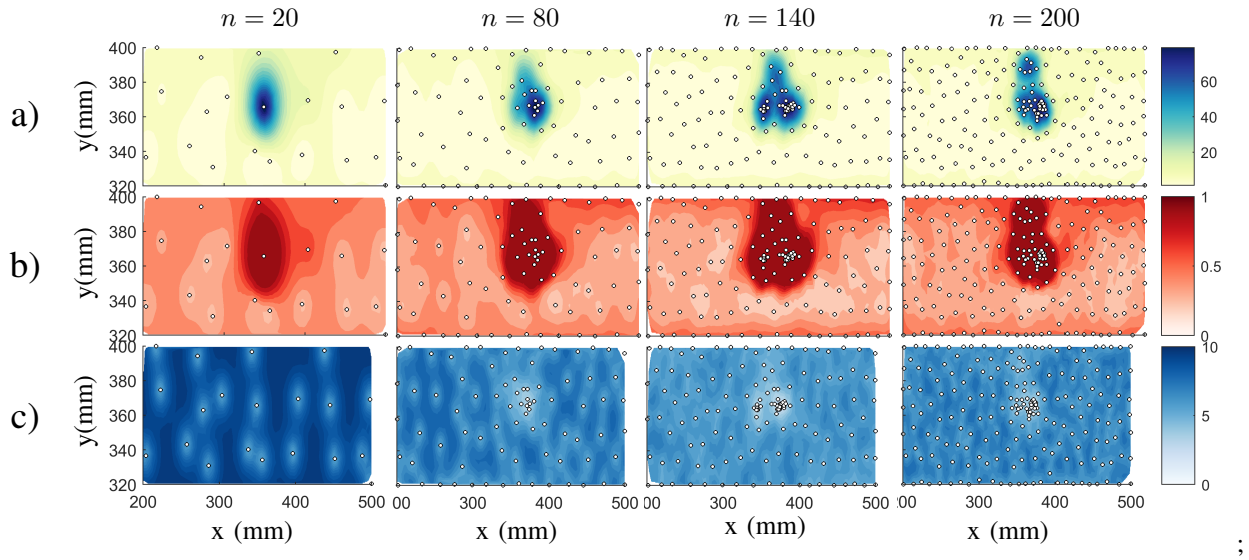

Fig. 8: a) Novelty scores (contour indicates GP mean while round markers denote measured locations), b) probability of damage, and c) GP standard deviation, from a two-dimensional subset NDT data from an aerospace specimen. Illustration shows evolution of predictions as observations are collected. n denotes the number of observations.

the same observation multiple times at the same location.

G. Convergence

An important question to ask is when should the algorhtm stop searching? This is non-trivial. If it were to be collecting observations forever, then it would bear no advantage against a simple grid search. All three hyperparameters affect convergence characteristics, each in a different way.

The lower bound on the variance given by σ_n^2 is clearly important from the point of view of convergence criteria. If all the regions of high function values have been explored and the uncertainty across the whole range of possible measurement locations has been shrunk to a level close to σ_n^2, then there is no more information to be gained from further observations, so the optimisation has converged and the search should be stopped. The noise variance is therefore a critical value and it is crucial that it be estimated from the actual data collected, so that it represents the true noise floor that the optimiser is working towards.

As intuition would now dictate, the length-scale controls the total number of observations required to achieve a given spatial resolution, and to shirnk the uncertainty elsewhere to the lower-bound. Short length-scales therefore lead to low convergence rates and vice-versa. This effect is illustrated in Figure 10, where the average variance and its 3σ bounds are shown for 800 iterations of Algorithm 3 over a two-dimensional sub-section of a composite specimen.

The process noise of the GP σ_f, has the ability to control the exploration-exploitation characteristics of the optimiser. Because it defines an upper limit on $\mathbf{v}_t$, if it is set too low, even with a low number of observations, most areas of the space will default to σ_f^2. The result is that the acquisition function (assuming EI, or a similar measure is used) will choose to only scan at points of high mean values, thus performing pure exploitation. This would lead to a high

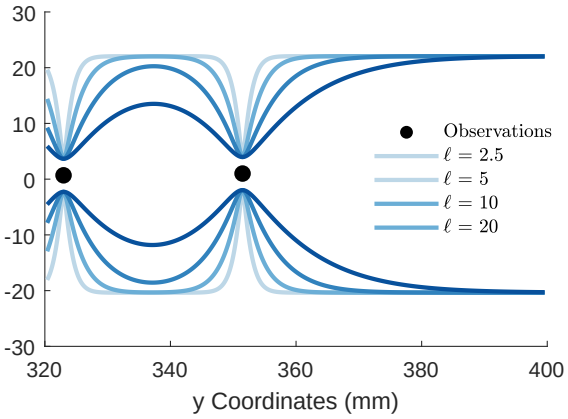

Fig. 9: Effect of length-scale hyperparameter on uncertainty predictions of the GP. Example shows 3σ bounds for four GP posteriors, conditioned on two observations, using increasing length-scales. A Matérn $3/2$ covariance function was used.

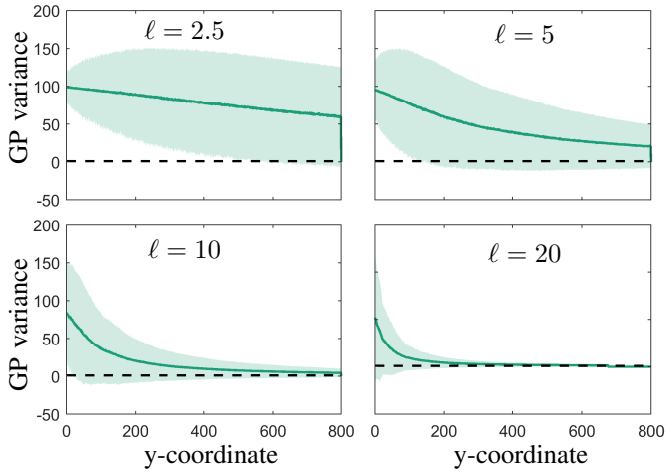

Fig. 10: Illustration of convergence rate with varying lengthscales. The solid line represents the average variance of the GP predictions in two-dimensional space, the shaded area represents the 3σ bounds for the GP variance (the variance of the variance).

resolution scan around areas were flaws are suspected (high $\mathbf{m}_t$), but an under-exploration of the rest of the space. On the other hand, setting a high value of σ_f^2 would result in pure exploitation, as it would lead to extremely high uncertainty in the space observations. A middle ground is clearly necessary, and one reasonable solution would be to scale the data to unit standard deviation, and set scaling hyper-parameter to a reasonable value such as $\sigma_f^2 = 1$. However, the data in this case are novelty scores, evaluated sequentially, so the re-scaling of the data would also need to happen sequentially, introducing another layer of complexity.

H. Hyperparameter selection

As discussed in Section IV-F, selection of ℓ should be done purely in terms of expected flaw size, so this should be user and problem dependent. The process and noise variances on the other hand, should be estimated from the data, as these control the optimiser behaviour and the convergence criteria. There are a number of methods available to do so, and a discussion of these is outside the scope of this paper. The view of the authors, however, is that some effort should be taken so that these two hyperparameters at least approximate the best fit of the data. In this paper, hyperparameters are learned using the data gathered at the initialisation stage, using Bayesian optimisation with the marignal likelihood of the GP as an objective function. More details on the general use of Bayesian optimisation for machine learning models is given in [56]. There are various ways of carrying out the hyper-parameter seleciton in the general case of GPs, some are discussed in [45]. Some interesting results using Sequential Monte Carlo (SMC) to update hyperparameter estimates in an optimisation setting are presented in [57].

I. Managing the computational burden of the GP

Note that in order to evaluate them, the matrix $\mathbf{P}$, defined in Equation (9) has to be inverted. The covariance matrix $K(\mathbf{x}_t, \mathbf{x}_t)$ will have a size of $n_t \times n_t$, where n_t is the number of training observations. Inversion of $\mathbf{P}$, therefore carries a computational complexity of $\mathcal{O}(n_t)^3$. As the number of training observations increase, the computational burden can be a problem for the GP model. Various methods exist in order to aleviate this, most of which leverage the idea of approximating the posterior over f using *inducing points*. Reviews of inducing-point approximations can be found in [58] and more recently in [59]. Applications of sparse GPs to modelling large quantities of spatial data can also be found in [60]. Note that while sparse approximations to GPs are practical in certain situations, to the authors' awareness, their application to the problem of Bayesian optimisation has not been investigated in detail, so any use of it should be carried out carefully.

In the context of inspection, the matrix inversion problem will only arise when scanning large areas with a small target minimum flaw size, as this would lead to a problematic number of large training points. A straight-forward way to deal with the computational burden would be to break the inspection area down to a smaller number of manageable sections, each with their own local GP.

Another source of computational burden is the evaluation of the covariance function between pairs of training and testing points, $K(\mathbf{x}, \mathbf{x}^\star)$ where, in the spatial domain, $\mathbf{x}^\star$ defines the field over which the GP posterior is to be evaluated. The number of training points, $\mathbf{x}$ should be small compared to the field over where a mean and variance of the GP are sought (if not then one may as well just do a high resolution C-scan instead). Due to the potentially large size of $\mathbf{x}$, the simple task of evaluating this covariance matrix could be fairly burdensome, to the point where computation at each step may take longer than the physical data collection.

This problem can be dealt with by evaluating only a small number of points in $\mathbf{x}$, chosen by sampling a uniform distribution over their indices, at every iteration of the Bayesian optimisation scheme. This would alleviate the computational

burden, but also ensure that in the limit of a large number of random draws from the physical coordinates, the evaluated field will approximate the true field of the GP.

V. Evaluation of algorithm performance

This section presents an experimental investigation of the application of the autonomous inspection scheme presented so far. The test specimen being presented is an aerospace composite panel manufactured from carbon fibre composite. The specimen contains at least two regions where delamination has occurred.

It is important to highlight that the algorithm is being applied to pre-recorded data. This allows for the investigation, presentation and discussion, to be of a general nature and focused on the statistical inference side of the algorithm, without having to consider the integration and constraints that would apply to any specific robotic interface.

A. Experimental set-up and data acquisition on composite wing panel

An 800mm $\times$ 550mm composite panel was scanned using a six-axis robotic head, with a water-coupled ultrasound probe. The probe consists of 64 transducers, each of which fires a 5MHz tone burst, and also acts as a receiver. The resolution of the scan can be adjusted, but for these results, the speed of the probe was adjusted to yield a spatial resolution of 0.8mm in the direction of the probe travel. The C-scan shown in Figure 2c was generated using this data set. Details of this experimental procedure have been published in [1], where the interested reader is referred to for further details.

In order to capture the range of different depths of this composite specimen, an acquisition time of 9.6μs was used. This equates to 480 samples at a sample rate of 50 MHz. The original ultrasound pulses collected through the water-coupled probe are susceptible to misalignment with respect to the arrival time of the first burst (the front wall reflection). This is evident, for example, in the illustrative B-scan shown in Figure 2b.

Two data features extracted from the ultrasound pulses were used: ToF and attenuation, between the front and back wall of the specimen. In order for the missaligment across B-scans not to affect the data features, the ToF and attenuation were both estimated using a Hilbert envelope of the autocorrelation of each pulse, where the location of the maximum lag specifies the ToF and the attenuation is extracted from the argument of the envelope at that time lag. Note that these computation could be performed directly in a compressed domain, thus requiring an even lesser quantity of data, adopting the techniques developed in [20].

Specimens containing different types of realistic damage are hard to come by, so in order to demonstrate the algorithm on a real specimen, the same specimen available was divided into a total of 8 different regions, some containing damage and some not. These regions are shown in Figure 11. Note that only regions from nominally similar areas of the composite specimen are considered here. These are the thin

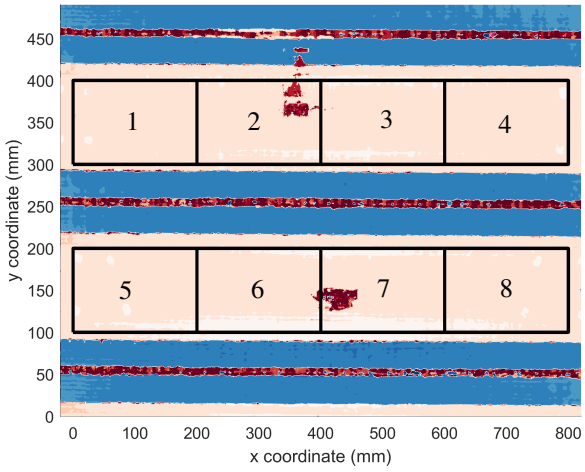

Fig. 11: Divisions of composite specimen used for experimental investigation. Colour-map indicates ultrasonic ToF.

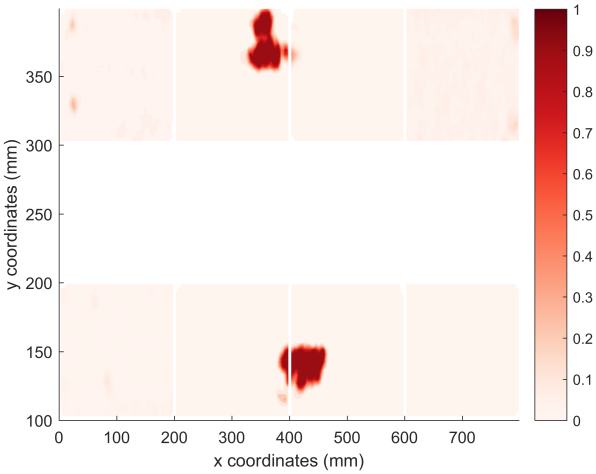

Fig. 12: Probability of damage contours resulting from the automated inspection algorithm, applied to eight local regions in the composite specimen.

wall sections, as there are two areas of known delamination in this region, as well as enough area free of damage in order to demonstrate the inspection results in both cases.

B. Results

The probability of damage resulting from the application of the autonomous inspection algorithm is shown in Figure 12, as contours overlaid over the spatial coordinates of each of the eight sub-sections analysed individualy. The damage probability contours, correctly localise the two known areas of de-lamination with a near 100% probability of damage. The two de-laminated regions have cross-sections of no less than 50mm. As such, the result is not too surprising, given that a length-scale of $\ell = 10mm$ was used for this search, so any regions with observations at a spacing much more than ℓ will develop high uncertainty, and thus be prioritised by the search. This makes the identification of damage in

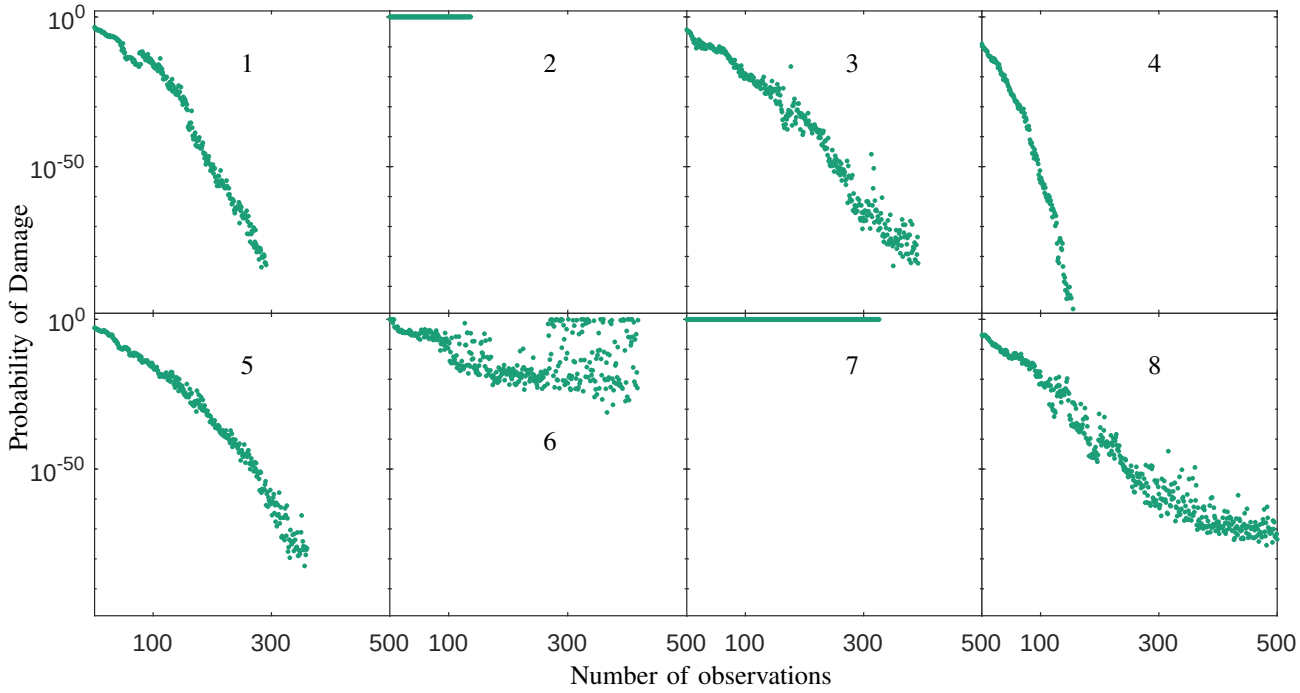

Fig. 13: Probability of damage evolution during sequential data acquisition, for the eight different regions analysed.

regions two and six, almost guaranteed. More interesting is the fact that the small area of de-lamination that was caught up in zone three has been detected with high probability, even though the maximum size of the de-laminated area, (as observed through the original high resolution C-scan) it has an overal maximum length of 5mm.

As observations are collected sequentially, one is interested in the evolution of the probability of damage for every region. This is shown in Figure 13. For each region of inspection, the interesting metric overal maximum observed damage, so in order to summarise this, the 99^{th} percentile of the estimated damage prbability per region is plotted in Figure 13. Note that this "test" was run with a maximum budget of 500 observations per region, although some terminated early, due to the lowest predicted GP variance reaching the estimated σ_n^2. In zones two and seven, where a probability of damage of 100% has been estimated, termination occured as the algorithm attempted to collect data at a previously observed location, of high novelty index value. This is an artifact of applying the algorithm to pre-recorded data, as there is a finite discretisation grid. The only region that did not converge in this case is region eight, albeit with a very small probability of damage of $10{\times}10^{-80}$.

During the first 100 observations, the damage probability tends to jitter. This is due to the fact that the robust statistics were re-estimated througout the first 100 observations, and held fixed afterwards. Also, in regions with small manufacturing imperfections (evident as small round features, predominantly in regions one and four), the 99^{th} percentile of the damage probability tends to jitter around, instead of converging down to a small value.

It is easy to note that in the regions where damage was found, this was found with a very low number of observations, and any more observations from that point onwards only help to confirm the presence, location and extent of damage (exploitation, in the Bayesian optimisation context), as well as exploring the remaining space, once the damage has been characterised well. Conversely, regions without damage take more observations to converge to a low damage probability, as it simply takes more observations to reduce the uncertainty along all spatial coordinates by means of exploring the space. It is also interesting to see that in zone six, where only a 5mm region of damage exists, the damage probability did not reach a level close to 100% until around 300 observations were acquired.

The advantage of applying this autonomous inspection scheme is made clear when one considers that the originald C-scan data for each zone consists of approximately 39000 observations (individual ultrasound pulses, or A-scans). Performing the data collection sequentially using the Bayesian optimisation scheme reduces this significantly, while at the same providing the required information output in the form of a damage probability.

VI. CONCLUSIONS

A fully data-driven framework for autonomous inspection based on a combination of Bayesian optimisation ideas, and robust outlier analysis has been presented, and demonstrated on an NDT data set consisting of ultrasound measurements on a high value aerospace composite specimen. The demonstration of the algorithm on this specific specimen was chosen as it contains damage-free regions as well as de-laminated sections. The algorithm is data-driven as it makes use of robust inference of the mean and covariance of the NDT data features, which means that outliers that are present in the data set, either because they represent

anomalies or simply bad data, are identified and exluded from the estimation of the data statistics. The inspection scheme presented here is probabilistic, and it is aimed at minimising risk against an expected flaw size. The risk minimisation is provided by utilising standard Bayesian optimisation techniques, which also allow this algorithm to run sequentially. The impact is that the damage inference can be performed on-line, and terminated either when no further information can be gained from the scanned, or when damage has been found. Bayesian optimisation is designed to find optimal points using the minimum number of observations possible. In the context of inspeciton, this implies that if a specimen does contain damage, or other anomalies, they will be found quickly, and the inspeciton will focus around this areas. If no damage is present, the inspection will focus on minimising risk, by means of exploring the space. This paper has focused on the presentation of the algorithm, with demonstrations on real-world robotic platforms left as motivation for future work.

ACKNOWLEDGMENT

This work has been funded by the UK Engineering and Physical Sciences Research Council (EPSRC), through the Autonomous Inspection in Manufacturing and Remanufacturing (AIMaReM) grant.

REFERENCES

[1] C. Mineo, C. MacLeod, M. Morozov, S. G. Pierce, R. Summan, T. Rodden, D. Kahani, J. Powell, P. McCubbin, C. McCubbin, G. Munro, S. Paton, and D. Watson, "Flexible integration of robotics, ultrasonics and metrology for the inspection of aerospace components," in *AIP Conference Proceedings*, vol. 1806, p. 020026, AIP Publishing LLC, feb 2017.

[2] C. Mineo, S. G. Pierce, P. I. Nicholson, and I. Cooper, "Robotic path planning for non-destructive testing - A custom MATLAB toolbox approach," *Robotics and Computer-Integrated Manufacturing*, vol. 37, pp. 1–12, feb 2016.

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

[4] R. Fuentes, *On Bayesian Networks for Structural Health Monitoring*. Doctoral thesis, The University of Sheffield, 2017.

[5] M. Friedrich, W. Galbraith, and G. Hayward, "Autonomous mobile robots for ultrasonic NDE," in *Proceedings - IEEE Ultrasonics Symposium*, vol. 1, pp. 898–901, IEEE, 2006.

[6] R. Bogue, "The role of robotics in nondestructive testing," *Industrial Robot: An International Journal*, vol. 37, pp. 421–426, aug 2010.

[7] J. T. Stetson and W. De Odorico, "Robotic inspection of fiber reinforced composites using phased array UT," in *AIP Conference Proceedings*, vol. 1581, pp. 1889–1895, American Institute of Physics, feb 2014.

[8] R. Summan, G. Dobie, J. Hensman, S. G. Pierce, K. Worden, D. O. Thompson, and D. E. Chimenti, "A probabilistic approach to robotic NDE inspection," in *AIP Conference Proceedings*, vol. 10, pp. 1999–2006, 2010.

[9] S. Doctor, T. Hall, and L. Reid, "SAFT the evolution of a signal processing technology for ultrasonic testing," *NDT International*, vol. 19, pp. 163–167, jun 1986.

[10] R. Drai, F. Sellidj, M. Khelil, and A. Benchaala, "Elaboration of some signal processing algorithms in ultrasonic techniques: application to materials NDT," *Ultrasonics*, vol. 38, pp. 503–507, mar 2000.

[11] F. Honarvar, H. Sheikhzadeh, M. Moles, and A. N. Sinclair, "Improving the time-resolution and signal-to-noise ratio of ultrasonic NDE signals," *Ultrasonics*, vol. 41, pp. 755–763, mar 2004.

[12] L. Udpa and S. Udpa, *IEE Proceedings. Part F. Radar and signal processing.*, vol. 138. Institution of Electrical Engineers, feb 1991.

[13] H. Sohn, D. Dutta, J. Y. Yang, M. DeSimio, S. Olson, and E. Swenson, "Automated detection of delamination and disbond from wavefield images obtained using a scanning laser vibrometer," *Smart Materials and Structures*, vol. 20, p. 045017, apr 2011.

[14] S. Sambath, P. Nagaraj, and N. Selvakumar, "Automatic defect classification in ultrasonic NDT using artificial intelligence," *Journal of Nondestructive Evaluation*, vol. 30, pp. 20–28, mar 2011.

[15] N. A. Akram, D. Isa, R. Rajkumar, and L. H. Lee, "Active incremental Support Vector Machine for oil and gas pipeline defects prediction system using long range ultrasonic transducers," *Ultrasonics*, vol. 54, pp. 1534–1544, aug 2014.

[16] T. M. Nunes, V. H. C. de Albuquerque, J. P. Papa, C. C. Silva, P. G. Normando, E. P. Moura, and J. M. R. Tavares, "Automatic microstructural characterization and classification using artificial intelligence techniques on ultrasound signals," *Expert Systems with Applications*, vol. 40, pp. 3096–3105, jun 2013.

[17] R. Marchant and F. Ramos, "Bayesian Optimisation for informative continuous path planning," in *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 6136–6143, IEEE, may 2014.

[18] G. Cardoso and J. Saniie, "Data compression and noise suppression of ultrasonic NDE signals using wavelets," in *IEEE Symposium on Ultrasonics, 2003*, vol. 16, pp. 250–253, IEEE, 2015.

[19] G. M. Zhang, C. Z. Zhang, and D. M. Harvey, "Sparse signal representation and its applications in ultrasonic NDE," *Ultrasonics*, vol. 52, no. 3, pp. 351–363, 2012.

[20] R. Fuentes, K. Worden, I. Antoniadou, C. Mineo, S. Pierce, and E. Cross, "Compressive sensing for direct time of flight estimation in ultrasound-based NDT," in *Structural Health Monitoring 2017: Real-Time Material State Awareness and Data-Driven Safety Assurance - Proceedings of the 11th International Workshop on Structural Health Monitoring, IWSHM 2017*, vol. 2, 2017.

[21] J.-F. Wang, A. Stein, B.-B. Gao, and Y. Ge, "A review of spatial sampling," *Spatial Statistics*, vol. 2, pp. 1–14, 2012.

[22] J. Li and A. D. Heap, "Spatial interpolation methods applied in the environmental sciences: A review," *Environmental Modelling & Software*, vol. 53, pp. 173–189, mar 2014.

[23] N. Best, S. Richardson, and A. Thomson, "A comparison of Bayesian spatial models for disease mapping," *Statistical Methods in Medical Research*, vol. 14, pp. 35–59, feb 2005.

[24] F. Schoefs, E. Bastidas-Arteaga, T. Tran, G. Villain, and X. Derobert, "Characterization of random fields from NDT measurements: A two stages procedure," *Engineering Structures*, vol. 111, pp. 312–322, mar 2016.

[25] M. L. Stein, *Interpolation of spatial data: some theory for kriging*. Springer Series in Statistics, 1999.

[26] C. Gomez-Cardenas, Z. Sbartaï, J. Balayssac, V. Garnier, and D. Breysse, "New optimization algorithm for optimal spatial sampling during non-destructive testing of concrete structures," *Engineering Structures*, vol. 88, pp. 92–99, apr 2015.

[27] H. Sang and A. E. Gelfand, "Continuous Spatial Process Models for Spatial Extreme Values," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 15, pp. 49–65, mar 2010.

[28] J. H. Kurz, A. Jüngert, S. Dugan, G. Dobmann, and C. Boller, "Reliability considerations of NDT by probability of detection (POD) determination using ultrasound phased array," *Engineering Failure Analysis*, vol. 35, pp. 609–617, dec 2013.

[29] B. Yee, F. Chang, J. Couchman, G. Lemon, and P. Packman, "Assessment of NDE reliability data," tech. rep., 1975.

[30] F. Schoefs, A. Clément, and A. Nouy, "Assessment of ROC curves for inspection of random fields," *Structural Safety*, vol. 31, pp. 409–419, sep 2009.

[31] N. Dervilis, E. Cross, R. Barthorpe, and K. Worden, "Robust methods of inclusive outlier analysis for structural health monitoring," *Journal of Sound and Vibration*, vol. 333, pp. 5181–5195, sep 2014.

[32] P. J. Rousseeuw, "Least median of squares regression," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 871–880, 1984.

[33] D. M. Hawkins and J. S. Simonoff, "High Breakdown Regression and Multivariate Estimation," *Journal of the Royal Statistical Society, Series C*, vol. 42, no. 2, pp. 423–432, 1993.

[34] D. M. Hawkins, "The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data," *Computational Statistics & Data Analysis*, vol. 17, pp. 197–210, feb 1994.

[35] M. Hubert and M. Debruyne, "Minimum covariance determinant," jan 2010.

[36] R. W. Butler, P. L. Davies, and M. Jhun, "Asymptotics for the Minimum Covariance Determinant Estimator," *The Annals of Statistics*, vol. 21, pp. 1385–1400, sep 1993.

[37] P. J. Rousseeuw and K. Van Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[38] Y. Humeida, P. D. Wilcox, M. D. Todd, and B. W. Drinkwater, "NDT & E International A probabilistic approach for the optimisation of ultrasonic array inspection techniques," *NDT and E International*, vol. 68, pp. 43–52, 2014.

[39] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[40] S. Streltsov and P. Vakili, "A Non-myopic Utility Function for Statistical Global Optimization Algorithms," *Journal of Global Optimization*, vol. 14, pp. 283–298, 1999.

[41] J. Mockus, *Bayesian Approach to Global Optimization*, vol. 37. 1989.

[42] J. Mockus, "Application of Bayesian approach to numerical methods of global and stochastic optimization," *Journal of Global Optimization*, vol. 4, pp. 347–365, jun 1994.

[43] A. O'Hagan and J. Kingman, "Curve Fitting and Optimal Design for Prediction," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 40, no. 1, pp. 1–42, 1978.

[44] A. Zilinskas, "On the use of statistical models of multimodal functions for the construction of the optimization algorithms," in *Lecture Notes in Control and Information Sciences*, ch. 23, 1980.

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

[46] H. J. Kushner, "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise," *Journal of Basic Engineering*, vol. 86, p. 97, mar 1964.

[47] D. R. Jones, "A Taxonomy of Global Optimization Methods Based on Response Surfaces," *Journal of Global Optimization*, vol. 21, no. 4, pp. 345–383, 2001.

[48] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[49] D. D. Cox and S. John, "A statistical method for global optimization," in *Multidisciplinary Design Optimization: State of the Art*, pp. pages 315–329, 1997.

[50] P. Hennig and C. J. Schuler, "Entropy Search for Information-Efficient Global Optimization," 2011.

[51] M. D. McKay, R. J. Beckman, and W. J. Conover, "Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, vol. 21, pp. 239–245, may 1979.

[52] M. Stein, "Large sample properties of simulations using latin hypercube sampling," *Technometrics*, vol. 29, no. 2, pp. 143–151, 1987.

[53] S. Raudys and A. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 252–264, mar 1991.

[54] P. Zwiernik, C. Uhler, and D. Richards, "Maximum likelihood estimation for linear Gaussian covariance models," *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, vol. 79, no. 4, pp. 1269–1292, 2017.

[55] F. Mustapha, K. Worden, S. G. Pierce, and G. Manson, "Damage Detection Using Stress Waves and Multivariate Statistics: an Experimental Case Study of an Aircraft Component," *Strain*, vol. 43, pp. 47–53, feb 2007.

[56] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian Optimization of Machine Learning Algorithms," in *Advances in Neural Information Processing Systems 25 (NIPS)*, pp. 2951–2959, 2012.

[57] R. B. Gramacy and N. G. Polson, "Particle learning of Gaussian process models for sequential design and optimization," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 102–118, 2011.

[58] J. Quinonero-Candela, J. M. De, C. E. Rasmussen, and C. M. De, "A Unifying View of Sparse Approximate Gaussian Process Regression," *Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[59] T. D. Bui, J. Yan, and R. E. Turner, "A Unifying Framework for Gaussian Process Pseudo-Point Approximations using Power Expectation Propagation," *Journal of Machine Learning Research*, vol. 18, pp. 1–72, 2017.

[60] C. Park, J. Z. Huang, and Y. Ding, "Domain Decomposition Approach for Fast Gaussian Process Regression of Large Spatial Data Sets," *Journal of Machine Learning Research*, vol. 12, no. May, pp. 1697–1728, 2011.